Generative modelling with diffusion models on Riemannian manifolds

Michael Hutchinson

Department of Statistics

University of Oxford

a dissertation submitted for

Doctor of Philosophy

October 2024

Declaration

This dissertation and all research contained in it are a product of my original work, except where indicated otherwise by explicit statement or reference. All ideas, quotations, and data originating from the works of others, published or otherwise, are fully acknowledged according to standard practices of the academic discipline.

Acknowledgments

This thesis is dedicated to all who have supported me in one way or another during the pursuit of the work contained in this thesis. The process of completing this degree has been one of the most challenging experiences of my life, and would not be here without their support. It has at the same time as being difficult been the most enjoyable, rewarding experience I have had, and I am grateful to everyone who has made this such a positive experience.

Firstly, I am profoundly grateful to all those who have supervised my work and helped me grow as a researcher. I thank my supervisor Yee Whye Teh for taking me on as a student and giving me the support and time to develop my own interests and skills. To say I was surprised to receive an offer to join his group is an understatement, and I am grateful for the chance taken on me to pursue research. I would also like to thank Max Welling for his support as a supervisor during these studies. I also thank Rich Turner, who started me on the path to pursuing machine learning academically with an ad-hoc tutorial on neural network in my first year of undergraduate studies, who went on to supervise my masters work, and encouraged me to pursue the opportunity to study with Yee Whye and Max.

Secondly, I would like to thank the academic communities I have encountered on my path. The Oxford Computational Statistics and Machine Learning group has been an ideal place to pursue research, and I am grateful for the wide range of personalities and interests that makes it such a special place. I thank Adam Foster, Adam Golinski, Emilien Dupont, Bryn Elesedy, Bobby He, Tom Rainforth, Dominic Richards, Mrinank Sharma, Jean-Francois Ton, Charline Le Lan and Sheheryar Zaidi for their support in the group. I am particularly grateful to the group of collaborators who worked on much of the work in this thesis with me, Emile Mathieu, Valentin De Bortoli, Leo Klarner, Nic Fishman, James Thornton, Vincent Dutordoir, and Arnaud Doucet. I would specifically like to thank Leo for help proof-reading this thesis, and Valentin for his answering of unending technical questions I posed to him while writing.

Away from Oxford I would like to thank Christos Louizos and Matthias Reisser for the time I spent working with them at Qualcomm, and Danilo Rezende and Ivo Danihelka while I was at DeepMind. I would like to thank Maurice Weiler, Roberto Bondesan, and all the people I met at Hotel Casa for making my time in Amsterdam such a refreshing experience. I would also like to thank the group of collaborators consisting of Alexander Terenin, Viacheslav Borovitskiy, So Takao and Marc Deisenroth for helping me pursue a research project well outside my comfort zone and teaching me a lot along the way. I am particularly grateful to

Alex for teaching me so much about how to present research beautifully, and for providing an inspiration for academic standards. I would also like to thank Peter Wirnsberger and Max Jaderberg for giving me the time and support to write up this thesis while I started working at Isomorphic Labs.

Besides research, I am indebted to those who have been integral parts of my life over the years and who have provided so much joy, advice, and perspective. Those who I have known my whole life, MJ, Alex, Dishey, Paul, Laura, Catherine, Nicola, Peter, Esme, Harry and River. Those who I met at school and who helped make that a positive experience, Finn, James, Hannah, Ed and Serena. Friends from Cambridge, Arthur, Georgia, Barney, Freddie, Adam, Arianna, Leah, Sathya, Seb, Rohan, Damien, Robin, Alex and Beki. And the people I spent so much timer with in Oxford, Becca, Tom, Julia, Alizeh, Jen, Morgan, Nina, Sahra, Clare, Imi, Julian, and all the members of University College WCR and UCBC, who have made studying for this degree so enjoyable.

Finally, I would like to thank my family. My parents Melanie and David have always encouraged me in pursuing what I find meaningful in life, and for that I am forever grateful. My siblings Harry and Susanna have been pillars of support and levity, and the best a brother could hope for.

Abstract

Generative modelling is a cornerstone approach of modern deep learning for modelling the distribution of a complex random variable from samples. It has seen applications in a wide range of settings, including the modelling of images, text, and video. Orthogonally, the field of geometric deep learning has developed to incorporate prior knowledge about the geometric structure of a problem into the solutions deep learning systems learn. The intersection of these domains has proved highly applicable to solving scientific modelling problems with deep learning.

In this thesis we study a particular class of generative models, known as diffusion models, or score-based generative models. Typical diffusion models model densities supported on a Euclidean space. We generalise this to a series of settings in Riemannian geometry.

Firstly, we extend the continuous-time diffusion modelling framework to model densities supported on Riemannian manifolds.

Secondly, we extend it to model densities supported on Riemannian manifolds with boundaries. As a means to accomplishing this we also introduce a new discretisation technique for reflected stochastic differential equations.

Finally, we extend it to model densities supported on the spaces of tensor fields on Riemannian manifolds, and on the spaces of paths on Riemannian manifolds. In tandem with this we introduce new techniques for conditionally sampling from diffusion models.

In total these contributions allow for the application of diffusion models to a wide range of scientific problems, and we demonstrate this applicability in numerous settings.

Table of Contents

1 Introduction . . . 1
2 Background . . . 9
2.1 Geometric machine learning . . . 9
2.1.1 Geometric priors . . . 10
2.1.2 Some types of geometric priors . . . 14
2.1.3 Manifolds . . . 15
2.1.4 Geometric deep learning on manifolds . . . 24
2.2 Deep generative modelling . . . 24
2.2.1 Score-Based Generative Modelling . . . 26
2.2.2 Deep generative modelling on manifolds . . . 37
3 Score-based modelling on Riemannian manifolds . . . 39
3.1 Introduction . . . 39
3.2 Riemannian Score-Based Generative Modelling . . . 40
3.2.1 Stochastic differential equations on manifolds . . . 41
3.2.2 Noising processes on manifolds . . . 42
3.2.3 Time-reversal on Riemannian manifolds . . . 44
3.2.4 Score approximation on Riemannian manifolds . . . 46
3.2.5 Likelihood computation . . . 48
3.2.6 Approximate sampling of diffusions . . . 49
3.2.7 Predictor-corrector scheme . . . 50
3.2.8 Parametric family of vector fields . . . 51
3.2.9 Riemannian score-based generative models . . . 55
3.2.10 Amortised conditional modelling . . . 56
3.2.11 Invariant distribution modelling . . . 56
3.3 Riemannian Score-Based Generative Modelling on Compact Manifolds . . . 57
3.3.1 Forward noising process . . . 57
3.3.2 Learning the score . . . 57
3.3.3 Convergence results in the compact setting . . . 59
3.3.4 Comparison of score-based model types . . . 60
3.4 Experiments . . . 60
3.4.1 Stereographic baseline method . . . 60
3.4.2 Geologic and weather datasets on the sphere . . . 61
3.4.3 Synthetic data on tori . . . 62
3.4.4 Synthetic data on the Special Orthogonal group . . . 63
3.4.5 Deeper comparison with Moser flows . . . 64
3.4.6 Synthetic data on hyperbolic space . . . 66

3.5 Conclusion . . . 66
4 Score-based modelling on constrained domains . . . 67
4.1 Introduction . . . 67
4.1.1 Technical setting . . . 69
4.2 Log-barrier diffusion models . . . 69
4.2.1 Hessian metrics . . . 71
4.2.2 Forward noising process . . . 71
4.2.3 Time-reversal . . . 71
4.2.4 Sampling . . . 72
4.2.5 Extending the technical results . . . 72
4.3 Reflected diffusion models . . . 72
4.3.1 Skorokhod problem . . . 73
4.3.2 Time-reversal . . . 74
4.3.3 Sampling . . . 75
4.3.4 Likelihood evaluation . . . 75
4.3.5 Extending the technical results . . . 76
4.4 Training and parametrising score functions for constrained diffusions . . . 77
4.5 Related work for sampling on constrained manifolds . . . 78
4.6 Experimental results . . . 78
4.6.1 Method characterisation on convex polytopes . . . 78
4.6.2 Modelling robotic arms under force and velocity constraints 79
4.6.3 Modelling protein loops with anchored endpoints . . . 81
4.7 Conclusion . . . 84
5 Efficient score-based modelling on constrained domains . . . 87
5.1 Introduction . . . 87
5.2 Diffusion models for constrained manifolds via Metropolis sampling . . . 88
5.2.1 Practical limitations of existing constrained diffusion models . . . 89
5.2.2 Metropolis approximation of reflected Brownian motion . . . 90
5.2.3 Convergence of the Metropolis sampler to the reflected Brownian motion . . . 90
5.3 Related work on approximations of reflected stochastic differential equations . . . 94
5.4 Experimental results . . . 95
5.4.1 Synthetic distributions on simple polytopes . . . 96
5.4.2 Modelling proteins and robotic arms under convex constraints . . . 96
5.4.3 Modelling geospatial data within non-convex country borders . . . 98
5.5 Conclusion . . . 100
6 Score-based modelling on geometric structures on Riemannian manifolds . . . 101
6.1 Introduction . . . 101
6.2 Geometric neural diffusion processes . . . 103
6.2.1 Euclidean tensor fields . . . 103

6.2.2 Continuous diffusion on function spaces 104
6.2.3 Invariant and equivariant geometric neural diffusion processes 106
6.2.4 Extension to tensor fields on manifolds 109
6.2.5 Extension to manifold-valued functions 111
6.3 Langevin conditional sampling of diffusion model 111
6.3.1 Noising schemes 113
6.3.2 Likelihood evaluation 114
6.4 Related work 115
6.5 Experimental results 116
6.5.1 1D regression over stationary scalar fields 116
6.5.2 Regression over Gaussian process vector fields 118
6.5.3 Global tropical cyclone trajectory prediction 119
6.6 Discussion 120
7 Conclusion 123
7.1 Contributions and applications 123
7.2 Concurrent work 123
7.3 Future work 124
7.3.1 Beyond diffusions: flow and bridge matching on manifolds 124
7.3.2 Energy-based training 125
A Manifolds 161
A.1 Topological Spaces 161
A.1.1 Continuity & Convergence 163
A.1.2 Goldilocks topologies 164
A.1.3 Constructing topologies 165
A.2 Topological Manifolds 169
A.2.1 Bundles and Sections 170
A.2.2 Charts and atlases 171
A.2.3 Partitions of unity 172
A.3 Differentiable and Smooth Manifolds 173
A.4 Tensors 174
A.4.1 Algebraic structures 174
A.4.2 Vector Spaces 175
A.4.3 Tensors 177
A.4.4 Bases 177
A.4.5 Einstein summation convention 179
A.4.6 Symmetric tensors 180
A.4.7 Alternating tensors 180
A.5 Tangent, cotangent, and tensor spaces 182
A.5.1 Tangent spaces and derivations 182
A.5.2 The differential 185
A.5.3 Submersions, immersions, and embeddings 185
A.5.4 Local coordinates 186
A.5.5 Cotangent and tensor spaces 189
A.6 Tensor Bundles 190
A.6.1 Tangent bundles and vector fields 190
A.6.2 Tensor bundles and fields 194
A.6.3 Lie derivatives 196

A.7 Lie Groups . . . 199
A.7.1 Homogeneous spaces . . . 200
A.7.2 Group actions and equivariant maps . . . 200
A.7.3 Semi-direct products . . . 201
A.7.4 Representations . . . 201
A.8 Riemannian Metrics . . . 201
A.8.1 Riemannian metrics . . . 202
A.8.2 Tangent-cotangent isomorphism . . . 203
A.8.3 Pullback metrics and submanifolds . . . 204
A.9 Differential Forms, orientations, and integration . . . 206
A.9.1 Exterior derivatives . . . 207
A.9.2 Orientations . . . 208
A.9.3 Integration on manifolds . . . 209
A.9.4 Measures induced by volume forms . . . 213
A.10 Connections, geodesics, and parallel transport . . . 213
A.10.1 Connections on the tangent bundle and tensor bundles . . . 214
A.10.2 Covariant derivatives . . . 215
A.10.3 Vector fields on curves . . . 216
A.10.4 Geodesics . . . 216
A.10.5 The exponential and logarithm maps . . . 217
A.10.6 Parallel transport . . . 217
A.10.7 The Levi-Civita connection . . . 218
A.10.8 The Laplace-Beltrami operator . . . 219
A.10.9 Eigenspectrum of the Laplace-Beltrami operator . . . 220
B Stochastic Differential Equations . . . 223
B.1 A review of measure theory based probability . . . 223
B.2 Stochastic processes . . . 228
B.3 Stochastic differential equations . . . 231
B.4 The Riemann-Stieltjes integral . . . 233
B.5 The Ito and Stratonovich integrals . . . 234
B.5.1 Advantages to different stochastic integrals . . . 236
B.5.2 Connections between different stochastic integrals . . . 238
B.5.3 Multi-dimension stochastic differential equations . . . 238
B.5.4 Solutions to stochastic differential equations . . . 239
B.5.5 Weak solutions . . . 240
B.6 The Kolmogorov equations . . . 241
B.6.1 The Kolmogorov backward equation . . . 241
B.6.2 The generator of a diffusion . . . 243
B.6.3 The Kolmogorov forward equation . . . 244
B.7 Langevin Dynamics . . . 246
B.8 Time-reversal results . . . 247
B.9 Connections to ordinary differential equations . . . 248
B.10 Approximate sampling from stochastic differential equations . . . 249
C Score-based modelling on Riemannian manifolds . . . 251
C.1 Notation . . . 251
C.2 Preliminaries on stochastic Riemannian geometry . . . 252
C.2.1 Tensor field, metric, connection and transport . . . 252
C.2.2 Stochastic Differential Equations on manifolds . . . 254

C.2.3 Brownian motion on manifolds . . . 255
C.3 Time-reversal formula: extension to Riemannian manifolds . . . 259
C.3.1 Informal derivation . . . 260
C.3.2 Proof of Theorem 3.6 . . . 262
C.4 Convergence of RSGM . . . 268
C.4.1 Main results . . . 268
C.4.2 Discretization bounds for GRW . . . 271
C.5 Proof of the implicit score-matching loss on manifolds, proposition 3.7 . . . 278
C.6 Experimental details . . . 279
C.6.1 Sphere . . . 280
C.6.2 Torus . . . 281
C.6.3 Special Orthogonal group . . . 282
D Score-based modelling on constrained domains . . . 285
D.1 Brownian motion in local coordinates . . . 285
D.2 Geodesic Brownian Motion . . . 286
D.3 Reflected Brownian Motion and Skorokhod problems . . . 288
D.4 Implicit Score Matching Loss . . . 290
D.4.1 Proof of ISM . . . 290
D.4.2 Importance of scaling function . . . 290
D.5 Proof of proposition 4.5 . . . 291
D.6 Time-reversal for reflected Brownian motion . . . 292
D.7 Experimental details . . . 294
D.7.1 Synthetic data on polytopes . . . 295
D.7.2 Constrained SPD matrices for robotic arms modelling . . . 297
D.7.3 Conformational modelling of polypeptide backbones under anchor point constraints . . . 297
E Efficient score-based modelling on constrained domains . . . 299
E.1 Convergence to the reflected process . . . 299
E.1.1 Properties of the tubular neighbourhood . . . 300
E.1.2 Technical lemmas . . . 303
E.1.3 Lower bound on the inside probability and control of moments of order two and higher . . . 305
E.1.4 Properties of large drift terms . . . 307
E.1.5 Convergence on compact sets . . . 309
E.1.6 Convergence on the boundary . . . 310
E.1.7 Submartingale problem and weak solution . . . 312
E.1.8 Extension to the Metropolis process . . . 313
E.2 Modelling geospatial data within non-convex boundaries . . . 315
E.2.1 Derivation of bounded geospatial dataset . . . 315
E.2.2 Point-in-spherical-polytope algorithms . . . 315
E.3 Supplementary Experimental Results . . . 316
E.3.1 Evaluating log-barrier models . . . 316
E.3.2 Experimental details . . . 317
E.3.3 Constrained Configurational Modelling of Robotic Arms . . . 318
E.3.4 Conformational Modelling of Protein Backbones . . . 320

F Score-based modelling on geometric structures on Riemannian manifolds . . . 323
F.1 Organisation of appendices . . . 323
F.2 Ornstein Uhlenbeck on function space . . . 323
F.2.1 Multivariate Ornstein-Uhlenbeck process . . . 323
F.2.2 Conditional score . . . 325
F.2.3 Several score parametrisations . . . 325
F.2.4 Exact (marginal) score in Gaussian setting . . . 327
F.2.5 Langevin dynamics . . . 327
F.2.6 Likelihood evaluation . . . 327
F.3 Invariant neural diffusion processes . . . 327
F.3.1 E(n)-equivariant kernels . . . 327
F.3.2 Proof of proposition 6.2 . . . 328
F.3.3 Equivariant posterior maps . . . 329
F.4 Experimental details . . . 335
F.4.1 Regression 1d . . . 335
F.4.2 Gaussian process vector fields . . . 337
F.4.3 Tropical cyclone trajectory prediction . . . 339

1 | Introduction

The last decade has seen an explosion in the research and application of machine learning algorithms to academic and real-world problems. The majority of this progress has been driven by the rise of *deep learning* as an effective tool to learn to approximate very complex functions from data via gradient-based optimisation. Although the roots of deep learning are much older, the beginning of the modern rise can be traced to the period 2010-2012. Existing algorithms were made significantly cheaper to run by reducing hardware costs and investing in engineering efforts to have them run on *graphics processing units* (GPUs).

Deep learning

Graphics processing units

This enabled a significant rise in the number of training iterations and data points that could be used in deep learning models. The increase in computation power, along with the flexibility of deep learning models, allowed them to surpass other classes of models. For example, Gaussian processes (Williams and Rasmussen, 2006) and support vector machines (Boser et al., 1992), for which training and inference scale poorly with the number of data points or there is a restriction on the class of learnable functions. The essay "The bitter lesson" (Sutton, 2019) reflects this evolution; general purpose methods with little built-in knowledge, given enough data and training resources, outperform more methods in which we try to encode partial human knowledge into the modelling solution.

A strong contender for the first moment of significant success in deep learning is the Alexnet convolutional architecture (Krizhevsky et al., 2012) outperforming classical methods in the ImageNet competition (Russakovsky et al., 2015). There are many other notable successes from recent years. Game-playing agents have matched or beaten top-ranked humans in 49 Atari games (Mnih et al., 2015), Go (Silver et al., 2016), Starcraft (Vinyals et al., 2019), DOTA 2 (Berner et al., 2019), and in a Diplomacy tournament against human players ((FAIR)† et al., 2022). Machine learning algorithms have made breakthrough progress on scientific problems, such the protein structure prediction (Abramson et al., 2024), and have started to show progress on algorithm finding (Fawzi et al., 2022) and controlling plasma instabilities in fusion reactors (Degrave et al., 2022). Machine learning models are producing realistic images and videos (Saharia et al., 2022a; Betker et al., 2023), and answering complex questions about text, images, and videos from users (OpenAI et al., 2024; Gemini-Team et al., 2024).

Early efforts in deep learning focused on for the most part *classification* and *regression* problems, where we attempt to associate with some input x some output y. Examples include sorting images into categories or predicting the price of a house given information about its location and features. In these tasks, from a

Classification

Regression

probabilistic point of view, we are aiming to learn the distribution $p(y|X = x)$ from a dataset of examples drawn randomly from $p(x, y)$.

Typically, x is of very high dimension, such as an image, and y, is a single scalar drawn from a categorical distribution, such as an image label. In addition, the distribution of $p(y|X = x)$ often either takes a single point value or is tightly clustered around one. We do not then need particularly complex methods to specify the output of the model we learn since the function we need to learn is relatively simple.

Deep generative modelling
Density estimation

Beyond the supervised tasks of classification and regression, *deep generative modelling* has become a mainstay of deep learning. The core task of generative modelling is similar to that of *density estimation.* Both focus on fitting a mode to samples we have obtained from a distribution p, $\{x_i\}_{i=1}^N$, $x_i \sim p$, and both aim to generate new samples $x \sim p(x)$, or to extract information about this distribution. Where they differ is that deep generative models use the flexibility and generalisation capabilities of deep learning to learn complex distributions. This leads to models of complex data, such as images or structured text, that have remained out of reach for non-deep learning approaches.

Deep generative modelling often aims to be able to sample the reverse of the distribution learnt in the classification or regression setting, $p(x|Y = y)$. While this seems like a semantic point, it flips the complexity of the problem. It takes us from predicting a low-dimension summary of a high-dimension object to predicting a distribution of high-dimension objects given a low-dimension summary. This allows generative modelling to answer some incredibly powerful questions. We can learn to generate images given a class or text prompt, to generate structured text that answers a question provided by a user, or to generate the three-dimensional structure of a protein given the DNA sequence that specifies it. Cornerstone approaches in deep generative modelling include generative adversarial networks (Goodfellow et al., 2020), variational auto-encoders (Kingma and Welling, 2013; Rezende et al., 2014), normalising flows (Rezende and Mohamed, 2015; Tabak and Vanden-Eijnden, 2010; Tabak and Turner, 2013), and continuous normalising flows (Chen et al., 2018).

Curse of dimensionality

The high dimension of the prediction targets and complexity and diversity of outputs corresponding to the same input to a model however makes the task of generative modelling significantly more difficult. This is a common feature of many problems, known as the *curse of dimensionality* (Bellman, 1958, page ix). As the dimension of the problem scales, even with large amounts of data available to the practitioner, the volume of the problem space grows exponentially. These large datasets can end up still being sparse in the observation space of the problem.

Similar to many classical approaches, one key way to surpass this curse is the exploitation of the *structure* of the problem at hand. Typically, the space of possible solutions to a problem can be reduced significantly by placing conditions on our solutions derived from prior knowledge about the problem.

Firstly we may consider certain symmetries that a problem possesses. In many image recognition problems, the output of a model should not change if we translate an image. Respecting this structure results in convolutional neural networks

(Fukushima, 1980; Waibel et al., 1989; LeCun et al., 1989). In time series problems temporal shift invariance leads to recurrent neural networks (Hochreiter and Schmidhuber, 1997; Rumelhart et al., 1986). On set-like inputs, for which the order in which elements are presented is not important, we arrive at transformers (Vaswani et al., 2017), graph neural networks (Sperduti, 1993) and other order invariant architectures.

Secondly, standard deep learning algorithms assume the data they process live in a standard Euclidean space. In many applications, however, particularly in physical and social science domains, data lives on non-flat manifolds, or discrete structures such as graphs. Examples include meteorological data living on the surface of the earth, a sphere, and social interaction networks being well represented by graphs. Extending algorithms to data that live in these non-standard spaces is key to learning well on these problems. While these two goals might seem initially disparate, they have significant overlap on many common spaces, particularly homogenous spaces of Lie groups, such as the sphere, torus, projective spaces and hyperbolic spaces.

Geometric deep learning

The field of *geometric deep learning* (Bronstein et al., 2017; Bronstein et al., 2021) aims to tackle these two problems. Firstly it aims to unify these seemingly disparate architectures that each encode some property of the data into a single framework, understanding the invariances in terms of group theory, and providing tools for building new architectures with the desired invariances. Secondly it considers these architectures in the light of non-Euclidean data, either on continuous spaces that have differentiable and geometric structures, but are not "flat" like Euclidean space, or on discrete spaces such as grids and graphs.

The intersection of generative modelling and geometric machine learning has been of significant interest as it provides a tool to tackle complex problems in various scientific domains, for example, protein structure prediction (Jumper et al., 2021), lattice quantum chromodynamics modelling (Abbott et al., 2024) and material discovery (Merchant et al., 2023). All of these methods exploit the symmetry of the physical problem at hand to drastically reduce the space of solutions, making the problem tractable.

Diffusion models
Score-based generative models

Recently the development of *diffusion models* or *score-based generative models* (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020) have shown to be an extraordinarily powerful class of deep generative models. They are however limited in their application to Euclidean data as the diffusion processes used in their construction are inherently Euclidean objects.

The overarching aim of this thesis is to generalise score-based models from Euclidean space to more general continuous geometries, namely Riemannian manifolds, Riemannian manifolds with boundaries, and more complex geometric objects on Riemannian manifolds. This allows for the application of these models to non-Euclidean data and for the incorporation of symmetries into the modelled densities.

The rest of this thesis is in the following format:

Chapter 2 introduces the topics of geometric machine learning, with a focus on learning on manifolds, and deep generative modelling, with a focus on score-based

generative models. We provide both a brief technical introduction to the areas and a literature review of the state of the field relevant to this thesis. In addition, appendices A and B contain more in-depth background material on the topics of Riemannian manifolds and stochastic differential equations, aimed to make the thesis self-contained and give a gentle introduction to the mathematics.

Chapter 3 investigates how one can define a score-based generative model for distributions supported on a Riemannian manifold. There are several motivating problems for this extension. For example, modelling data that lives on Lie groups, such as the poses of objects, states of quantum systems, and internal angles of molecules and proteins covers a wide range of scientific problems. On more general manifolds, modelling geospatial data such as the distribution of natural disasters and weather patterns on the surface of the earth, and modelling tree-like hierarchies in hyperbolic space are other interesting applications. The core contributions of this chapter are extending the theoretical framework of score-based models to Riemannian manifolds by proving new score-matching loss and time-reversal results extended to the case of manifolds. We also provide practical choices for the various components of score-based models on manifolds and demonstrate the applicability of these methods on a series of problems.

Chapter 4 investigates how to define a score-based model supported on a Riemannian manifold which is not geodesically complete. This lack of completeness comes in one of two ways. Firstly properties of the solution desired may place constraints on the support of the density modelled. Alternatively, the manifold that the data lives on may come with a natural boundary, forming a manifold with boundary. Motiving modelling problems include modelling data that lives inside a bounded range, such as pixel intensity, modelling data that lives on the simplex, the manifold containing categorical distributions, modelling paths of robotic movements with constraints on the speed of their movement, and modelling rigid link loops, such as cyclic peptides. The core contributions of this chapter are the proposal of two new noising processes for score-based models that remain contained inside a boundary specified, namely reflected Brownian motion and log-barrier diffusion processes. We prove new score-matching results and time reversal results applicable to these new noising processes, and provide practical methods for learning score functions near boundaries. Finally, we demonstrate the capabilities of these methods on a series of problems.

Chapter 5 revisits the reflected Brownian motion scheme of chapter 4. The discretisation schemes typically used in discretising reflected Brownian motion are computationally expensive, and limit the speed of diffusion model training. In this chapter we propose a new discretisation scheme for reflected Brownian motion, a Metropolised scheme, that significantly speeds up the discrete approximation of reflected stochastic differential equations. We prove that this Metropolised scheme is a valid discretisation of the reflected Brownian motion and demonstrate its practical performance relative to the regular discretisation of reflected stochastic differential equations when used in diffusion models on constrained domains.

Chapter 6 goes beyond modelling densities on manifolds to model densities over more complex geometric structures on manifolds. We tackle two main classes. Firstly, spaces of sections of vector bundles on manifolds, which includes smooth

functions on manifolds and vector fields on manifolds. Motivating problems in this case include modelling weather data, such as temperature, pressure, and wind velocity. Secondly, we model distributions on paths on manifolds. Motivating problems in this case include modelling the trajectory of storms over the surface of the earth, and modelling path planning for robots with rotating joints. Core contributions of this chapter are providing a framework to model these complex data types, and introducing methods to incorporate symmetries of distributions into the models. We also propose a new Langevin sampling based conditional sampling scheme for score-based models to provide an alternative existing methods. Unlike most other conditional sampling schemes for score-based models, we show that this method in a limit will sample the correct conditional distribution of the score-based model. We finally demonstrate the empirical performance of the method on a number of tasks.

Publications in this thesis The work in this thesis is based on the following publications:

V. De Bortoli*, E. Mathieu*, M. J. Hutchinson*, J. Thornton, Y. W. Teh, and A. Doucet. Riemannian score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2022.

N. Fishman, L. Klarner, V. De Bortoli, E. Mathieu, and M. J. Hutchinson. Diffusion Models for Constrained Domains. *Transactions on Machine Learning Research*, 2023.

N. Fishman, L. Klarner, E. Mathieu, M. J. Hutchinson, and V. De Bortoli. Metropolis Sampling for Constrained Diffusion Models. In *Advances in Neural Information Processing Systems*, 2023.

E. Mathieu*, V. Dutordoir*, M. J. Hutchinson*, V. De Bortoli, Y. W. Teh, and R. Turner. Geometric neural diffusion processes. In *Advances in Neural Information Processing Systems*, 2024.

* indicates equal contribution.

Additional publications In addition to these I have worked on a number of other projects during the course of DPhil studies in the space of geometrics machine learning and generative modelling. These projects resulted in the following publications:

M. J. Hutchinson*, C. L. Lan*, S. Zaidi*, E. Dupont, Y. W. Teh, and H. Kim. LieTransformer: Equivariant Self-Attention for Lie Groups. In *International Conference on Machine Learning*, 2021.

P. Holderrieth*, M. J. Hutchinson*, and Y. W. Teh. Equivariant Learning of Stochastic Fields: Gaussian Processes and Steerable Conditional Neural Processes. In *International Conference on Machine Learning*, 2021.

M. J. Hutchinson, M. Reisser, and C. Louizos. Federated Functional Variational Inference. In *Bayesian Deep Learning workshop, Advances in Neural Information Processing Systems*, 2021.

M. J. Hutchinson, A. Terenin, V. Borovitskiy, S. Takao, Y. Whye Teh, and M. P. Deisenroth. Vector-valued Gaussian Processes on Riemannian Manifolds via Gauge Equivariant Projected Kernels. In *Advances in Neural Information Processing Systems*, 2021.

J. Thornton, M. J. Hutchinson, E. Mathieu, V. De Bortoli, Y. W. Teh, and A. Doucet. Riemannian Diffusion Schrödinger Bridge. In 2022. Cited on page 125.

A. Phillips*, T. Seror*, M. J. Hutchinson, V. De Bortoli, A. Doucet, and E. Mathieu. Spectral Diffusion Processes. In *Advances in Neural Information Processing Systems*, 2022.

A. Phillips, H.-D. Dau, M. J. Hutchinson, V. De Bortoli, G. Deligiannidis, and A. Doucet. Particle Denoising Diffusion Sampler. In *Forty-first International Conference on Machine Learning*, 2024. Cited on page 126.

V. De Bortoli, M. J. Hutchinson, P. Wirnsberger, and A. Doucet. Target Score Matching. *arXiv preprint arXiv:2402.08667*, 2024. Cited on page 126.

* indicates equal contribution.

Work on COVID-19 During the period of the COVID-19 pandemic I also spent time working on various statistical modelling efforts related to the pandemic. This work was completed as part of the Imperial COVID-19 Response team, the Royal Society DELVE Initiative action team, and the Turing-RSS Health Data Lab. Publications from work in this period that I was involved in are:

T. A. Mellan, H. H. Hoeltgebaum, S. Mishra, C. Whittaker, R. P. Schnekenberg, A. Gandy, H. J. T. Unwin, M. A. Vollmer, H. Coupland, I. Hawryluk, et al. Subnational analysis of the COVID-19 epidemic in Brazil, 2020.

M. Monod, A. Blenkinsop, X. Xi, D. Hebert, S. Bershan, V. C. Bradley, Y. Chen, H. Coupland, S. Filippi, J. Ish-Horowicz, and others (14th/29). Age groups that sustain resurging COVID-19 epidemics in the United States. *Science*, 2021.

M. A. Vollmer, S. Mishra, H. J. T. Unwin, A. Gandy, T. A. Mellan, V. Bradley, H. Zhu, H. Coupland, I. Hawryluk, M. J. Hutchinson, et al. Report 20: Using mobility to estimate the transmission intensity of COVID-19 in Italy: a subnational analysis with future scenarios. *MedRxiv*, 2020.

H. J. T. Unwin, S. Mishra, V. C. Bradley, A. Gandy, T. A. Mellan, H. Coupland, J. Ish-Horowicz, M. A. Vollmer, C. Whittaker, S. L. Filippi, et al. State-level tracking of COVID-19 in the United States. *Nature communications*, 2020.

B. He*, S. Zaidi*, B. Elesedy*, M. J. Hutchinson*, A. Paleyes, G. Harling, A. Johnson, and Y. W. Teh. Technical Document 3: Effectiveness and Resource Requirements of Test, Trace and Isolate Strategies. *Royal Society DELVE Initiative, Report on Test, Trace, Isolate Systems*, 2020.

Y. W. Teh, A. Bhoopchand*, P. Diggle*, B. Elesedy*, B. He*, M. J. Hutchinson*, U. Paquet*, J. Read*, N. Tomasev*, and S. Zaidi*. Efficient Bayesian Inference of Instantaneous Re-production Numbers at Fine Spatial Scales, with an Application to Mapping and Nowcasting the Covid-19 Epidemic in British Local Authorities. *Joural of the Royal Statistical Society: Series A*, 2021.

* indicates equal contribution.

2 | Background

In this chapter we provide an overview of two areas of machine learning. In section 2.1 we focus on geometric machine learning and in particular its application to Riemannian manifolds (section 2.1.3). In section 2.2 we focus on deep generative modelling, with a focus on diffusion models (section 2.2.1).

2.1. Geometric machine learning

Geometric machine learning is a field of machine learning concerned with producing models that preserve the geometric structure of the space they are applied to.

This is achieved by carefully considering the space in question, the structure it has that we wish to preserve, and the symmetries that lead to this. This approach has two main benefits.

Firstly, it greatly reduces the problem space of possible solutions. Typical deep learning models can learn to approximate any function, proved by ubiquitous *universal approximation theorem*s (Cybenko, 1989; Leshno et al., 1993; Ismailov, 2023; Maiorov and Pinkus, 1999). However, the *curse of dimensionality* (Bellman, 1958) means that as we expand the dimension of a problem space, the number of data samples needed to accurately estimate a function in reasonable classes, such as Lipschitz functions or Sobelov spaces, grows exponentially fast with the dimension of the space (Tsybakov, 2009).

Universal approximation theorem

Curse of dimensionality

This poses a problem for the generalisation of machine learning models in high dimension to data not previously seen. *Function regularisation* schemes can help significantly with regularising the learnt function and improving generalisation, either explicitly via a penalty, e.g. L_2 weight decay (Krogh and Hertz, 1991), or implicitly via the mechanics of stochastic gradient descent (Vardi and Shamir, 2021), but this is not always enough. Enforcing that models respect the structure of a given space greatly reduces the space of solutions, which in turn reduces the quantity of data needed to achieve a given accuracy.

Function regularisation

For example, consider a simple *image recognition* task. We aim to recognise if a cat or a dog is in the image. If we apply a typical *multilayer perceptron* to this image, it will need to explicitly learn what a cat or dog is in every position in the image. This is because it treats all pixels uniquely, preserving all the structure of the input. On the other extreme, we can apply a network that is invariant to the position of pixels, treating their intensities as a histogram, such as a *transformer*. This erases

Image recognition

Multilayer perceptron

Transformer

all spatial structure in the input. This will fail to learn however as it cannot identify how pixels are arranged with respect to one another, preventing the recognition of relative multi-pixel structures such as ears and eyes. In the middle of these is the standard *convolutional neural network* (LeCun et al., 1989). This respects that images can be translated up and down, left and right, and remain the same, but that if pixels are rearranged more than this, then it is a different image. This strikes the right balance between preserving all structure of the input and preserving none. As a result, a convolutional neural network takes significantly fewer parameters and computation to produce the same or better performance than a multilayer perceptron, and significantly outperforms a set-based model on this task.

Convolutional neural network

The second result of this consideration is the application of machine learning models to *non-Euclidean data*. Typical models make an implicit assumption that the data they are dealing with are vectors living in Euclidean space. This is convenient as Euclidean vector spaces come with many mathematical tools. There exists a significant number of problems that do not natively lie in Euclidean space, however. By considering the inputs more carefully, we can expand the input spaces of models to a wide variety of *geometric spaces*, such as sets, graphs, grids, and manifolds. This opens up the application of machine learning to a wide range of new problem domains, for example applications to social networks, molecular science, quantum physics, and climate sciences.

Non-Euclidean data

Geometric spaces

Overall, geometric machine learning aims to view the space that particular data lives on through its geometric structure and the symmetries they possess. It allows us to unify common machine learning models such as convolutional neural networks, recurrent neural networks, graph neural networks, and many others into a single theory. This also allows us to quickly design networks for new types of geometric spaces and transfer tools between them.

2.1.1. Geometric priors. Let us consider a set X on which we are going to place a machine learning model. X may come with additional information.

Geometric structure

Set of symmetries

Firstly it may come with a *geometric structure*, a set of relations telling us about how the points in the set are related to each other. Secondly it may come with a *set of symmetries*, telling us how the elements can be permuted while preserving the geometric structure the space.

Functions between geometric spaces

Two modelling scenarios arise in geometric deep learning. Firstly, for geometric spaces X and Y each with their own geometric structure we may be interested in learning *functions between geometric spaces* of the form $f : X \to Y$. For example, consider a two dimension grid, $X = \mathbb{Z}_n \times \mathbb{Z}_n$. The grid has a natural structure of points and their neighbours. The structure is preserved if we translate the grid up/down or left/right, if we rotate the grid, or if we mirror the grid through an axis. The grid structure therefore encodes translation, rotation and reflection symmetries, and tells us that models on a grid should preserve such symmetries.

Signal on a space

Secondly, we may be interested in learning functions of signals on X. A *signal on a space* is a function that assigns a value in some space Z to every point in the space X, a function $f : X \to Z$. Continuing the grid example, a signal $f : \mathbb{Z}_n \times \mathbb{Z}_n \to [0, 1]^3$ assigning three intensity values to each point on the grid makes up an image. We

can denote the space of signals on X taking values in Z as $\mathcal{X}(X, Z)$. A model that takes as input an image should respect the symmetries of the underlying grid.

Symmetry groups

Symmetries of spaces are encoded naturally in mathematical group. The basic principle of a group of symmetries is that symmetries can be composed, i.e. that application of one symmetry after another is itself a symmetry, that symmetries can be undone, i.e. they are operations that preserve the information of the object, and that there is a symmetry that leaves the object unchanged.

Formally this can be summarised as

Group

DEFINITION 2.1. *A group $(G, \cdot)$ is a set G along with an operation $\cdot : G \times G \to G$ such that it has the following properties:*

ASSOCIATIVITY *For element $f, g, h, \in G$, $f \cdot (g \cdot h) = (f \cdot g) \cdot h$.*

IDENTITY *There exists an element $e \in G$ such that for every element $g \in G$, $e \cdot g = g \cdot e = g$.*

INVERTIBILITY *There exists for every element $g \in G$ and element $g' \in G$ such that $g \cdot g' = g' \cdot g = e$. This is typically denoted as G^{-1}.*

CLOSURE *For every $g, h \in G$, $g \cdot h \in G$.*

Groups can have significant extra structure in addition to this basic definition. For example some group actions are commutative, $g \cdot h = h \cdot g$, known as *Abelian group*, some are continuous functions, known as *topological groups*, and some have Riemannian manifold structure, known as *Lie groups*. Often the group operation, like the multiplication symbol, is omitted, and instead operations are written as $g \cdot h = gh$.

Abelian group
Topological groups
Lie groups

For example, the *cyclic group* of order n is defined by element $\{0, ..., n\}$, has group composition $\cdot : (g, h) \mapsto g + h \mod n$, and has identity element 0. Taking a product of this group, $C_n \times C_n = C_n^2$ defines a group of cycles in two directions. The *dihedral group* D_n represents cycles with an additional flip operation. It has elements $(1, 1), ..., (1, n), (-1, 0), ..., (-1, n)$ where the first index indicates if there is a flip or not, and the second a cycle of order n. It has group composition $\cdot : (a, g), (b, h) \mapsto (a * b, g + h \mod n)$ and identity element $(1, 0)$.

Cyclic group
Dihedral group

Group actions and representations

Groups in the abstract are often not the objects we want to study. Instead, we are interested in their action on some space. The *(left) action of a group on a space X* is a function $\triangleright : G \times X \to X$ that is compatible with the group operation, that is

(left) action of a group on a space

$$g \triangleright (h \triangleright x) = (g \cdot h) \triangleright x \tag{2.1}$$

for $g, h \in G$, $x \in X$.

To continue the grid example, the group C_n^2 defines the translation symmetries of a two dimension grid of size n, wrapping around at the edges. The group D_4 represents the rotation and reflection symmetry of the grid. Their *semi-direct*

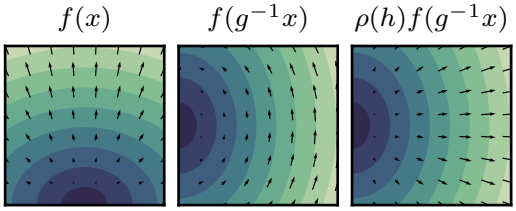


Figure 2.1. Illustration of a vector field $f : \mathbb{Z}^2 \to \mathbb{R}^2$ being transformed by a group element $g = (r, h) \in C_n^2 \rtimes D_4$.

Semi-direct product

product[1] $C_n^2 \rtimes D_4$ represents the combined translation, rotation, and reflection symmetries of the grid.

Representation of a group

A *representation of a group* is a specific type of action of a group acting linearly on a vector space. A (real) representation of a group is a map $\rho : G \to \mathrm{GL}(n)$, a map from the group into the space of invertible $n \times n$ matrices. This map must obey $\rho(g \cdot h) = \rho(g)\rho(h)$ where the right-hand side is standard matrix multiplication. Their action on $\mathbb{R}^n$ and the group composition operations are given by the usual matrix multiplication. There also exist complex representations, action on complex vector spaces. Representations that map to orthogonal or unitary matrices are called orthogonal or unitary representations. An introductory text to linear representations is Serre (1977).

Actions on signals

Representations are useful in building *actions on signals*. Let us consider a signal on X taking values in $\mathbb{R}^n$, $f \in \mathcal{X}(X, \mathbb{R}^n)$. If we have a group action G on X and a representation of X acting on $\mathbb{R}^n$, then an action of the group G on this signal can be defined by

$$\left(g \triangleright_f f\right)(x) = \rho(g) f(g^{-1} \triangleright_X x) \tag{2.2}$$

for a representation of G, ρ. This states that if we transform our signal by g, then this involves mapping the underlying space X under the group action, and also mapping the observations under the same group action. This whole operation can be seen as a single group action, and so again we can view the space of signals just as a geometric space in its own right.

Trivial representation

To again continue the image grid example, the action of C_n^2 will and translate an image, and the action of D_4 will rotate and flip it. The representation action on the colour channels of the image however is simply the identity action, leaving the channels unchanged. This representation is the *trivial representation*, $\rho(g) \to I$.

Faithful representation

To see the action of a more complex representation, consider different data attached at each point on the grid, for example a vector representing the wind direction at each point. In this case, we might expect that when we rotate the grid, the wind vector will rotate with it. This is given by the *faithful representation*, a representation that is a bijection, of D_4. If we let $h = (a, r) \in D_4$, with a representing the flip component and $r \in [0, 2\pi]$ the rotation component, then a faithful representation is given by

[1] A type of group product that means any element of the group can be decomposed as successive actions of one group and then the next

$$\rho(h) = \begin{bmatrix} \cos r & -a \sin r \\ \sin r & a \cos r \end{bmatrix}. \tag{2.3}$$

If we place the action of this representation on the wind vectors attached to each point, then for $g = (t, h) \in C_n^2 \rtimes D_4$, the action on the wind field is given by

$$(g \triangleright f)(x) = \rho(h) f(g^{-1} \triangleright x). \tag{2.4}$$

This is illustrated in figure 2.1.

Symmetric functions

Two classes of function are of interest to us in geometric deep learning. Consider a space X with a symmetry group G, and a space Y with the same symmetry group. X and Y could be underlying geometric spaces, or spaces of signals.

Invariant functions are functions $f : X \to Y$ such that Invariant functions

$$f(g \triangleright_X x) = f(x). \tag{2.5}$$

In the image analysis case an example of this is assigning a label to an image regarding its contents, for example "cat" or "dog". The label does not change if the image is translated, rotated or reflected.

Equivariant functions are functions $f : X \to Y$ such that Equivariant functions

$$f(g \triangleright_X x) = g \triangleright_Y f(x). \tag{2.6}$$

An example of this might be locating where in the image a cat or a dog is. If we translate or rotate the image, the location of the cat or dog will also translate or rotate.

The goal of many geometric deep learning architectures is to produce models that respect either the invariance or equivalence of the problem at hand. In doing so, we automatically allow models to generalise to examples that they have seen before, but under a different group symmetry.

This is exactly how we can use geometric principles to reduce the size of the hypothesis space of functions that we try to learn from. If we consider a hypothesis space of functions $\mathcal{H}$ between some spaces X and Y with a shared symmetry group G, then the invariant and equivariant hypothesis spaces are

$$\mathcal{H}_{\text{invariant}} = \{f : X \to Y \ : \ f(g \triangleright_X x) = f(x) \ \forall x \in X, \ g \in G\} \subset \mathcal{H}, \tag{2.7}$$

and

$$\mathcal{H}_{\text{equivariant}} = \{f : X \to Y \ : \ f(g \triangleright_X x) = g \triangleright_Y f(x) \ \forall x \in X, \ g \in G\} \subset \mathcal{H}. \tag{2.8}$$

Depending on the original hypothesis space of functions and the group symmetry G, this can be a significant reduction in the hypothesis space. There exist various learning theory (Vapnik, 2000; Hastie et al., 2009) results that guarantee that under the assumption that the problem at hand does possess the specified symmetry, a model that respects this symmetry exhibit guaranteed better generalisation performance (Elesedy and Zaidi, 2021; Elesedy, 2021; Lyle et al., 2020; Behboodi et al., 2022).

2.1.2. Some types of geometric priors. Here we briefly discuss a number of common geometric priors used in geometric deep learning. We discuss the case of sets, graphs, grids, homogenous spaces and manifolds.

Sets

The simplest of geometric architectures concerns sets. Typically, we only consider finite sets. With no additional geometric structure, any network processing sets should not take into account the order in which elements are presented. In building networks for sets then, we need architectures with *permutation equivariant* layers.

Permutation equivariant

DeepSets (Zaheer et al., 2017) and transformers (Vaswani et al., 2017) are the two most common architectures for learning on sets.

Graphs

Graphs bring one additional element over sets. Graphs consist of a set of nodes $\mathcal{V}$ and a set of edges connecting some nodes together, $\mathcal{E}$. Similar to sets, the goal is to create an architecture that is invariant to permutations of the inputs. Each node and each edge may carry some information in the way of features.

Given the size of graphs typically processed, architectures aim not to use a global permutation equivariant function as it would be too expensive. Instead, they use successive layers that apply local operations to nodes, their neighbours, and the edges connecting them. These layers, combined with non-linearities and sometimes pooling operations, form the backbone of graph neural networks.

Convolutional graph neural networks

Convolutional graph neural networks apply a convolution to the features of the local neighbourhood to produce updated features. Examples of this include Kipf and Welling (2016), Defferrard et al. (2016), and Wu et al. (2019).

Attentive graph neural networks

Attentive graph neural networks apply attention between nodes and their neighbours' features. Examples of this include Velickovic et al. (2017), Monti et al. (2017), and Zhang et al. (2018a).

Message-passing graph neural networks

Message-passing graph neural networks construct messages between pairs of nodes from the features of each node, and aggregate them to produce updated features. Example of this include Gilmer et al. (2017) and Battaglia et al. (2018). These kinds of networks are useful for modelling structures social networks and other interactions.

Geometric graphs

In addition to basic graphs, a common extension is *geometric graphs*. These are graphs that in addition to node features and edges include a spatial position at each node. In addition to being equivariant to permutations of the nodes, we require that networks processing this kind of graph are also equivariant to global transforms of the locations of each node. Examples of approaches to this type of structure include Satorras et al. (2021) and Fuchs et al. (2020), taking message passing and attention-based approaches respectively. Such models are useful for example in models of physics for modelling systems of particles interacting with one another.

Grids

Grids are a regular lattice of points in n dimensions, where each point is connected to its neighbours. This can be considered a restriction of a graph which has a specific structure.

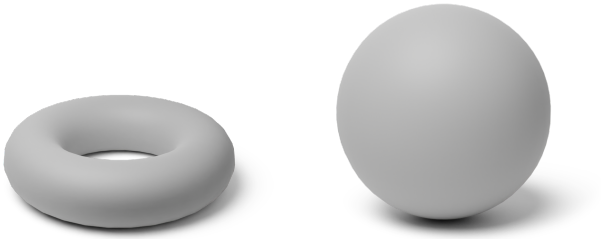


Figure 2.2. Common manifolds: The torus, T^2, and sphere, $\mathcal{S}^2$.

Most commonly in machine learning we deal with one dimension grids, representing *time series* of points, and two dimension grids representing *images*. Higher dimension grids are also possible to represent three dimension structures and more.

Time series
Images

The primary natural group action on grids is translation. In each dimension of the grid, we can translate the coordinates, moving each point to the point next to it. This additional structure allows us to require *translation* equivariance of maps on grids, and this greatly reduces the space of linear operators to that of *discrete convolutions*.

Translation
Discrete convolutions

In one dimension, recurrent neural networks exploit the grid structure to process each element in the grid in turn (Elman, 1990; Jordan, 1997), including extensions such as long-short term memory networks (Hochreiter and Schmidhuber, 1997; Rumelhart et al., 1986) and gated recurrent units (Cho et al., 2014).

This structure is also exploited in convolutional neural networks (Fukushima, 1980; Waibel et al., 1989; LeCun et al., 1989) and residual networks (He et al., 2016), along with a multitude of other works.

Beyond the translation symmetry of grids (Cohen and Welling, 2016; Cohen and Welling, 2017) extend the capabilities of two dimension convolutional neural networks to be equivariant to 90^o rotations and reflections of the grid as well.

2.1.3. Manifolds.

Topological manifolds

We now give a more in-depth look at Riemannian manifolds and geometric machine learning on them, as they are the main setting of study in this thesis. In this section we aim to give an introduction to the core concepts without delving into the details. A more complete introduction can be found in appendix A.

A useful introductory series of textbooks on the subject are Lee (2010), Lee (2013), and Lee (2006) and the lecture notes Schuller (2013). Other useful texts are Kolár et al. (2013) and Thorne et al. (2000). Bronstein et al. (2021) also contains a machine learning oriented introduction to Riemannian manifolds.

Like other geometric priors manifolds are sets equipped with extra geometric structure. There are three main levels that we discuss. Topological manifolds admit topological properties such as continuous functions. Smooth manifolds in addition gain differentiable structure, allowing for the construction of differentiation and

linear structures on the manifold. Finally, the addition of a Riemannian metric adds usual geometric notation like distance, angles, and straight lines.

Topological space

Open sets

Homeomorphisms

A *topological space* is a tuple $(X, \mathcal{O}_X)$. This is a set X along with a *topology* $\mathcal{O}_X$, a collection of subsets of X closed under pairwise intersections and arbitrary unions, called *open sets*. Topologies allow for the definition of notions of continuity, convergence of sequences and locality, amongst others. Maps between topological spaces that preserve the topology are *homeomorphisms*.

Topological manifold

A d-dimension *topological manifold* is a topological space with additional conditions. It must *locally* be homeomorphic to d-dimension Euclidean space, and it must be *second-countable* and *Hausdorff*. These properties make the topology of the manifold non-pathological. Amongst other things, the Hausdorff property implies that the limits of sequences are unique, and the second countable property makes the notion of topological convergence and sequential convergence coincide.

Charts

Atlas

Transition functions

On topological manifolds we may define *charts*. Charts are tuples (U, ϕ), where U is an open set in $\mathcal{O}_X$ and $\phi : U \to \mathbb{R}^d$ is a map from U into an open subset of $\mathbb{R}^n$. A collection of charts such that the open sets cover the manifold is an *atlas* of X, $\mathcal{A}_X$. An atlas that contains all possible charts is a *maximal atlas*. *Transition functions* are defined on the region where for two charts (U, ϕ), (V, ψ) overlap. The transition function $\phi \circ \psi^{-1} : \mathbb{R}^d \to \mathbb{R}^d$ is a map between open sets of Euclidean space.

Local coordinates

Coordinate-free

Unlike Euclidean space, manifolds do not come with a canonical choice of coordinates. We can induce a system of *local coordinates* through a chart, but since there are no canonical charts, there are no canonical coordinates on manifolds. For this reason when working with manifolds we must define our various operations in a *coordinate-free* manner. This is a large part of the difficulty presented when working with manifolds.

Manifold with boundary

A d-dimension topological *manifold with boundary* is defined the same as a topological manifold, except that it may be locally homeomorphic to the Euclidean half-plane. This notion encodes the idea that a manifold may not continue forever in a given direction. The points on the manifold homeomorphic to the edge of the half-plane are known as the *boundary* of the manifold, denoted ∂X, and themselves form a manifold.

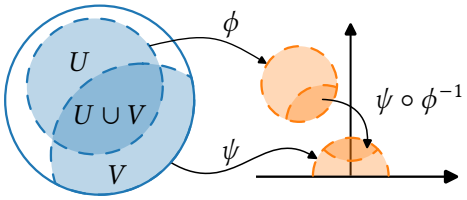


Figure 2.3. Two charts of the atlas $\mathcal{A}_X$ of X, the unit disk, a manifold with boundary. (U, ϕ), a chart mapping into a subset of $\mathbb{R}^n$, and (V, ψ), a chart mapping into a subset of the upper half-space $\mathbb{H}^2$.

Figure 2.4. The projection of the earth onto a flat map is the prototypical example of a chart. While the Mercator projection it looks like the chart covers the whole globe, but mathematically it is in fact missing the poles and a line connecting them as we need the set to be an open one.

Smooth manifolds

A *smooth manifold* or *differentiable manifold* $(X, \mathcal{O}_X, \mathcal{A}_X)$ is a topological manifold $(X, \mathcal{O}_X)$ along with an atlas for which all the transition maps between charts are infinitely differentiable. This allows us to define notions of differentiation on the manifold. Maps between manifolds that preserve the smooth structure of the manifold are called *diffeomorphisms*.

Smooth manifold

Diffeomorphisms

If we can restrict the atlas of a smooth manifold to charts where the determinant of the Jacobian is positive for all pairs we call this an *orientable* manifold as it is allows for a consistent choice of handedness of local bases on the manifold.

Orientable

On smooth manifolds we can define *smooth functions* $f \in C^\infty(X)$ from the manifold to the real numbers as functions such that for any chart in the atlas, (U, ϕ), the composition function $f \circ \phi^{-1} : \mathbb{R}^d \to \mathbb{R}$ is infinitely differentiable.

Smooth functions

The *tangent space* of a point, p, on a manifold, denoted T_pX, can be thought of as a local linear approximation of the manifold, like placing a plane tangent to the surface of the manifold. Formally it is made up of the vector space of *directional derivatives* of smooth functions at the point on the manifold. Tangent spaces are isomorphic to $\mathbb{R}^d$. Unlike $\mathbb{R}^d$ however, it does not come with a canonical choice of basis, again necessitating working in a coordinate-free manner.

Tangent space

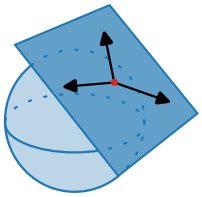


Figure 2.5. The tangent space of the point $\bullet \in \mathcal{S}_2$ with example tangent vectors.

The *cotangent space* at a point on a manifold, T_p^*X is the vector space dual of the tangent space. By taking arbitrary tensor products of the tangent and cotangent spaces we can construct *tensor spaces* at each point on the manifold. These are denoted T_q^pX for the tensor product of p copies of tangent space and q copies of the cotangent space.

Cotangent space

Tensor spaces

The *differential* of a smooth function between manifolds $F : X \to Y$ is a linear map

Differential

Pushforward map

$F_* : T_pX \to T_{F(p)}Y$ that connects the tangent space T_pX to $T_{F(p)}Y$. The *pushforward map* denoted F_*, is another name for the differential. The differential of a smooth function $f \in C^\infty(X)$ is a notion similar to the gradient of function in standard multivariate calculus, except it maps a function to a cotangent vector, not a tangent vector as one might expect.

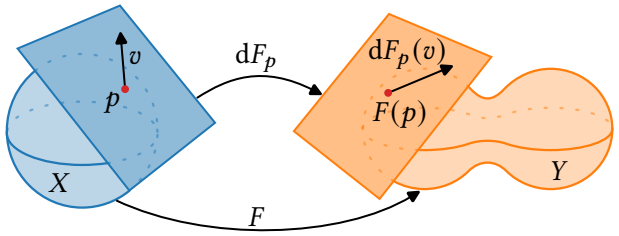


Figure 2.6. A smooth map between smooth manifolds $F : X \to Y$ induces a natural map between the tangent spaces of p and $F(p)$, the differential $\mathrm{d}F_p$.

Tangent bundle
Vector field

The union of all tangent spaces on the manifold, $TX = \bigcup_{p\in X}\{(p, v) : v \in T_pX\}$, combined with the projection operation $\mathrm{proj}_X : (x, v) \mapsto x$, and the manifold itself gives rise to the *tangent bundle* of the manifold, (TX, proj_X, X). This is itself a smooth manifold with a structure inherited from X. A *vector field* is a map $\sigma : X \to TX$ such that the composition $\mathrm{proj}_X \circ \sigma = \mathrm{id}_X$. In other words, it assigns a tangent vector at a given point on the manifold to every point. We denote the space of vector fields as $\Gamma(TX)$. While each tangent space on the manifold is isomorphic to $\mathbb{R}^d$, a vector field cannot be expressed as a function $\sigma : X \to \mathbb{R}^d$. This is down to issues with choosing a globally smooth basis on $\Gamma(TX)$.

Fibre bundle
Fibre
Section
Vector bundle

The tangent bundle is an example of a *fibre bundle*. This is a triple (E, proj_B, B), consisting of a *total space* manifold E, a *base space* manifold B and a projection $\mathrm{proj}_B : E \to B$. The *fibre* of a point $b \in B$ is defined as $\mathrm{proj}_B^{-1}(b)$. The fibres of all points in B are isomorphic to one another. A vector field is an example of a *section* of a fibre bundle, a map that assigns every point in B a point in its respective fibre. Stronger than this, it is an example of a *vector bundle*, a fibre bundle where the fibre is a vector space.

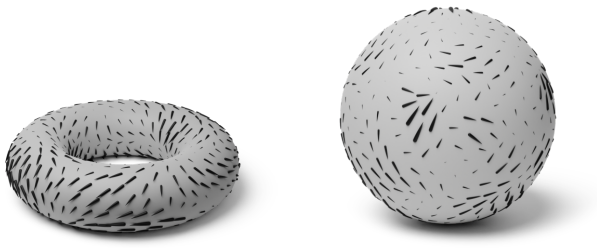


Figure 2.7. Examples of a vector fields on the torus and sphere.

Cotangent bundle
Tensor bundles
Covector fields
(p, q)-tensor fields

From the cotangent and general tensor spaces at a point on a manifold we can construct other vector bundles, namely the *cotangent bundle* and arbitrary *tensor bundles*. Sections of these are known as *covector fields* and (p, q)*-tensor fields* respectively. The spaces of these fields are denoted $\Gamma(T^*X)$ and $\Gamma\left(T^p_qX\right)$ respectively. The study of more advanced topics on manifolds involve the study of tensor fields.

As the vector space dual of tangent spaces, the elements of the cotangent space define linear maps over elements of the tangent space. Covector fields similarly define linear maps over vector fields, where the linear map is defined pointwise on the manifold. This operation is denoted by the *interior product*, $\lrcorner : \Gamma(TX) \times \Gamma(T^*X) \to C^\infty(X)$. Tensor fields are multilinear maps over combinations of vector and covector fields, and we can generalise the interior product to this case.

Interior product

Differential forms are $(0, k)$-tensor fields that at each point on the manifold are *alternating tensors*, also known as k-forms. As it turns out differential forms are particularly suitable to analysis. The *exterior derivative* is a natural notion of differentiation that maps k-forms to $(k+1)$-forms. On smooth functions, which are 0-forms, it agrees with the differential.

Differential forms

Exterior derivative

A nowhere vanishing d-form on a d-dimension manifold is known as a *volume form*. The existence of these is guaranteed on an orientable manifolds. These allow us to define a notion of the volume spanned by a set of vectors. Using this we can define a consistent notion of *integration of smooth functions* on manifolds. The choice of volume form is not unique on smooth manifolds however, and different choices of volume form lead to different integrals of the same smooth function. Via the Riesz-Markov-Kakutani representation theorem we can connect a volume form to a measure-theoretic Radon measure on the manifold, a measure compatible with the topology of the manifold, known as the *volume measure*. This generalises the *Lebesgue measure* to manifolds.

Volume form

Integration of smooth functions

Volume measure

Lebesgue measure

It should be noted it is possible to define a notation of integration on non-orientable manifolds also via *densities* on the manifold.

Densities

Riemannian manifolds

Riemannian manifolds add one additional point of structure over a smooth manifold. They are a quadruple $(X, \mathcal{O}_X, \mathcal{A}_X, g)$, containing a topology, smooth atlas and a *Riemannian metric*.

Intuitively a *Riemannian metric* is a smooth choice of inner product on each tangent space. More precisely is a smooth non-vanishing symmetric $(0, 2)$-tensor field. These are guaranteed to exist for all smooth manifolds. The metric also defines a *Riemannian norm* $\|v\|_g = \sqrt{g(v, v)}$ on the tangent space.

Riemannian metric

Riemannian norm

A choice of Riemannian metric allows us to canonicalise a series of choices we had to make on smooth manifolds. Such choices as called *intrinsic choices*. For example, it allows us to define a unique volume form on the manifold, vol_g, giving a unique notion of integration on Riemannian manifolds.

Intrinsic choices

This unique choice of integration also lets us define an *inner product on smooth functions* on the manifold, $\langle f, h \rangle_g = \int_X f(p)h(p) \, \text{d}\,\text{vol}_g$ for $f, h \in C^\infty(X)$. Using this inner product we can define the Hilbert space $L^2(X)$ of *L^2-integrable functions* on the manifold as the completion of $C^\infty(X)$ under the inner product. We can extend this to vector bundles on the manifold, for example for vector fields by $\langle V, W \rangle = \int_X g_p(V(p), W(p)) \, \text{d}\,\text{vol}_g$ for $V, W \in \Gamma(TX)$.

Inner product on smooth functions

L^2-integrable functions

By the Riesz-Fréchet representation theorem, which states that every dual vector can be expressed as an inner product with a vector, we can define a *tangent-*

Tangent-cotangent isomorphism

cotangent isomorphism. We identify the covector $\omega \in T_p^*X$ with the vector $v \in T_pX$ if

$$\omega(w) = g(v, w) \quad \forall\ w \in T_pX. \tag{2.9}$$

Riemannian gradient

Using this we can define the *Riemannian gradient* of a smooth function $f \in C^\infty(X)$ as the unique map $\nabla_g : C^\infty(X) \to \Gamma(TX)$ such that

$$\mathrm{d}f \lrcorner V = g\big(\nabla_g f, V\big) \quad \forall\ V \in \Gamma(TX). \tag{2.10}$$

This is also denoted grad_g. This definition of the gradient is now more similar to the usual multivariate calculus case as it results in a vector field, not a covector field.

Divergence of a vector

We can also define the *divergence of a vector* on a Riemannian manifold via its volume form. It is the unique map $\operatorname{div}_g : \Gamma(TX) \to C^\infty(X)$ such that

$$\mathrm{d}(V \lrcorner \operatorname{vol}_g) = \operatorname{div}_g \operatorname{vol}_g V \quad \forall\ V \in \Gamma(TX). \tag{2.11}$$

The divergence is also notated as ∇_g^*, as it is the dual of the Riemannian gradient in the sense that for $V \in \Gamma(TX)$ and $f \in C^\infty(X)$

$$\big\langle \nabla_g f, V \big\rangle_g = \big\langle f, \nabla_g^* V \big\rangle_g. \tag{2.12}$$

Stoke's theorem

This definition of divergence lets us state a version of *Stoke's theorem* on a manifold with boundary.

$$\int_X \operatorname{div}_g V \,\mathrm{d}\operatorname{vol}_g = \int_{\partial X} g(V, N) \,\mathrm{d}\operatorname{vol}_g, \tag{2.13}$$

where N is a vector field normal to the boundary. If the manifold has no boundary then the right-hand side is zero. This has the same intuitive meaning as in the Euclidean case. It tells us that the integral of the divergence of a vector field inside a volume is equal to the flux passing through the boundary of the volume.

Connection

To define further geometric concepts on Riemannian manifolds we need another construction, a *connection.* A connection is a map $\nabla : \Gamma(TX) \times \Gamma(TX) \to \Gamma(TX)$ that is linear, distributive in the second argument and compatible with scaling. It defines a differentiation of one vector field with respect to another. As such, the operation $\nabla(V, W)$ is usually written $\nabla_V W$.

Levi-Civita connection

On smooth manifolds we can choose many connections. With a metric however we can make a canonical choice, the *Levi-Civita connection*, that is compatible with the metric. We will assume to always be using this connection.

Curve on a manifold

Velocity of a curve

A *curve on a manifold* is a map $\gamma : I \to X$, where I is a connected subset of $\mathbb{R}$. The *velocity of a curve* is given by $\gamma' = \frac{\mathrm{d}\gamma}{\mathrm{d}t}$. We say that a vector field is *parallel* to a curve if $\nabla_{\gamma'} V = 0$.

Geodesic

Length of a curve

A *geodesic* generalises the notion of straight lines to manifolds. We say that a curve is a geodesic if its velocity is parallel to the curve, that is $\nabla_{\gamma'}\gamma' = 0$. We define the *length of a curve* by the integral $L(\gamma) = \int_I \|\gamma'(t)\|_g \,\mathrm{d}t$. A geodesic passing through two points on a manifold is provably the shortest path between those two points.

Geodesics distance

This allows us to define a *geodesics distance* on the manifold as

$$\mathrm{d}_g(p, q) = \arg\min_\gamma L(\gamma),\ s.t.\ \gamma(0) = p,\ \gamma(1) = q,\ \gamma : [0, 1] \to \mathcal{M}. \tag{2.14}$$

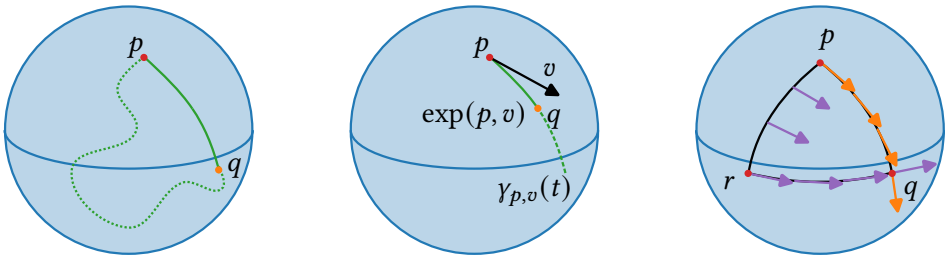


Figure 2.8. *Left*: Geodesic (solid) and non-geodesic (dotted) paths from p to q. *Middle:* The exponential map of $v \in T_pX$ at $p \in X$ along the geodesic $\gamma_{p,v}$ to q. *Right:* The parallel transport of a tangent vector is path dependant. Transporting a vector at p to q directly (orange) vs via r (purple) results in a different vector at r.

The *exponential map* at a point is a map $\exp : TX \to X$ that maps $(p, v) \mapsto \gamma_{p,v}(1)$ where $\gamma_{p,v}$ is the unique geodesic with $\gamma(0) = p$ and $\gamma'(0) = v$. The inverse of the exponential map is the *logarithm map*, $\log : X \times X \to TX$ that maps $(p, q) \mapsto v$ such that $\exp(p, v) = q$. Exponential map Logarithm map

In order to be able to compare tangent vectors in different tangent spaces we need a notion of *parallel transport*. For a path γ and $(p, v) \in TX$, there exists a unique vector field $V \in \Gamma(TX)$ such that $\nabla_{\gamma'} V = 0$ and $V(p) = v$. We say that the parallel transport of v along γ is simply $V(\gamma(t))$. Note that parallel transport is entirely dependent on the path, and as such there is no canonical way to compare tangent vectors in different tangent spaces. The norm of a vector is preserved by parallel transport. If we parallel transport two tangent vectors along the same path, the Riemannian inner product of them is preserved along the path also. Parallel transport

On manifolds with boundary we can define a *boundary intersection* operation that defines when a geodesic will hit the boundary of the manifold. For a point $p \in X$ and a unit tangent vector $v \in T_pX$ we can define this as the function $i : X \times TX \to \mathbb{R}^+$ such that Boundary intersection

$$i(p, \boldsymbol{v}) = \arg\min_{t \in \mathbb{R}^+} \exp_p(t\boldsymbol{v}) \in \partial X. \tag{2.15}$$

Finally, can define a notion of *reflection of tangent vectors* in a plane. For a point $p \in X$, a vector to be reflected $v \in T_pX$, and a unit vector normal to the plane of reflection $n \in T_pX$, we can define the reflection of v by the plane of n as Reflection of tangent vectors

$$v - 2g(v, n)n. \tag{2.16}$$

This is particularly useful for reflecting tangent vectors on the boundary of a manifold.

The Laplace-Beltrami operator

Fourier analysis and its analogues are key components in creating geometry-aware models on standard geometric structures such as graphs, grids, and homogenous spaces. As such, we want to develop a similar analysis on Riemannian manifolds.

The *Laplace-Beltrami* operator is a key differential operator on Riemannian manifolds allowing this, and generalises the standard Euclidean Laplace operator. Laplace-Beltrami

It is defined as the operator

$$\Delta_g : C^\infty(X) \to C^\infty(X) \qquad \Delta_g = \nabla_g^* \nabla_g = \operatorname{div}_g \operatorname{grad}_g . \tag{2.17}$$

This can also be defined via its expression in local coordinates, or via the trace of the second covariant derivative on the manifold. This operator is coordinate-independent, and importantly commutes with isomorphisms of Riemannian manifolds, making it ideal for use in models respecting such isometries (Canzani, 2013, p. 46).

The operator self-adjoint in the sense that for $f \in C^\infty(X)$

$$\left\langle \nabla_g f, \nabla_g f \right\rangle_g = \left\langle \Delta_g f, f \right\rangle_g = \left\langle f, \Delta_g f \right\rangle_g. \tag{2.18}$$

This can be extended further to L^2-integrable functions on the manifold, and also to L^2-integrable vector fields on the manifold where $-\Delta_g$ is a self-adjoint unbounded positive-definite operator (Strichartz, 1983, theorem 2.4).

Since these are Hilbert spaces, we can leverage spectral theory to decompose functions in the Hilbert space over eigenvalues and eigenfunctions of $-\Delta_g$. In particular for compact manifolds, the spectrum of $-\Delta_g$ is countable.

Strun-Liouville decomposition

THEOREM 2.2. *Let (X, g) be a compact manifold. Then there exists an orthonormal basis of $L^2(X)$, $f_n \in L^2(X)$, $n \in \mathbb{Z}^+$, such that*

$$-\Delta_g f_n = \lambda_n \tag{2.19}$$

and that the basis is ordered in the sense that $0 \leq \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$, and $\lambda_n \to \infty$ as $n \to \infty$.

Proof. Chavel (1984) ∎

This means we can write any function $f \in L^2(X)$ on a compact manifold as the infinite sum $f = \sum_{i=1}^{\infty} \langle f, f_n \rangle_g f_n$. This basis expansion can be truncated to a finite length $f_N \approx \sum_{i=1}^{N} \langle f, f_n \rangle_g f_n$. The error of this basis truncation can be bounded.

Bounded error of the Sturm-Liouville decomposition

THEOREM 2.3. *Let (X, g) be a compact Riemannian manifold, and $(f_n, \lambda_n)_{n=0}^{\infty}$ be the eigen-spectrum of $-\Delta_g$. Let $f \in L^2(X)$. Let f_N be the truncated expansion of f. Then*

$$\|f - f_N\|_g^2 \leq \frac{\left\|\nabla_g f\right\|^2}{\lambda_{N+1}}. \tag{2.20}$$

Proof. (Aflalo and Kimmel, 2013) ∎

It can also be shown that this approximation is optimal in the sense that no other basis will obtain a better error bound than the truncated Sturm-Liouville approximation (Aflalo et al., 2015).

Riemannian isometries

For two Riemannian manifolds (X, g^x) and (Y, g^y) and a map between them $F : X \to Y$, we define the *pullback metric* $F^* g_y$, a metric on X point wise via the pullback

Pullback metric

$$F^* g^y_p(v, w) = g^y_{F(p)}(\mathrm{d}F_p(v), \mathrm{d}F_p(w)) \quad \forall\ (p, v), (p, w) \in TX. \tag{2.21}$$

We call a map F such that $g^x = F^* g^y$ a *Riemannian isometry*.

Riemannian isometry

Such isometries and particularly the group of isometries of a manifold, denoted $\mathrm{Isom}_g(X)$, onto itself, are key symmetries of study when producing geometry-aware deep learning models.

The Nash embedding theorem

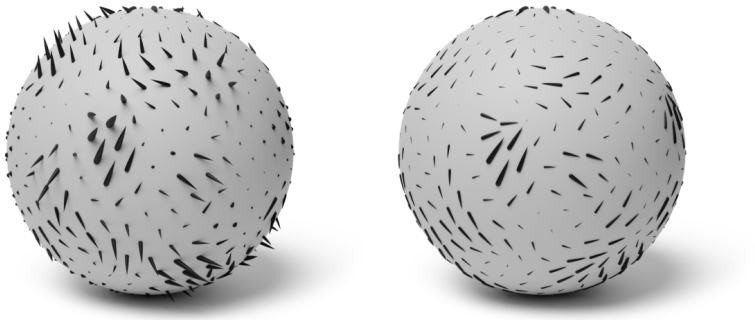


Figure 2.9. A sphere isometrically embedded into Euclidean space. An ambient Euclidean vector field (left) projected onto the surface of the sphere (right).

THEOREM 2.4. *Given a d-dimension smooth Riemannian manifold (X, g), there exists a smooth embedding $F : X \to \mathbb{R}^n$ such that*

Nash embedding theorems

$$g = F^* g_{\mathbb{R}^n}, \tag{2.22}$$

where $g_{\mathbb{R}^n}$ is the standard Euclidean metric, for $n \leq \frac{d(d+1)(3d+11)}{2}$.

Proof. (Nash, 1956) ∎

The Nash embedding theorem is a cornerstone result of Riemannian geometry. It allows us to take any manifold and embed it into Euclidean space while preserving the geometric properties of the manifold.

From a technical point of view embeddings are avoided as there is no canonical choice. However, it is useful in a few ways in machine learning.

Firstly for smoothly parametrising machine learning models, it provides a global way to smoothly specify the inputs to a function on a manifold via Euclidean coordinates, real numbers, and therefore a way to parametrise globally smooth functions on the manifold in a computer.

It also allows for an easy way to specify smooth vector fields on manifolds. We cannot express a vector field on a manifold as a function $V : X \to \mathbb{R}^d$, see section 3.2.8.

Finally, it lets us compute differential operators on parametrised functions more easily using automatic differentiation. For example, for a vector field on an isometrically embedded manifold the ambient Euclidean divergence coincides with the intrinsic divergence (Rozen et al., 2021).

2.1.4. Geometric deep learning on manifolds. There are several distinct challenges when constructing deep learning models that act on manifolds.

Firstly, due to the lack of a canonical choice of coordinates for the manifold and for the tangent space, it is key to develop models that are independent of a specific parametrisation of the manifold and tangent space.

Secondly, we may want a method to respect the group of self-isomorphisms of a Riemannian manifold. These symmetries are typically not global symmetries for most manifolds, and a large class of real-world manifolds do not have any symmetries.

Lastly, and most rarely, we may wish to develop methods that are invariant to *diffeomorphisms* of manifolds, ignoring the metric structure and looking simply at the shape of the manifold.

Most of the architectures developed for manifolds deal with signals supported on the manifold, scalar functions on the manifold or sections of tensor bundles. These approaches are based on manifold-aware convolutions. Weiler et al. (2023) treats this topic in great detail, with a comprehensive development of these methods, and a deep coverage of the literature on this topic.

There is also some literature on methods for learning functions of the manifold itself. Methods have been developed for functions of Lie groups (Huang et al., 2017), Grassmann manifolds (Huang et al., 2018; Zhang et al., 2018b), and general manifolds (Fang et al., 2023). There is also significant literature for learning functions on meshes, discretised two dimension manifolds, for the purpose of representing textures (Koestler et al., 2022; Baatz et al., 2022; Chibane and Pons-Moll, 2020; Oechsle et al., 2019) and displacement maps (Yifan et al., 2022).

2.2. Deep generative modelling

Kernel density estimation

Deep generative modelling aims to combine the flexibility of deep learning with classical density estimation. The aim is to overcome the rigidity of classical methods such as *kernel density estimation* (Rosenblatt, 1956; Parzen, 1962), and to overcome the curse of dimensionality present in these estimation problems.

The most common application of deep generative models include images (Saharia et al., 2022a; Ramesh et al., 2021), text (Team et al., 2023; OpenAI et al., 2024), and video (Brooks et al., 2024), and these application spaces are beginning to see widespread deployment in the real world.

In addition, there has been recently strong interest in applying deep generative modelling to scientific problems. Applications have arisen in molecular physics (Zheng et al., 2024; Klein et al., 2024), particle physics (Louppe et al., 2019; Paganini et al., 2018), quantum mechanics (Cheng et al., 2024; Pfau et al., 2020; Pfau et al.,

2024), statistical mechanics (Noé et al., 2019), protein modelling (Abramson et al., 2024; Krishna et al., 2024).

At a high level, the deep generative modelling task has access to a series of samples $\{x_i\}_{i=1}^n$, $x_i \sim p_{\text{data}}$ from some ground truth data distribution, to which we have no access. We aim to learn a parametrised model, q_θ, which is close to p_{data} in some sense. This closeness is measured by a statistical divergence, $\mathcal{D}$, and the goal is therefore

$$\arg\min_{\theta\in\Theta} \mathcal{D}(p_{\text{data}} \,||\, q_\theta). \tag{2.23}$$

Since we do not have access to p_{data}, this is approximated by the *empirical distribution*

Empirical distribution

$$p_{\text{data}} \approx \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}, \tag{2.24}$$

replacing the data distribution by a mix of delta distributions of the data we have available. Typical statistical divergences include the Kullback-Leibler divergence (Kullback and Leibler, 1951) (which can be seen as performing maximum likelihood inference), f-divergences (Rényi, 1961), the Wasserstein distance (Kantorovich, 1960), maximum mean discrepancy (Gretton et al., 2012) and integral probability metrics (Müller, 1997).

Typically, we want to be able to generate new samples from this distribution and evaluate the likelihood of samples under our model. In addition, we may want to perform *conditional generation* from our models. If our data distribution is a joint distribution $p_{\text{data}}(x, y)$ where x is some high dimension observation such as an image, and y is a low dimension label or summary of x, such as a text description, then we may want our models to be able to generate samples from a distribution $q_\theta(x|Y = y)$, trying to match as closely as possible $p_{\text{data}}(x|Y = y)$.

Conditional generation

There exist a wide variety of deep generative models. Almost all of them exploit a strategy of taking random noise in some base space, $z \sim p_{\text{prior}}$ and transforming it through some function to produce samples.

These can be categorised into a series of main methods. *Generative adversarial networks* (Goodfellow et al., 2014) train two models, one generator G to create samples from noise, and one discriminator D which tries to distinguish between model samples and real data samples. The two models are jointly trained by minimising

Generative adversarial networks

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_{\text{prior}}}[\log(1 - D(G(z)))], \tag{2.25}$$

where p_z is typically Gaussian noise. These models are known to be unstable to train (Wiatrak et al., 2019) however, and do not allow for likelihood evaluations.

Variational autoencoders (Kingma and Welling, 2013; Rezende et al., 2014) apply a probabilistic perspective to autoencoders (Rumelhart et al., 1986). The data is assumed to be the marginal of some joint distribution $p_{data}(x) \int p_{\text{joint}}(x, z)\, \mathrm{d}z$, where again for a sample z_i is a low dimension summary of x_i. A prior is placed over the latent space, $p_{\text{prior}}(z)$. The decoder, $q_\theta(x \mid z)$ attempts to model the conditional distribution $p_{decoder}(x \mid z)$, such that $q_\theta(x \mid z)p_{\text{prior}}(z) = p_{\text{joint}}(x, z)$. The encoder, $q_\phi(z \mid x)$ aims to match the reverse distribution, $p_{\text{encoder}}(z|x)$ so that

Variational autoencoders

$q_\phi(z \mid x)p_{\text{data}}(x) = p_{\text{joint}}(x, z)$. One observed drawback of variational autoencoders is that they tend to produce fuzzy or blurry samples, particularly in the image domain.

Normalising flows

Normalising flows (Rezende and Mohamed, 2015; Tabak and Vanden-Eijnden, 2010; Tabak and Turner, 2013) exploit the fact that for an invertible function f and a distribution $p_{\text{prior}}(z)$, the density of the pushforward of $p_{\text{prior}}(z)$ through f, $p_{\text{model}} = f^* p_{\text{prior}}$, is given by

$$\log p_{\text{model}}(x) = \log p_{\text{prior}}(f^{-1}(x)) + \log\left|\det \frac{\mathrm{d}f^{-1}(x)}{\mathrm{d}x}\right|. \tag{2.26}$$

f is designed to be parametrisable, invertible, and have a fast to compute the log-det-Jacobian. With this we can compute the likelihood of the data under this model quickly and therefore train it via maximum likelihood. The requirement that f be invertible places significant difficulty on designing flexible f. To add complexity to the model, we can stack multiple transformations on top of one another. However, this can be increasingly expensive and normalising flows can suffer from expressiveness issues and poor training dynamics.

Continuous normalising flows

Continuous normalising flows (Chen et al., 2018; Grathwohl et al., 2019) are an extension of normalising flows that replace the transformation f with a neural ordinary differential equation (Chen et al., 2018), an ordinary differential equation with a vector field parametrised by a neural network. This creates a model that continuously deforms noise into samples from the model. They are also trained by maximum likelihood, and requires the computation of the change in likelihood induced by an ordinary differential equation. This step can be expensive, and so continuous normalising flows can struggle to scale to higher dimensions.

Diffusion models
Score-based models

Diffusion models or *score-based models* (Song and Ermon, 2019; Sohl-Dickstein et al., 2015; Ho et al., 2020) are a class of model that define a forward noising process that corrupts samples from p_{data} under a specific nosing process. These models then learn to reverse this noising process through a denoising process in order to generate samples. We will explore these models in more detail in section 2.2.1. Diffusion models have proven to be very effective. They are fast to train, produce sharp, high quality samples, and place very few restrictions on the types of neural networks that can be used to produce them. One drawback is that computing likelihoods under these models is either impossible, or very slow and non-exact.

2.2.1. Score-Based Generative Modelling.

In this section we give a presentation of the score-based generative modelling framework.

There are two perspectives on diffusion models; discrete-time diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) and continuous-time diffusion models (Song et al., 2020b). These are related, and the discrete-time models can be seen as a discretisation of the continuous-time framework. As such, this presentation will focus only on the continuous-time version. Throughout we will work with a state space of dimension d, such that a data point is an element $\boldsymbol{X} \in \mathbb{R}^d$.

Stochastic differential equations

The core of diffusion models are *stochastic differential equations*. These are random variables, denoted for example $(\boldsymbol{X}_t)_{t\in[0,T]}$, that place a distribution on time-indexed

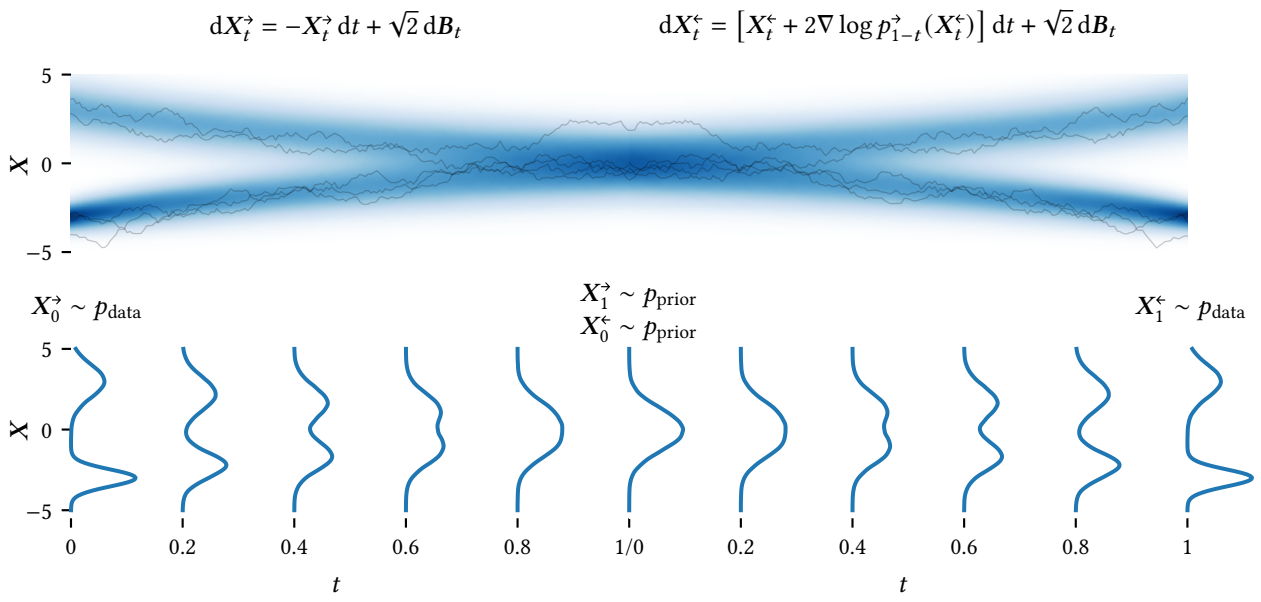


Figure 2.10. An illustrative diagram of the main concepts of a diffusion model. *Top:* Histogram plots of the path density of the evolution of the stochastic differential equations, with example trajectories. *Bottom*: Time-marginal densities, $p_t^{\rightarrow}$ and $p_t^{\leftarrow}$ of the path densities. *Left:* The forward noising stochastic differential equation. *Right:* The reverse denoising stochastic differential equation.

The forward noising process is initialised with $p_0^{\rightarrow} = p_{\text{data}} = \frac{1}{2}\mathcal{N}(3, 1^2) + \frac{1}{2}\mathcal{N}(-3, 0.5^3)$ and converges to $p_1^{\rightarrow} = p_{\text{data}} = \mathcal{N}(0, 1^2)$ via the stochastic differential equation $\mathrm{d}X_t^{\rightarrow} = -X_t^{\rightarrow}\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$ over the time interval $[0, 1]$. The reverse denoising process is initialised with $p_0^{\leftarrow} = p_1^{\rightarrow} = p_{\text{prior}}$ and converges to $p_1^{\leftarrow} = p_0^{\rightarrow} = p_{\text{data}}$ via the stochastic differential equation $\mathrm{d}X_t^{\leftarrow} = \left[X_t^{\leftarrow} + 2\nabla \log p_{1-t}^{\rightarrow}(X_t^{\leftarrow})\right]\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$ over the time interval $[0, 1]$. Note how $p_t^{\rightarrow} = p_{1-t}^{\leftarrow}$.

paths through a state space. They are driven by an instantaneous change formula, in a common form given by

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \Sigma(t, \boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{B}_t. \tag{2.27}$$

Drift term

Diffusion term

$\boldsymbol{b} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is a *drift term*, like an ordinary differential equation, and $\Sigma(t, \boldsymbol{X}_t) : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is a *diffusion term* controlling the instantaneous noise added into the stochastic differential equation, and $\boldsymbol{B}_t$ is a Brownian motion. A technical introduction to measure theory, stochastic differential equations and various tools useful in the study of score-based models can be found in appendix B.

A summary of the diffusion modelling strategy is:

Forward stochastic differential equation

1. First we evolve p_{data} under a stochastic differential equation, which we shall term the noising or *forward stochastic differential equation*. We require that this stochastic differential equation converges to a known reference distribution, p_{prior} at large time scales. This gives us a path between the data distribution and the reference distribution.

2. Using a time reversal result, we can obtain a stochastic differential equation that has the same sample paths as our forward noising stochastic differential equation, but reversed in time. Using this denoising or *reverse stochastic dif-*

Reverse stochastic differential equation

ferential equation we can create a map from p_{prior} to p_{data}, giving a generative model.

3. All the terms in the reverse stochastic differential equation are known, except for the "score" term, $\nabla \log p_t$, of the stochastic differential equation. We therefore need a method to approximate or learn this term. Fortunately, with the large *score matching* toolbox we have available to us there are several options for learning this score.

Score matching

Discrete sampling schemes

4. Using the standard toolkit of *discrete sampling schemes* for stochastic differential equation we can sample the reverse stochastic differential equation in practise.

Likelihood of a sample

5. Using the connection between a stochastic differential equation and an associated ordinary differential equations we can evaluate the *likelihood of a sample* under the model, using standard ordinary differential equation solving toolkits.

See figure 2.10 for an illustration of this summary. Let us in detail look at each of these steps.

The forward noising stochastic differential equation

The first stage of a diffusion model is to corrupt the data towards a known and tractable noise distribution, p_{prior}. The point of this step is to define a path between p_{data} and p_{prior}, with the intent later on of following this path in the reverse direction.

Langevin dynamics

Langevin dynamics (see appendix B.7) are a type of stochastic differential equation. If we have access to a density of the form $p(x) \propto \exp(-U(x))$, then the stochastic differential equation

$$\mathrm{d}\boldsymbol{X}_t = \nabla U(\boldsymbol{X})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t \tag{2.28}$$

will in the limit of infinite time converge to p, regardless of the distribution we initialise the stochastic differential equation with.

We can also rescale this stochastic differential equation by a time dependant noise factor, $\beta(t)$, and obtain the same result, giving a stochastic differential equation of the form

$$\mathrm{d}\boldsymbol{X}_t = \beta(t)\nabla U(\boldsymbol{X})\,\mathrm{d}t + \sqrt{2\beta(t)}\,\mathrm{d}\boldsymbol{B}_t. \tag{2.29}$$

This time rescaling can be useful in diffusion models for optimising the efficiency of the discretised model rollouts and focusing the majority of the learning power of the score network on important regions of the stochastic differential equation.

Langevin dynamics gives us an option for our stochastic differential equation that converges to a known distribution. Choosing the unit Gaussian as our reference distribution, $p_T(\boldsymbol{X}) \propto \exp\left(-\frac{1}{2}\boldsymbol{X}^T\boldsymbol{X}\right)$, we arrive at $U(\boldsymbol{X}) = -\boldsymbol{X}$, and thus a forward noising stochastic differential equation of the form

$$\mathrm{d}\vec{\boldsymbol{X}}_t = -\vec{\boldsymbol{X}}_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t. \tag{2.30}$$

Ornstein–Uhlenbeck process

This is the well known *Ornstein–Uhlenbeck process* which converges exponentially quickly towards the unit Gaussian. In the literature this is typically known as

the *variance preserving stochastic differential equation* or "VP-SDE". A property that will be useful later is that the transition density in the forward direction, $p_{s|t}(\boldsymbol{X}_s \mid \boldsymbol{X}_t),\ s > t$ has analytic density and sampling. For a stochastic differential equation with noising schedule $\beta(t)$ this is given by

Variance preserving stochastic differential equation

$$p_{s|t}(\boldsymbol{X}_s \mid \boldsymbol{X}_t) = \mathcal{N}\left(\boldsymbol{X}_s \;\middle|\; e^{-\int_t^s \beta(\tau)\,\mathrm{d}\tau}\boldsymbol{X}_t, \left(1 - e^{-2\int_t^s \beta(\tau)\,\mathrm{d}\tau}\right)\boldsymbol{I}\right) \tag{2.31}$$

Another other common alternative is the *variance exploding stochastic differential equation* or "VE-SDE". In this we simply apply Brownian motion as the noising stochastic differential equation,

Variance exploding stochastic differential equation

$$\mathrm{d}\boldsymbol{X}_t = \mathrm{d}\boldsymbol{B}_t. \tag{2.32}$$

The transition density of this stochastic differential equation is given by

$$p_{s|t}(\boldsymbol{X}_s \mid \boldsymbol{X}_t) = \mathcal{N}\left(\boldsymbol{X}_s \;\middle|\; \boldsymbol{X}_t, \textstyle\int_t^s \beta(\tau)\,\mathrm{d}\tau \boldsymbol{I}\right). \tag{2.33}$$

The variance of this will continue to grow with time and eventually the variance of this transition density will dominate the variance of p_{data}. The distribution

$$p_T(\boldsymbol{X}_T) = \int p_{T|0}(\boldsymbol{X}_T \mid \boldsymbol{X}_0) p(\boldsymbol{X}_0)\,\mathrm{d}\boldsymbol{X}_0 \tag{2.34}$$

will therefore be well approximated by $\mathcal{N}\left(\boldsymbol{X}_T \;\middle|\; \boldsymbol{0}, \int_t^s \beta(\tau)\,\mathrm{d}\tau \boldsymbol{I}\right)$, and we can use this as the reference distribution.

A final alternative way to choose the forward stochastic differential equation is to start with the forward transition density and work backwards to the coefficients of the stochastic differential equation, which we will call the $\alpha - \sigma$ *parametrisation*. We restrict ourselves to stochastic differential equations with transition densities of the form

$\alpha - \sigma$ parametrisation

$$p_{t|0}(\boldsymbol{X}_t|\boldsymbol{X}_0) = \mathcal{N}\left(\boldsymbol{X}_t; \alpha_t \boldsymbol{X}_0, \sigma_t^2\right). \tag{2.35}$$

Let $\alpha_t : \mathbb{R} \to \mathbb{R}$ be a differentiable mean function, with the restriction that $\alpha_0 = 1$. Let $\sigma_t : \mathbb{R} \to \mathbb{R}$ be a differentiable noise function, with the restriction that $\sigma_0 = 0$. Then the stochastic differential equation that has this forward transition density is

$$\mathrm{d}\boldsymbol{X}_t = \frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}\,\mathrm{d}t + \left[\frac{\mathrm{d}\sigma_t^2}{\mathrm{d}t} - \frac{\mathrm{d}\log\alpha_t}{\mathrm{d}t}\,\sigma_t^2\right]\mathrm{d}\mathbf{B}_t. \tag{2.36}$$

If we pick α_t and σ_t such that as $t \to T$, $\alpha_t \to 0$, then the process will converge to the distribution $\mathcal{N}\left(0, \sigma_T^2\right)$, giving us a reference distribution to work with.

The VP-SDE and VE-SDE fit into this framework, with

$$\alpha_t = \exp\left(-\int_0^t \beta(\tau)\,\mathrm{d}\tau\right) \qquad \sigma_t = 1 - \exp\left(-2\int_0^t \beta(\tau)\,\mathrm{d}\tau\right) \qquad \text{VP-SDE} \tag{2.37}$$

$$\alpha_t = 1 \qquad \sigma_t = \int_0^t \beta(\tau)\,\mathrm{d}\tau \qquad \text{VE-SDE.} \tag{2.38}$$

It also allows us to easily explore other schedules such as the cosine schedule where we set $\alpha_t = \sin\left(\frac{t\pi}{2T}\right)$, $\sigma_t = \cos\left(\frac{t\pi}{2T}\right)$, or even linear schedules where we set $\alpha_t = \frac{T-t}{T}$, $\sigma_t = \frac{t}{T}$.

The choice of forward noising stochastic differential equation can have a significant impact on performance of a diffusion model. Interesting empirical studies can be seen for example in Karras et al. (2022), Karras et al. (2024), Nichol and Dhariwal (2021), and Hoogeboom et al. (2023).

Applying a time reversal result

The crucial step to be able to follow backwards the path defined by our forward noising stochastic differential equation is the application of a time reversal result.

We prove a simple result showing that we can indeed create such a stochastic differential equation in the following, leveraging the *Fokker-Plank equation.*

The Fokker-Plank eqaution

THEOREM 2.5. *Let $(\boldsymbol{X}_t)_{t\geq 0}$ be stochastic differential equal, with initial distribution $\boldsymbol{X}_0 \sim p_0$ and equation*

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \Sigma(t, \boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{B}_t. \tag{2.39}$$

Let $p(t,x)$ denote the density of $\boldsymbol{X}_t$ at time t. Then

$$\begin{aligned} \frac{\partial p}{\partial t} &= L^* p \\ p(0,x) &= p_0(x) \end{aligned} \tag{2.40}$$

where $L^* f = -\operatorname{div}(\boldsymbol{b} f) + \frac{1}{2}\Sigma\Sigma^T \Delta f$ and Δ is the Laplace operator.

Proof. See theorem B.12. ∎

The crux of these results is that for a stochastic differential equation, we can find another stochastic differential equation with the same time-marginal distributions, but reversed in time.

Fokker-Plank time reversal

THEOREM 2.6. *Let a stochastic differential equation, $\left(\boldsymbol{X}_t^{\rightarrow}\right)_{t\in[0,T]}$, called the forward noising stochastic differential equation, be given by*

$$\mathrm{d}\boldsymbol{X}_t^{\rightarrow} = \boldsymbol{b}(\boldsymbol{X}_t^{\rightarrow}, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0^{\rightarrow} \sim p_{data} \tag{2.41}$$

with time-marginal densities $p_t^{\rightarrow}(\boldsymbol{X})$. Let another stochastic differential equation, $\left(\boldsymbol{X}_t^{\leftarrow}\right)_{t\in[0,T]}$, called the reverse-time stochastic differential equation, be given by

$$\mathrm{d}\boldsymbol{X}_t^{\leftarrow} = \left[-\boldsymbol{b}(\boldsymbol{X}_t^{\leftarrow}, t) + \sigma(t)^2 \nabla \log p_t(\boldsymbol{X}_t^{\leftarrow})\right]\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0^{\leftarrow} \sim p_T^{\rightarrow}, \tag{2.42}$$

with time-marginal densities $p_t^{\leftarrow}(\boldsymbol{X})$.

Then the time-marginals are equal but reversed in time, in the sense that $p_t^{\rightarrow} = p_{T-t}^{\leftarrow}\ \forall\, t \in [0,T]$.

Proof. The Fokker-Plank equation for the forward stochastic differential equation is given by

$$\frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t} = -\operatorname{div}\left(\boldsymbol{b}(\boldsymbol{X}, t) p_t^{\rightarrow}(\boldsymbol{X})\right) + \tfrac{1}{2}\sigma^2(t)\Delta p_t^{\rightarrow}(\boldsymbol{X}) \tag{2.43}$$

Looking at $\Delta p_t^{\rightarrow}(\boldsymbol{X})$ we see that

$$\begin{aligned}\Delta p_t^{\rightarrow}(\boldsymbol{X}) &= \operatorname{div} \nabla p_t^{\rightarrow}(\boldsymbol{X}) && (2.44)\\ &= \operatorname{div}\left[\frac{p_t^{\rightarrow}(\boldsymbol{X})}{p_t^{\rightarrow}(\boldsymbol{X})}\nabla p_t^{\rightarrow}(\boldsymbol{X})\right] && (2.45)\\ &= \operatorname{div}\left[p_t^{\rightarrow}(\boldsymbol{X})\nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right] && (2.46)\\ &= \operatorname{div}\left(p_t^{\rightarrow}(\boldsymbol{X})\nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right) && (2.47)\end{aligned}$$

Adding and subtracting this identity we get

$$\begin{aligned}&\frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t} && (2.48)\\ &= -\operatorname{div}\left(\boldsymbol{b}(\boldsymbol{X},t)p_t^{\rightarrow}(\boldsymbol{X})\right) + \tfrac{1}{2}\sigma^2(t)\Delta p_t^{\rightarrow}(\boldsymbol{X}) - \sigma(t)^2 \operatorname{div}\left(p_t^{\rightarrow}(\boldsymbol{X})\nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right) && (2.49)\\ &= -\operatorname{div}\left(\left[\boldsymbol{b}(\boldsymbol{X},t) - \sigma(t)^2\nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right]p_t^{\rightarrow}(\boldsymbol{X})\right) - \tfrac{1}{2}\sigma^2(t)\Delta p_t^{\rightarrow}(\boldsymbol{X}) && (2.50)\end{aligned}$$

Taking the negative of this last expression we define

$$\frac{\partial p_t^{\leftarrow}(\boldsymbol{X})}{\partial t} = \operatorname{div}\left(-\left[-\boldsymbol{b}(\boldsymbol{X},t) + \sigma(t)^2\nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right]p_t^{\rightarrow}(\boldsymbol{X})\right) + \tfrac{1}{2}\sigma^2(t)\Delta p_t^{\rightarrow}(\boldsymbol{X}) \quad (2.51)$$

By definition then

$$\frac{\partial p_t^{\leftarrow}(\boldsymbol{X})}{\partial t} = -\frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t}. \quad (2.52)$$

If we assume that for some $t \in [0,t]$ $p_t^{\rightarrow} = p_{T-t}^{\leftarrow}$, then by the *Picard–Lindelöf theorem* (see appendix B.3) we have that $p_t^{\rightarrow} = p_{T-t}^{\leftarrow}$ for all $t \in [0,T]$. Since eq. (3.32) is the form of the Fokker-Plank equation of the reverse stochastic differential equation specified, the claim follows. ∎

This proof is inspired by one found in Nelson (1967, chapter 13).

This theorem states only that the time-marginals the of the forward and reverse distributions match. For generative modelling this is all that matters as we only care that we can recover p_{data}.

More commonly cited in the score-based modelling literature are stronger results that guarantee the whole distribution on *paths* of the forward and reverse stochastic differential equations match. A discussion of commonly cited time reversal results can be found in appendix B.8. See figure 2.10 for the difference between the matching of time marginals and the path distributions.

Learning the score

The next step in producing a diffusion model is learning the score term that appears in the reverse stochastic differential equation, $\nabla_{\boldsymbol{X}_t} \log p_t(\boldsymbol{X}_t)$. At first glance this may seem difficult, but a strong literature exists on learning this term, known as *score matching*. In order to estimate this score function, we will use an approximation taking inputs $\boldsymbol{X}_t$ and t. We will denote this $\boldsymbol{s} : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$. We will also use the notion $\boldsymbol{s}_t : \mathbb{R}^d \to \mathbb{R}^d$ to refer to this function partially evaluated at time t, $\boldsymbol{s}_t(\cdot) = \boldsymbol{s}(\cdot, t)$. Typically, we will make this a neural network of some kind, and look to optimise it via gradient-based learning.

Score matching

Hyvärinen (2005) proves the following result:

Implicit score matching (ISM)

THEOREM 2.7. *Consider a density $p_a(A)$ supported on $\mathbb{R}^d$ relative to the Lebesgue measure. Then,*

$$\nabla_A \log p_a(A) = \underset{g \text{ measurable in } L^2}{\arg\min} \; \mathbb{E}_A\left[\tfrac{1}{2}\|g(A)\|^2 + \operatorname{div}(g)(A)\right]. \tag{2.53}$$

Using this we can learn the score of a distribution p_a if we just have access to good samples from p_a as we can approximate this expectation with Monte Carlo estimation.

We can apply this to our stochastic differential equation setting to say that the score at a given time is given by

$$\nabla_{X_t} \log p_t(\boldsymbol{X}_t) = \arg\min_{\boldsymbol{s}_t} \mathbb{E}_{\boldsymbol{X}_t}\left[\tfrac{1}{2}\|\boldsymbol{s}_t(\boldsymbol{X}_t)\|^2 + \operatorname{div}(\boldsymbol{s}_t)(\boldsymbol{X}_t)\right]. \tag{2.54}$$

We denote then this loss function as ℓ_t^{ism}.

In order to draw samples from $p_t(\boldsymbol{X}_t)$ we can utilise the fact that

$$p_t(\boldsymbol{X}_t) = \int p_{t|0}(\boldsymbol{X}_t) p_0(\boldsymbol{X}_0)\, \mathrm{d}\boldsymbol{X}_0. \tag{2.55}$$

We can sample then from this joint density as we have samples from $p_0 = p_{\text{data}}$, and we have access to sampling from the transition density $p_{t|0}$ from the forward noising stochastic differential equation. We then discard the samples from p_0 to obtain samples from p_t.

A limitation of this objective is that it requires computation of the divergence of the score. While this is achievable using automatic differentiation packages, it is expensive in high dimension, scaling in cost with $O(d^2)$.

It is possible to apply to this the Hutchinson trace trick estimator (Hutchinson, 1989) to stochastically approximate the divergence to get a loss function that scales with $O(d)$ at the cost of higher variance, as detailed in Song et al. (2020a), where it is called *sliced score matching (SSM)*. We denote this loss function ℓ_t^{ssm}. In practise this additional variance does not seem to be too important, as the variance of the Monte Carlo estimation of the integral over p_t is larger.

Sliced score matching (SSM)

Denoising score matching (DSM)

Denoising score matching (DSM) (Vincent, 2011) is an alternative approach to implicit score matching that typically has lower variance. However, applying this objective requires additional tractability on the forward noising process. Recall that the definition of the conditional distribution states that for L^2-integrable random variables A and B, the conditional expectation map

$$g^* : b \mapsto \mathbb{E}_A[A \mid B = b] \tag{2.56}$$

is given by the minimiser

$$g^* = \underset{g \text{ measurable in } L^2}{\arg\min} \; \mathbb{E}_{A,B}\left[\|A - g(B)\|^2\right] \tag{2.57}$$

where $\mathbb{E}_A[A \mid B = b]$ is the conditional expectation of A given a value of B. We apply this by putting $A = \nabla \log p_{c|d}(C \mid D)$ and $B = C$ for new random variables C and D, assuming they have densities relative to the Lebesgue measure, and so

$$g(C) = \mathbb{E}_D\left[\nabla \log p_{c|d}(D \mid C) \,\middle|\, C\right]. \tag{2.58}$$

We can then apply the Tweedie identity.

LEMMA 2.8. The Tweedie identity

$$\mathbb{E}_D\left[\nabla_C \log p_{c|d}(C \mid D) \,\middle|\, C\right] = \nabla_C \log p_c(C) \tag{2.59}$$

Proof.

$$\begin{aligned}
&\mathbb{E}_D\left[\nabla_C \log p_{c|d}(C \mid D) \,\middle|\, C\right] \\
&= \int_{\mathbb{R}^d} \nabla_C \log p_{c|d}(C \mid D) p_{d|c}(D|C)\, \mathrm{d}D && \text{By definition} \\
&= \int_{R^d} \frac{\nabla_C p_{c|d}(C \mid D)}{p_{c|d}(C \mid D)} \frac{p_{c|d}(C \mid D) p_d(D)}{p_t(C)}\, \mathrm{d}D && \text{By the gradient of log and Bayess rule} \\
&= \frac{1}{p_t(C)} \int_{\mathbb{R}^d} p_d(D) \nabla_C p_{c|d}(C \mid D)\, \mathrm{d}D && \text{Cancel terms and remove non-integrands} \\
&= \frac{1}{p_t(C)} \nabla_C \int_{\mathbb{R}^d} p_d(D) p_{c|d}(C \mid D)\, \mathrm{d}D && \text{By linearity of integrals} \\
&= \frac{\nabla_C p_t(C)}{p_c(C)} = \nabla_C \log p_c(C) && \text{By the gradient of log}
\end{aligned}$$

■

Combining this we see that

$$\begin{aligned}
\underset{g \text{ measurable in } L^2}{\arg\min}\ \mathbb{E}_{C,D}\left[\left\|\nabla_C \log p_{c|d}(C \mid D) - g(C)\right\|^2\right] &= \mathbb{E}_D\left[\nabla_C \log p_{c|d}(C \mid D) \,\middle|\, C\right] \\
&= \nabla_C \log p_c(C), \qquad (2.60)
\end{aligned}$$

and therefore we can learn the score for $\nabla_C \log p_c(C)$ if we have access to

1. samples from the joint distribution $p_{c,d}(C, D)$ and
2. access to the analytic conditional score $\nabla_C \log p_{c|d}(C \mid D)$.

To apply this to our stochastic differential equation setting, we pick $C = \boldsymbol{X}_t$ and $D = \boldsymbol{X}_s$, with $s < t$. We have access to samples from $p_{t,s}(\boldsymbol{X}_t, \boldsymbol{X}_s)$ by using the fact that

$$\begin{aligned}
p_{t,s}(\boldsymbol{X}_t, \boldsymbol{X}_s) &= \int_{\mathbb{R}^d} p_{t|s,0}(\boldsymbol{X}_t \mid \boldsymbol{X}_s, \boldsymbol{X}_0) p_{s|0}(\boldsymbol{X}_s \mid \boldsymbol{X}_0) p_0(\boldsymbol{X}_0)\, \mathrm{d}\boldsymbol{X}_0 && (2.61) \\
&= \int_{\mathbb{R}^d} p_{t|s}(\boldsymbol{X}_t \mid \boldsymbol{X}_s) p_{s|0}(\boldsymbol{X}_s \mid \boldsymbol{X}_0) p_0(\boldsymbol{X}_0)\, \mathrm{d}\boldsymbol{X}_0 && (2.62)
\end{aligned}$$

where the second line follows as stochastic differential equation transitions satisfy the Markov property. We have access to samples from p_0 from our data, and the stochastic differential equation tells us how to sample from the transitions. We also need access to the analytic form of $\nabla_{\boldsymbol{X}_t} \log p_{t|s}(\boldsymbol{X}_t \mid \boldsymbol{X}_s)$. For most stochastic differential equations used in the literature, such as the variance exploding and preserving, this is available. If it is not available, often good approximations are available over for short time steps between s and t.

The most common form of this loss in the literature is to set $s = 0$. This simplifies the sampling as then we only need to sample a single transition. This is the denoising score matching described in Song et al. (2020b). We describe the more

general scheme here as it is useful when we only have access to an approximation of $p_{t|s}$ over short times.

We shall label this loss $\ell_t^{t|s}$ in the general case and $\ell_t^{t|0}$ when using $s = 0$.

To summarise:

Loss	Requires		Complexity	Variance
	Samples from $p_{t\|s}$	Analytic $\nabla \log p_{t\|s}$		
ℓ_t^{ssm}	✓	✗	$O(d)$	Highest
ℓ_t^{ism}	✓	✗	$O(d^2)$	Middle
$\ell_t^{t\|s}$	✓	$t - s$ small	$O(d)$	Lowest
$\ell_t^{t\|0}$	✓	✓	$O(d)$	Lowest

Table 2.1. A comparison of score-matching objectives for diffusion models

The objectives detailed match the score at given time t only. In order to match the score at all times, we integrate this objective over all times $t \in [0, T]$.

$$\ell(\boldsymbol{s}) = \int_{[0,T]} \lambda(t) \ell_t(\boldsymbol{s}_t) \, \mathrm{d}t. \tag{2.63}$$

Notice there is a weighting function applied over all t. While this weighting does not affect the minima of the objective, it does rebalance where a network will prioritise learning the score most. Many choices for $\lambda(t)$ exist and can be tuned to optimise different objectives.

A common feature of all of these objectives is that they are

1. stable regression objectives and
2. require evaluation of the score network at a given time t only.

Compare this to the objectives of normalising flow models, VAE models, or GANs. These objective functions require us to run a full generation of samples from the model in order to evaluate the objective. In contrast, the score matching objectives only need us to evaluate the generative model at one point in the generative process.

Additionally, the score matching objectives at a given time are stable with respect to other times. That is, the minima at a given time will not change as we optimise the score function at other times. This is in comparison to most objectives of other deep generative models, where changing the weights at one point in the network changes the optimum weights at other points in the network are. This is exacerbated in GANs, as changing the discriminator function changes the objective the generator function is trying to learn, leading to well known instabilities. This property ensured by the fact that in diffusion models we specify the path we wish the model to take from noise to data. In other models it is left to the model to find this path.

There is one drawback to score-matching objectives however, and that is the Monte Carlo sampling required to evaluate them makes them particularly noisy. This can

be mitigated however by taking an average of the weights throughout training. Typically, exponential moving averages have been used, and the smoothing rate requires tuning. Recently, Karras et al. (2024) proposed new averaging schemes, and importantly a method to perform hyperparameter optimisation of the smoothing rates post training, requiring only a single training run of the model.

Likelihood training of score-based models

Typically, generative models are trained by minimising the forward Kullback-Leibler divergence between the data distribution and the model. That is they minimise the following objective

$$\begin{aligned}
\mathcal{D}_{KL}(p_{\text{data}} \mid q_\theta) &= \int \log\left(\frac{p_{\text{data}}}{q_\theta}\right) \mathrm{d}p_{\text{data}} && (2.64)\\
&= \int \log p_{\text{data}} \, \mathrm{d}p_{\text{data}} - \int \log q_\theta \, \mathrm{d}p_{\text{data}} && (2.65)\\
&= -\mathbb{E}_{p_{\text{data}}}[\log q_\theta] + \text{constant} && (2.66)\\
&\geq -\mathbb{E}_{p_{\text{data}}}[q_\theta] \qquad \text{by Jensen's inequality.} && (2.67)
\end{aligned}$$

This objective can therefore be seen also as maximising the average log-likelihood of the model under the data, or maximising a lower bound on the likelihood of the model under the data. It may be that we want to train a score-based model via maximum likelihood. It is not immediately obvious how to connect score-matching to this objective, however it is possible to do so.

Let us denote the measure placed on the space of diffusion paths defined by the data distribution and the forward noising stochastic differential equation as $P_{\text{data}} = p_{\text{data}} Q^{\rightarrow}_{\cdot|0}$ where $Q^{\rightarrow}_{\cdot|0}$ is the conditional path measure given an initial data point in the diffusion. Let us also denote $Q^\theta = p_{\text{prior}} Q^{\leftarrow}{}_\theta(\cdot|T)$ as the measure on paths defined by initialising the parametrised reverse stochastic differential equation with the reference distribution the forward noising process converges to. Then we can write

$$\mathcal{D}_{KL}(p_{\text{data}} \mid q_\theta) \leq \mathcal{D}_{KL}(P_{\text{data}} \mid Q_\theta) \tag{2.68}$$

by the data-processing inequality. Huang et al. (2021) and Song et al. (2021) show that by applying the chain rule for path measures (Léonard, 2014, theorem 2.4) and Girsanov's theorem (Øksendal, 2003, section 8.6), for a forward noising process with diffusion term $\sigma(t)$,

$$\begin{aligned}
\mathcal{D}_{KL}(P_{\text{data}} \mid Q_\theta) =& \mathcal{D}_{KL}\left(p_T \,\middle|\, p_{\text{prior}}\right) && (2.69)\\
&+ \frac{1}{2}\int_0^T \sigma(t)^2 \, \mathbb{E}_{p_t}\left[\left\| \nabla_{\boldsymbol{X}_t} \log p_t(\boldsymbol{X}_t) - \boldsymbol{s}_t^\theta(\boldsymbol{X}_t) \right\|_2^2\right] \mathrm{d}t, && (2.70)
\end{aligned}$$

showing that optimising any score-matching objective with time weighting $\lambda(t) = \sigma(t)^2$ maximises the same lower bound on the maximum likelihood of the data as other generative models.

Likelihood ordinary differential equations

The final aspect that we are missing is the ability to compute the likelihood of a sample from a diffusion model. In order to do this we will leverage the *Fokker-Plank*

Fokker-Plank equation

equation again (theorem B.12). Recall that for a stochastic differential equation of the form

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)\,\mathrm{d}t + \Sigma(\boldsymbol{X}_t, t)\,\mathrm{d}\boldsymbol{B}_t, \tag{2.71}$$

the time-evolution of the density is given by

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial \boldsymbol{X}_i}\left[\boldsymbol{b}(\boldsymbol{X}, t) p_t \boldsymbol{X}\right] + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2}{\partial \boldsymbol{X}_i \partial \boldsymbol{X}_j}\left[\Sigma(\boldsymbol{X}, t)\Sigma(\boldsymbol{X}, t)^T p_t(\boldsymbol{X})\right]. \tag{2.72}$$

Using this, we can obtain the following result:

The likelihood flow ordinary differential equation

THEOREM 2.9. *For a stochastic differential equation of the form*

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)\,\mathrm{d}t + \Sigma(\boldsymbol{X}_t, t)\,\mathrm{d}\boldsymbol{B}_t \tag{2.73}$$

initialised with $\boldsymbol{X}_0 \sim p_0$, the ordinary differential equation

$$\mathrm{d}\boldsymbol{X}_t = \left[\boldsymbol{b}(\boldsymbol{X}_t, t) - \frac{1}{2}\left[\operatorname{div}\left[\Sigma(\boldsymbol{X}, t)\Sigma(\boldsymbol{X}, t)^T\right] + \Sigma(\boldsymbol{X}, t)\Sigma(\boldsymbol{X}, t)^T \nabla_{\boldsymbol{X}_t} \log p_t(\boldsymbol{X}_t)\right]\right] \mathrm{d}t$$

initialised with $\boldsymbol{X}_0 \sim p_0$ has the same marginal density at all times $t \in [0, T]$

Proof. Due to Maoutsa et al. (2020) and Song et al. (2020b, appendix D). ∎

In diffusion models we typically use $\Sigma(\boldsymbol{X}, t) = \sigma(t)I$, a simple time-dependant scalar noise. With this, the likelihood ordinary differential equation simplifies to

$$\mathrm{d}\boldsymbol{X}_t = \left[\boldsymbol{b}(\boldsymbol{X}_t, t) - \frac{1}{2}\sigma(t)^2 \nabla_{\boldsymbol{X}_t} \log p_t(\boldsymbol{X}_t)\right] \mathrm{d}t. \tag{2.74}$$

Fortunately there already exist tools to allow us to compute the change in likelihood induced by an ordinary differential equation trajectory, assuming the ordinary differential equation describes the flow of a density.

Instantanious change of likelihood of an ordinary differential equation

THEOREM 2.10. *Let $(\boldsymbol{X}_t)_{t\in[0,T]}$ be an ordinary differential equation with governing equation*

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)\,\mathrm{d}t \tag{2.75}$$

and time marginal densities $p_t(\boldsymbol{X}_t)$. Assume that $\boldsymbol{b}$ is uniformly Lipshitz continuous in $\boldsymbol{X}_t$ and continuous in t.

Then, the instantaneous change in log probability is given by

$$\frac{\partial \log p_t(\boldsymbol{X}_t)}{\partial t} = -\operatorname{tr}\left(\frac{\mathrm{d}\boldsymbol{b}}{\mathrm{d}\boldsymbol{X}_t}\middle|(\boldsymbol{X}_t, t)\right) = -\operatorname{div}\boldsymbol{b}(\boldsymbol{X}_t, t) \tag{2.76}$$

Proof. Proved in Chen et al. (2018, appendix A). ∎

Using this, we can evaluate the likelihood of a point under our diffusion model at $t = 0$. For a given point $\boldsymbol{X}_0$ is equal to the change in likelihood over the course of the ordinary differential equation, plus the likelihood of the point at the other end,

$$\log p_0(\boldsymbol{X}_0) = \int_T^0 \operatorname{div}\boldsymbol{b}(\boldsymbol{X}_t, t)\,\mathrm{d}t + \log p_T(\boldsymbol{X}_T) \tag{2.77}$$

where $\boldsymbol{X}_t$ is the path the point $\boldsymbol{X}_0$ follows under the ordinary differential equation and p_T is the reference distribution of the diffusion model.

In practice, we solve these as a joint ordinary differential equation to take advantage of error tolerant solvers on the joint system,

$$\mathrm{d}\begin{bmatrix}\boldsymbol{X}_t \\ \log p(\boldsymbol{X}_t)\end{bmatrix} = \begin{bmatrix}\boldsymbol{b}(\boldsymbol{X}_t, t) \\ \operatorname{div}\boldsymbol{b}(\boldsymbol{X}_t, t)\end{bmatrix} \tag{2.78}$$

Continuous normalising flows (CNFs)

This is in fact how *continuous normalising flows (CNFs)* (Chen et al., 2018) work. The model is defined by a parametrised drift function $\boldsymbol{b}_\theta$, and is trained by maximum likelihood by backpropagating through the ordinary differential equation solver (Chen et al., 2018). There exist a number of strategies for backpropagating through ordinary differential equation solvers. These are discussed well in Kidger (2021), which has accompanying it the python library `diffrax` for handling ordinary differential equations and other differential equations easily.

https://github.com/patrick-kidger/diffrax

To apply this scheme for computing the likelihood to diffusion models we simply use our modified drift function in eq. (2.74) and use the reference distribution from our diffusion model for p_T.

This also gives us an additional method to sample from our diffusion model. Instead of solving the stochastic differential equation approximately (appendix B.10) we can sample from p_T and solve this ordinary differential equation in reverse to draw samples from the model. This can outperform sampling the stochastic differential equation in certain scenarios (Karras et al., 2022; Song and Ermon, 2019).

2.2.2. Deep generative modelling on manifolds. Finally, we discuss prior work on deep generative modelling on manifolds.

We discuss two settings of generative modelling on manifolds. Namely, the setting of modelling a density on the manifold itself, and also the setting of modelling a density on *signals* on the manifold. Here we cover methods for which the manifold structure is prescribed, in contrast with methods that jointly learn the manifold structure and density (e.g. Brehmer and Cranmer, 2020; Caterini et al., 2021).

Densities on a manifold

Simple classical methods exist for fitting mixture of distribution models on manifolds, for example (Peel et al., 2001) fit mixture of Kent distributions and Mardia et al. (2007) fit mixtures of von Mises distributions. These models struggle with representation power and generalisation in high dimensions.

A series of methods have been proposed that modelled densities on manifolds by defining them as a pushforward of densities modelled on Euclidean space. For this to work exactly we require a homeomorphism $\phi : \mathbb{R}^n \to X$. This can only exist if the manifold is topologically homeomorphic to $\mathbb{R}^n$, and so is quite limiting. Falorsi et al. (2019) propose a method of this form for reparametrising densities on Lie groups. Bose et al. (2020) propose a method for modelling densities on hyperbolic space using Euclidean normalising flows pushed onto hyperbolic space. Gemici et al. (2016) propose a similar method for generic manifolds. In addition to the homeomorphism restriction, these methods do not respect the underlying

geometry of the space well. In order to map Euclidean space onto these manifolds, a significant distortion of the metric is required. This can cause regions of the space that are significantly stretched or squashed by the homeomorphism to be difficult to model accurately.

Methods based on normalising flows have become a popular choice for intrinsically modelling densities on manifolds. Rezende et al. (2020) produce a method for discrete normalising flows that operate on tori and spheres directly. Cohen et al. (2021) and Rezende and Racanière (2021) build normalising flows on manifolds via optimal transport concave potential maps. Schemes for general equivariant normalising flows on manifolds have also been developed (Katsman et al., 2021). A significant line of work has been developed for normalising flows on products of the group SU(3) (Abbott et al., 2022a; Abbott et al., 2022b; Abbott et al., 2023), generic SU(n) (Boyda et al., 2021) and U(1) (Kanwar et al., 2020). One drawback to normalising flow methods is the difficulty in designing invertible layers for normalising flows. This makes designing transferable methods for generic manifolds difficult.

Methods based on continuous normalising flows are much easier to design for generic manifolds. They remove the need to design specific invertible layers to build the flow models, and instead simply require the specification of a time-varying vector field. Methods extending continuous normalising flows to Riemannian manifolds have been developed by **lou2020eural**; Falorsi and Forré (2020), Falorsi (2021), and Mathieu and Nickel (2020). A significant drawback of continuous normalising flows is the need to solve an ordinary differential equation and back-propagate through it at each step of training since this is computationally expensive. Rozen et al. (2021) develop an extension of Riemannian continuous normalising flows by placing a divergence free condition on the vector field. This allows them to circumvent the need to compute the ordinary differential equation at the cost of enforcing the divergence free condition via a regulariser.

Signals on manifolds

While the literature for modelling densities on Riemannian manifolds is reasonably developed, and there exists significant literature on generative modelling of signals in Euclidean space (Goodfellow et al., 2014; Kingma and Welling, 2013; Rezende et al., 2014; Rezende and Mohamed, 2015; Tabak and Vanden-Eijnden, 2010; Tabak and Turner, 2013; Chen et al., 2018; Grathwohl et al., 2019; Song et al., 2020b; Sohl-Dickstein et al., 2015; Ho et al., 2020) there exists little work on generative modelling of signals on manifolds.

There is a line of work in Gaussian process modelling that can model scalar fields on manifolds (Borovitskiy et al., 2020; Azangulov et al., 2023a; Azangulov et al., 2023b) and vector fields (Hutchinson et al., 2021c). These are limited however in their expressive power and scalability due to the nature of Gaussian process models.

Holderrieth et al. (2021a) develop an equivariant neural process model on signals on Euclidean space, and could be extended to modelling signals on manifolds with the kernels defined in Hutchinson et al. (2021c).

3 | Score-based modelling on Riemannian manifolds

Work in this chapter is based on

V. De Bortoli*, E. Mathieu*, M. J. Hutchinson*, J. Thornton, Y. W. Teh, and A. Doucet. Riemannian score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2022.

and has been rewritten for this thesis with additional material.

Personal contributions:

1. Project conception with Emile.
2. Development of practical approach with Emile, Valentin.
3. Development of the code base with Emile.
4. Running of experiments: Earth and climate science, synthetic data on tori, and synthetic data on hyperbolic space experiments.
5. Mathematical results: The simplified time-reversal result on manifolds, theorem 3.5, and the implicit score matching on Riemannian manifolds, proposition 3.7.
6. Original manuscript writing with Emile and Valentin.

3.1. Introduction

In the background we introduced a type of generative model called score-based generative models, also called diffusion models (Song and Ermon, 2019; Song et al., 2020b; Ho et al., 2020; Dhariwal and Nichol, 2021). They formulate generative modelling as a denoising process. Noise is incrementally added to data using a diffusion process until it becomes approximately Gaussian. The generative model is then obtained by simulating an approximation of the corresponding time-reversal process, which progressively denoises a Gaussian sample to obtain a data sample. This process is also a diffusion whose drift depends on the log gradients of the noised data densities, i.e. the Stein scores, estimated using a neural network via score matching (Hyvärinen, 2005; Vincent, 2011).

Score-based generative models have been primarily applied to data living on Euclidean spaces, i.e. manifolds with flat geometry. However, in a large number of scientific domains the distributions of interest are supported on Riemannian manifolds. These include, to name a few, protein modelling (Shapovalov and Dunbrack Jr, 2011), cell development (Klimovskaia et al., 2020), image recognition (Lui, 2012), geological sciences (Karpatne et al., 2018; Peel et al., 2001), graph-structured and hierarchical data (Roy et al., 2007; Steyvers and Tenenbaum, 2005), robotics (Feiten et al., 2013; Senanayake and Ramos, 2018) and high-energy physics (Brehmer and Cranmer, 2020).

Riemannian score-based generative models

In this chapter we develop *Riemannian score-based generative models*, an extension of score-based generative models to Riemannian manifolds. The objective is to incorporate the geometry a given problem into the model. This requires constructing a noising process on the manifold that converges to an easy-to-sample reference distribution. We establish that, as in the Euclidean case, the corresponding time-reversal process is also a diffusion whose drift includes the Stein score which is intractable but can similarly be estimated via score matching. Methodological extensions are required as in most cases the transition kernel of the noising process cannot be sampled exactly. For example on compact manifolds it is typically only available as an infinite sum through the Sturm–Liouville decomposition (Chavel, 1984). To this end, we develop non-standard techniques for score estimation and rely on the use of geodesic random walks for sampling (Jørgensen, 1975). We provide theoretical convergence bounds for Riemannian score-based generative models on compact manifolds and demonstrate our approach on a range of manifolds and tasks, including modelling a number of natural disaster occurrence datasets collected by Mathieu and Nickel (2020). We show that Riemannian score-based generative models achieve better performance than recent baselines (Mathieu and Nickel, 2020; Rozen et al., 2021) and scale better to higher dimensions than these approaches also.

3.2. Riemannian Score-Based Generative Modelling

In this section, we introduce new tools and techniques to generalise score-based models from Euclidean space to Riemannian manifolds (without boundary).

More specifically our setting will be a *complete* (see appendix A.10.5), *orientable* (see appendix A.9.2), *connected* (see appendix A.1), *Riemannian* (see definition A.36) manifold *without boundary* (see definition A.5), $(\mathcal{M}, g)$.

Four components are required to extend score based generative models to this setting:

i) a forward noising process on $\mathcal{M}$ which converges to an easy-to-sample reference distribution,

ii) a time-reversal formula on $\mathcal{M}$ which defines a backward generative process,

iii) a method for approximating samples of stochastic differential equations on manifolds,

iv) a method to efficiently perform score matching on Riemannian manifolds.

We provide these tools, and detail them now.

If we assume that a density of a measure exists at any point, we assume that it exists relative to the *Riemannian volume measure* (see appendix A.9.4), vol_g, the measure induced by the Riemannian volume form on the manifold. This is the generalisation of the Lebesgue measure on Euclidean space to manifolds.

3.2.1. Stochastic differential equations on manifolds. The core technical tool of diffusion modelling in Euclidean space are stochastic differential equations. Appendix B gives an introduction to this topic. These tools do not immediately generalise to manifolds, the core obstacle being that general manifolds do not have vector space structure, and therefore the notion of an expectation is not defined for manifold-valued random variables. The general strategy for studying such random variables is to view them through test functions, functions that map the random variables on a Euclidean space where expectations can be defined. By choice of the class of test function, and showing that certain properties hold for all test functions, we can make conclusions about the random variables under study.

On Euclidean space *semimartingales* (see appendix B.2) form the core object of study. To extend this definition we view manifold-valued random processes through test functions.

Semimartingale on a manifold

DEFINITION 3.1. *Let $\mathcal{M}$ be a d-dimension differentiable manifold and $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with a filtration of $\mathcal{F}$, $(\mathcal{F}_t)_{t\in\mathbb{R}^+}$. A continuous $\mathcal{M}$-valued stochastic process $(\boldsymbol{X}_t)_{t\in[0,T]}$, $T \in \mathbb{R}^+$, is an $\mathcal{M}$-valued semimartingale if $f(\boldsymbol{X}_t)$ is a semimartingale with respect to the filtration for all $f \in C^\infty(\mathcal{M})$.*

$\mathcal{M}$-valued stochastic differential equation

To define an *$\mathcal{M}$-valued stochastic differential equation* we take inspiration from eq. (B.60). For a collection of ℓ time-dependent vector fields on $\mathcal{M}$, $V_1, ..., V_l$, and an ℓ-dimension Euclidean semimartingale we write

$$\mathrm{d}\boldsymbol{X}_t = V_i(\boldsymbol{X}_t) \circ \mathrm{d}Z_t^i \tag{3.1}$$

as the evolution of the stochastic differential equation, with the *Einstein summation* (see appendix A.4.5) over i implicit. We define the meaning of this through the definition of the solutions to this stochastic differential equation on a manifold.

Solution to a stochastic differential equation on a manifold

DEFINITION 3.2. *An $\mathcal{M}$-valued semimartingale X is a solution to this stochastic differential defined by vector fields $V_1, ..., V_\ell$ and $\mathbb{R}^\ell$-valued semimartingale Z if for all $f \in C^\infty(M)$ and X_0 a random variable taking values in $\mathcal{M}$,*

$$f(\boldsymbol{X}_t) \overset{d}{=} f(\boldsymbol{X}_0) + \int_0^t (V_i f)(\boldsymbol{X}_s, s) \circ \mathrm{d}Z_s^i. \tag{3.2}$$

$V_i f$ is the directional derivative of f with respect to V_i and the equality is an equality as distributions.

We typically use the Stratonovich integral in the definition of manifold-valued stochastic differential equations as it obeys a simpler chain rule than the Itô integral, making working with test functions and charts much simpler (see appendix B.5.1).

Using these definitions we can extend the tools presented in appendix B to the manifold setting, replacing Euclidean components with their geometric counterparts. Most importantly we can generalise *diffusion processes* (see appendix B.6) to manifolds, allowing us to connect manifold-valued stochastic differential equations with differential operators.

Diffusion processes on manifolds

DEFINITION 3.3. *For an $\mathcal{M}$-valued stochastic process $\boldsymbol{X}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and adapted to a filtration $(\mathcal{F}_t)_{t\in[0,T]}$ of $\mathcal{F}$, we say it is a diffusion process generated by L if*

$$\mathbb{E}[f(\boldsymbol{X}_t)] = \mathbb{E}[f(\boldsymbol{X}_s)] + \mathbb{E}\left[\int_s^t Lf(\boldsymbol{X}_\tau)\,\mathrm{d}\tau\right] \tag{3.3}$$

for all $f \in C^2(\mathcal{M})$, $t > s \in [0, T]$, where L is a smooth second order elliptic operator on $\mathcal{M}$.

We can then work with this generator and the Kolmogorov forward and backward equations to analyse manifold-valued stochastic differential equations.

Brownian motion on a manifold

The generator of a manifold-valued diffusion process gives us one way to define a *Brownian motion on a manifold*, $\boldsymbol{B}^{\mathcal{M}}$, as a stochastic process generated by $\frac{1}{2}\Delta_g$, where Δ_g is the *Laplace-Beltrami operator* (see appendix A.10.8) (Hsu, 2002, chapter 3).

For the rest of the chapter we will work with stochastic differential equations of the form

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t^{\mathcal{M}} \tag{3.4}$$

where $\boldsymbol{b}$ is a time-varying smooth vector field on the manifold and σ a scalar noise term. This is a diffusion, and is generated by the operator L defined by

$$Lf = \left\langle \boldsymbol{b}, \nabla_g f \right\rangle_g + \tfrac{\sigma^2}{2}\Delta_g f \quad \forall\, f \in C^2(\mathcal{M}) \tag{3.5}$$

with adjoint defined by

$$L^* f = -\operatorname{div}_g(\boldsymbol{b} f) + \tfrac{\sigma^2}{2}\Delta_g f \quad \forall\, f \in C^2(\mathcal{M}). \tag{3.6}$$

For a complete treatment of stochastic processes on manifolds, and in depth detail of analysis techniques, we refer to appendix C.2 and Hsu (2002).

3.2.2. Noising processes on manifolds. The first necessary component is a suitable generic noising process on manifolds that will converge to a convenient stationary distribution. In the background material we showed how the common noising processes used in Euclidean diffusion models are instances of *Langevin dynamics* (see appendix B.7). Recall these are given for a stochastic differential equation in Euclidean space by

Langevin dynamics

$$\mathrm{d}\boldsymbol{X}_t = -\nabla U(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t \tag{3.7}$$

where $U : \mathbb{R}^d \to \mathbb{R}$ with some regularity. This stochastic differential equation will converge in distribution to a density, relative to the Lebesgue measure, proportional to $\exp(-U(\boldsymbol{X}_t))$ (Roberts and Tweedie, 1996, theorem 2.1).

On Riemannian manifolds, we have very similar results for the equivalent Langevin stochastic differential equation.

$$\mathrm{d}\boldsymbol{X}_t = -\nabla_g U(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t^{\mathcal{M}} \tag{3.8}$$

where ∇_g is the Riemannian gradient and $\mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}$ is a Brownian motion on the manifold. For example, we can prove that a very similar stationary distribution exists.

Stationary distribution of Langevin dynamics on Riemannian manifolds

THEOREM 3.4. *For a stochastic differential equation of the form*

$$\mathrm{d}\boldsymbol{X}_t = -\nabla_g U(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}, \tag{3.9}$$

the distribution μ_U with Radon-Nikodyom derivate with respect to the Riemannian volume measure

$$\frac{\mathrm{d}\mu_U}{\mathrm{dvol}_g} = \exp(-U) = p_U \tag{3.10}$$

is invariant under the evolution of the dynamics.

Proof. Applying the adjoint of the generator of this diffusion to the *Fokker-Plank equation* (see theorem B.12) we see that

$$\begin{aligned} L^* p_U &= -\operatorname{div}_g\big(-\nabla_g U \exp(-U)\big) + \Delta_g \exp(-U) && (3.11)\\ &= -\operatorname{div}_g\big(-\nabla_g U \exp(-U)\big) + \operatorname{div}_g \nabla_g \exp(-U) \quad \text{using eq. (A.188)} && (3.12)\\ &= -\operatorname{div}_g\big(-\nabla_g U \exp(-U)\big) + \operatorname{div}_g\big(-\nabla_g U \exp(-U)\big) = 0. && (3.13) \end{aligned}$$

■

Further, results such as Gatmiry and Vempala (2022, theorem 1) show that the stochastic differential equation converges in distribution to the desired density. Immediately then, this motivates using a form of Langevin dynamics as our noising process. The difficulty then is how to choose a U to suit our needs. In the Euclidean setting the Gaussian distribution is the natural choice given its many useful properties. We therefore look at a number of ways of adapting the Gaussian to manifold.

Heat kernel on a manifold

One option is the *heat kernel on a manifold*, the solutions to

$$\frac{\partial K}{\partial t} = \Delta_g K, \quad K : \mathbb{R}^+ \times \mathcal{M} \times \mathcal{M}. \tag{3.14}$$

Unfortunately there is not typically a closed from solution for the heat kernel on most manifolds, making accessing U difficult. In general there is also not a cheap algorithm to sample from this distribution. Instead, more expensive MCMC, such as Langevin dynamics, must be used to sample from this distribution.

Riemannian normal

Instead, we could target the *Riemannian normal* distribution (Pennec, 2006). We achieve this by setting

$$U(x) = d_g(x, \mu)^2/(2\gamma^2), \tag{3.15}$$

where d_g is the *geodesic distance* (see appendix A.8.3) under the metric g and $\mu \in \mathcal{M}$ is an arbitrary mean location. This induces the drift

$$\nabla U(\boldsymbol{X}_t) = -\exp_{\boldsymbol{X}_t}^{-1}(\mu)/\gamma^2 \tag{3.16}$$

where $\exp_x : \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ denotes the *exponential map* (see appendix A.10.5) on the manifold. Again however, in general there is not a cheap algorithm to sample from this distribution, and so we have to resort to MCMC methods.

Exponential wrapped Gaussian

Finally, we can target the *exponential wrapped Gaussian*, under the assumption that the manifold in question is *geodesically complete* (see appendix A.10.5). This is the pushforward of a Gaussian distribution in the tangent space of an arbitrarily chosen mean location $\mu \in \mathcal{M}$ through the exponential map, giving

$$p(x) = \exp_\mu^* \mathcal{N}(0, \sigma^2) \tag{3.17}$$

for a chosen variance σ^2. Sampling this distribution is very simple, we sample a regular Gaussian and pass the samples through the exponential map.

Computing the density, and therefore the required U is only simple in the case where the exponential map is a bijection between the tangent space at a point and the manifold. This is true for some non-compact manifolds such as Euclidean space and hyperbolic space, but not others such as the circle, $\mathcal{S}_2$. In this case we can write the density as

$$p(x) = \mathcal{N}\left(\exp_\mu^{-1}(x); 0, \sigma^2\right) \quad x \in \mathcal{M}. \tag{3.18}$$

Applying the change of variable formula, we arrive at a potential given by

$$\begin{aligned} U(x) &= d_{\mathcal{M}}(x, \mu)^2/(2\gamma^2) + \log|\det|D \exp^{-1} \mu(x) && (3.19)\\ &= d_{\mathcal{M}}(x, \mu)^2/(2\gamma^2) + 1/\log|\det|D \exp \mu(x), && (3.20) \end{aligned}$$

Jacobian

where D denotes the *Jacobian* (see appendix A.5.4) of the log map.

Where the exponential map is not a bijection, but the manifold is still geodesically complete, the exponential map multiple points in the tangent space to the same point in the manifold. The inverse exponential map is not well-defined in this case, but we can instead replace this with summation over preimage of the exponential maps. Let us denote this as $\mathcal{P}(x) = \left\{\exp_\mu(y) = x \ : \ y \in T_\mu\mathcal{M}\right\}$. Then the density is given by

$$p(x) = \sum_{y \in \mathcal{P}(x)} \mathcal{N}\left(y; 0, \sigma^2\right) \quad x \in \mathcal{M}. \tag{3.21}$$

One recovers the standard Ornstein–Uhlenbeck noising process (Song et al., 2020b) for all three of the options for U when $\mathcal{M} = \mathbb{R}^d$ and $\mu = 0$ since then the drift $b(t, \boldsymbol{X}_t) = \frac{1}{2} \exp_{\boldsymbol{X}_t}^{-1}(0) = -\frac{1}{2} \boldsymbol{X}_t$.

3.2.3. Time-reversal on Riemannian manifolds. Next we show that time reversal results can be extended to the case of Riemannian manifolds. We present here two theorems.

The first is a simple proof in the spirit of theorem 2.6, which shows that the time-marginals of a reverse process match those the forward process. This is all we need for generative modelling as we only care that the initial distribution and final time distributions match

Fokker-Plank time reversal on a manifold

THEOREM 3.5. *Let $(\mathcal{M}, g)$ be a Riemannian manifold. Let $\boldsymbol{B}_t^{\mathcal{M}}$ be a Brownian motion on $\mathcal{M}$.*

Let the manifold-valued stochastic differential equation $\boldsymbol{X}^{\rightarrow}$, called the forward stochastic differential equation, be given by

$$\mathrm{d}\boldsymbol{X}_t^{\rightarrow} = \boldsymbol{b}(\boldsymbol{X}_t^{\rightarrow}, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}, \quad \boldsymbol{X}_0^{\rightarrow} \sim \pi, \quad t \in [0, T], \tag{3.22}$$

with time-marginal distributions $p_t^{\rightarrow}(\boldsymbol{X})$. Let the manifold-valued stochastic differential equation $\boldsymbol{X}_t^{\leftarrow}$, called the reverse stochastic differential equation, be given by

$$\begin{aligned}\mathrm{d}\boldsymbol{X}_t^{\leftarrow} = \left[-\boldsymbol{b}(\boldsymbol{X}_t^{\leftarrow}, T-t) + \sigma(T-t)^2 \nabla \log p_{T-t}^{\rightarrow}(\boldsymbol{X}_t^{\leftarrow})\right]\mathrm{d}t + \sigma(T-t)\,\mathrm{d}\boldsymbol{B}_t^{\mathcal{M}},\\ \boldsymbol{X}_0^{\leftarrow} \sim p_T^{\rightarrow}, \quad t \in [0, T],\end{aligned} \tag{3.23}$$

with time-marginal distributions $p_t^{\leftarrow}(\boldsymbol{X})$.

Then the time-marginals of these two stochastic differential equations match, but reversed in time, in the sense that $p_t^{\rightarrow} = p_{T-t}^{\leftarrow}\ \forall\ t \in [0, T]$.

Proof. The Fokker-Plank equation for the forward stochastic differential equation is given by

$$\frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t} = -\operatorname{div}_g\big(\boldsymbol{b}^{\rightarrow}(\boldsymbol{X}, t) p_t^{\rightarrow}(\boldsymbol{X})\big) + \tfrac{1}{2}\sigma^2(t)\Delta_g p_t^{\rightarrow}(\boldsymbol{X}) \tag{3.24}$$

Looking at $\Delta_g p_t^{\rightarrow}(\boldsymbol{X})$ we see that

$$\begin{aligned}\Delta_g p_t^{\rightarrow}(\boldsymbol{X}) &= \operatorname{div}_g \nabla_g p_t^{\rightarrow}(\boldsymbol{X}) && (3.25)\\ &= \operatorname{div}_g\left[\frac{p_t^{\rightarrow}(\boldsymbol{X})}{p_t^{\rightarrow}(\boldsymbol{X})}\nabla_g p_t^{\rightarrow}(\boldsymbol{X})\right] && (3.26)\\ &= \operatorname{div}_g\left[p_t^{\rightarrow}(\boldsymbol{X})\nabla_g \log p_t^{\rightarrow}(\boldsymbol{X})\right] && (3.27)\\ &= \operatorname{div}_g\big(p_t^{\rightarrow}(\boldsymbol{X})\nabla_g \log p_t^{\rightarrow}(\boldsymbol{X})\big) && (3.28)\end{aligned}$$

Adding and subtracting this identity we get

$$\begin{aligned}\frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t} &= -\operatorname{div}_g\big(\boldsymbol{b}(\boldsymbol{X}, t) p_t^{\rightarrow}(\boldsymbol{X})\big) - \tfrac{1}{2}\sigma^2(t)\Delta_g p_t^{\rightarrow}(\boldsymbol{X}) + \sigma^2(t)\Delta_g p_t^{\rightarrow}(\boldsymbol{X}) && (3.29)\\ \frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t} &= -\operatorname{div}_g\big(\boldsymbol{b}(\boldsymbol{X}, t) p_t^{\rightarrow}(\boldsymbol{X})\big) - \tfrac{1}{2}\sigma^2(t)\Delta_g p_t^{\rightarrow}(\boldsymbol{X})\\ &\quad + \sigma(t)^2 \operatorname{div}_g\big(p_t^{\rightarrow}(\boldsymbol{X})\nabla_g \log p_t^{\rightarrow}(\boldsymbol{X})\big) && (3.30)\\ &= -\operatorname{div}_g\big(\left[\boldsymbol{b}(\boldsymbol{X}, t) - \sigma(t)^2\nabla_g \log p_t^{\rightarrow}(\boldsymbol{X})\right] p_t^{\rightarrow}(\boldsymbol{X})\big) - \tfrac{1}{2}\sigma^2(t)\Delta_g p_t^{\rightarrow}(\boldsymbol{X}) && (3.31)\end{aligned}$$

Taking the negative of the right-hand side of this equation we define,

$$\frac{\partial p_t^{\leftarrow}(\boldsymbol{X})}{\partial t} = -\operatorname{div}_g\big(\left[-\boldsymbol{b}(\boldsymbol{X}, t) + \sigma(t)^2\nabla_g \log p_t^{\leftarrow}(\boldsymbol{X})\right] p_t^{\leftarrow}(\boldsymbol{X})\big) + \tfrac{1}{2}\sigma^2(t)\Delta_g p_t^{\leftarrow}(\boldsymbol{X}). \tag{3.32}$$

By definition then

$$\frac{\partial p_t^{\leftarrow}(\boldsymbol{X})}{\partial t} = -\frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t}. \tag{3.33}$$

If we assume that for some $t \in [0, t]$ $p_t^{\rightarrow} = p_{T-t}^{\leftarrow}$, then by the *Picard–Lindelöf theorem* (see appendix B.3) we have that $p_t^{\rightarrow} = p_{T-t}^{\leftarrow}$ for all $t \in [0, T]$. Since eq. (3.32) is the form of the Fokker-Plank equation of the reverse stochastic differential equation specified, the claim follows. ∎

Next we also provide a generalisation of stronger results available in the Euclidean case, see appendix B.8 for an explanation of these. The main difference between these stronger results and theorem 3.5 above is that theorem 3.5 only guarantees that the *marginal* distributions, the distribution of the stochastic differential equation at given times, match. The stronger results tell us that the distribution on the whole space of *paths*, the joint distribution of the value of the stochastic process at all times, of the forward and reverse stochastic differential equations match. This result is in the spirit of Cattiaux et al. (2023), and we also provide an alternative in the spirit of Haussmann and Pardoux (1986).

Manifold time-reversed process

THEOREM 3.6. *Let $\left(\boldsymbol{B}_t^{\mathcal{M}}\right)_{t\geq 0}$ be a Brownian motion on $\mathcal{M}$ such that $\boldsymbol{B}_0^{\mathcal{M}}$ has distribution the volume form* vol_g[1]. *Let T be a time $T > 0$. Let $\left(\boldsymbol{X}_t^{\rightarrow}\right)_{t\in[0,T]}$ be the stochastic differential equation governed by*

$$\mathrm{d}\boldsymbol{X}_t^{\rightarrow} = \boldsymbol{b}(\boldsymbol{X}_t^{\rightarrow})\,\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}. \tag{3.34}$$

Let $\left(\boldsymbol{X}_t^{\leftarrow}\right)_{t\in[0,T]} = \left(\boldsymbol{X}_{T-t}^{\rightarrow}\right)_{t\in[0,T]}$. Assume that $KL(\mathbb{P}\,\mathbb{Q}) < +\infty$, where $\mathbb{Q}$ is the distribution on the paths of $(\boldsymbol{B}_t^{\mathcal{M}})_{t\in[0,T]}$ and $\mathbb{P}$ the distribution of $(\boldsymbol{X}_t^{\rightarrow})_{t\in[0,T]}$. In addition, assume that $\mathbb{P}_t = \mathcal{L}(\boldsymbol{X}_t^{\rightarrow})$, the distribution of $\boldsymbol{X}_t^{\rightarrow}$, admits a smooth positive density p_t with respect to vol_g *for any $t \in [0,T]$. Then, $\left(\boldsymbol{X}_t^{\leftarrow}\right)_{t\in[0,T]}$ is associated with the stochastic differential equation*

$$\mathrm{d}\boldsymbol{X}_t^{\leftarrow} = \left[-\boldsymbol{b}(\boldsymbol{X}_t^{\leftarrow}) + \nabla \log p_{T-t}(\boldsymbol{X}_t^{\leftarrow})\right]\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}. \tag{3.35}$$

Proof. Delayed to appendix C.3. ∎

This result is stronger than necessary for the purposes of generative modelling, but allows for more developed analysis to be carried out on a diffusion model. Theorem 3.5 is however a clear corollary of theorem 3.6 as the time marginal distributions are simply projections of the distribution on paths at a given time.

3.2.4. Score approximation on Riemannian manifolds. As with the time reversal in the Euclidean setting, the reverse process from eq. (3.35) involves the Stein score $\nabla \log p_t$ of the stochastic differential equation. We will need to again approximate this quantity through a score matching loss. This quantity at a specific time t is a map from a point on the manifold to a tangent vector at that point on the manifold, making it a vector field. As a function of time and a point on the manifold, we can see this as a *time-dependent* vector field on the manifold.

In this section we will assume that we have an $\mathcal{M}$-valued stochastic process $(\boldsymbol{X}_t)_{t\in[0,T]}$ defined by

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim p_0. \tag{3.36}$$

We denote

- $\mathbb{P}_t$ as the distribution of $\boldsymbol{X}_t$, with density p_t.
- $\mathbb{P}_{s,t}$ as the joint distribution of $\boldsymbol{X}_s, \boldsymbol{X}_t$, with density $p_{s,t}$.

[1] Note that in the case of a non-compact manifold vol_g is only a measure and not a probability measure.

- $\mathbb{P}_{t|s}$ as the *conditional distribution* (see appendix B.1) of $\boldsymbol{X}_t$ given $\boldsymbol{X}_s$, with density $p_{t|s}$.

We assume all the densities exist relative to the *Riemannian volume measure* (see appendix A.9.4), vol_g. We will let $\boldsymbol{s} : [0,T] \times \mathcal{M} \to T\mathcal{M}$ be an arbitrary smooth time-varying vector field on $\mathcal{M}$, representing the approximation we learn to the score function. We discuss methods of parametrising this vector field in section 3.2.8.

The *denoising score matching* (see section 2.2.1) loss carries over naturally to the manifold setting, nothing about its derivation presents difficulties when transiting to more general geometries. In the manifold setting the denoising score matching loss $\ell_{t|s}$ is given by

$$\ell_{t|s}(\boldsymbol{s}) = \int_{\mathcal{M}^2} \left\| \boldsymbol{s}(t, \boldsymbol{X}_t) - \nabla \log p_{t|s}(\boldsymbol{X}_t \mid \boldsymbol{X}_s) \right\|_g^2 \mathrm{d}\,\mathbb{P}_{s,t} \tag{3.37}$$

where $\|\cdot\|_g$ is the *Riemannian norm* (see appendix A.8.1). Unfortunately however, on general geometries, for the forward noising processes given by taking Langevin dynamics with either the Riemannian normal distribution or the wrapped Gaussian distribution the forward transition kernels, $p_{t|0}$, do not have analytic forms. This makes application of the denoising score matching loss on general geometries difficult unless the specific manifold and forward noising process in question has a tractable form.

We instead turn to *implicit score matching* (see theorem 2.7) as this avoids needing access to the forward transition kernel. The derivation of the implicit score matching loss does not carry over to the manifold setting as it uses calculus results specific to the Euclidean setting. Here we prove a version of the implicit score matching result suitable for manifolds.

Manifold implicit score matching

PROPOSITION 3.7. *Let $t, s \in (0,T]$ with $t > s$. Under sufficient regularity conditions of the product $p_{t|s}(\boldsymbol{X}_t|\boldsymbol{X}_s)\boldsymbol{s}(t, \boldsymbol{X}_t)$,*

$$\ell_{t|s}(\boldsymbol{s}_t) = 2\ell^{ism}(\boldsymbol{s}_t) + \int_{\mathcal{M}^2} \left\| \nabla_{\boldsymbol{X}_t} \log p_{t|s}(\boldsymbol{X}_t|\boldsymbol{X}_s) \right\|_g^2 \mathrm{d}\,\mathbb{P}_{s,t} \tag{3.38}$$

where $\ell^{ism}(\boldsymbol{s}_t)$ is given by

$$\ell^{ism}(\boldsymbol{s}_t) = \int_{\mathcal{M}} \left[\frac{1}{2} \|\boldsymbol{s}_t(\boldsymbol{V}_t)\|_g^2 + \mathrm{div}_g(\boldsymbol{s}_t)(\boldsymbol{X}_t) \right] \mathrm{d}\,\mathbb{P}_{s,t} \tag{3.39}$$

where $\|\cdot\|_g$ is the Riemannian norm and div_g the Riemannian divergence.

Proof. Appendix C.5 ∎

The main difference from the Euclidean case is the application of manifold divergence theorems on non-compact manifolds (Gaffney, 1954). The regularity conditions on $p_{t|s}(\boldsymbol{X}_t|\boldsymbol{X}_s)\boldsymbol{s}(t, \boldsymbol{X}_t)$ require that this product, its absolute value, and its determinant are L_1 integrable, a very mild condition usually satisfied by the field decaying suitably fast away from some point.

Since the second term is independent of $\boldsymbol{s}_t$ for any $t \in (0,T]$, the minimizers of the loss ℓ_t^{ism} on $\Gamma(T\mathcal{M})$ are the same as the ones for $\ell_{t|s}$.

This loss requires computing the Riemannian divergence of a vector field. How we compute this depends on the parametrisation of the vector field we use, and this is discussed in section 3.2.8.

Again as in the Euclidean case and following Song and Ermon (2020) and Nichol and Dhariwal (2021), these losses match the score for a single time t. The losses are therefore integrated over time and can be weighted with a term $\lambda_t > 0$ to produce a loss function that matches the score over all times.

3.2.5. Likelihood computation. Similarly to Song et al. (2020b), with access to $\nabla \log p_t$ we can compute the likelihood a Riemannian score-based diffusion model places on a particular sample.

Appendix B.9 shows how we can relate a stochastic differential equation with general coefficients,

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim p_0 \tag{3.40}$$

, to an ordinary differential equation with the same time-marginal distributions given by

$$\mathrm{d}\boldsymbol{X}_t = \left[\boldsymbol{b}(t, \boldsymbol{X}_t) - \frac{1}{2}\sigma(t)^2 \nabla \log p_t(\boldsymbol{X}_t)\right] \mathrm{d}t, \quad \boldsymbol{X}_0 \sim p_0. \tag{3.41}$$

Section 2.2.1 shows how we can also compute the instantaneous change in likelihood of an ordinary differential equation trajectory, defined by drift field $\boldsymbol{b} : \mathbb{R}^n \times [0, T]$, via

$$\frac{\mathrm{d}\log p_t(\boldsymbol{X}_t)}{\mathrm{d}t} = -\operatorname{div}(\boldsymbol{b})(\boldsymbol{X}_t, t). \tag{3.42}$$

Combining these two ordinary differential equations we can produce an ordinary differential equation that when run forward in time allows us to compute the likelihood of a given point, and when run backwards in time allows us to sample from the score based diffusion model.

We can quickly generalise this to the manifold setting. Previous works, such as Mathieu and Nickel (2020, proposition 2) for example have already shown that the manifold equivalent of eq. (3.42) is given by

$$\frac{\mathrm{d}\log p_t(\boldsymbol{X}_t)}{\mathrm{d}t} = -\operatorname{div}_g(\boldsymbol{b})(\boldsymbol{X}_t, t). \tag{3.43}$$

We can then also prove the following

Manifold time-marginal ordinary differential equation

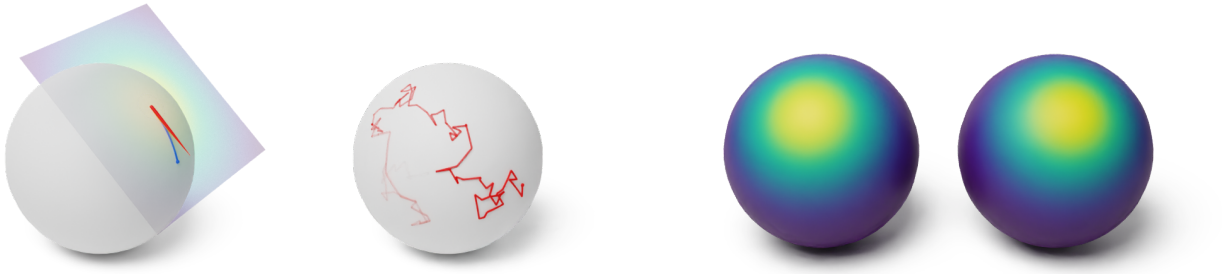
THEOREM 3.8. *For a manifold-valued stochastic differential equation*

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim p_0 \tag{3.44}$$

with time-marginal distributions $p_t(\boldsymbol{X})$, the ordinary differential equation

$$\mathrm{d}\boldsymbol{X}_t = \left[\boldsymbol{b}(t, \boldsymbol{X}_t) - \frac{1}{2}\sigma(t)^2 \nabla_g \log p_t(\boldsymbol{X}_t)\right] \mathrm{d}t, \quad \boldsymbol{X}_0 \sim p_0 \tag{3.45}$$

has the same time-marginal distributions $p_t(\boldsymbol{X})$ as the forward-time stochastic differential equation.



(a) A single step of a Geodesic Random Walk.

(b) Many steps yield an approximate trajectory.

(c) Gaussian random walk [Left] and the Brownian motion density [Right] agree well for small time steps.

Figure 3.1. Geodesic random walks can be used to approximate Brownian motion and more generally stochastic differential equations on manifolds. (a) At each step, tangential noise is sampled (red), which is added the drift term (not pictured). This tangent vector is then pushed through the exponential map to produce a geodesic step on the manifold (blue). (b) Iterating this procedure yield approximate sample paths from the process.

Proof. Apply the same method as theorem B.16, by applying the identity

$$\Delta_g f(\boldsymbol{X}) = \operatorname{div}(\nabla_g f(\boldsymbol{X})) = \operatorname{div}(f(\boldsymbol{X})\nabla_g \log f(\boldsymbol{X})) \tag{3.46}$$

from theorem 3.5. ∎

Using this we arrive at a *likelihood ordinary differential equation for Riemannian score-based generative models* by solving the joint equation

Likelihood ordinary differential equation for Riemannian score-based generative models

$$\begin{bmatrix} \mathrm{d}\boldsymbol{X}_t \\ \mathrm{d}\log p_t(\boldsymbol{X}_t) \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}(t, \boldsymbol{X}_t) - \frac{1}{2}\sigma(t)^2 \nabla \log p_t(\boldsymbol{X}_t) \\ -\operatorname{div}_g(\boldsymbol{b})(\boldsymbol{X}_t, t) \end{bmatrix} \mathrm{d}t. \tag{3.47}$$

3.2.6. Approximate sampling of diffusions. Obtaining samples from stochastic differential equations on a manifold is non-trivial in general. As in the Euclidean case, we cannot sample the continuous time dynamics of a manifold valued stochastic process, and so must resort to finite time step approximations.

The simplest approach would be to isometrically embed the manifold into Euclidean space and apply the Euclidean space techniques of appendix B.10 to sample the resulting Euclidean stochastic differential, an *extrinsic* approach. Unfortunately with a discrete step size, the discretisation is guaranteed almost surely to leave the surface of the manifold, where the process becomes undefined. While this can be corrected by taking small step sizes and projecting samples back onto the surface of the manifold, these projections can be expensive and the procedure inaccurate.

Here instead we consider an *intrinsic* approach based on geodesic random walks, see Jørgensen (1975) for a review of their properties. Geodesic random walks can approximate *any* well-behaved diffusion on $\mathcal{M}$. Hence, we introduce them in a general framework and consider discretisations of any $\mathcal{M}$-valued stochastic differential equation. This generalisation is key to sampling the backward diffusion process defined in theorem 3.6.

Geodesic random walk

DEFINITION 3.9. *Let $\boldsymbol{X}_0^\gamma \sim p_0$ be an $\mathcal{M}$-valued random variable. Let $\boldsymbol{b} : \mathbb{R}^+ \times \mathcal{M} \to T\mathcal{M}$ be a drift function and $\sigma : \mathbb{R}^+ \to \mathbb{R}^+$ a noise term. Let $\boldsymbol{X}_t$ be a manifold-valued stochastic differential equation governed by*

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}. \tag{3.48}$$

For any $\gamma > 0$, we define $(\boldsymbol{X}_k^\gamma)_{k\in\mathbb{N},k>0}$ such that

$$\boldsymbol{X}_{k+1}^\gamma = \exp_{\boldsymbol{X}_k^\gamma}\left[\gamma b(k\gamma, \boldsymbol{X}_k^\gamma) + \sqrt{\gamma}\boldsymbol{V}_{k+1}\right], \tag{3.49}$$

where $(\boldsymbol{V}_k)_{k\in\mathbb{N}}$ is a sequence of $T_{\boldsymbol{X}_k^\gamma}\mathcal{M}$-valued random variables such that for any $k \in \mathbb{N}$,

$$\mathbb{E}[\boldsymbol{V}_{k+1} \mid \mathcal{F}_n] = 0 \text{ and } \mathbb{E}\left[\boldsymbol{V}_{k+1}\boldsymbol{V}_{k+1}^\top \,\middle|\, \mathcal{F}_k\right] = \sigma\sigma^\top, \tag{3.50}$$

where $\mathcal{F}_k$ is the filtration generated by $\left\{\boldsymbol{X}_k^\gamma\right\}_{k=0}^n$.

We say that the $\mathcal{M}$-valued process $(\boldsymbol{X}_k^\gamma)_{k\in\mathbb{N}}$ is a geodesic random walk of eq. (3.48).

This definition gives us a clear blueprint to sample $\mathcal{M}$-valued stochastic differential equations. We initialise the geodesic random walk with $\boldsymbol{X}_0^\gamma \sim p_0$, and then iterate the geodesic random walk using the drift and noise coefficients of the stochastic differential equation.

Algorithm 3.10 summarises the implementation of this algorithm for sampling the stochastic differential equation. Figure 3.1 provides a graphical illustration of this procedure.

Importantly, there are results available for this scheme that guarantee a certain level of accuracy between the discretised process and the continuous process. For example, see Kuwada (2012) and Cheng et al. (2022) for quantitative error bounds in the time-homogeneous case. These error bounds depend on a number of constants encoding the curvature of the manifold, but importantly in order to ensure that the expected distance between the approximate dynamics and the continuous time dynamics are bounded by ϵ, one needs to use a number of discretisation steps that scales with $\frac{1}{\epsilon^2}$. This rate is the same order as required to achieve the same accuracy with Euler-Maruyama discretisation of Euclidean stochastic differential equations, and so we do not lose any sampling efficiency by being in the manifold setting. We can therefore, with increasing numbers of steps, obtain arbitrary precision for the use case we have.

In addition, we extend these results of Cheng et al. (2022) to cover the time-inhomogeneous setting that we use in diffusion modelled. See appendix C.4.2 for these results.

3.2.7. Predictor-corrector scheme.

Predictor-corrector scheme

In addition to presenting geodesic random walks as a general framework for discretising stochastic differential equations on manifolds, in this section we present a *predictor-corrector scheme*, adapting the techniques of Allgower and Georg (2012) and Song et al. (2020b) to the manifold setting.

For a given step in the geodesic random walk approximation of a stochastic differential equation, the variable $\boldsymbol{X}_k^\gamma$ may not be exactly sampled from the distribution

Algorithm 3.10 Geodesic Random Walk

Require: $T, K, X_0^\gamma, \boldsymbol{b}, \sigma, \mathrm{P}$
1: $\gamma = T/K$ ▷ Step-size
2: **for** $k \in \{0, \dots, K-1\}$ **do**
3: $\quad Z_{k+1} \sim \mathrm{N}(0, \mathbf{I})$ ▷ Sample a Gaussian in the tangent space of X_k^γ
4: $\quad W_{k+1} = \gamma \boldsymbol{b}(k\gamma, X_k^\gamma) + \sqrt{\gamma}\sigma(k\gamma, X_k^\gamma) Z_{k+1}$ ▷ Compute the Euler–Maruyama step on tangent space
5: $\quad X_{k+1}^\gamma = \exp_{X_k^\gamma}[W_{k+1}]$ ▷ Move along the geodesic defined by W_{k+1} and X_k^γ on $\mathcal{M}$
6: **return** $\{X_k^\gamma\}_{k=0}^K$

of its continuous time counterpart, $X_{\gamma k}$. The aim of the corrector step is to try and reduce this error.

Let $k \in \mathbb{N}$ be a step in a geodesic random walk with step size γ. Let X_k^γ be a sample from the geodesic random walk, and approximate sample of $X_{k\gamma}$. Let $p_{k\gamma}$ be the density of $X_{k\gamma}$. We introduce a second geodesic random walk defined by

$$X_{k,k'+1}^{\gamma,\gamma'} = \exp_{X_{k,k'}^{\gamma,\gamma'}}\left[\gamma' \nabla_g \log p_{k\gamma}(k\gamma, X_{k,k'}^{\gamma,\gamma'}) + \sqrt{2\gamma'} Z_{k'}\right], \tag{3.51}$$

where $Z_{k'}$ is a unit variance Gaussian vector.

Letting $\gamma' \to 0$, we obtain that under mild assumptions, see (Kuwada, 2012, Theorem 3.1), $\left(X_{k,k'}^{\gamma,\gamma'}\right)_{k\in\mathbb{N}}$ converges to the stochastic differential equation

$$\mathrm{d}X_s^{k\gamma} = \tfrac{1}{2}\nabla \log p_{k\gamma}(X_s^{k\gamma})\,\mathrm{d}s + \mathrm{d}B_s^{\mathcal{M}} \quad X_0^{k\gamma} = X_k^\gamma. \tag{3.52}$$

This is a Langevin dynamics with stable distribution $p_{k\gamma}$, as desired. By running this geodesic random walk for small step sizes between steps of the main geodesic random walk, we can improve the overall sampling quality. Typically, we might run a single step of correction for every step of the geodesic random walks. This scheme is the manifold equivalent of the predictor-corrector scheme found in Song et al. (2020b).

3.2.8. Parametric family of vector fields.. Finally, we need to choose a method to approximate $(\nabla \log p_t)_{t\in[0,T]}$ by a parametrised function $\mathbf{s}_\theta : [0,T]\times\mathcal{M} \to \Gamma(T\mathcal{M})$, that is differentiable with respect to its parameters.

Coordinate vector fields

One strategy is to parametrise the score function in a chart to give *coordinate vector fields*. If we choose a chart (ϕ, U) we can define a set of local vector fields $\{E_i\}_{i=1}^d = \{\partial_i \phi\}_{i=1}^d$. These span the tangent bundle inside the chart, but not outside it. Coordinate vector fields

We can then define a score function on the chart as

$$\boldsymbol{s}_\theta(t, \boldsymbol{X}) = \sum_{i=1}^d \bar{\boldsymbol{s}}_\theta^i(t, \boldsymbol{X}) E_i(\boldsymbol{X}) \tag{3.53}$$

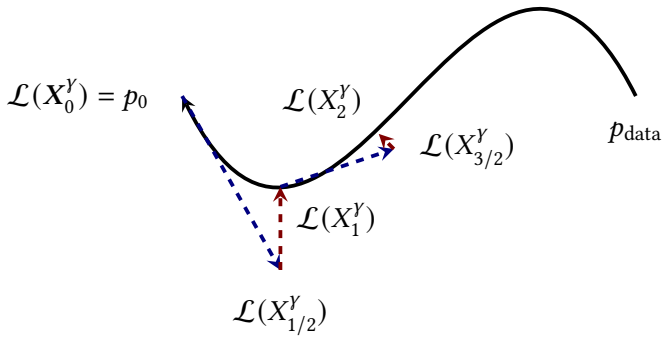


Figure 3.2. Illustration of the effect of a corrector step on stochastic differential equation rollouts. The black line corresponds to the dynamics of the noising process $(p_t)_{t\in[0,T]}$. The blue dashed lines correspond to the predictor step and the red dashed lines correspond to the corrector step (projecting back onto the initial dynamics). Steps from whole integer to half integers are prediction steps, and from half to whole correction steps.

where $\bar{\mathbf{s}}^i_\theta(t, \boldsymbol{X}) : [0,T] \times \mathcal{M} \to \mathbb{R}^d$ is a parametrised neural network.

To compute the divergence of this we can use the coordinate expression for the divergence in a chart (ϕ, U), see eq. (A.155),

$$\operatorname{div} V = \frac{1}{\sqrt{\det g}} \partial\phi_i \left(\sqrt{\det g}\, V_i \right). \tag{3.54}$$

We can use an automatic differentiation package to compute this divergence. This is achieved by interpreting $\partial\phi_i \left(\sqrt{\det g}\, V_i \right)$ as the Euclidean divergence of the vector field

$$V(\phi_1, ..., \phi_n) = \begin{bmatrix} \sqrt{\det g(\phi_1, ..., \phi_n)}\, V_1(\phi_1, ..., \phi_n) \\ \vdots \\ \sqrt{\det g(\phi_1, ..., \phi_n)}\, V_n(\phi_1, ..., \phi_n) \end{bmatrix}, \tag{3.55}$$

and this can be computed by an automatic differentiation package if we have the expression for the metric and the components of the vector field in the chart.

As in the Euclidean case the computation of a divergence in an automatic differentiation package requires d Jacobian-vector calls. This can again be sped up with Hutchinson's trace trick (Hutchinson, 1989) by applying it to the computation of the Euclidean divergence component of eq. (3.55).

The main issue with this approach is patching together different charts. Ideally we would have a chart that covers the manifold except for a set of measure zero.

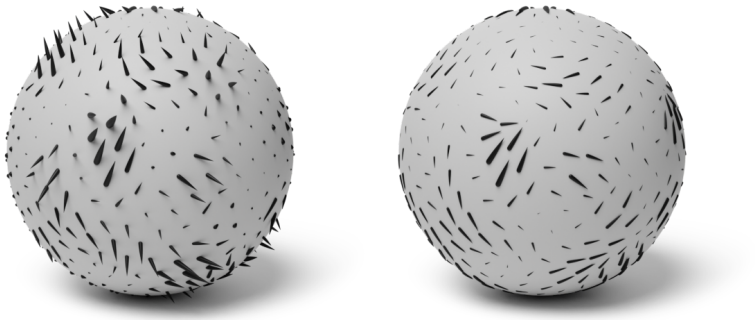


Figure 3.3. A sphere isometrically embedded into Euclidean space. An ambient Euclidean vector field (left) projected onto the surface of the sphere (right).

Examples of charts that could be used on the sphere in this way are latitude-longitude charts on the sphere, or polar coordinates. Even with such a chart however the vector fields will be non-smooth around the seams of the charts.

Projected vector fields

A second approach is to use *projected vector fields*. Let $\mathcal{M}$ be a d-dimension Riemannian manifold and let Projected vector fields

$$F : \mathcal{M} \to \mathbb{R}^n \qquad \mathrm{d}F_p : T_p\mathcal{M} \to T\mathbb{R}^n \tag{3.56}$$

be an *isometric embedding* (see appendix A.8.3) into $\mathbb{R}^n$, guaranteed by the *Nash embedding theorem* (see theorem A.37). Choosing a basis $e_i(p)$ on $T_p\mathcal{M}$, and using the canonical basis on $T_{F(p)}\mathbb{R}^n$, ϵ_i, we can create a *projection matrix*, $P_F : \mathcal{M} \to \mathbb{R}^{d\times n}$, defined by Projection matrix

$$P_F(p)^i_j = \left\langle \mathrm{d}F_p(e_i), \epsilon_j \right\rangle. \tag{3.57}$$

This projects a tangent vector in $T_{F(p)}\mathbb{R}^n$, $v = \sum_{i=1}^n v^i\epsilon_i$ into a tangent vector in $w \in T_p\mathcal{M}$, $w = \sum_{i=1}^d w^i e_i(p)$, via

$$w^i = \sum_{j=1}^n P_F^T(p)^i_j v^j. \tag{3.58}$$

The transpose of this map,

$$P_F(p)^i_j = \left\langle \mathrm{d}F_p(e_j), \epsilon_i \right\rangle, \tag{3.59}$$

allows us to project a tangent vector $w \in T_p\mathcal{M}, w = \sum_{i=1}^d w^i e_i(p)$ to $v \in \mathrm{d}F(T_p\mathcal{M}) \subset T_{F(p)}\mathbb{R}^n$, $v = \sum_{i=1}^n v^i\epsilon_i$, via

$$v^i = \sum_{j=1}^n P_F^T(p)^i_j w^j. \tag{3.60}$$

We can additionally use this two maps to compute the *tangential projection* of a vector $v \in T_{F(p)}\mathbb{R}^n$ into the linear subspace $\mathrm{d}F_p(T_p\mathcal{M}) \subset T_{F(p)}\mathbb{R}^n$, via the coordinate expression Tangential projection

$$\tilde{v}^i = \sum_{j=1}^d \sum_{k=1}^n P_F^T(p)_{ij} P_F(p)_{jk} v^k. \tag{3.61}$$

This construction is in fact independent of the chosen basis on $T_p\mathcal{M}$ and so gives a coordinate-free projection, see figure 3.3.

We can construct a score network via

$$\boldsymbol{s}_\theta(t, \boldsymbol{X}) = \sum_{i=1}^{d} \sum_{j=1}^{n} P_F(\boldsymbol{X})^i_j \bar{\boldsymbol{s}}^j_\theta(t, \boldsymbol{X}) e_i(\boldsymbol{X}) \tag{3.62}$$

where $\bar{\boldsymbol{s}}^j_\theta(t, \boldsymbol{X}) : [0, T] \times \mathcal{M} \to \mathbb{R}^n$ is a parametrised neural network specifying a vector field in the ambient Euclidean space on the surface of the manifold.

We can compute the divergence of this vector field by computing the Euclidean divergence of this vector field projected back into Euclidean space via the projection

$$\sum_{k=1}^{n} \sum_{i=1}^{d} \sum_{j=1}^{n} P_F^T(p)^k_i P_F(\boldsymbol{X})^i_j \bar{\boldsymbol{s}}^j_\theta(t, \boldsymbol{X}) epsilon_k(F(\boldsymbol{X})) \tag{3.63}$$

as these divergences coincide (Huang et al., 2022).

The main drawback of this approach is that it relies on a known isometric embedding of the manifold, which may not always be available.

Global basis fields

Global basis fields

A final most general strategy is to parametrise a vector field as a finite sum of *global basis fields* weighted by smooth functions on the manifold. That is,

$$V = f^i E_i, \quad f_i \in C^\infty(\mathcal{M}),\ E_i \in \Gamma(T\mathcal{M}). \tag{3.64}$$

Parallelisable

In d-dimension Euclidean space there exists a set of d smooth basis fields $(E_i)_{i=1}^d$ which space the tangent space, as it is *parallelisable* (see appendix A.6.1), creating a bijection between the tangent bundle and $C^\infty(\mathcal{M}, \mathbb{R}^d)$. We can then express the score function as

$$\boldsymbol{s}_\theta(t, \boldsymbol{X}) = \sum_{i=1}^{d} \bar{\boldsymbol{s}}^i_\theta(t, \boldsymbol{X}) E_i(\boldsymbol{X}) \tag{3.65}$$

where $\bar{\boldsymbol{s}}^j_\theta(t, \boldsymbol{X}) : [0, T] \times \mathcal{M} \to \mathbb{R}^n$ is a parametrised neural network. This is not true in general for manifolds, most are *non-parallelisable*, and the space of vector fields forms a module, not a vector space (see appendix A.6.1) meaning we do not have a Hammel basis for the space of vector fields.

Finitely generated

Fortunately however the space of vector fields is *finitely generated* (see definition A.24) for any Riemannian manifold, meaning that can find a set of fields that do span the tangent space, but with some redundancy. We can rely on this larger set of smooth vector fields $\{E_i(x)\}_{i=1}^n$ with $n > d$ to parametrise sections of the tangent bundle. This does not create a bijection between the tangent bundle and $C^\infty(\mathcal{M}, \mathbb{R}^n)$, but does create an injection, so any vector field can be represented. Then it suffices to construct a smooth neural network $\bar{\boldsymbol{s}}_\theta : [0, T] \times \mathcal{M} \to \mathbb{R}^n$ to parametrise the score network as

$$\boldsymbol{s}_\theta(t, x) = \sum_{i=1}^{n} \bar{\boldsymbol{s}}^i_\theta(t, \boldsymbol{X}) E_i(x). \tag{3.66}$$

The divergence of such a vector field can be computed by

$$\operatorname{div}_g V = \operatorname{div}_g \sum_{i=1}^{n} f_i E_i \tag{3.67}$$

$$= \sum_{i=1}^{n} E_i(f_i) + \sum_{i=1}^{n} f_i \operatorname{div}_g E_i \tag{3.68}$$

$$= \sum_{i=1}^{n} \mathrm{d}f_i(E_i) + \sum_{i=1}^{n} f_i \operatorname{div}_g(E_i). \tag{3.69}$$

$\mathrm{d}f_i$ can be computed with automatic differentiation, although it requires n backward passes, and we look for sets of vector fields where $\operatorname{div}_g E_i$ is simple to compute (Falorsi and Forré, 2020, chapter 5).

Divergence-free vector fields

We use one example of this approach for for Lie groups. For any Lie group G we can use a set of *divergence-free vector fields*. We can immediately use a global basis of smooth frames of size d as Lie groups are parallelisable. This is useful as we can define a smooth score function using d basis fields.

To construct this let

$$L_g : G \to G, \quad L_g(g') = g \cdot g' \tag{3.70}$$

the left translation of a Lie group. We call a vector field $V \in \Gamma(TG)$ on a Lie group left-invariant if

$$(\mathrm{d}L_g)_{g'}(V_{g'} = V_{g \cdot g'}). \tag{3.71}$$

For any vector $v \in T_eG$, where $e \in G$ is the identity element, there exists a left-invariant vector field $V \in \Gamma(TG)$ (Lee, 2013, Chapter 8). We can therefore use any basis of a single tangent space T_eG, $(v_i)_{i=1}^n$, to create a left-invariant global basis of $\Gamma(TG)$, $(E_i)_{i=1}^d$ such that $E_i(e) = v_i$.

As a result of the left invariance of the basis fields, we have that for any $i \in \{1, \dots, d\}$, $\operatorname{div}_g(E_i) = 0$, simplifying the divergence computation in eq. (3.69). Note that this simplified divergence computation can be extended to any homogeneous space of G by projecting the divergence free vector fields from the Lie group onto the homogeneous space, where they remain divergence free.

Further examples of this approach can be found in Falorsi and Forré (2020, chapter 5).

3.2.9. Riemannian score-based generative models. Combining the parametrisation of the score function in section 3.2.8 with the score matching losses of section 3.2.4, the time-reversal formula of theorem 3.6 and the sampling of forward and backward processes described in section 3.2.6, we define our Riemannian score-based generative modelling algorithm in algorithm 3.11. This algorithm can also benefit from a predictor-corrector scheme as in Song et al. (2020b), detail in section 3.2.7.

In the next two sections we propose two simple additions to this model.

Riemannian score-based generative models

Algorithm 3.11

Require: $\varepsilon, T, N, \{X_0^m\}_{m=1}^M$, loss, $\mathbf{s}$, θ_0, N_{iter}, p_{inv}, P

/// TRAINING ///
for $n \in \{0, \dots, N_{\text{iter}} - 1\}$ **do**
  $X_0 \sim (1/M) \sum_{m=1}^M \delta_{X_0^m}$ ▷ Random mini-batch from dataset
  $t \sim U([\varepsilon, T])$ ▷ Uniform sampling between ε and T
  $\boldsymbol{X}_t = \text{GRW}(t, N, X_0, b, \mathbf{I}, \text{P})$ ▷ Approximate forward diffusion with algorithm 3.10
  $\ell(\theta_n) = \ell_t(T, N, X_0, \boldsymbol{X}_t, \text{loss}, \mathbf{s}_{\theta_n})$ ▷ Compute score matching loss from table 3.1
  $\theta_{n+1} = \texttt{optimizer\_update}(\theta_n, \ell(\theta_n))$ ▷ ADAM optimizer step
$\theta^\star = \theta_{N_{\text{epoch}}}$
/// SAMPLING ///
$Y_0 \sim p_{\text{inv}}$ ▷ Sample from uniform distribution
$b_\theta^\star(t, x) = \mathbf{s}_{\theta^\star}(T - t, x)$ for any $t \in [0, T]$, $x \in \mathcal{M}$ ▷ Reverse process drift
$\{Y_k\}_{k=0}^N = \text{GRW}(T, N, Y_0, b_{\theta^\star}, \mathbf{I}, \text{P})$ ▷ Approximate reverse diffusion with Algorithm 3.10
return $\theta^\star$, $\{Y_k\}_{k=0}^N$

3.2.10. Amortised conditional modelling. We can easily extend Riemannian score-based generative models to the conditional sampling setting. By amortising the conditioning of score-based models with respect to an observation y it is possible to approximately sample from a given posterior distribution.

In the Euclidean setting this idea has been successfully applied for several image processing problems such as de-blurring, denoising or in-painting (see for instance Kawar et al., 2021a; Kawar et al., 2021b; Lee et al., 2022; Sinha et al., 2021; Batzolis et al., 2021; Chung et al., 2022).

Similarly, RSGM can be amortised to handle such situations in the case where the underlying posterior distribution is supported on a manifold. Practically, this requires for the score network takes an additional input, i.e $\mathbf{s}_\theta(t, x; y)$, and we simply train this score function with standard score-matching losses.

3.2.11. Invariant distribution modelling. We can also extend Riemannian score-based generative models for modelling probability distributions with known invariances. Let G be a subgroup of the Riemannian isometries of the manifold being modelled on, Isom_g. We assume that a given data distribution, π is invariant under the action of this group. That is

$$\pi(g \cdot x) = \pi(x) \tag{3.72}$$

for all $g \in G$ and $x \in \mathcal{M}$. Following Köhler et al. (2020), we have that if p_{prior}, a reference distribution, is invariant with respect to G and $\phi : \mathcal{M} \to \mathcal{M}$ is equivariant with respect to G, then the pushforward probability density $p = p_{\text{prior}} \circ \phi^{-1}$ is invariant with respect to G.

Let us consider the probability flow ϕ associated with the reverse diffusion, eq. (3.35),

$$\mathrm{d}X_t^\leftarrow = \left[-\boldsymbol{b}(X_t^\leftarrow) + \nabla \log p_{T-t}(X_t^\leftarrow)\right] \mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}. \tag{3.73}$$

We know this is also given by the solution of the following ordinary differential

equation, see section 3.2.5,

$$\mathrm{d}\boldsymbol{X}_t = \left[\boldsymbol{b}(t, \boldsymbol{X}_t) - \frac{1}{2}\sigma(t)^2 \nabla \log p_t(\boldsymbol{X}_t)\right] \mathrm{d}t. \tag{3.74}$$

In order to make the solutions to this ordinary differential equation equivariant it is sufficient to make the drift of the ordinary differential equation equivariant (Köhler et al., 2020). This invariance of the modelled density transfers to the reverse diffusion map as it shares the same time-marginals as the likelihood flow.

To enforce this we require that the drift term $\boldsymbol{b}$ and the Stein score approximation with a score network $\boldsymbol{s}_\theta(t, \cdot)$ are both equivariant. If we are using a Langevin dynamics with an invariant target as the noising process then by construction the drift term will be equivariant to the same transformations as the targeted distribution is invariant. It is sufficient then to parametrise the score network so that it is equivariant with respect to its manifold valued argument.

3.3. Riemannian Score-Based Generative Modelling on Compact Manifolds

Assuming compactness of the manifold $\mathcal{M}$, we can leverage a number of special properties to implement a specific case of our algorithm.

3.3.1. Forward noising process. In the compact case we can choose $U(x) = 0$, giving Brownian motion,

$$\mathrm{d}\boldsymbol{X}_t = \mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}. \tag{3.75}$$

This will converge to a density proportional to the Riemannian volume form, simply the *uniform distribution on a compact manifold* given by $\mu = \frac{\mathrm{vol}_g}{\int_{\mathcal{M}} \mathrm{d}\,\mathrm{vol}_g}$. This is well-defined as a compact manifold has finite volume. Regardless of the starting distribution the stochastic differential equation will converge to the uniform distribution exponentially quickly (Baudoin, 2013).

Uniform distribution on a compact manifold

3.3.2. Learning the score. Targeting the uniform distribution is additionally useful as the transition kernel of the resulting Langevin stochastic differential equation is the heat kernel.

Contrary to the Gaussian transition density of Brownian motion in the Euclidean setting, it is typically only available as an infinite series on compact manifolds.

In order to circumvent this issue we consider two techniques: i) a truncation of the Sturm–Liouville decomposition, ii) a Taylor expansion around $t = 0$ called a Varadhan asymptotics.

Sturm–Liouville decomposition

In the case of compact manifolds the heat kernel can be represented by a *Sturm–Liouville decomposition*, see theorem A.58. For any $t > 0$ and $x_0, x_t \in \mathcal{M}$ the

Sturm–Liouville decomposition

expansion of the heat kernel is given by

$$p_{t|0}(x_t|x_0) = \sum_{j\in\mathbb{N}} \mathrm{e}^{-\lambda_j t}\phi_j(x_0)\phi_j(x_t). \tag{3.76}$$

The convergence occurs in $\mathrm{L}^2(\mathrm{vol}_g \otimes \mathrm{vol}_g)$, and $(\lambda_j)_{j\in\mathbb{N}}$ and $(\phi_j)_{j\in\mathbb{N}}$ are the eigenvalues, respectively the eigenvectors of $-\Delta_g$, the *Laplace-Beltrami operator* (see appendix A.10.8) on the manifold (Saloff-Coste, 1994, Section 2). When the eigenvalues and eigenvectors are known, we rely on an approximation of the logarithmic gradient of $p_{t|0}$ by truncating the sum in eq. (3.76) with $J \in \mathbb{N}$ terms to obtain for any $t > 0$ and $x_0, x_t \in \mathcal{M}$

$$\nabla_{x_t} \log p_{t|0}(x_t|x_0) \approx S_{J,t}(x_0, x_t) \triangleq \nabla_{x_t} \log \sum_{j=0}^{J} \mathrm{e}^{-\lambda_j t}\phi_j(x_0)\phi_j(x_t). \tag{3.77}$$

Under regularity conditions on $\mathcal{M}$ it can be shown that for any $x, y \in \mathcal{M}$ and $t \geq 0$, $\lim_{J\to+\infty} S_{J,t}(x_0, x_t) = \nabla_{x_t} \log p_{t|0}(x_t|x_0)$ (Jones et al., 2008, Lemma 1). For many compact manifolds then eigenvalues and eigenvectors are computable, see appendix A.10.9 for examples of d-dimension tori and sphere.

Varadhan's asymptotics

Varadhan's asymptotics

When the eigenvalues and eigenvectors are unknown or not tractable, we can still derive an approximation of the heat kernel for small times t. Using *Varadhan's asymptotics*, see Bismut (1984, Theorem 3.8) or Chen et al. (2023, Theorem 2.1), for any $x, y \in \mathcal{M}$ with $y \notin \mathrm{Cut}(x)$ (where $\mathrm{Cut}(x)$ is the cut-locus of x in $\mathcal{M}$ (Lee, 2018, Chapter 10)) we have that

$$\lim_{t\to 0} t\nabla_{x_t} \log p_{t|0}(x_t|x_0) = \exp_{x_t}^{-1}(x_0). \tag{3.78}$$

It should be noted that Varadhan's asymptotics are only accurate over small time scales, and therefore should only be used with $\ell_{t|s}$ for small $t - s$ to preserve accuracy.

Using the previously defined score-matching losses for generic manifolds and the approximations to the heat kernel in this section, we highlight the various methods to compute $\nabla \log p_t$ in table 3.1, along with the requirements to use the losses and their computational complexity.

Mixture of approximations

One final improvement we can make to these approximations is to use a suitable mixture of approximations. As demonstrated in figure 3.4, at small diffusion times Varadhan's asymptotics are significantly more accurate than the Sturm–Liouville decomposition with a small truncation. At larger times, the Sturm–Liouville decomposition is accurate even with a small truncation, but Varadhan's asymptotics are quite inaccurate. Using these opposing accuracies, we can first choose a number of basis functions we wish to use to truncate the heat kernel with, and then numerically find the diffusion time where Varadhan's asymptotics becomes more accurate than the expansion. We can then choose during each training step the more accurate expression of the heat kernel for the given time step of the diffusion.

Loss	Approximation	Loss function	Requirements $p_{t\|0}$	Requirements $\exp_{X_t}^{-1}$	Complexity
$\ell_{t\|0}$ (DSM)	None	$\frac{1}{2}\mathbb{E}\left[\|\mathbf{s}(X_t) - \nabla \log p_{t\|0}(X_t\|X_0)\|^2\right]$	✓	✗	$O(1)$
	Truncation (3.77)	$\frac{1}{2}\mathbb{E}\left[\|\mathbf{s}(X_t) - S_{J,t}(X_0, X_t)\|^2\right]$	asymptotic expansion	✗	$O(1)$
	Varhadan (3.78)	$\frac{1}{2}\mathbb{E}\left[\|\mathbf{s}(X_t) - \exp_{X_t}^{-1}(X_0)/t\|^2\right]$	✗	✓	$O(1)$
$\ell_{t\|s}$ (DSM)	Varhadan (3.78)	$\frac{1}{2}\mathbb{E}\left[\|\mathbf{s}(X_t) - \exp_{X_t}^{-1}(X_s)/(t-s)\|^2\right]$	✗	✓	$O(1)$
ℓ_t^{im} (ISM)	Deterministic	$\mathbb{E}\left[\frac{1}{2}\|\mathbf{s}(X_t)\|^2 + \operatorname{div}(\mathbf{s})(X_t)\right]$	✗	✗	$O(d)$
	Stochastic	$\mathbb{E}\left[\frac{1}{2}\|\mathbf{s}(X_t)\|^2 + \varepsilon^\top \partial\mathbf{s}(X_t)\varepsilon\right]$	✗	✗	$O(1)$

Table 3.1. Computational complexity of score matching losses with respect to score network forward and backward passes. ε is a random variable on $\mathrm{T}_{X_t}\mathcal{M}$ such that $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon\varepsilon^\top] = \mathbf{I}$.

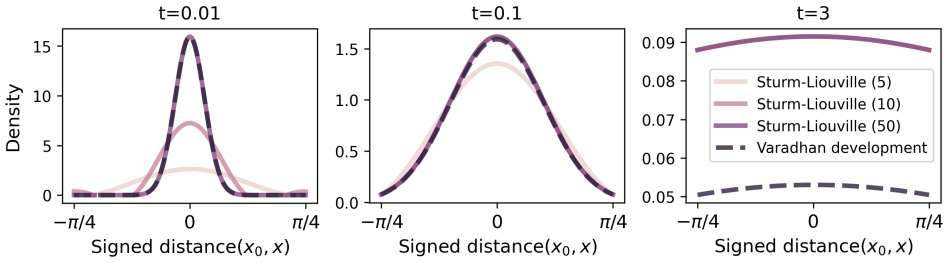


Figure 3.4. Slice of heat kernel $p_{t|0}(x_t|x_0)$ on $\mathbb{S}^2$ for different approximations.

3.3.3. Convergence results in the compact setting. We now provide a theoretical analysis of Riemannian score-based generative models under the assumption that $\mathcal{M}$ is compact. The following result ensures that Riemannian score-based generative models generates samples whose distribution is close to the data distribution p_0. Let us denote $\{X_k^Y\}_{n\in\{0,\dots,N\}}$ the sequence generated by Algorithm 3.11. This result relies on the following assumption, which is satisfied for a large class of manifolds $\mathcal{M}$ such as the d-dimensional sphere and torus, compact matrix groups and products of these manifolds.

ASSUMPTION 3.12. *There exist $C, \alpha > 0$ such that for any $t \in (0, 1]$ and $x \in \mathcal{M}$, $p_{t|0}(x|x) \leq Ct^{-\alpha/2}$, where $p_{t|0}(\cdot|x_0)$ is the density of the heat kernel, i.e. the density of $\boldsymbol{B}_t^{\mathcal{M}}$ with initial condition x_0* [2].

Error bound on the learnt density of a Riemannian socre-based gerneative model

THEOREM 3.13. *Assume assumption 3.12, that p_0 is smooth and positive and that there exists $\mathtt{M} \geq 0$ such that for any $t \in [0, T]$ and $x \in \mathcal{M}$, $\|\mathbf{s}_{\theta^\star}(t, x) - \nabla \log p_t(x)\| \leq \mathtt{M}$, with $\mathbf{s}_{\theta^\star} \in \mathrm{c}([0, T], \mathcal{X}(\mathcal{M}))$. Then if $T > 1/2$, there exists $C \geq 0$ independent on T such that*

$$\boldsymbol{W}_1\big(\mathcal{L}(X_N^Y), p_0\big) = C(\mathrm{e}^{-\lambda_1 T} + \sqrt{T/2}\mathtt{M} + \mathrm{e}^T\gamma^{1/2}), \tag{3.79}$$

where $\boldsymbol{W}_1$ is the Wasserstein distance of order one on the probability measures on $\mathcal{M}$.

The proof is postponed to Appendix C.4. In particular, for any $\varepsilon > 0$, choosing

[2] The diagonal upper-bound is implied by Sobolev inequalities which control of the growth of some functions by the growth of their gradient. Assumption 3.12 is satisfied in our experiments, see Saloff-Coste (1994) and Gross (1992).

Ingredient \ Space	Euclidean	'Generic' Manifold	Compact Manifold
Forward process $\mathrm{d}\boldsymbol{X}_t =$	$-\frac{1}{2}\boldsymbol{X}_t \mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}$	$-\frac{1}{2}\nabla_{\boldsymbol{X}_t} U(\boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}$	$\mathrm{d}\boldsymbol{B}_t^{\mathcal{M}}$
Easy-to-sample distribution	Gaussian	Wrapped Gaussian	Uniform
Time reversal	Cattiaux et al. (2023)	Theorem 3.6	
Sampling forward process	Direct	Geodesic Random Walk (Algorithm 3.10)	
Sampling backward process	Euler–Maruyama	Geodesic Random Walk (Algorithm 3.10)	

Table 3.2. Differences between SGM on Euclidean spaces and Riemannian score-based generative models on Riemannian manifolds.

$T > 0$ large enough, $\mathtt{M}$ small enough (which can be achieved using the universal property of neural networks) and γ small enough, we get that $\boldsymbol{W}_1\big(\mathcal{L}(\boldsymbol{X}_N^{\leftarrow}), p_0\big) \leq \varepsilon$. This result might seem weaker than the result obtained for Moser flows in Rozen et al. (2021, Theorem 3), but we emphasize that our bound takes into account the time-discretization contrary to Rozen et al. (2021) which considers the continuous-time flow. If we consider the time-reversed continuous-time stochastic differential equation then we recover a bound in total variation distance, see Appendix C.4. Note that the upper bound $\mathtt{M}$ encompasses both the bias introduced by the use of a neural network and the bias introduced by the use of an approximation of the score.

3.3.4. Comparison of score-based model types. Finally, in table 3.2 we summarised the main differences between score based models on Euclidean space, general Riemannian manifolds, and compact Riemannian manifolds.

3.4. Experiments

In this section we benchmark the empirical performance of Riemannian score-based generative models along with other manifold-valued methods, Peel et al. (2001), Rozen et al. (2021), and Mathieu et al. (2019). We also compare to a 'stereographic' score-based model, introduced in section 3.4.1. First, we assess their modelling capacity on earth and climate science spherical data. Then, we test the methods' scalability with respect to manifold dimensions with a synthetic experiment on the torus $\mathbb{T}^d$. Finally, we evaluate the models' regularity and time complexity with a synthetic $\mathrm{SO}_3(\mathbb{R})$ target. Additional experimental details are provided in Appendix C.6. The code used to run the experiments can be found at https://github.com/oxcsml/riemannian-score-sde.

Stereographic baseline model

3.4.1. Stereographic baseline method. Here we briefly present the *stereographic baseline model* used in our experiments. This is a novel method, and an alternative to a fully intrinsic Riemannian score based method, instead relying on a non-bijective mapping between the manifold and Euclidean space, where we can use typical score-based modelling methods. The purpose is to demonstrate the benefits of incorporating geometry in to score-based models correctly.

We use as an example here for presentation the sphere, but the method can be used

METHOD	TRAINING	LIKELIHOOD EVALUATION	SAMPLING
RCNF	Solving ODE $O(dN)$	Solving augmented ODE $O(dN)$	Solving ODE $O(N)$
MOSER FLOW	Computing div $O(dk)$ or $O(k)$	Solving augmented ODE $O(dN)$	Solving ODE $O(N)$
RSGM	Score matching $O(d)$ or $O(1)$	Solving augmented ODE $O(dN)$	Solving SDE $O(N^*)$

Table 3.3. Summary of computational complexity (with respect to neural network forward and backward passes) for different methods. d is the manifold dimension, k the number of Monte Carlo batches in Moser flow's regulariser, N is the number of steps in the (adaptive) ODE solver, whereas N^* is the number of steps in the SDE Euler-Maruyama solver– which can usually be lower than N. Moser flow and Riemannian score-based generative models training complexity varies if the Hutchinson stochastic estimator is used. See table 3.1 for score matching losses' complexity.

on other manifolds.

In general these models work as follows:

1. Project the data points from the manifold, up to a set of measure zero, A, to Euclidean space through an invertible function $f : \mathcal{M} \setminus A \to \mathbb{R}^d$. This function should cover Euclidean space and be a smooth function. For example for the sphere we use the stereographic projection of the earth onto the plane, which misses out a single point at the pole.
2. Train a Euclidean score-based generative model on the data points projected to Euclidean space, giving a density p_θ on $\mathbb{R}^d$ (where θ are the parameters of the density).
3. Define the density on the manifold as the pushforward of the density in Euclidean space under the inverse of the bijection, $P_{\theta,\mathcal{M}} = f_*^{-1} p_\theta$.

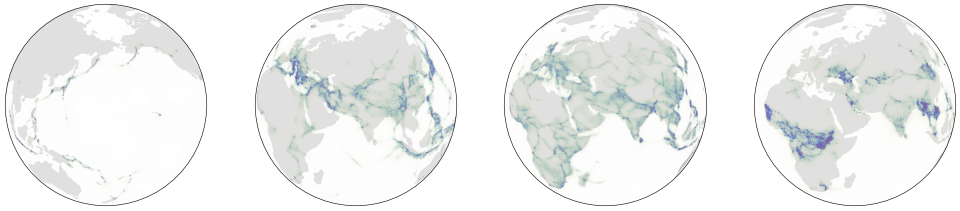
One can also apply these models to tori. By using the bijection $f : \theta \mapsto \tan(\theta)$ we can project each coordinate of a torus onto the real line.

This method is inspired by the approach taken in Gemici et al. (2016) to place normalizing flows on spheres and tori.

In general, we found that these models perform less well than their intrinsic counterparts. In order to map density near the seams of the bijection, it requires the model to send data points off to infinity in the Euclidean space. This is numerically challenging and leaves artefacts in the pushforward density on the manifold. In addition, the models tend to under and overestimate the density of densities in regions warped by the map onto Euclidean space.

3.4.2. Geologic and weather datasets on the sphere. We start by evaluating Riemannian score-based generative models on a collection of simple datasets, each containing an empirical distribution of occurrences of geologic and weather events on the surface of the earth. These events are: volcanic eruptions ((NGDC/WDS), 2022b), earthquakes ((NGDC/WDS), 2022a), floods (Brakenridge, 2017) and wildfires (EOSDIS, 2020).

We compare to previous baseline methods: Riemannian Continuous Normalizing Flows (Mathieu and Nickel, 2020), Moser Flows (Rozen et al., 2021), a mixture of Kent distributions (Peel et al., 2001), and our stereographic baseline method.



(a) Volcano (b) Earthquake (c) Flood (d) Fire

Figure 3.5. Trained score-based generative models on earth sciences data. The learned density is coloured green-blue. Blue and red dots represent training and testing data points, respectively.

We evaluate the log-likelihood of each model, extending to the manifold setting the likelihood computation techniques of score based generative models, see Section 3.2.5.

Table 3.4 shows that all benchmarked methods have comparable performance when evaluated on these simple tasks with Riemannian score-based generative models performing marginally better on most datasets. However, we empirically notice that Moser flows are slow to train and additionally that both Moser flows and stereographic score based generative models are computationally expensive to evaluate. Figure 3.5 visually compares the fits of Riemannian score based generative models to the data.

METHOD	VOLCANO	EARTHQUAKE	FLOOD	FIRE
MIXTURE OF KENT	$-0.80_{\pm 0.47}$	$0.33_{\pm 0.05}$	$0.73_{\pm 0.07}$	$-1.18_{\pm 0.06}$
RIEMANNIAN CNF	$\mathbf{-6.05_{\pm 0.61}}$	$0.14_{\pm 0.23}$	$1.11_{\pm 0.19}$	$\mathbf{-0.80_{\pm 0.54}}$
MOSER FLOW	$-4.21_{\pm 0.17}$	$\mathbf{-0.16_{\pm 0.06}}$	$\mathbf{0.57_{\pm 0.10}}$	$\mathbf{-1.28_{\pm 0.05}}$
STEREOGRAPHIC SCORE-BASED	$-3.80_{\pm 0.27}$	$\mathbf{-0.19_{\pm 0.05}}$	$\mathbf{0.59_{\pm 0.07}}$	$\mathbf{-1.28_{\pm 0.12}}$
RIEMANNIAN SCORE-BASED	$-4.92_{\pm 0.25}$	$\mathbf{-0.19_{\pm 0.07}}$	$\mathbf{0.45_{\pm 0.17}}$	$\mathbf{-1.33_{\pm 0.06}}$
DATASET SIZE	827	6120	4875	12809

Table 3.4. Negative log-likelihood scores for each method on the earth and climate science datasets. Bold indicates best results (up to statistical significance). Means and confidence intervals are computed over 5 different runs. Novel methods are shown with blue shading.

3.4.3. Synthetic data on tori. We now move to another manifold, that is the torus $\mathbb{T}^d = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$, to assess the scalability of the different methods with respect to the dimension d.

The main method we compare to is Moser flows (Rozen et al., 2021) as it was the most competitive method in the previous experiment. As a target distribution we consider a wrapped Gaussian on $\mathbb{T}^d$ with a random mean and unit variance. Moser flows' (Rozen et al., 2021) loss involves a regularization term which involves an integral over the manifold, approximated by a Monte Carlo estimator with uniform proposal. This term regularizes Moser flows towards probability measures i.e. with unit volume. We therefore expect Moser flows to fail in high-dimension as the number of samples K required for the MC estimator to be accurate will grows

METHOD	$M = 16$		$M = 32$		$M = 64$	
	LOG-LIKELIHOOD	NFE	LOG-LIKELIHOOD	NFE	LOG-LIKELIHOOD	NFE
MOSER FLOW	$0.85_{\pm 0.03}$	$2.3_{\pm 0.5}$	$0.17_{\pm 0.03}$	$2.3_{\pm 0.9}$	$\mathbf{-0.49_{\pm 0.02}}$	$7.3_{\pm 1.4}$
EXP-WRAPPED SGM	$\mathbf{0.87_{\pm 0.04}}$	$0.5_{\pm 0.1}$	$0.16_{\pm 0.03}$	$0.5_{\pm 0.0}$	$-0.58_{\pm 0.04}$	$0.5_{\pm 0.0}$
RSGM	$\mathbf{0.89_{\pm 0.03}}$	$\mathbf{0.1_{\pm 0.0}}$	$\mathbf{0.20_{\pm 0.03}}$	$\mathbf{0.1_{\pm 0.0}}$	$\mathbf{-0.49_{\pm 0.02}}$	$\mathbf{0.1_{\pm 0.0}}$

Table 3.5. Test log-likelihood and associated number of function evaluations (NFE) in 10^3 on the synthetic mixture distribution with M components on $\mathrm{SO}_3(\mathbb{R})$. Bold indicates best results (up to statistical significance). Means and standard deviations are computed over 5 different runs. Novel methods are shown with blue shading.

as $O(e^d)$. Additionally, the memory required to compute this estimator grows either in $O(Kd)$ for exact divergences or $O(K)$ for approximated divergences (see table 3.3).

In figure 3.6, we observe that Riemannian score-based generative models are able to fit the target distribution well even in high dimension, with a linear or constant computational cost—depending on the divergence estimator. In contrast, Moser flows scale poorly with the dimension, to the extent that we are unable to train them for $d \geq 10$. This is due to the combination of the complexity which grows linearly with both the dimension d and the number of MC samples K, which itself ought to grow exponentially with d—as discussed in the previous paragraph. This is illustrated by the gap between the 'Moser' and 'ODE' likelihoods which increases with the manifold dimension (see left Figure 3.6).

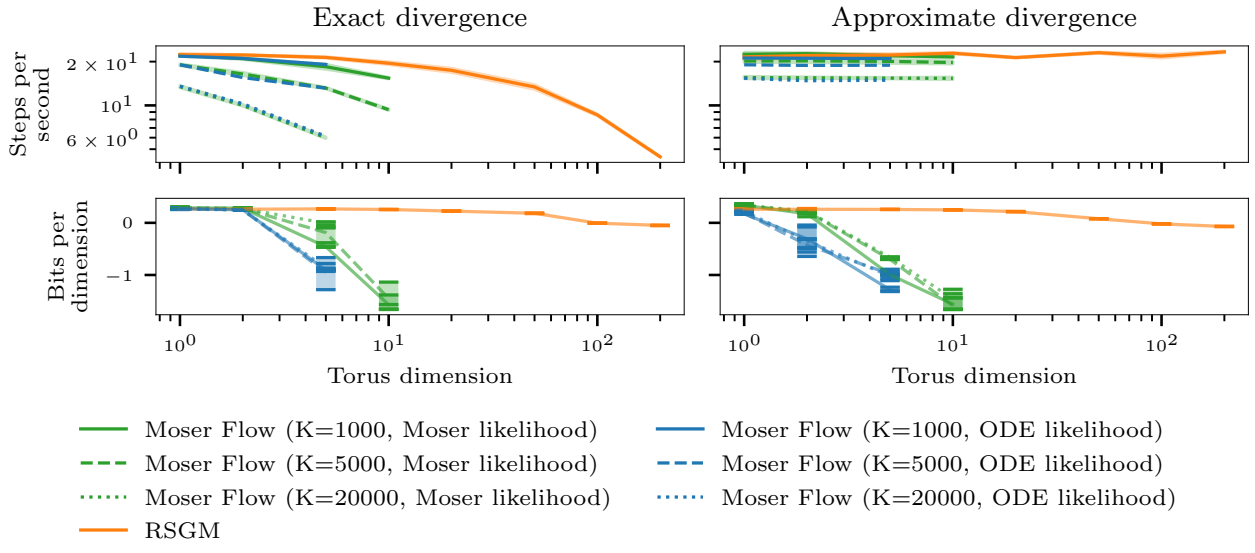
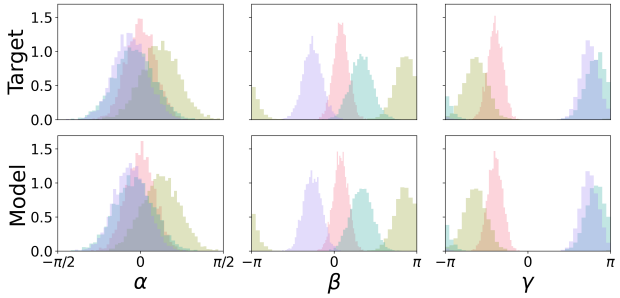
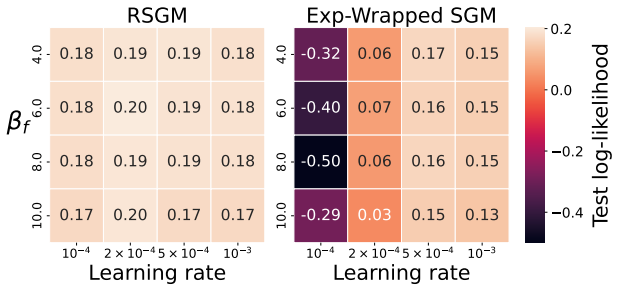


Figure 3.6. Comparison of Moser flows and Riemannian score-based generative models training speed and performance on the synthetic high-dimension torus task. Moser flows trained with $\lambda_{\min} = 1$. We report two likelihoods, the 'Moser' closed form density—not guaranteed to be normalized—and the 'ODE' likelihood given by solving an augmented ODE (as in CNFs) with the vector field induced by the Moser flow density—which is guaranteed to have unit volume.



(a) Histograms of $\mathrm{SO}_3(\mathbb{R})$ samples from a target mixture distribution with $M = 4$ components, represented via their Euler angles.



(b) Riemannian score-based generative models are much more robust to hyperparameters than Exp-wrapped SGMs. The diffusion coefficient is given by $\sigma(t, \boldsymbol{X}_t) = \sqrt{\beta(t)}$, $\beta(t) = \beta_0 + (\beta_f - \beta_0)t$.

Figure 3.7. Trained score-based generative models on synthetic $\mathrm{SO}_3(\mathbb{R})$ data.

3.4.4. Synthetic data on the Special Orthogonal group. In order to demonstrate the broad range of applicability of our model we now turn to the task of density estimation on the special orthogonal group

$$\mathrm{SO}(d) = \left\{ \mathrm{Q} \in \mathbb{R}^{d \times d} \;:\; \mathrm{Q}\mathrm{Q}^\top = \mathbf{I},\ \det(\mathrm{Q}) = 1 \right\}. \tag{3.80}$$

We consider the synthetic dataset consisting of samples in $\mathrm{SO}_3(\mathbb{R})$ from a mixture of wrapped normal distributions with M components. We compare Riemannian score-based generative models against Moser flows and a wrapped-exponential baseline inspired by Falorsi et al. (2019)—where we parametrise a standard Euclidean SGM on $\mathfrak{so}(3)$ that is then pushed-forward on $\mathrm{SO}(d)$. Riemannian score-based generative models are trained using the $\ell_{t|0}$ (DSM) loss with the Varadhan approximation (see table 3.1).

In addition to demonstrate the conditional modelling extension described in section 3.2.10, we model the distribution conditioned on the wrapped normal modes that a sample came from.

From table 3.5 we observe that, Riemannian score-based generative models perform consistently, whether the target distribution has few or many mixture components M, as opposed to Exp-wrapped SGMs and Moser flows which only perform well in some range of M. Similarly to section 3.4.3, we find Moser flows to be much slower to train due to the large number of Monte Carlo samples needed in the regulariser ($K = 10^4$). We also note from table 3.5 that the number of score network evaluations (NFE) is significantly lower for Riemannian score-based generative models, and is particularly detrimental for Moser flows ($\gg 10^3$).

3.4.5. Deeper comparison with Moser flows. Having observed the empirical poorer performance of Moser flows (Rozen et al., 2021) in the previous experiments, sections 3.4.3 and 3.4.4, we now analyse this discrepancy from a theoretical angle. In Moser flows and score-based models we are aiming to interpolate between two distributions, p_{data} and p_{prior}. In Moser flows this is given by the linear interpolation,

$$p_t^{\text{moser}} = t p_{\text{prior}} + (1 - t) p_{\text{data}}, \tag{3.81}$$

assuming we are interpolating over the time interval $t \in [0, 1]$. By contrast in score-based models the interpolated density is given by

$$p_t^{\text{diffusion}} = \int p_{\text{data}}(\boldsymbol{X}) p_{t|0}(\cdot|\boldsymbol{X}) \, \mathrm{d}\boldsymbol{X}. \tag{3.82}$$

These interpolations are compared in figure 3.8.

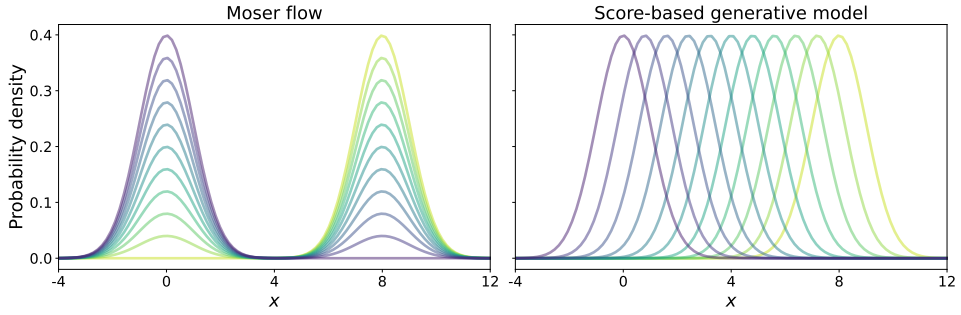


Figure 3.8. Interpolated density between the reference $p_{\text{prior}} = \mathrm{N}(0, 1)$ and target $p_{\text{data}} = \mathrm{N}(8, 1)$ distributions.

The difference between these two approaches is that Moser flows perform interpolate in *density space*, whereas score-based models interpolate in *sample space*. Interpolation in the *density space* results in spontaneous creation of density, whereas interpolation in *sample space* corresponds to a displacement of the density. In that respect, Moser flows can be seen as performing *vertical displacement* whereas score-based models correspond to *horizontal displacement*, see Santambrogio (2017).

There is a significant drawback with the 'spontaneous creation of density' of Moser flows. When solving trajectories in *sample space*— as we typically do for sampling or likelihood evaluation purposes—the Stein score's amplitude can get extremely high where the prior and target distributions have little overlap as shown on figure 3.9. This results in very stiff differential equations to solve, resulting in high

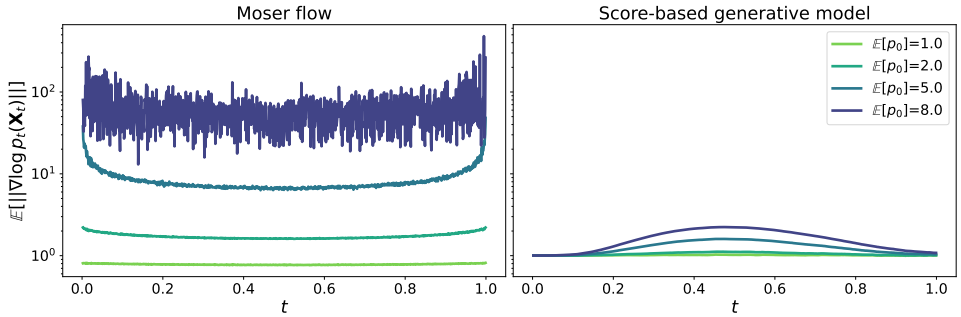
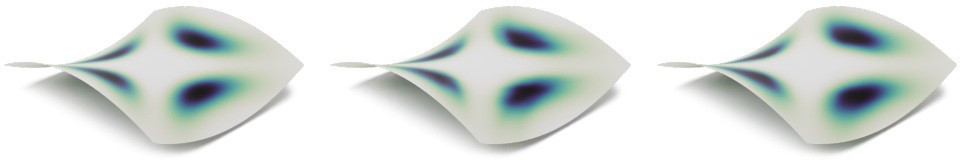


Figure 3.9. Expected norm of the Stein score along trajectories interpolating between reference and target $p_0 = \mathrm{N}(a, 1)$ distributions for different target mean.

numbers of function evaluations needed and poorer rollout accuracy. Additionally, it makes learning the vector field parametrising the flow more challenging.

This gives us an additional reason, beyond the difficulty of enforcing the divergence free vector field condition required by Moser flows, to prefer the framework of score-based models.

3.4.6. Synthetic data on hyperbolic space. Finally, we demonstrate Riemannian score-based generative models on a non-compact manifold: the two-dimension hyperbolic space $\mathbb{H}^2$, which is defined as the simply connected space of constant negative curvature. We use Langevin dynamics as the noising process and target a wrapped Gaussian as the invariant distribution. We again consider a synthetic dataset of samples from a mixture of exp-wrapped normal distribution. From figure 3.10, we can qualitatively see that both score-based models are able to fit the target distribution.



(a) Target distribution. (b) Exponential wrapped score-based generative models. (c) Riemannian score-based generative models.

Figure 3.10. Samples from different probability distributions on $\mathbb{H}^2$ coloured with respect to their density.

3.5. Conclusion

In this chapter we introduced Riemannian score-based generative models, a class of deep generative models that represent target densities supported on Riemannian manifolds, as the time-reversal of Langevin dynamics. The main benefits of our method stems from its scalability to high dimensions, its applicability to a broad class of manifolds due to the diversity of available loss functions, and crucially its capacity to model complex datasets while incorporating the geometry of the data. We also provided theoretical guarantees on the convergence of these models. There is a large class of manifold that we cannot target with this method however, namely *manifolds with boundary* (see definition A.5). The next chapter of this thesis focuses on this unresolved case.

4 | Score-based modelling on constrained domains

Work in this chapter is based on

N. Fishman, L. Klarner, V. De Bortoli, E. Mathieu, and M. J. Hutchinson. Diffusion Models for Constrained Domains. *Transactions on Machine Learning Research*, 2023.

and has been rewritten for this thesis with additional material.

Personal contributions:

1. Project conception with Valentin, Nic.
2. Development of practical approaches with Nic, Leo, Valentin, including the problem-specific details.
3. Development of the code with Nic and Leo.
4. Running experiments: Robotics experiment.
5. Supervision of the project.

4.1. Introduction

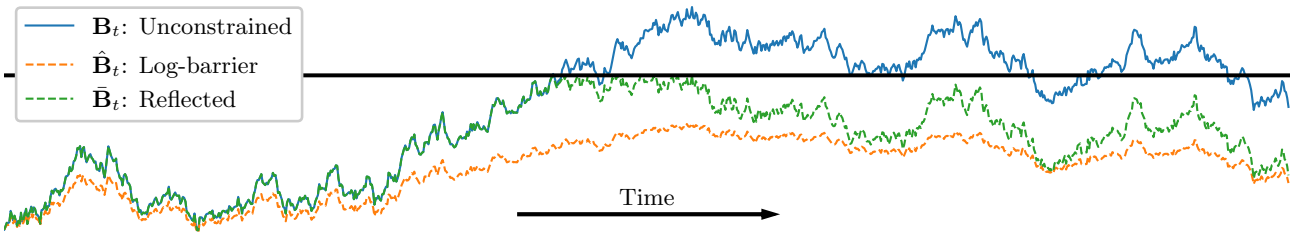


Figure 4.1. The behaviour of different types of noising processes considered in this chapter defined on the unit interval. $\mathbf{B}_t$: Euclidean (unconstrained) Brownian motion. $\hat{\mathbf{B}}_t$: Log barrier forward noising process. $\bar{\mathbf{B}}_t$: Reflected Brownian motion. All sampled with the same initial point and driving noise. Black line indicates the boundary.

A key assumption made in the previous chapter when generalising score-based models to Riemannian manifolds is that the stochastic processes considered as noising processes are defined *for all times*. While this holds for a large class of

stochastic processes and manifolds, it is not the case for most manifolds with boundary defined via a set of inequality constraints. The key assumption that allowed this to hold was that the manifolds were *geodesically complete* (see appendix A.10.5).

Geodesically complete

Hypercube

For instance, in the case of the *hypercube*, $(-1, 1)^d$ equipped with the Euclidean metric, the Riemannian Brownian motion coincides with the Euclidean d-dimensional Brownian motion $(\boldsymbol{B}_t)_{t\in\{0,T\}}$ as long as $\boldsymbol{B}_t \in (-1, 1)$. However, with probability one $(\boldsymbol{B}_t)_{t\geq 0}$ escapes from $(-1, 1)^d$, meaning that the Riemannian Brownian motion is not defined for times after this escape, and the framework of the chapter 3 does not apply.

Constrained manifolds

Such *constrained manifolds* comprise a wide variety of settings—including polytopes and convex sets of Euclidean spaces—and are studied across a wide range of disciplines, ranging from computational statistics (Morris, 2002), robotics (Han and Rudolph, 2006) quantum physics (Lukens et al., 2020), and computational biology (Thiele et al., 2013). Deriving principled diffusion models that are able to operate directly on these manifolds is thus of significant practical importance. This enables generative modelling in data-scarce and safety-critical settings in which constraints on the modelled domain may reduce the number of degrees of freedom or prevent unwanted behaviour.

As sampling problems on such manifolds are important, a flurry of Markov chain based methods have been developed to sample from unnormalised densities (Kook et al., 2022; Heirendt et al., 2019). Successful algorithms include the *reflected Brownian motion* (Williams, 1987; Petit, 1997; Shkolnikov and Karatzas, 2013), *log-barrier methods* (Kannan and Narayanan, 2009; Lee and Vempala, 2017; Noble et al., 2023; Kook et al., 2022; Gatmiry and Vempala, 2022; Lee and Vempala, 2018) and hit-and-run approaches in the case of polytopes (Smith, 1984; Lovász and Vempala, 2006).

Reflected Brownian motion

Log-barrier methods

In this chapter, we study the generative modelling counterparts of these algorithms through the lens of diffusion models. Among existing methods for statistical sampling on constrained manifolds, the geodesic Brownian motion (Lee and Vempala, 2017) and the reflected Brownian motion (Williams, 1987) are continuous stochastic processes, and thus well suited for extending the continuous Riemannian diffusion framework developed in the previous chapter. In particular, we introduce two principled diffusion models for generative modelling on constrained domains based on

(i) the geodesic Brownian motion, leveraging tools from the log-barrier methods, discussed in section 4.2, and

(ii) the reflected Brownian motion, discussed in section 4.3.

In both cases, we show how one can extend the ideas of time-reversal and score matching to these settings. We demonstrate the practical utility of these methods on a range of tasks defined on convex polytopes and the space of symmetric positive definite matrices, including the constrained conformational modelling of proteins and robotic arms. While both models extend classical diffusion models to inequality-constrained settings, they exhibit very different behaviours. We discuss their practical differences in section 4.6.

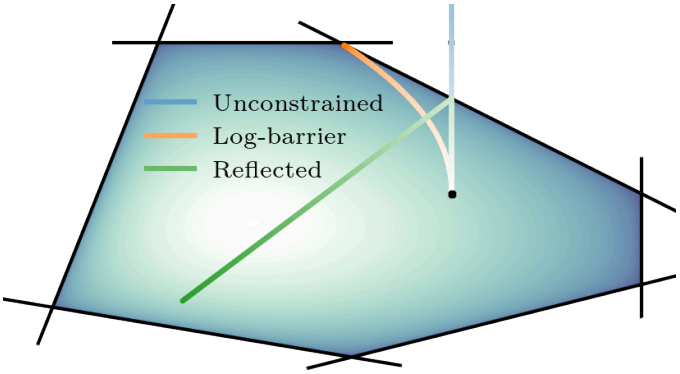


Figure 4.2. A convex polytope defined by six constraints $\{f_i\}_{i\in\mathcal{I}}$, along with the log barrier potential, and geodesics under the log-barrier metric and under the Euclidean metric with and without reflection at the boundary.

4.1.1. Technical setting. In this chapter, we are concerned with what we shall term *constrained* manifolds. More precisely, given a Riemannian manifold $(\mathcal{N}, h)$, we consider a family of real functions $\{f_i : \mathcal{N} \to \mathbb{R}\}_{i\in\mathcal{I}}$ indexed by $\mathcal{I}$. We then define

$$\mathcal{M} = \{x \in \mathcal{N} \ : \ f_i(x) < 0, \ i \in \mathcal{I}\}, \tag{4.1}$$

and derive models such that their densities are constrained to be in the set $\mathcal{M}$.

In this scenario, $\{f_i\}_{i\in\mathcal{I}}$ are interpreted as a set of constraints on $\mathcal{N}$. For example, choosing $\mathcal{N} = \mathbb{R}^d$ and affine constraints $f_i(x) = \langle a_i, x\rangle - b_i$, $x \in \mathbb{R}^d$, we get that $\mathcal{M}$ is an open polytope as illustrated in figure 4.2. This setting naturally appears in many areas of engineering, biology, and physics (Boyd et al., 2004; Han and Rudolph, 2006; Lukens et al., 2020).

This setting is very similar to that of a *manifold with boundary* (see definition A.5). If the interior of the constraints form an open set on the manifold then we can consider this as exactly as a manifold with boundary. We consider the constraints in the way set out as it allows us to consider a slightly larger setting of constraints. In light of this we term the set

Manifold with boundary

$$\partial\mathcal{M} = \{x \in \mathcal{M} \ : \ f_i(x) = 0 \text{ for any } i \in I \text{ and } f_i(x) \ngeq 0 \text{ for any } i \in I\} \tag{4.2}$$

the boundary, although this slightly abuses the terminology of a manifold with boundary.

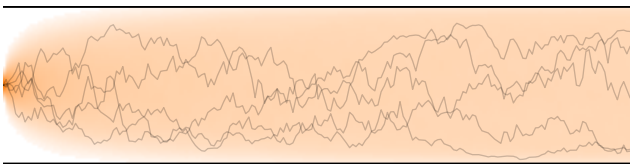
While the two methods we introduced in this chapter can be applied to arbitrary constraints, in our applications we focus on two specific settings:

(a) *polytopes*—$\mathcal{N} = \mathbb{R}^d$ with linear boundaries forming a convex set, — Polytopes

(b) *symmetric positive definite matrices* (SPD) under maximum trace conditions—$\mathcal{N} = \mathcal{S}^d_{++}$, a quadratic boundary on the elements of the matrix. — Symmetric positive definite matrices

4.2. Log-barrier diffusion models

First we consider a method based on *barrier methods*. Barrier methods work by constructing a smooth potential $\phi : \mathcal{M} \to \mathbb{R}$ such that it blows up on the boundary — Barrier methods



$$\mathrm{d}\boldsymbol{X}_t = \tfrac{1}{2}\operatorname{div}(g^{-1})(\boldsymbol{X}_t)\,\mathrm{d}t + g(\boldsymbol{X}_t)^{-1/2}\,\mathrm{d}\boldsymbol{B}_t$$

Figure 4.3. Convergence of the barrier Langevin dynamics on the unit interval to the uniform distribution.

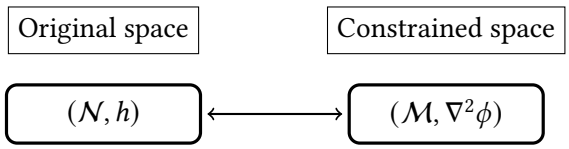


Figure 4.4. Illustrative diagram of the barrier method and the change of metric.

of a desired set, see Nesterov et al. (2018). Such potentials are at the basis of interior methods in optimisation (Boyd et al., 2004).

For a single boundary, $\partial\mathcal{M}$ to define a barrier function it is sufficient to be able to measure the distance to the boundary,

$$d(x, \partial\mathcal{M}) = \min_{y\in\partial\mathcal{M}} d(x, y), \tag{4.3}$$

and a monotone decreasing function $\phi : \mathbb{R}^+ \to \mathbb{R}$ such that $\lim_{z\to 0}\phi(z) = \infty$ (plus some technical conditions, see (Nesterov and Nemirovskii, 1994)). We can then define the barrier function as $\phi(d(x, \partial M))$.

In the setting with multiple constraints we can define the barrier as

$$\phi(\min_{i\in I} d(x, f_i)), \tag{4.4}$$

where f_i is the set defined by $f_i(y) = 0, y \in \mathcal{N}$. We can also define the barrier as

$$\sum_{i\in I} \phi(d(x, f_i)). \tag{4.5}$$

If ϕ is smooth this guarantees the barrier function to be smooth, unlike the first construction.

Log barrier methods

Of the functions we can choose for ϕ, $\phi(x) = -\log(x)$, is the most common, and methods based off this are known as *log barrier methods* (Lee and Vempala, 2017).

In general solving the minimum distance to the boundary of a set is a highly non-trivial optimisation problem. Linear boundaries are one case which admits a simple closed form. For a convex polytope $\mathcal{M}$ defined by the constraints $\mathrm{A}x < b$, with $\mathrm{A} \in \mathbb{R}^{m\times d}$ and $b \in \mathbb{R}^m$, giving m constrains, the distance to i^{th} boundary is given by

$$d(x, f_i) = \langle A_i, x\rangle - b_i. \tag{4.6}$$

The log barrier $\phi : \mathcal{M} \to \mathbb{R}_+$ is therefore given by

$$\phi(x) = -\sum_{i=1}^{m} \log(\langle \mathrm{A}_i, x\rangle - b_i), \tag{4.7}$$

assuming that $\|\mathrm{A}_i\| = 1$.

The core idea of barrier methods is to 'warp the geometry' of the constrained space, stretching it as the boundary is approached so it is never hit. This bypasses the need to explicitly deal with the boundary.

4.2.1. Hessian metrics. Let us assume that $\mathcal{N} = \mathbb{R}^d$, that the constraints $(f_i)_{i\in\mathcal{I}}$ form a convex compact set, and that we have a barrier function ϕ which is strictly convex and smooth on $\mathcal{M}$, the interior of the constraints.

Assuming this, the Hessian $\nabla^2\phi$ is positive definite and thus defines a valid Riemannian metric on $\mathcal{M}$. We endow $\mathcal{M}$ with $g = \nabla^2\phi$ as a Riemannian metric, making it into a *Hessian manifold* (Shima and Yagi, 1997). It can be proved that for such a metric the resulting exponential map is defined on the whole tangent space for all points and therefore the manifold $(\mathcal{M}, \nabla^2\phi)$ is a manifold *without* boundary (Lee and Vempala, 2017). Hessian manifold

In the special case where the barrier is given by eq. (4.7), we get for any $x \in \mathcal{M}$

$$g(x) = \mathrm{A}^\top S^{-2}(x)\mathrm{A} \quad \text{with} \quad S(x) = \mathrm{diag}(b_i - \langle \mathrm{A}_i, x\rangle)_i. \tag{4.8}$$

While developed and most commonly used in optimisation, barrier methods can also be used for sampling (Lee and Vempala, 2017).

4.2.2. Forward noising process. Let us consider a compact set of Euclidean space equipped with a Euclidean metric, a set of constraints, and a Hessian non-Euclidean metric described above.

We consider the following Langevin dynamics, relative to the Euclidean metric, as a forward process

$$\mathrm{d}\boldsymbol{X}_t = \tfrac{1}{2}\mathrm{div}(g^{-1})(\boldsymbol{X}_t)\mathrm{d}t + g(\boldsymbol{X}_t)^{-\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t, \tag{4.9}$$

where the divergence of the metric $\mathrm{div}\, g^{-1}$ is given by the *generalisation of the divergence to tensor fields* (see appendix A.10.2). On Euclidean space this is given by the column-wise divergence of g^{-1}.

Under mild assumptions on $\mathcal{M}$, $\mathrm{div}(g^{-1})$ and g^{-1} we get that $(\boldsymbol{X}_t)_{t\geq 0}$ is well-defined and for any $t \geq 0$, $\boldsymbol{X}_t \in \mathcal{M}$. In particular, for any $t \geq 0$, $\boldsymbol{X}_t$ does not reach the boundary.

In addition, $(\boldsymbol{X}_t)_{t\geq 0}$ is irreducible as a Markov process. Hence, assuming that $\mathcal{M}$ is compact, the uniform distribution on $\mathcal{M}$ is the unique invariant measure of the process $(\boldsymbol{X}_t)_{t\geq 0}$ and $(\boldsymbol{X}_t)_{t\geq 0}$ converges to the uniform distribution as $t \to \infty$, giving us a tractable reference distribution for score-based modelling. We refer the reader to appendix D.2 for a proof of these results.

This stochastic process was first proposed by Lee and Vempala (2017) in the context of efficient sampling from the uniform distribution over a polytope. Under similar conditions, $\boldsymbol{X}_t$ admits a density p_t with respect to the Lebesgue measure, and we have that

$$\partial_t p_t = \tfrac{1}{2}\mathrm{Tr}(g^{-1}\nabla^2 p_t). \tag{4.10}$$

4.2.3. Time-reversal. Assuming that g^{-1} and its derivative are bounded on $\mathcal{M}$, the time-reversal of eq. (4.9) is given by theorem 3.5, in particular we have

$$\mathrm{d}\boldsymbol{X}^{\leftarrow}_t = \left[-\frac{1}{2}\mathrm{div}(g^{-1}) + \mathrm{div}(g^{-1}) + g^{-1}\nabla \log p_{T-t}(\boldsymbol{X}^{\leftarrow}_t)\right]\mathrm{d}t + g(\boldsymbol{X}^{\leftarrow}_t)^{-\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t \tag{4.11}$$

$$= \left[\frac{1}{2}\mathrm{div}(g^{-1}) + g^{-1}\nabla \log p_{T-t}\right](\boldsymbol{X}^{\leftarrow}_t)\mathrm{d}t + g(\boldsymbol{X}^{\leftarrow}_t)^{-\frac{1}{2}}\mathrm{d}\boldsymbol{B}_t. \tag{4.12}$$

$\boldsymbol{X}_0^\epsilon$ is initialised with the uniform distribution on $\mathcal{M}$ (which is close to the one of $\boldsymbol{X}_T$ for large T).

We refer to section 4.4 for details on the training and parametrisation of the score function.

4.2.4. Sampling. Sampling from the forward eq. (4.9) and backward eq. (4.12) processes, once the score is learnt, requires a discretisation scheme. We use *geodesic random walks* (see section 3.2.6) for this purpose, introduced in section 3.2.6.

Geodesic random walks

However, contrary to the previous chapter, we will not have access explicitly to the exponential map of the Hessian manifold. Instead, we rely on an approximation, using a *retraction* in place of the explicit exponential map. A retraction takes a step in a given direction with a first order approximation of the exponential map, and if the sample ends outside the constraints it is projected back inside the constraints. See Absil and Malick (2012) and Boumal (2023) for a deeper discussion and alternatives.

4.2.5. Extending the technical results. In this section we have made two strong assumptions: that the underlying manifold is a Euclidean space, and that the set of constrains form a compact, convex set.

Relaxing the assumption that the underlying manifold is a Euclidean space seems possible as there exists results on Hessian metrics for non-flat space, e.g. see Sustretov (2022), but these results are not well-developed in a general setting.

Relaxing the compact assumption is possible if we can factorise the space into a compact set and an unconstrained set. We can then apply the method in this section to the constrained part and the method from chapter 3 to the unconstrained part. Relaxing the compact assumption when this is not the case would be challenging as the definition of a suitable reference distribution would become difficult.

Relaxing the convexity assumption is likely to be difficult, but possibly if one can find a suitable way of defining a Hessian metric and proving the geodesic completeness of such a manifold. It is likely other assumptions would be required on the set, such as being star-shaped.

4.3. Reflected diffusion models

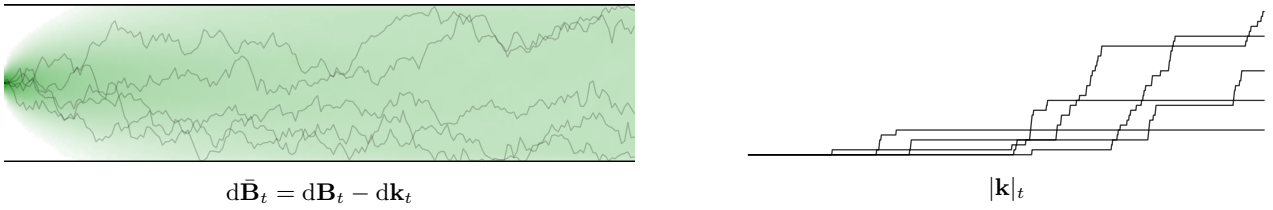


Figure 4.5. *Left:* Convergence of the reflected Brownian motion on the unit interval to the uniform distribution. *Right:* Value of $|\boldsymbol{k}|_t$ for the trajectory samples on the left through time.

Another approach to deal with the geometry of $\mathcal{M}$ is to use the standard metric h and forward dynamics of $\mathcal{N}$, constraining it to $\mathcal{M}$ by *reflecting* the process

whenever it would encounter a boundary. This results in a *reflected stochastic differential equation*. We will first assume that $\mathcal{M}$ is compact and convex. To simplify the presentation we focus on the Euclidean case $\mathcal{N} = \mathbb{R}^d$ with smooth boundary ∂M[1]. Reflected stochastic differential equation

The key difference between this approach and the barrier approach is that in the reflected case we leave the geometry unchanged. All we need to do is show that the dynamics induced by reflecting the forward process whenever it hits the boundary leads to an invariant distribution and admit a time-reversal.

It is worth noting that while barrier approaches have received considerable theoretical attention in the sampling literature (Lee and Vempala, 2017; Noble et al., 2023), reflected methods have remained comparatively undeveloped from the methodological and practical point of view.

In the next section, we recall the technical basics of reflected stochastic processes.

4.3.1. Skorokhod problem. We define a reflected Brownian motion as the solution to the Skorokhod problem. A solution to the *Skorokhod problem* (Skorokhod, 1961) consists into two coupled processes $(\bar{\boldsymbol{B}}_t, \boldsymbol{k}_t)_{t\geq 0}$. $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ acts *locally* as a Euclidean Brownian motion, $(\boldsymbol{B}_t)_{t\geq 0}$, and $(\boldsymbol{k}_t)_{t\geq 0}$ compensates for the excursion of $(\boldsymbol{B}_t)_{t\geq 0}$ outside the boundary so that $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ remains in $\mathcal{M}$. Skorokhod problem

We say that $(\bar{\boldsymbol{B}}_t, \boldsymbol{k}_t)_{t\geq 0}$ is a solution to a specific Skorokhod problem if $(\boldsymbol{k}_t)_{t\geq 0}$ and $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ are two processes satisfying mild conditions, see appendix D.3, such that for any $t \geq 0$,

$$\bar{\boldsymbol{B}}_t = \bar{\boldsymbol{B}}_0 + \boldsymbol{B}_t - \boldsymbol{k}_t \in \mathcal{M}, \tag{4.13}$$

and

1. $|\boldsymbol{k}|_t = \int_0^t \mathbf{1}_{\bar{\boldsymbol{B}}_s \in \partial\mathcal{M}} \mathrm{d}|\boldsymbol{k}|_s$, where $\mathbf{1}$ is the indicator function, and
2. $\boldsymbol{k}_t = \int_0^t \boldsymbol{n}(\bar{\boldsymbol{B}}_s)\mathrm{d}|\boldsymbol{k}|_s$,

where $(|\boldsymbol{k}|_t)_{t\geq 0}$ is the total variation of $(\boldsymbol{k}_t)_{t\geq 0}$ and $\boldsymbol{n}$ is the outward normal vector field to $\mathcal{M}$[2].

When $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ hits the boundary, the first condition, $\boldsymbol{k}_t = \int_0^t \boldsymbol{n}(\bar{\boldsymbol{B}}_s)\mathrm{d}|\boldsymbol{k}|_s$, tells us that $-\boldsymbol{k}_t$ "compensates" for $\bar{\boldsymbol{B}}_t$ by pushing the process back into $\mathcal{M}$ along the inward normal $-\boldsymbol{n}(\bar{\boldsymbol{B}}_s)$. The second condition, $|\boldsymbol{k}|_t = \int_0^t \mathbf{1}_{\bar{\boldsymbol{B}}_s \in \partial\mathcal{M}} \mathrm{d}|\boldsymbol{k}|_s$ can be interpreted as $\boldsymbol{k}_t$ being constant when $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ does not hit the boundary. As a result $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ can be understood as the continuous-time counterpart to a reflected Gaussian random walk. The process $(\boldsymbol{k}_t)_{t\geq 0}$ can be related to the notion of *local time* (Revuz and Yor, 2013) and quantifies the amount of time $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ spends at the boundary $\partial\mathcal{M}$.

Lions and Sznitman (1984, theorem 2.1) ensure the existence and uniqueness of a solution to the Skorokhod problem. One key observation is that the event $\{\bar{\boldsymbol{B}}_t \in \partial\mathcal{M}\}$ has probability zero (Harrison and Williams, 1987, Section 7, Lemma 7).

[1] We refer to appendix D.6 for a definition of smooth boundary.

[2] We extend the normal $\boldsymbol{n}$ to the whole space by letting $\boldsymbol{n}(x) = 0$ if $x \notin \partial\mathcal{M}$.

As in the *unconstrained* setting, one can describe the dynamics of the density of $\bar{\boldsymbol{B}}_t$.

Time evolution of the density of a Skorokhod problem

PROPOSITION 4.1. *For any $t > 0$, the distribution of $\bar{\boldsymbol{B}}_t$ admits a density with respect to. the Lebesgue measure denoted p_t. In addition, we have for any $x \in \mathrm{int}(\mathcal{M})$ and $x_0 \in \partial\mathcal{M}$*

$$\partial_t p_t(x) = \tfrac{1}{2}\Delta p_t(x), \quad \partial_{\boldsymbol{n}} p_t(x_0) = 0, \tag{4.14}$$

where we recall that $\mathbf{n}$ is the outward normal to $\mathcal{M}$.

Proof. Burdzy et al. (2004, theorem 2.2) ∎

Neumann boundary conditions

Note that contrary to the unconstrained setting, the heat equation has *Neumann boundary conditions*, that the change in density on the boundary is tangential to the boundary.

As a consequence of proposition 4.1, similar to the compact Riemannian setting (Saloff-Coste, 1994), it can be shown that the reflected Brownian motion converges to the uniform distribution on $\mathcal{M}$ exponentially fast (Loper, 2020; Burdzy et al., 2006), see section 4.3. Hence, $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ is a candidate for a forward noising process in the context of diffusion models.

4.3.2. Time-reversal. In order to extend the diffusion model approach to the reflected setting, we need to derive *time-reversal* for $(\bar{\boldsymbol{B}}_t)_{t\in[0,T]}$. Namely, we need to characterise the evolution of $(\boldsymbol{X}^{\leftarrow}_t)_{t\in[0,T]} = (\bar{\boldsymbol{B}}_{T-t})_{t\in[0,T]}$. It can be shown that the time-reversal of $(\bar{\boldsymbol{B}}_t)_{t\in[0,T]}$ is also the solution to a Skorokhod problem.

Time reversal of Brownian Shorokhod problems

THEOREM 4.2. *There exist $(\boldsymbol{k}^{\leftarrow}_t)_{t\geq 0}$ a bounded variation process and a Brownian motion $(\boldsymbol{B}_t)_{t\geq 0}$ such that*

$$\boldsymbol{X}^{\leftarrow}_t = \boldsymbol{X}^{\leftarrow}_0 + \boldsymbol{B}_t + \int_0^t \nabla \log p_{T-s}(\boldsymbol{X}^{\leftarrow}_s)\mathrm{d}s - \boldsymbol{k}^{\leftarrow}_t. \tag{4.15}$$

In addition, for any $t \in [0,T]$ we have

$$|\boldsymbol{k}|^{\leftarrow}_t = \int_0^t \mathbf{1}_{\boldsymbol{X}^{\leftarrow}_s \in \partial\mathcal{M}} \mathrm{d}|\boldsymbol{k}|^{\leftarrow}_s, \quad \boldsymbol{k}^{\leftarrow}_t = \int_0^t \boldsymbol{n}(\boldsymbol{X}^{\leftarrow}_s)\mathrm{d}|\boldsymbol{k}|^{\leftarrow}_s. \tag{4.16}$$

Proof. Appendix D.6. ∎

This proof follows Petit (1997) which provides a time-reversal in the case where $\mathcal{M}$ is the positive orthant. It is based on an extension of Haussmann and Pardoux (1986) to the reflected setting, with a careful handling of the boundary conditions. In particular, contrary to Petit (1997), we do not rely on an explicit expression of p_t but instead use the intrinsic properties of $(\boldsymbol{k}_t)_{t\geq 0}$.

Informally, theorem 4.2 means that the process $(\boldsymbol{X}^{\leftarrow}_t)_{t\in[0,T]}$ satisfies

$$\mathrm{d}\boldsymbol{X}^{\leftarrow}_t = \nabla \log p_{T-t}(\boldsymbol{X}^{\leftarrow}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t - \mathrm{d}\boldsymbol{k}^{\leftarrow}_t, \tag{4.17}$$

which echoes the usual time-reversal formula of theorem 3.5. In practice, in order to sample from $(\boldsymbol{X}^{\leftarrow}_t)_{t\in[0,T]}$, one needs to consider the *reflected* version of the *unconstrained* dynamics $\mathrm{d}\boldsymbol{X}^{\leftarrow}_t = \nabla \log p_{T-t}(\boldsymbol{X}^{\leftarrow}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t$.

Algorithm 4.3 Discretisation of the stochastic differential equation $\mathrm{d}\boldsymbol{X}_t = b(t, \boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{B}_t - \mathrm{d}\boldsymbol{k}_t$. Reflected random walk

Require: T (simulation time), N (number of steps), X_0^γ (initial point), $\{f_i\}_{i\in\mathcal{I}}$ (boundary functions)
$\gamma = T/N$
for $k \in \{0, \dots, N-1\}$ **do**
  $Z_{k+1} \sim \mathrm{N}(0, \mathbf{I})$
  $X_{k+1}^\gamma = \mathrm{ReflectedStep}[X_k^\gamma, \sqrt{\gamma} Z_{k+1}, \{f_i\}_{i\in\mathcal{I}}]$ ▷ See algorithm 4.4.
return $\{X_k^\gamma\}_{k=0}^N$

Algorithm 4.4 The algorithm operates by repeatedly taking geodesic steps until one of the constraints is violated or the step is fully taken. Upon hitting the boundary we parallel transport the tangent vector to the boundary and then reflect it against the boundary. We then start a new geodesic from this point in the new direction. The $\arg\mathrm{intersect}_t$ function computes the distance one must travel along a geodesic in direction $\boldsymbol{s}$ until constraint f_i is intersected. For a discussion of boundary intersection, paralleltransport, $\exp_g$ and reflect please see section 2.1.3. Reflected step algorithm

Require: $x \in \mathcal{M}$, $\boldsymbol{v} \in \mathrm{T}_x\mathcal{M}$, $\{f_i\}_{i\in\mathcal{I}}$
  $\ell \leftarrow \|\boldsymbol{v}\|_g$
  $\boldsymbol{s} \leftarrow \boldsymbol{v}/\|\boldsymbol{v}\|_g$
  while $\ell \geq 0$ **do**
    $d_i = \arg\min_{t\in\mathbb{R}^+} \exp_p(t\boldsymbol{v}) \in \partial f_i$ ▷ Boundary intersection.
    $i \leftarrow \arg\min_i \ d_i \ s.t. \ d_i > 0$ ▷ Select minimum boundary distance.
    $\alpha \leftarrow \min(d_i, \ell)$ ▷ Minimum distance to boundary or end of transport.
    $x' \leftarrow \exp_g(x, \alpha\boldsymbol{s})$ ▷ Exponential map step.
    $\boldsymbol{s} \leftarrow \mathrm{paralleltransport}_g(x, \boldsymbol{s}, x')$ ▷ Transport tangent vector to the boundary.
    $\boldsymbol{s} \leftarrow \mathrm{reflect}(\boldsymbol{s}, f_i)$ ▷ Reflect on the boundary.
    $\ell \leftarrow \ell - \alpha$ ▷ Reduce step size by distance travelled.
    $x \leftarrow x'$
  return x

4.3.3. Sampling. In practice, we approximately sample the reflected dynamics by considering the Markov chain given by algorithm 4.3. We refer to Pacchiarotti et al. (1998) and Bossy et al. (2004) for weak convergence results on this numerical scheme in the Euclidean setting. This makes use of a discretised reflection step, detailed in algorithm 4.4.

4.3.4. Likelihood evaluation. It is possible to extend the likelihood ordinary differential equation approach to computing the likelihood of a model sample in the reflected Brownian motion case, similar to section 3.2.5.

PROPOSITION 4.5. *Assume that $\mathcal{M} \subset \mathbb{R}^d$ is a bounded open set with smooth boundary. Assume that $(t, x) \mapsto \nabla \log p_t(x)$ is smooth on $[0, +\infty) \times \partial\mathcal{M}$.* Likelihood ordinary differential equation for refelcted Brownian motion

Let $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ be the reflected Brownian motion with $\bar{\boldsymbol{B}}_0 \sim p_0$ smooth and supported in $\mathcal{M}$. Let $(\boldsymbol{X}_t)_{t\geq 0}$ be given for any $t \geq 0$ by

$$\mathrm{d}\boldsymbol{X}_t = \tfrac{1}{2}\nabla \log p_t(\boldsymbol{X}_t)\mathrm{d}t \tag{4.18}$$

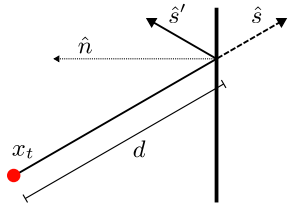


Figure 4.6. Reflection against a linear boundary. For a step s with magnitude $|s|$ and direction $\hat{s}$ the distance to the boundary described by the normal $\hat{n}$ and offset b is $d = \frac{\langle \hat{s}, x_t \rangle - b}{\langle \hat{s}, \hat{n} \rangle}$. The reflected direction is given by $\hat{s}' = \hat{s} - 2\langle \hat{s}, \hat{n} \rangle \hat{n}$.

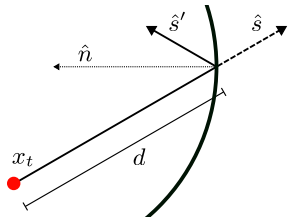


Figure 4.7. Reflection against a spherical boundary. For a sphere of radius r, the distance to the boundary from x_t in direction s is given by $d = \frac{1}{2}(\langle \hat{s}, x_t \rangle^2 + 4(r^2 - \|x_t\|^2))^{1/2} - \frac{1}{2}\langle \hat{s}, x_t \rangle$. The normal at the intersection can be computed as the unit vector in the direction $-2(d\hat{s} + x_t)$, and then $\hat{s}'$ as above.

and $\mathbf{X}_0 \sim p_0$, where p_t denotes the density of $\bar{\mathbf{B}}_t$ with respect to. the Lebesgue measure for any $t > 0$. Then for any $t \in \{0, T\}$, $\bar{\mathbf{B}}_t$ and $\mathbf{X}_t$ have the same distribution.

Proof. Appendix D.5 ∎

4.3.5. Extending the technical results. In this section we have made three strong assumptions: that the underlying manifold is a Euclidean space, and that the set of constrains form a compact, convex set, and that the boundary of the set is smooth.

Relaxing the assumption that the underlying manifold is a Euclidean space seems possible as the definition of Skorokhod problems extends naturally to the Riemannian manifold setting. The main difficulty in analysing Skorokhod problems is careful analysis of the behaviour on the boundary, which should be possible on manifolds. There exists little literature on this topic at present, however.

Similar to the barrier methods, relaxing the compact assumption is possible if we can factorise the space into a compact set and an unconstrained set.

Relaxing the convexity assumption is likely to be difficult, and would require significant new results on the convergence of reflected Brownian motion on non-convex sets. This may be possible with other assumptions on the set, such as being star-shaped.

Relaxing the smoothness assumption is also likely to be very difficult as the technical complexity of dealing with such boundaries becomes high. We note however

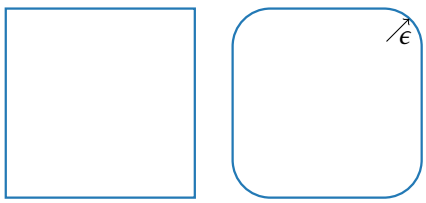


Figure 4.8. *Left:* A square with non-smooth corners. *Right:* A square with $\epsilon-$smoothed corners.

that this is not a significant practical limitation. For any non-smooth regions of a boundary, for example a corner, we can apply an $\epsilon-$smoothing to the corner to make the boundary smooth, resulting in extremely minor practical differences, see figure 4.8.

4.4. Training and parametrising score functions for constrained diffusions

In order to train log-barrier and reflected diffusion models we need a *tractable* score matching loss on constrained manifolds. We will prove that the implicit score-matching loss can be extended to the constrained manifold setting so long as we enforce that the score is zero on the boundary. This proof holds for both the log-barrier and the reflected process.

Implicit score-matching on constrained domains

Proposition 4.6. *Let $s \in \mathrm{C}^\infty(\{0,T\} \times \mathbb{R}^d, \mathbb{R}^d)$ such that for any $x \in \partial\mathcal{M}$ and $t \geq 0$, $\langle s_t(x), \boldsymbol{n}(x)\rangle = 0$. Then, there exists $C > 0$ such that*

$$\mathbb{E}\left[\|\nabla \log p_t - s_t\|^2\right] = \mathbb{E}\left[\|s_t\|^2 + 2\,\mathrm{div}(s_t)\right] + C, \tag{4.19}$$

where $\mathbb{E}$ is taken over $\boldsymbol{X}_t \sim p_t$ and $t \sim \mathcal{U}([0,T])$.

Proof. Appendix D.4.1 ∎

This result immediately implies we can optimise the score network approximate the score function using the implicit score matching loss function so long as we enforce a Neumann boundary condition, which naturally arises from proposition 4.5 and so is a constraint on the score function that does not reduce the modelling capacity of the model.

Following Liu et al. (2022a) we can accommodate this boundary condition in our score parametrisation by scaling the score output of the neural network by a monotone function $h(d(x,\partial\mathcal{M}))$ where d is the distance from x to the boundary, where $h(0) = 0$. In particular, we use a clipped ReLU function: $s_\theta(t,x) = \min(1, \mathrm{ReLU}(d(x,\partial\mathcal{M}) - \delta)) \cdot \mathrm{NN}_\theta(t,x)$ with $\delta > 0$ a threshold so the model is forced to be zero "close" to the boundary as well as exactly on the boundary. The inclusion of this scaling function is necessary to produce reasonable results as we show in appendix D.4.2.

While this scheme enforces a stronger condition on the score function then Neumann conditions, that it is zero on the boundary, not just tangent to the boundary, we find in practise that this works well, and is cheaper than enforcing Neumann conditions only.

The forward processes (eq. (4.9)) and (eq. (4.15)) for the barrier and reflected methods cannot be sampled in closed form, so at training time samples from the conditional distributions $p(\boldsymbol{X}_t|\boldsymbol{X}_0)$ are obtained by discretising these processes. To take the most of this computational overhead, we use several samples from the discretised forward trajectory $(\boldsymbol{X}_{t_1}, \cdots, \boldsymbol{X}_{t_k}|\boldsymbol{X}_0)$ instead of only using the last sample.

4.5. Related work for sampling on constrained manifolds.

Sampling from a distribution on a space defined by a set of constraints is an important ingredient in several computational tasks, such as computing the volume of a polytope (Lee and Vempala, 2017). Incorporating such constraints inside MCMC algorithms while preserving fast convergence properties is an active field of research (Kook et al., 2022; Lee and Vempala, 2017; Noble et al., 2023). In this work, we are interested in sampling from the uniform distribution defined on the constrained set in order to define a proper *forward process* for our diffusion model. Log-barrier methods such as the Dikin walk or Riemannian Hamiltonian Monte Carlo (Kannan and Narayanan, 2009; Lee and Vempala, 2017; Noble et al., 2023) change the geometry of the underlying space and define stochastic processes which never violate the constraints, see (Kannan and Narayanan, 2009; Noble et al., 2023) for instance. If we keep the Euclidean metric, then *unconstrained* stochastic processes might not be well-defined for all times. To counter this effect, it has been proposed to *reflect* the Brownian motion (Williams, 1987; Petit, 1997; Shkolnikov and Karatzas, 2013). Finally, we also highlight hit-and-run approaches (Smith, 1984; Lovász and Vempala, 2006) which generalise Gibbs' algorithm and enjoy fast convergence properties provided that one knows how to sample from the one-dimensional marginals.

4.6. Experimental results

To demonstrate the practical utility of the constrained diffusion models introduced in sections 4.2 and 4.3, we evaluate them on a series of increasingly difficult synthetic tasks on convex polytopes, including the hypercube and the simplex, in section 4.6.1. We then highlight their applicability to real-world settings by considering two problems from robotics and protein design. In particular, we show that our models are able to learn distributions over the space of $d \times d$ symmetric positive definite (SPD) matrices S_{++}^d under trace constraints in section 4.6.2. This setting that is essential to describing and controlling the motions and exerted forces of robotic platforms (Yoshikawa, 1985). In section 4.6.3, we use the parametrisation introduced in Han and Rudolph (2006) to map the problem of modelling the conformational ensembles of loops of proteins with fixed endpoints to the product manifold of a convex polytope and a torus.

We refer to appendix D.7 for more details on the experimental setup. The code used for the experiments in this chapter can be found at https://github.com/oxcsml/constrained-diffusion.

4.6.1. Method characterisation on convex polytopes. First, we aim to assess the empirical performance of our methods on constrained manifolds of increasing dimension. For this, we focus on polytopes and construct synthetic datasets that represent bimodal distributions. In particular, we investigate two specific instances of polytopes: hypercubes and simplices. In appendix D.7.1 we also present results on the Birkhoff polytope. We quantify the performance of each model via the

Table 4.1. MMD metrics between samples from synthetic distributions and trained constrained and unconstrained (Euclidean) diffusion models. Means and confidence intervals are computed over 5 different runs.

Space	d	Log-barrier		Reflected		Euclidean	
		MMD	% in $\mathcal{M}$	MMD	% in $\mathcal{M}$	MMD	% in $\mathcal{M}$
$[-1,1]^d$	2	.066 ± .006	100.0	.055 ± .015	100.0	.062 ± .011	98.8
	3	.209 ± .077	100.0	.080 ± .004	100.0	.076 ± .004	98.5
	10	.330 ± .004	100.0	.313 ± .048	100.0	.081 ± .005	96.4
Δ^d	2	.050 ± .012	100.0	.043 ± .002	100.0	.055 ± .013	96.4
	3	.238 ± .009	100.0	.181 ± .007	100.0	.068 ± .014	96.3
	10	.275 ± .015	100.0	.290 ± .009	100.0	.060 ± .003	92.6

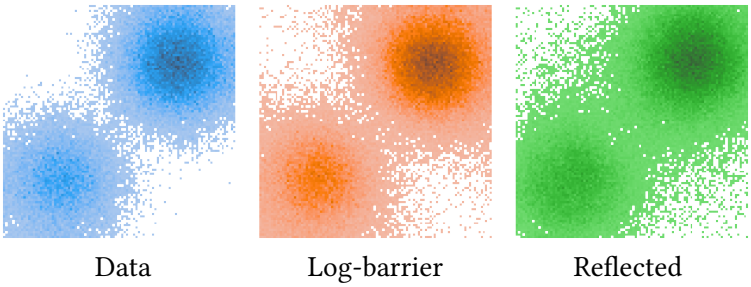
Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), which is a kernel-based metric between distributions. We present a qualitative comparison of the log barrier and reflected Brownian motion models in figure 4.9, and observe that both methods recover the data distribution on the two-dimensional hypercube and simplex, although the reflected method produces a better fit. In table 4.1, we report the MMD between the data distribution and the learnt diffusion models across a range of dimensions. Here, we similarly observe that the reflected method consistently yields better results than the log-barrier one.

In addition, we compare both of these models to unconstrained Euclidean diffusion models. We note that they are outperformed by the constrained models in lower dimensions, but generate better results in higher dimensions. In contrast, however we do see that the unconstrained models do not keep their modelled density inside the boundary, and this gets worse with increasing dimension. There are a number of potential explanations for this: First, we note that the mixture of Normal distributions we use as a synthetic data-generating process places significantly less probability mass near the boundary as its dimension increases, more closely resembling an unconstrained mixture distribution that is easier for the Euclidean diffusion models to learn, while posing a challenge to the log-barrier and reflected diffusion models that initialise at the uniform distribution within the constraints. Additionally, we note that the design space and hyperparameters used for all experiments were informed by best practices for Euclidean models that may be suboptimal for the more complex dynamics of constrained diffusion models.

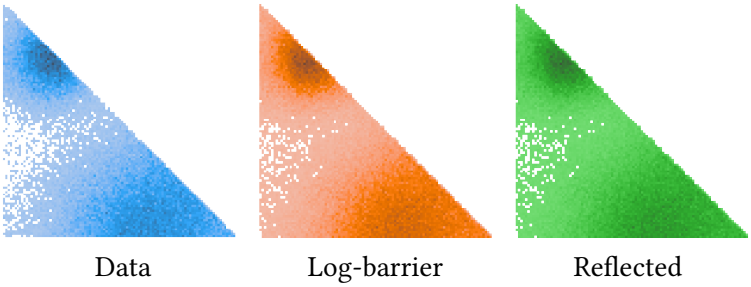
4.6.2. Modelling robotic arms under force and velocity constraints. Accurately determining and controlling the movement and exerted forces of robotic platforms is a fundamental problem in many real-world robotics applications. A kinetostatic descriptor that is commonly used to quantify the ability of a robotic arm to move and apply forces along certain dimensions is the so-called *manipulability ellipsoid* $E \in \mathbb{R}^d$ (Yoshikawa, 1985) which is naturally described as a symmetric positive definite (SPD) matrix $\mathrm{M} \in \mathbb{R}^{d\times d}$ (Jaquier et al., 2021). The manifold of such $d \times d$ SPD matrices, denoted as S^d_{++}, is defined as the set of matrices $\{x^\top \mathrm{M} x \geq 0,\ x \in \mathbb{R}^d : \mathrm{M} \in \mathbb{R}^{d\times d}\}$.

Manipulability ellipsoid

In many practical settings, it may be desirable to constrain the volume of E to retain flexibility or limit the magnitude of an exerted force, which can be expressed



(a) 2D square data.



(b) 2D Dirichlet data.

Figure 4.9. Histograms of samples from the data distribution and from trained constrained diffusion models.

as an upper bound on the trace of M, i.e. as the inequality constraint $\sum_{i=1}^{d} \mathrm{M}_{ii} < C$ with $C > 0$. Constraining the rest of the entries of the matrix to ensure it is SPD is non-trivial.

Alternatively, we can parametrise the SPD matrices via their Cholesky decomposition. Each SPD matrix has a unique decomposition of the form $\mathrm{M} = \mathrm{L}\mathrm{L}^\top$, where L is a lower triangular matrix with strictly positive diagonal (Golub and Van Loan, 2013, p.143). Constraining the entries of this matrix simply requires ensuring the diagonal is positive. The trace of the SPD matrix is given by $\mathrm{Tr}(\mathrm{M}) = \mathrm{Tr}(\mathrm{L}\mathrm{L}^\top) = \sum_{ij} \mathrm{L}_{ij}^2$, and results in the constraint on the entries of L to live in a ball of radius C. We additionally model the two-dimensional position of the arm.

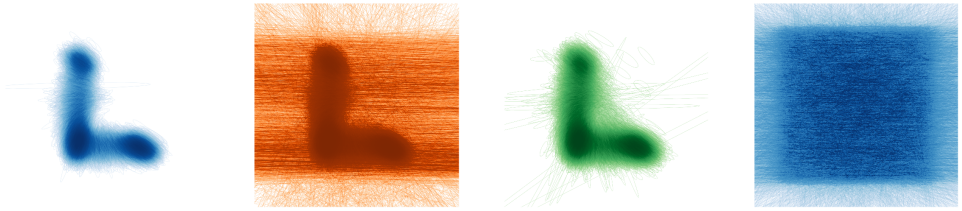
In summary, the space over which we parametrise the diffusion models is defined as

$$\left[\mathrm{L} \in \mathbb{R}^{d(d+1)/2} : \mathrm{L}_{i,i} > 0, \sum_{i,j} \mathrm{L}_{i,j}^2 < C \right] \times \mathbb{R}^2. \tag{4.20}$$

Under the Euclidean metric, we can apply both our log-barrier and reflected approaches. The positive diagonal (linear) constraint is handled similarly to the polytope setting. The reflection on the ball boundary is defined and illustrated on figures 4.6 and 4.7.

Using this framework, we model the datasets presented in Jaquier et al. (2021). This consists of demonstrations of a robotic arm drawing different letters in the plane, providing the respective positional trajectories ($\mathbb{R}^2$) and velocity manipulability ellipsoids (S_{++}^2).

We use the processing routines provided by Jaquier et al. (2021) to interpolate



(a) Samples from the data distribution.

(b) Samples from our log-barrier diffusion model.

(c) Samples from our reflected diffusion model.

(d) Samples from the uniform distribution.

Figure 4.10. Samples in $S^2_{++} \times \mathbb{R}^2$ from (a) the data distribution, (b) our log-barrier diffusion model, (c) our reflected diffusion model and (d) the uniform distribution. Each sample is visualised as the manipulability ellipsoid encoded by the SPD matrix $M \in S^2_{++}$ placed at the corresponding location in $\mathbb{R}^2$. Additional results and full correlation plots are postponed to appendix D.7.2.

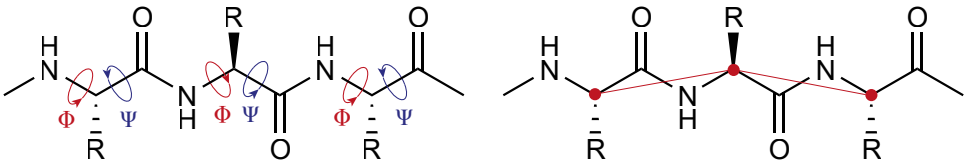
the trajectories into 10^4 distinct points, for each of which we derive the position $\mathbf{x} \in \mathbb{R}^2$ and the PSD matrix $e \in S^2_{++}$. parametrising the velocity manipulability ellipsoid $M \in \mathbb{R}^2$. The resulting data is split into training, validation, and test sets by trajectory. We add a small amount of Gaussian noise to these trajectories.

We visualise samples from this distribution as ellipsoids respecting the SPD matrix placed at a point in the plane. The dataset, samples from the trained log-barrier and reflected diffusion models, and the uniform distribution are shown in figure 4.10. We qualitatively observe that the reflected method is able to fit the joint data distribution better than the log-barrier method. Quantitatively, the reflected method achieves an MMD of $0.161_{\pm 0.003}$ with respect to the dataset as opposed to $0.247_{\pm 0.012}$d for the log-barrier method.

4.6.3. Modelling protein loops with anchored endpoints. Polypeptides and proteins constitute an important class of biogenic macromolecules that underpin most aspects of organic life. Accurately modelling their conformational ensembles,i.e. the set of three-dimensional structures they assume under physiological conditions, is essential to both understanding the biological function of existing and designing the enzymatic or therapeutic properties of novel proteins (Lane, 2023). Motivated by the success of diffusion models in computer vision and natural language processing, there has been considerable interest in applying them to learn and sample from distributions over the conformational space of protein structures (Watson et al., 2022; Trippe et al., 2022; Wu et al., 2024a).

Problem parameterisation

Proteins are biopolymers in which a sequence of N amino acids is joined together through $N-1$ peptide bonds, resulting in a so-called polypeptide backbone with protruding amino acid residues. As the deviation of chemical bond lengths and angles from their theoretical optimum is generally negligible, the problem of modelling the three-dimensional structure of this polypeptide chain is often reframed in the space of the internal torsion angles Φ and Ψ, see figure 4.11a for an illustration, which can be modelled on a $(2N-2)$-dimensional torus $\mathbb{T}^{2N-2}$.



(a) A commonly-used approximate parametrisation of backbone geometry only considers the C_α torsion angles Φ and Ψ.

(b) As peptide bond orientations can be inferred relatively reliably, researchers often only model the C_α traces.

Figure 4.11. Standard approaches to modelling the conformations of polypeptide backbones.

In many data-scarce settings such as antibody or enzyme design, it is often unnecessary or even undesirable to model the structure of an entire protein, as researchers are primarily interested in specific functional sites with distinct biochemical properties. However, generating conformational ensembles for such substructural elements necessitates positional constraints on their endpoints to ensure that they can be accommodated by the remaining protein scaffold. While it is conceivable that a diffusion model could derive such constraints from limited experimental data, we argue that it is much more efficient and precise to encode them explicitly.

For this purpose, we adopt the distance constraint formulation from Han and Rudolph (2006) and interpret the backbone as a spatial chain with N spherical joints and fixed-length links, see figure 4.12a for an illustration. After selecting a suitable anchor point, the geometry of the polypeptide chain is fully specified by (a) the set of link lengths $\ell = \{\ell_j\}_{j=1}^{N}$, (b) the set of vectors $\mathbf{r} = \{r(1,j)\}_{j=2}^{N}$ between the anchor point and each atom in the chain, and (c) the set of dihedral angles

$$\mathbf{T} = \left\{ \arccos\left(\frac{|\langle r(1,j) \times r(1,j+1), r(1,j+1) \times r(1,j+2) \rangle|}{|r(1,j) \times r(1,j+1)||r(1,j+1) r(1,j+2)|} \right) \right\}_{j=2}^{N-2} \in \mathbb{T}^{N-3} \tag{4.21}$$

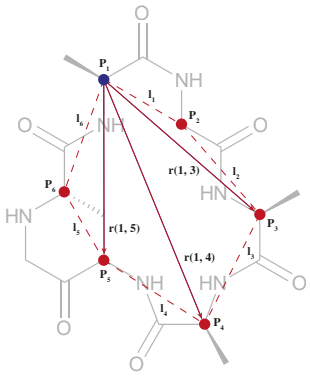
between any three consecutive vectors. After specifying the fixed bond lengths ℓ, including an arbitrary anchor point distance $d_{\text{anchor}} = \ell_N = r(1,N)$, the set of valid vectors $\mathbf{r}$ is given by the convex polytope $\mathbb{P} \subseteq \mathbb{R}^{N-3}$ defined by the following linear constraints (see figure 4.12b for an illustration):

$$\begin{aligned} r(1,3) &\leq \ell_1 + \ell_2, \\ -r(1,3) &\leq -|\ell_1 - \ell_2|, \end{aligned} \tag{4.22}$$

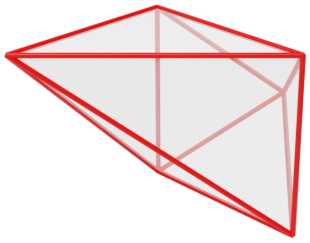
$$\left.\begin{aligned} r(1,j) - r(1,j+1) &\leq \ell_j, \\ -r(1,j) + r(1,j+1) &\leq \ell_j, \\ -r(j) - r(j+1) &\leq -\ell_j, \end{aligned}\right\} \; 3 \leq j \leq N-2, \tag{4.23}$$

$$\begin{aligned} r(1,N-1) &\leq \ell_{N-1} + d_{\text{anchor}}, \\ -r(1,N-1) &\leq -|\ell_{N-1} - d_{\text{anchor}}|. \end{aligned} \tag{4.24}$$

This means that the set of all valid polypeptide backbone conformations is defined by the product manifold $\mathbb{P} \times \mathbb{T}^{N-3}$, enabling us to train diffusion models that exclusively generate conformations with a fixed anchor point distance d_{anchor}.



(a) An illustrative diagram of the proposed parametrisation for modelling the C_α trace geometry of the cyclic peptide c-AAGAGG.



(b) The convex polytope constraining the diagonals of the triangles for the given bond lengths in the illustrated molecule. The total design space is the product of this polytope with the 4D flat torus.

Figure 4.12. Parametrising the conformational space of polypeptide backbones under anchor point distance constraints.

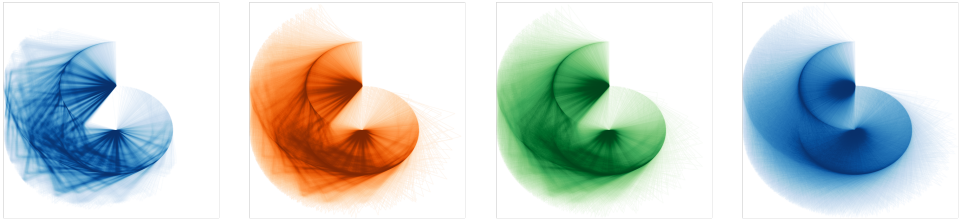
Data generation

As a proof-of-concept for the practicality of our methods, we chose to model the conformational distribution of the cyclic peptide c-AAGAGG. Cyclic peptides are an increasingly important drug modality with therapeutic uses ranging from antimicrobials to oncology, exhibiting circular polypeptide backbones (i.e. $d_{\text{anchor}} = 0$) that confer a range of desirable pharmacodynamic and pharmacokinetic properties (Dougherty et al., 2019). To reduce the dimension of the problem, we only consider the C_α traces (with fixed C_α-C_α link distances of 3.6 Å) instead of the full polypeptide backbone (see figure 4.11b), although we note that our framework is applicable to both settings.

To derive a suitable dataset, the product manifold $\mathbb{P} \times \mathbb{T}^3$ describing the conformations of cyclic C_α traces of length $N = 6$ was constructed, see figures 4.12a and 4.12b for an illustration, and used to generate 10^7 uniform samples satisfying the anchor point distance constraint $d_{\text{anchor}} = 0$. Subsequently, an estimate of the free energy E_i of each sample i was obtained by (1) reconstructing the full-atom peptide from each C_α trace using the PULCHRA algorithm (Rotkiewicz and Skolnick, 2008), (2) relaxing all non-C_α backbone and side-chain atoms (keeping the C_α positions fixed), and (3) quantifying the potential energy of each of the resulting conformations using the OpenMM suite of molecular dynamics tools (Eastman et al., 2017), and the AMBER force field (Hornak et al., 2006). These free energy estimates were then used to approximate the Boltzmann distribution over conformational states

$$p_B(i) \propto \exp\left(-\frac{E_i}{k_B T}\right), \tag{4.25}$$

where temperature was set to $T = 273.15\,\text{K}$ and $k_B = 1.380\,649 \times 10^{-23}\,\text{J}\,\text{K}^{-1}$ is the Boltzmann constant. We then apply a very minor amount of smoothing to the resulting distribution by running forward Brownian motion on both the polytope



(a) Samples from the data distribution.

(b) Samples from our log-barrier diffusion model.

(c) Samples from our reflected diffusion model.

(d) Samples from the uniform distribution.

Figure 4.13. Planar projection of the modelled C_α chains from (a) the training dataset, (b) our log-barrier diffusion model, (c) our reflected diffusion model and (d) the uniform distribution. Additional results and full correlation plots are postponed to appendix D.7.3.

and the torus for 10 steps, using a small step size of 5×10^{-3} and the respective metrics. Finally, a subsample of 10^6 C_α traces was drawn from this distribution and used for training and evaluating our models.

Modelling the problem

To learn a distribution over this space, we leverage the methodology introduced in sections 4.2 and 4.3 for the polytope component $\mathbb{P}^3$ and chapter 3 for the torus component $\mathbb{T}^4$.

A qualitative comparison of samples from the data distribution, our log-barrier and reflected diffusion models, and the uniform distribution is presented in figure 4.13. For enhanced visual clarity, we project the modelled spatial chain onto the 2D plane by removing the (unconstrained) torus component of the product manifold and only plotting the planar chains encoded by the (constrained) polytope component (a correlation plot of the full product manifold is presented in figure D.8).

It is immediately obvious that the data distribution is highly multimodal, encompassing a large number of locally optimal conformational clusters. Nevertheless, both our reflected and log-barrier diffusion models are able to robustly approximate this challenging energetic landscape, producing samples that reflect key conformational states and yielding similar MMD metrics of $0.032_{\pm 0.021}$ and $0.032_{\pm 0.001}$, respectively. As a point of comparison, the uniform distribution on the polytope-torus product has an MMD of $0.112_{\pm 0.001}$. The improved relative performance of the log-barrier method seems likely due to the fact that the data for this experiment is largely in the interior of the constrained set, not close to the boundary.

4.7. Conclusion

Learning complex distributions supported on constrained spaces is a crucial task in many natural and engineering sciences, including computational statistics (Morris, 2002), robotics (Han and Rudolph, 2006), quantum physics (Lukens et al., 2020) and computational biology (Thiele et al., 2013). In this chapter we extend continuous diffusion models to this setting, proposing two complementary approaches—one

based on log-barrier methods and the other on the reflected Brownian motion. For both approaches, we derive the time-reversal formula, propose discretisation schemes and extend the score-matching toolbox. We demonstrated the utility of our methods on a range of synthetic and real-world tasks. This including the constrained conformational modelling of proteins and robotic arms, and find that reflected methods, while enjoying fewer theoretical guarantees than their log-barrier counterparts, often yield preferable results.

We conclude by highlighting important future directions. First, the computational cost of performing the reflection when discretising the reflected Brownian motion is high. Finding numerically efficient approximations of the reflected process is therefore necessary to extend this methodology to high dimension settings. Second, the retraction used in place of the exponential map for the barrier method leads to a high number of discretisation steps to ensure a good approximation. Designing a faster forward process for the log-barrier method is key to targeting more complex distributions.

5 | Efficient score-based modelling on constrained domains

Work in this chapter is based on

N. Fishman, L. Klarner, E. Mathieu, M. J. Hutchinson, and V. De Bortoli. Metropolis Sampling for Constrained Diffusion Models. In *Advances in Neural Information Processing Systems*, 2023.

and has been rewritten for this thesis with additional material.

Personal contributions:

1. Project conception with Valentin, Nic, Leo.
2. Development of practical approaches with Nic, Leo, Valentin, including the problem-specific details.
3. Development of the code with Nic and Leo.
4. Running experiments: Robotics and convergence speed experiments.
5. Supervision of the project.

5.1. Introduction

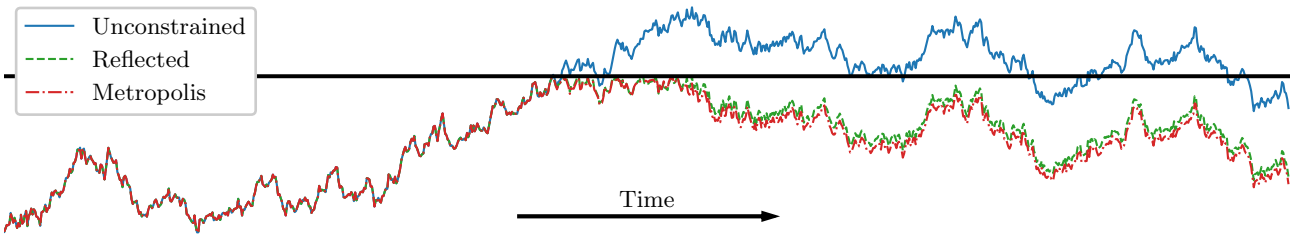


Figure 5.1. Visual comparison of a discretisation of the unconstrained Brownian motion, two discretisations of the reflected Brownian motion: one based on a reflection scheme and the other based on our Metropolis sampler. The Metropolised trajectory is very close to that of the reflected one, while being significantly easier to implement and cheaper to compute.

IN THE PREVIOUS CHAPTER we extended the applicability of diffusion models to inequality-constrained manifolds by investigating the generative modelling applications of classic sampling schemes based on log-barrier methods (Kannan and Narayanan, 2009; Lee and Vempala, 2017; Noble et al., 2023; Kook et al., 2022; Gatmiry and Vempala, 2022; Lee and Vempala, 2018) and the reflected Brownian motion (Williams, 1987; Petit, 1997; Shkolnikov and Karatzas, 2013). While empirically promising, the proposed algorithms can be computationally and numerically burdensome, and require bespoke implementations for different manifolds and constraints.

In this chapter, we propose a new method for generative modelling on constrained manifolds based on a Metropolis discretisation of the reflected Brownian motion. This Metropolised discretisation has two main advantages:

1. From a technical perspective it is lightweight: the only additional requirement over those outlined in chapter 3 that is needed to implement a constrained diffusion model is a binary function that indicates whether any given point is within the constrained set. This is in contrast to the methods of chapter 4 which require significantly more components.

2. From a computational perspective it is much cheaper to compute, and more numerically stable than the methods proposed in chapter 4.

Our core theoretical contribution is to show that this new discretisation converges to the reflected stochastic differential equation by using the invariance principle for stochastic differential equations with boundary (Stroock and Varadhan, 1971).

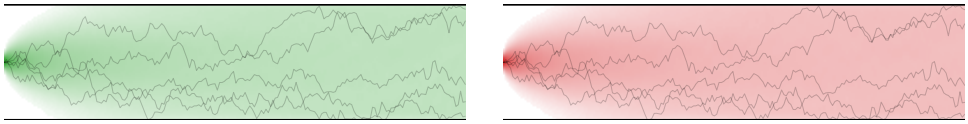
To the best of our knowledge, this is the first time that such a process has been investigated in a machine learning context, or elsewhere. We demonstrate that our method attains improved empirical results on manifolds with convex and non-convex constraints by applying it to a range of problems from geospatial modelling, robotics and protein design.

5.2. Diffusion models for constrained manifolds via Metropolis sampling

In section 5.2.1, we highlight the practical limitations of existing constrained diffusion models in section 5.2.2 propose an alternative Metropolis sampling-based approach. In section 5.2.3, we outline our proof that this process corresponds to a valid discretisation of the reflected Brownian motion, justifying its use in diffusion models. An overview of the samplers, and the various operations required to run them, we cover in this section are presented in table 5.1. Table 5.2 compares the computational cost of training various instantiations of constrained diffusion models, and the geometric domains they cover.

The technical setting we consider in this chapter is the same as in chapter 4. Briefly, given a Riemannian manifold $(\mathcal{N}, h)$, we consider a family of real functions $\{f_i : \mathcal{N} \to \mathbb{R}\}_{i \in \mathcal{I}}$ indexed by $\mathcal{I}$. We then define

$$\mathcal{M} = \{x \in \mathcal{N} \ : \ f_i(x) < 0, \ i \in \mathcal{I}\}, \tag{5.1}$$



(a) Reflected Brownian motion

(b) Metropolised approximation of Brownian motion

Figure 5.2. Evolution of the density of the reflected Brownian motion and its Metropolis sampling-based approximation on the unit interval starting from a delta mass.

and derive models such that their densities are constrained to be in the set $\mathcal{M}$. A more detailed description of the setting can be found in section 4.1.1.

5.2.1. Practical limitations of existing constrained diffusion models.

Barrier metrics. In the barrier approach, section 4.2, the constrained manifold is transformed into an unconstrained space via a barrier metric. This metric is defined by

$$g = \nabla^2 \phi(x), \quad \phi(x) = \sum_{i \in \mathcal{I}} \phi(d(x, f_i)), \tag{5.2}$$

where $d(x, f_i)$ is the minimum distance from the point x to the set defined by $f_i(x) = 0$ and ϕ_i is a monotone decreasing function such that $\lim_{z \to 0} \phi_i(z) = \infty$. Under additional regularity assumptions, ϕ is called a *barrier function*, see (Nesterov and Nemirovskii, 1994). This metric replaces the original metric on the space, warping the geometry. This definition ensures that the barrier function induces a well-defined exponential map on the manifold, making the Riemannian diffusion model framework chapter 3 apply.

In practice, evaluating ϕ requires computing $d(x, \partial f_i)$ (and its derivatives), which can be prohibitively expensive. Furthermore, since the exponential under the induced manifold is not easy to compute, in the barrier method it is approximated by projecting the exponential of the original metric back into the constrained set. This incurs additional bias and necessitating a step projecting samples back inside the constraints, which can itself be computationally intractable, see Absil and Malick (2012) and Boumal (2023).

Mirror diffusion models

Mirror diffusion models (Tae, 2023; Liu et al., 2024) work on similar principles to the barrier methods in section 4.2. Instead of modelling on the original space, they project the data space under the map $\nabla\phi : \mathcal{M} \to \mathbb{R}^d$. For convex $\mathcal{M} \subset \mathbb{R}^d$ this is a bijection. This allows for the placing of a diffusion model on Euclidean space and pulling it back onto the original constrained space, a procedure reminiscent of the baseline introduced in section 3.4.1. It suffers from the same issue as the barrier method, warping the underlying geometry of $\mathcal{M}$, but does enjoy faster training.

Reflected stochastic processes. The alternative approach explored in chapter 4 is based on reflected Brownian motion. One possible discretisation of the reflected SDE is to (i) consider a classical step of the Euler-Maruyama discretisation (or the Geodesic Random Walk in the Riemannian setting) and (ii) reflect this step

according to the boundary defined by $\partial\mathcal{M}$. To compute the reflection, one must check whether the step crosses the boundary. If it does, the point of intersection needs to be calculated, the ray reflected at this point, and the step continued in the reflected direction. This can require an arbitrarily large number of reflections depending on the step size, the geodesic on the manifold, and the geometry of the bounded region within the manifold. We refer to section 4.3 and algorithm 4.4 for the detail of the typical reflection discretisation.

Lou and Ermon (2023) also consider a reflected Brownian motion for diffusion modelling. By considering a restricted setting of the hypercube they obtain an analytic form of sampling for the forward noising process and denoising score matching. However, to model more complex constraints they require a step to warp the constraints into the hypercube, modifying the geometry of the problem.

An alternative approach to discretising a reflected stochastic differential equation is to replace the reflection with a projection, see (Słomiński, 1994) for instance. However, the projection still requires the most expensive part of the reflection algorithm: computing the intersection of the geodesic with the boundary.

5.2.2. Metropolis approximation of reflected Brownian motion. As outlined in the previous section, both of the existing approaches for constrained diffusion models require manifold- and constraint-specific implementations, and become computationally intractable as the complexity and dimension of the modelled geometry increases.

This limits their practical usefulness even for relatively simple manifolds with well-defined exponential maps and linear inequality constraints (such as e.g. polytopes).

In algorithm 5.1 we introduce a method that aims to solve both of these problems.

Metropolis approximation of reflected Brownian motion step

Algorithm 5.1

Require: $p \in \mathcal{M}$, $\{f_i\}_{i\in\mathcal{I}}$
  Sample $\boldsymbol{v} \sim \mathrm{N}(0, \mathbf{I}) \in \mathrm{T}_p\mathcal{M}$
  $p' \leftarrow \exp_p(\boldsymbol{v})$
  if $f_i(p') < 0 \;\forall\; i$ **then**
    $p \leftarrow p'$
  return p

The discretisation step sampler we propose is similar to a classical Euler-Maruyama discretisation of the Brownian motion, except that, whenever a step would carry the Brownian motion outside the constrained region, we reject it. This is a *Metropolised* version of the usual discretisation and is almost trivial to implement compared to the existing barrier, reflection, and projection methods. Hence, this method enables the principled extension of diffusion models to arbitrarily constrained domains at virtually *no added implementation complexity or computational expense.*

5.2.3. Convergence of the Metropolis sampler to the reflected Brownian motion. In this section, we prove that the proposed Metropolis sampling-based process (algorithm 5.1) corresponds to a valid discretisation of the reflected process, eq. (4.15),

Table 5.1. Comparing the requirements of different constrained diffusion models. Distance to boundary and intersection operations are typically significantly more expensive to compute, and in some setting intractable. Set indicator functions are typically more lightweight and more commonly available on any given domain. Methods that rely on set indicator functions only are therefore computationally efficient.

Sampling Method	Set Indicator Function	Intersection Operator	Distance to Boundary
Barrier method (Chapter 4)	✓	✓	✓
Mirror diffusion (Liu et al., 2024)	✓	✓	✓
Reflected method (Chapter 4, Lou and Ermon (2023))	✓	✓	✗
Projection sampler (Lou and Ermon, 2023)	✓	✓	✗
Metropolis sampler (This chapter)	✓	✗	✗

Table 5.2. Comparison of the advantages and disadvantages of the different constrained (Riemannian) diffusion models covered in Section 5.2.1 with regard to training the score function and preserving the geometry of the domain. The methods proposed in chapter 4 and this chapter can be applied to arbitrary manifolds while preserving the underlying geometry of the manifold. However, this comes at the cost of making denoising score matching intractable and falling back on implicit score matching. By contrast the methods of Lou and Ermon (2023) and Liu et al. (2024) allow for the use of denoising score matching, but are limited to being applied ot Euclidean space, or warping the underlying geometry of a space. It should be noted that if the reflected method of chapter 4 is applied to the same setting as Lou and Ermon (2023) then we can make use of denoising score matching in the same way.

	Both required for fast DSM loss			
Diffusion Model	Tractable conditional score	Simulation-free forward sampling	Modelling domain	Preserves metric of $\mathcal{M}$
Reflected Diffusions				
Lou and Ermon (2023)	✓	✓	convex $\subset \mathbb{R}^d$	✗
chapter 4	✗	$O(d^2)$	convex $\subset$ any $\mathcal{M}$	✓
This chapter	✗	$O(d)$	convex $\subset$ any $\mathcal{M}$	✓
Barrier Diffusions				
chapter 4	✗	✗	convex $\subset$ any $\mathcal{M}$	✗
Liu et al. (2024)	✓	✓	convex $\subset \mathbb{R}^d$	✗

justifying its use in diffusion models. Here we focus on a concise presentation of the core concepts and the main results. A full proof can be found in appendix E.1.

Technical setting of the result

For simplicity, we restrict ourselves to the Euclidean setting with smooth convex boundaries. More specifically, for a function $\Phi : \mathbb{R}^n \to \mathbb{R}$ defining the constrained set, $\mathcal{M} = \{x \in \mathbb{R}^d : \Phi(x) > 0\}$, we assume that this set is bounded and that $\Phi \in C^2(\mathbb{R}^d, \mathbb{R})$ concave. In addition, we assume that for any $x \in \partial\mathcal{M}$, $\|\nabla\Phi(x)\| = 1$. These assumptions match those Stroock and Varadhan (1971) use for their study of the existence of solutions to the reflected Brownian motion.

While it seems possible to relax the *global* existence of Φ to a *local* one, i.e. local concavity, the global regularity assumption of the domain, $\Phi \in C^2(\mathbb{R}^d, \mathbb{R})$, is key. This regularity is essential to establish proposition 5.7 and the associated

geometrical result on tubular neighbourhoods. The smoothness of the domain boundary is central in the results of Kang and Ramanan (2017) on the equivalence of two definitions of reflected Brownian motion which we rely on. As in chapter 4, we note it is possibly to infinitesimally smooth non-smooth boundaries to match the technical setting with very minimal practical impact.

It also seems possible to extend these results to the more general manifold setting, but this is technically challenging and outside the scope of this work.

In section 5.4.3 we show however that the proposed algorithm does appear to practically work on a non-Euclidean manifold with non-convex, non-smooth boundaries.

High level proof

We begin with a definition of the Metropolis approximation of reflected Brownian motion.

Metropolis approximation of reflected Brownian motion

DEFINITION 5.2. *For any $\gamma > 0$ and $k \in \mathbb{N}$, let $X_0^\gamma \in \mathcal{M}$ and*

$$X_{k+1}^\gamma = \begin{cases} X_k^\gamma + \sqrt{\gamma} Z_k^\gamma & \text{if} \quad X_k^\gamma + \sqrt{\gamma} Z_k^\gamma \in \mathcal{M} \\ X_k^\gamma & \text{else} \end{cases} . \tag{5.3}$$

For $Z_k \sim \mathcal{N}(0, I)$. The sequence $(X_k^\gamma)_{k \in \mathbb{N}}$ is called the METROPOLIS APPROXIMATION OF REFLECTED BROWNIAN MOTION.

For any $\gamma > 0$, we consider $(\boldsymbol{X}_t^\gamma)_{t \geq 0}$, the linear interpolation of $(X_k^\gamma)_{k \in \mathbb{N}}$ such that for any $k \in \mathbb{N}$, $\boldsymbol{X}_{k\gamma}^\gamma = X_k^\gamma$. The following result is the main theoretical contribution of this section.

Convergence of the Metropolis approximation to the reflected Brownian motion

THEOREM 5.3. *Under assumptions on $\mathcal{M}$, for any $T \geq 0$, $(\boldsymbol{X}_t^\gamma)_{t \in \{0,T\}}$ weakly converges to the reflected Brownian motion $(\bar{\boldsymbol{B}}_t)_{t \in \{0,T\}}$ as $\gamma \to 0$.*

The rest of the section is devoted to a high level presentation of the proof of theorem 5.3.

It is theoretically impractical to work directly with the Metropolis approximation of reflected Brownian motion. Instead, we introduce an auxiliary process, show this converges to the reflected Brownian motion, and finally prove that the convergence of the auxiliary process implies the convergence of our Metropolis discretisation. We define this auxiliary process now.

Rejection approximation of reflected Brownian motion

DEFINITION 5.4. *For any $\gamma > 0$ and $k \in \mathbb{N}$, let $\hat{X}_0^\gamma \in \mathcal{M}$ and $\hat{X}_{k+1}^\gamma = \hat{X}_k^\gamma + \sqrt{\gamma} Z_k^\gamma$ with Z_k^γ a Gaussian random variable conditioned on $\hat{X}_k^\gamma + \sqrt{\gamma} Z_k^\gamma \in \mathcal{M}$. The sequence $(\hat{X}_k^\gamma)_{k \in \mathbb{N}}$ is called the* REJECTION APPROXIMATION OF REFLECTED BROWNIAN MOTION.

We call this process the *rejection approximation of reflected Brownian motion* since in practice, Z_k^γ can be sampled using rejection sampling, see algorithm 5.5.

Algorithm 5.5 Rejection approximation of reflected Brownian motion

Require: $p \in \mathcal{M}$, $\{f_i\}_{i \in \mathcal{I}}$
  Sample $\boldsymbol{v} \sim \mathrm{N}(0, \mathbf{I}) \in \mathrm{T}_p\mathcal{M}$
  $p' \leftarrow \exp_p(\boldsymbol{v})$
  while $f_i(p') \geq 0$ for any i **do**
    Sample $\boldsymbol{v} \sim \mathrm{Id}(0, 1) \in \mathrm{T}_p\mathcal{M}$
    $p' \leftarrow \exp_p(\boldsymbol{v})$
  return p'

The key difference between the rejection and Metropolis approaches is that where the Metropolis approach will leave the point stationary if the update falls outside the boundary, the rejection approach will resample the update until it lies inside the boundary. This leaves the transition measures of these two processes the same, except the Metropolis version contains a Dirac component at the initial point. While it is possible to implement the rejection approach practically, it is less desirable as it results in loops of unknown length in the sampling.

For any $\gamma > 0$, we also consider $(\hat{\boldsymbol{X}}_t^\gamma)_{t \geq 0}$, the linear interpolation of $(\hat{X}_k^\gamma)_{k \in \mathbb{N}}$ such that for any $k \in \mathbb{N}$, $\hat{\boldsymbol{X}}_{k\gamma}^\gamma = \hat{X}_k^\gamma$. In appendix E.1, we prove the following result.

THEOREM 5.6. *Under assumptions on $\mathcal{M}$, for any $T \geq 0$, $(\hat{\boldsymbol{X}}_t^\gamma)_{t \in \{0,T\}}$ weakly converges to the reflected Brownian motion $(\bar{\boldsymbol{B}}_t)_{t \in \{0,T\}}$ as $\gamma \to 0$.* Convergence of the rejection approximation to the reflected Brownian motion

Proof. For the full proof see appendix E.1. Here we give some elements of the proof.

Our approach is based on the invariance principle of Stroock and Varadhan (1971). More precisely, we show that we can compute an equivalent 'drift' and 'diffusion matrix' for the interpolated discretised process and that, as $\gamma \to 0$, the drift converges to zero and the diffusion matrix converges to $\mathbf{I}$.

In the Euclidean setting, this result, accompanied by mild regularity and growth assumptions, ensures that the discretisation weakly converges to the original stochastic differential equation. However, the case with boundary is much more complicated, primarily because the approximate drift might explode near the boundary, thus we need to verify exactly how the drift behaves as $\gamma \to 0$ and as the process approaches the boundary.

We show that the *normalised* drift converges to the inward normal near the boundary. This ensures that

(a) in the interior of $\mathcal{M}$ the drift converges to zero, i.e. locally in the interior of $\mathcal{M}$ the Brownian motion and the reflected Brownian motion coincide,

(b) on the boundary, the drift pushes the samples inside the manifold according to the inward normal, mimicking $(\boldsymbol{k}_t)_{t \geq 0}$ in eq. (4.15).

Finally, with results from Stroock and Varadhan (1971) and Kang and Ramanan (2017), we show the convergence to the reflected Brownian motion. ∎

Having established that the rejection process converges to the reflected Brownian motion, our next step is to show that the approximate drift and diffusion matrix of the Metropolised process are related to the approximate drift and diffusion terms of the rejection process via multiplication by the term

$$\alpha^\gamma(x) = \mathbb{P}\left[x + \sqrt{\gamma}Z\right] \in \mathcal{M} \tag{5.4}$$

for $Z \sim \mathcal{N}(0, \mathbf{I})$. Having established this, we can conclude the proof by noting that the approximate drift and the diffusion matrix in the rejection and Metropolis case coincide as $\gamma \to 0$. This is enough to apply the same results as in the rejection process, giving the desired convergence.

This requires the following result.

PROPOSITION 5.7. *Under assumptions on $\mathcal{M}$, $\forall\ \varepsilon > 0$, $\exists\ \bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma})$ and for any $x \in \mathcal{M}$, we have $\mathbb{P}(x + \sqrt{\gamma}Z \in \mathcal{M}) \geq 1/2 - \varepsilon$, with $Z \sim \mathrm{N}(0, \mathbf{I})$.*

Proposition 5.7 tells us that *locally* the boundary looks like a half-space when integrating with respect to. a Gaussian measure. The assumption on the smoothness of the boundary is key in proving this result. A corollary is that, for $\gamma > 0$ small enough and for any $k \in \mathbb{N}$, the probability that $X_{k+1}^\gamma = X_k^\gamma$ is upper bounded *uniformly* with respect to. $X_k^\gamma \in \mathcal{M}$. The proof of proposition 5.7 uses theorem E.2 whose proof relies on the concept of *tubular neighbourhoods* (Lee, 2013).

5.3. Related work on approximations of reflected stochastic differential equations

Several schemes have been introduced to approximately sample from reflected stochastic differential equations. They can be interpreted as modifications of classical Euler-Maruyama schemes used to discretise stochastic differential equations without boundary. One of the most common approaches is to use the Euler-Maruyama discretisation and project the solution onto the boundary if it escapes from the domain $\mathcal{M}$. In this case, mean-square error rates of order *almost* $1/2$ have been proven under various conditions (Liu, 1995; Chitashvili and Lazrieva, 1981; Pettersson, 1995; Słomiński, 1994). Concretely this means that $\mathbb{E}[\|\bar{\boldsymbol{B}}_t - X_n^{t/n}\|^2] = O(n^{-1+\varepsilon})$ with $\varepsilon > 0$ arbitrary small and where $(X_k^\gamma)_{k\in\mathbb{N}}$ is the projection scheme.

The rate $1/2$ is tight (Pacchiarotti et al., 1998). It is possible to use the Euler-Peano method to get slight improvements for a mean-square error rate of order of $1/2$, but this is impractical as it assumes that one can solve a (simplified) Skorokhod problem, which is usually intractable.

Liu (1993) introduced a *penalised* method which pushes the solution away from the boundary and shows a mean-square error of order $1/4$, see also (Pettersson, 1997). Weak errors of order 1 have been obtained in Bossy et al. (2004) and Gobet (2001) by introducing a reflection component in the discretisation or using some local approximation of the domain to a half-space. We refer to Pilipenko (2014) for an introduction to the discretisation of reflected SDEs. Closer to our work,

Table 5.3. Log-likelihood (↑) of a held-out test set from a synthetic bimodal distribution over convex subsets of $\mathbb{R}^d$ bounded by the hypercube $[-1, 1]^d$ and unit simplex Δ^d. Means and standard deviations are computed over 3 different runs. Training time in hours is listed in parentheses.

CONSTRAINT	d	REFLECTED *(HOURS)*	METROPOLIS *(HOURS)*
$[-1, 1]^d$	2	2.25 ± .01 *(8.95)*	**2.32 ± .05** *(0.72)*
	3	3.77 ± .13 *(8.97)*	**4.15 ± .15** *(0.71)*
	10	7.42 ± .77 *(10.15)*	**10.80 ± .34** *(0.90)*
Δ^d	2	1.01 ± .01 *(9.17)*	**1.06 ± .02** *(0.82)*
	3	2.64 ± .01 *(9.43)*	**3.23 ± .17** *(0.78)*
	10	7.00 ± .13 *(10.53)*	**7.81 ± .20** *(0.97)*

Burdzy and Chen (2008) consider three different methods to approximate reflected Brownian motions on general domains (two based on discrete methods and one based on killed diffusions). Only qualitative results are provided. To the best of our knowledge, no previous work in the probability literature has investigated the *Metropolised* scheme we propose.

Our Metropolis scheme is also related to the ball walk (Applegate and Kannan, 1991), which replaces the Gaussian random variable with a uniform over the ball (or the Dikin ellipsoid). Applegate and Kannan (1991) and Lovász and Vempala (2007) have studied the asymptotic convergence rate of the ball walk, but, to the best of our knowledge, its limiting behaviour when the step size goes to zero has not been investigated.

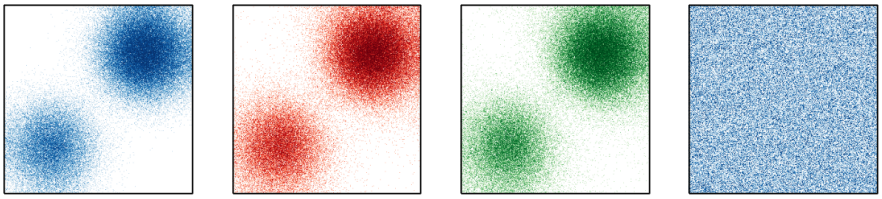
5.4. EXPERIMENTAL RESULTS

To demonstrate the practical utility and empirical performance of the proposed Metropolis diffusion models, we conduct a comprehensive evaluation on a range of synthetic and real-world tasks.

In section 5.4.1, we assess the scalability of our method by applying it to synthetic distributions on hypercubes and simplices of increasing dimension, similar to section 4.6.1. In section 5.4.2, we extend the evaluation to real-world tasks on manifolds with convex constraints by applying our method to the robotics and protein design datasets presented in sections 4.6.2 and 4.6.3.

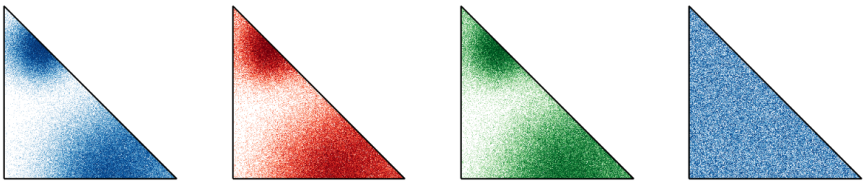
In section 5.4.3, we additionally demonstrate that our method extends to constrained manifolds with highly *non-convex* boundaries—a setting that is intractable with existing approaches. As we found—in line with chapter 5—that log-barrier diffusion models perform strictly worse than reflected approaches across all experimental settings, we focus on a more detailed comparison with the latter here and postpone additional empirical results to appendix E.3.1.

The code required to run the experiments in this chapter can be found HTTPS://GITHUB.COM/OXCSML/SCORE-SDE/TREE/METROPOLIS. Details of the experimental set up can be found in appendix E.3.2.



(a) Data distribution (b) Metropolis model (c) Reflected model (d) Uniform distribution

Figure 5.3. Qualitative comparison of samples from the data distribution, our Metropolis model, a Reflected model and the uniform distribution for a synthetic bimodal distribution on $[-1, 1]^2$.



(a) Data distribution (b) Metropolis model (c) Reflected model (d) Uniform distribution

Figure 5.4. Qualitative comparison of samples from the data distribution, our Metropolis model, a Reflected model and the uniform distribution for a synthetic bimodal distribution on Δ^2.

5.4.1. Synthetic distributions on simple polytopes. In this section, we investigate the scalability of the proposed Metropolis diffusion models by applying them to synthetic bimodal distributions over the d-dimensional hypercube $[-1, 1]^d$ and unit simplex Δ^d. A quantitative comparison of the log-likelihood of a held-out test set is presented in table 5.3, while visual comparisons of the fitted distributions are shown in figures 5.3 and 5.4.

We find that our Metropolis models outperform reflected approaches across all dimensions and constraint geometries by a substantial and statistically significant margin while training in one tenth of the time.

The degree of improvement seems to scale with the dimension of the problem; the larger the dimension of the experiment, the larger the gain in performance from using our proposed Metropolis scheme.

We observe a similar difference in the scaling properties of reflected and Metropolis models when measuring the convergence times of the respective forward noising processes to the uniform distribution on hypercubes $[-1, 1]^d$ and simplices Δ^d of increasing dimension. The results are presented in figure 5.5 and show that the convergence time of the Metropolis process scales linearly in the dimension, while the reflected process scales quadratically.

5.4.2. Modelling proteins and robotic arms under convex constraints. In addition to illustrating our method's scalability on high-dimensional synthetic tasks, we

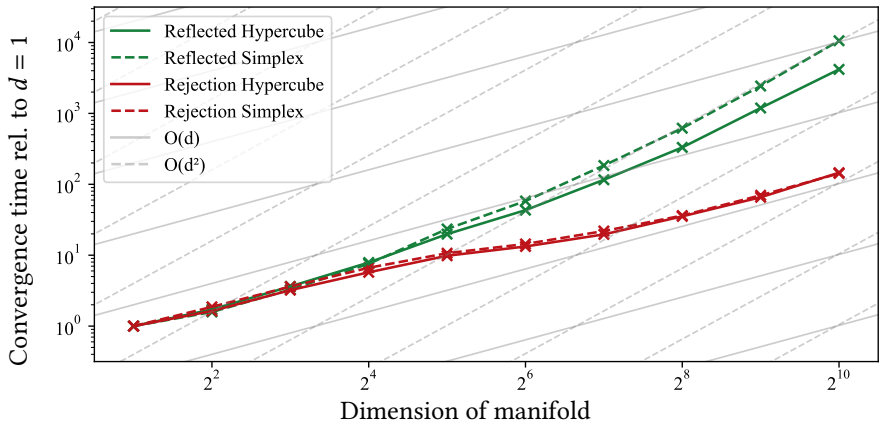
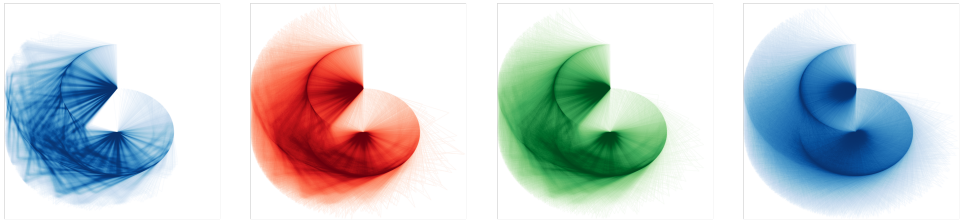


Figure 5.5. Convergence time of the Reflected (green) and Metropolis (red) forward noising processes to the uniform distribution on the hypercube $[-1, 1]^d$ and unit simplex Δ^d. The lines indicate functions fit with the PySR symbolic regression package (Cranmer, 2023) and correspond to empirical runtime complexities of $O(d^2)$ and $O(d)$, respectively, matching the superimposed scaling law isocontours.



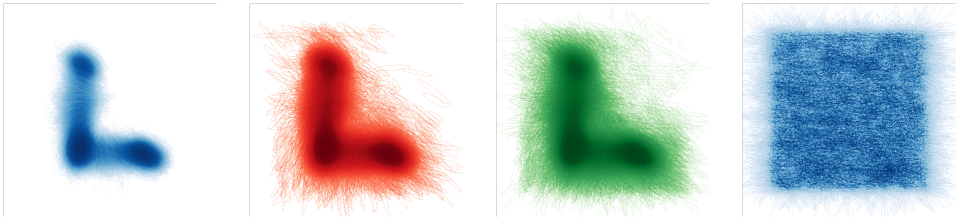
(a) Data distribution. (b) Metropolis samples. (c) Reflected samples. (d) Uniform distribution.

Figure 5.6. A qualitative comparison of 10^5 samples from the data distribution, our Metropolis diffusion model, a reflected diffusion model and the uniform distribution for the constrained conformational modelling of cyclic peptide backbones. For visual clarity, the figures only show the constrained planar projections encoded by $\mathbb{P} \subset \mathbb{R}^3$.

follow the experimental setup from sections 4.6.2 and 4.6.3 to demonstrate its practical utility and favourable empirical performance on two real-world problems from robotics and protein design. We follow an identical problem set up, data, and model hyperparameters.

We quantify the empirical performance of different methods by evaluating the log-likelihood of a held-out test set and present the resulting performance metrics in table 5.4.

Again, we find that our Metropolis model outperforms the reflected approach by a statistically significant margin while training 7-8 times as fast. Qualitative visual comparisons of samples from the true distribution, the trained diffusion models and the uniform distribution are presented in figures 5.6 and 5.7, with full univariate marginal and pairwise bivariate correlation plots postponed to appendices E.3.3 and E.3.4.



(a) Data distribution. (b) Metropolis samples. (c) Reflected samples. (d) Uniform distribution.

Figure 5.7. A qualitative comparison of 10^5 samples from the data distribution, our Metropolis diffusion model, a reflected diffusion model and the uniform distribution for the constrained modelling of manipulability ellipsoids of robotic arms. For visual clarity, the figures show the SPD matrices plotted.

Table 5.4. Log-likelihood (↑) of a held-out test set for the robotics and protein applications. Means and standard deviations are computed over 3 different runs. Training time in hours is listed in parentheses.

DATASET	DOMAIN	REFLECTED *(HOURS)*	METROPOLIS *(HOURS)*
ROBOTICS	$\mathcal{S}^2_{++} \times \mathbb{R}^2$	8.39 ± .06 *(9.52)*	**9.13 ± .03** *(1.36)*
PROTEINS	$\mathbb{P} \subset \mathbb{R}^3 \times \mathbb{T}^4$	15.20 ± .06 *(24.80)*	**15.33 ± .02** *(3.12)*

5.4.3. Modelling geospatial data within non-convex country borders. Motivated by the strong empirical performance of our approach on tasks with challenging convex constraints, we investigate its ability to approximate distributions whose support is restricted to manifolds with highly non-convex boundaries—a setting that is intractable with existing approaches. To this end, we derived a geospatial dataset based on wildfire incidence rates within the continental United States and train a Metropolis diffusion model constrained by the corresponding country borders on the sphere $\mathcal{S}^2$.

Dataset

To create the dataset we retrieved the rasterised version of the wildfire data provided by Welty and Jeffries (2020). This was converted it to a spherical geodetic coordinate system using the CARTOPY library (of the United Kingdom, 2015), and drew a weighted subsample of size 1×10^6. We then retrieved the country borders of the continental United States from Natural Earth (2023) and mapped them to the same geodetic reference frame as the wildfire data. A visualization of the resulting dataset are presented in figure 5.8a.

Point-in-spherical-polytope algorithms

The support of the data-generating distribution we aim to approximate is restricted to a highly non-convex set on the sphere, $\mathbb{P} \in \mathcal{S}^2$, given by the country borders of the continental United States. In order to apply our Metropolis method we need an efficient binary function to tell us if a point on the sphere lies inside this set.

To do so we adapt an efficient reformulation of the point-in-spherical-polygon

algorithm (Bevis and Chatelain, 1989) presented in (Ketzner et al., 2022). The algorithm requires the provision of a reference point $r \in \mathcal{S}^2$ known to be located in $\mathbb{P}$ and determines whether q is inside or outside the polygon by checking whether the geodesic between r and q crosses the polygon an even or odd number of times.

Letting $\hat{x} \in \mathbb{R}^3$ denote the Cartesian coordinates of a point $x \in \mathcal{S}^2$ embedded in Euclidean space. (Ketzner et al., 2022) rely on a Cartesian reference coordinate system $\hat{Q}$ (with its z-axis given by $\hat{r}$) and the corresponding spherical coordinate system Q to decompose the edge-crossing condition of Bevis and Chatelain (1989) into two efficiently computable parts. That is, the geodesic between q and r crosses an edge $e_i = (v_i, v_j)$ of the polygon if:

(i) the longitude of q in Q is bounded by the longitudes of v_i and v_j in Q, i.e.

$$\phi_Q(q) \in \big\{\min\big(\phi_Q(v_i), \phi_Q(v_j)\big), \max\big(\phi_Q(v_i), \phi_Q(v_j)\big)\big\}, \tag{5.5}$$

(ii) the plane specified by the normal vector $\hat{p}_i = \hat{v}_i \times \hat{v}_j$ represents an equator that separates q and r into two different hemispheres, i.e.

$$\operatorname{sign}(\langle \hat{p}_i, \hat{r} \rangle \cdot \langle \hat{p}_i, \hat{q} \rangle) = -1.$$

Especially when $\mathbb{P}$ is fixed and the corresponding coordinate transformations and normal vectors can be precomputed for each edge, this algorithm affords an efficient and parallelisable approach to determining whether any given point on $\mathcal{S}^2$ is contained by a spherical polytope.

In order to discretise Brownian motion on the manifold we use modified geodesic random walks, section 3.2.6, by including the Metropolis step, algorithm 5.1, to account for the reflection of the process.

Results

A qualitative visual comparison of samples from the true distribution, our model, and the uniform distribution is presented in figure 5.8 and a quantitative comparison to an unconstrained Riemannian score-based model from chapter 3 is given in table 5.5. This demonstrates that our approach is able to successfully model challenging multimodal and sparse distributions on non-flat constrained manifolds with highly non-convex boundaries.

Table 5.5. MMD (↓) of a held-out test set for the geospatial modelling dataset. Means and standard deviations are computed over 3 different runs. Average training time is provided in hours.

MODEL	DOMAIN	MMD	RUNTIME	% IN BOUNDARY
Unconstrained	$\mathcal{S}^2$	$0.1567 \pm_{0.013}$	**0.81**	63.3
Metropolis	$\mathbb{P} \subset \mathcal{S}^2$	$\mathbf{0.1388 \pm_{0.015}}$	3.86	**100.0**

(a) Data distribution. (b) Metropolis samples. (c) Uniform distribution.

Figure 5.8. Orthographic projection of 10^5 samples from (a) the data distribution, (b) our Metropolis diffusion model, and (c) the uniform distribution, for geospatial data (wildfire incidence rates) within a non-convex boundary (the continental United States). The projections are aligned with the geometric centre of the boundary and zoomed in ten-fold for visual clarity.

5.5. Conclusion

Accurately modelling distributions on constrained Riemannian manifolds is a challenging problem with a range of practical applications. In this chapter, we have proposed a mathematically principled and computationally scalable extension of the existing diffusion model methodology to this setting. Based on a *Metropolisation* of random walks in Euclidean spaces and on Riemannian manifolds, we have shown that our approach corresponds to a valid discretisation of the reflected Brownian motion, justifying its use in diffusion models. To demonstrate the practical utility of our method, we have performed an extensive empirical evaluation, showing that it outperforms existing constrained diffusion models on a range of synthetic and real-world tasks defined on manifolds with convex boundaries, including applications from robotics and protein design. Leveraging the flexibility and simplicity of our method, we have also demonstrated that it extends beyond convex constraints and is able to successfully model distributions on manifolds with highly non-convex boundaries.

While we found our method to perform well across the synthetic and real-world applications we considered, we expect it to perform poorly on certain constraint geometries. For instance, the current implementation relies on an isotropic noise distribution which could impede its performance on exceedingly narrow constraint geometries, even with correspondingly small step sizes. In this context, an important direction of future research would be to investigate whether we can instead sample from more suitable distributions, e.g. a Dikin ellipsoid, while maintaining the simplicity and efficiency of the Metropolis approach. More general topics of future work include the derivation of quantitative weak and mean square errors of our proposed discretisation scheme, as well as its application to more semantic constraints, for example with the objective of imposing sparsity in a basis or fixing the colour or style of an image.

6 | Score-based modelling on geometric structures on Riemannian manifolds

Work in this chapter is based on

E. Mathieu*, V. Dutordoir*, M. J. Hutchinson*, V. De Bortoli, Y. W. Teh, and R. Turner. Geometric neural diffusion processes. In *Advances in Neural Information Processing Systems*, 2024.

and has been rewritten for this thesis with additional material.

Personal contributions:

1. Project conception with Emile.
2. Development of geometric nerual diffusion process framework with Emile, Vincent.
3. Development of code with Emile, Vincent.
4. Running of experiments: Global tropical cyclone trajectory prediction experiment.
5. Developed the conditional sampling approaches.
6. Mathematical results: Those regarding the invariance/equivariance of the proposed models, and the results regarding the equivariance of the posterior maps of invariant measures.
7. Original manuscript writing with Emile and Vincent.

6.1. Introduction

Traditional denoising diffusion models are defined on finite-dimension Euclidean spaces (Song and Ermon, 2019; Song et al., 2020b; Ho et al., 2020; Dhariwal and Nichol, 2021). Extensions have recently been developed for more exotic distributions, such as those supported on Riemannian manifolds, developed in the previous chapters, and on function spaces of the form $f : \mathbb{R}^n \to \mathbb{R}^d$ (Dutordoir et al., 2023; Kerrigan et al., 2023; Lim et al., 2023a; Pidstrigach et al., 2023; Franzese et al., 2024; Bond-Taylor and Willcocks, 2023) i.e. modelling stochastic processes.

In this chapter, we extend diffusion models to further deal with distributions over functions that incorporate non-Euclidean geometry in two different ways. This investigation of geometry also naturally leads to the consideration of symmetries in these distributions, and as such we also present methods for incorporating these into diffusion models.

Tensor fields

Firstly, we look at *tensor fields* (see appendix A.6.2). Tensor fields are geometric objects that assign to all points on some manifold a value that lives in some vector space V, approximately functions $f : \mathcal{M} \to V$. These objects are central to the study of physics as they form a generic mathematical framework for modelling natural phenomena. Common examples include the pressure of a fluid in motion as $f : \mathbb{R}^3 \to \mathbb{R}$, representing wind over the Earth's surface as $f : \mathcal{S}^2 \to T\mathcal{S}^2$, $f \in \Gamma(T\mathcal{S}^2)$, sections of the tangent bundle, or modelling the stress in a deformed object as $f : \text{Object} \to T\mathbb{R}^3 \otimes T\mathbb{R}^3$, where $\otimes$ is the *tensor product* (see appendix A.4.3) of the tangent spaces. Given the inherent symmetry in the laws of nature, these tensor fields can transform in a way that preserves these symmetries. Leveraging this symmetry can drastically reduce the amount of data required to learn from and reduce training time (Elesedy and Zaidi, 2021; Elesedy, 2021; Lyle et al., 2020; Behboodi et al., 2022).

Tensor product

Paths on manifolds

Secondly, we look at functions with manifold codomain, and in particular, at functions of the form $f : \mathbb{R} \to \mathcal{M}$, representing *paths on manifolds*. The challenge in dealing with manifold-valued output, arises from the lack of vector-space structure. In applications, these functions typically appear when modelling processes indexed by time that take values on a manifold. Examples include tracking the eye of cyclones moving on the surface of the Earth (section 6.5.3), or modelling the joint angles of a robot as it performs tasks (Jaquier et al., 2021).

Gaussian processes

The lack of data or noisy measurements in the physical process of interest motivates a *probabilistic* treatment of such phenomena, in addition to its functional nature. Arguably the most important framework for modelling stochastic processes are *Gaussian processes* (Rasmussen, 2003), as they allow for exact or approximate posterior prediction (Titsias, 2009; Rahimi and Recht, 2007; Wilson et al., 2020). These can be viewed as generalisations of the typical multivariate Gaussian distribution, specified by a mean and covariance kernel. In particular, when choosing equivariant mean and kernel functions, Gaussian processes are invariant to the symmetries their mean and kernel function encode (this is commonly referred to as 'stationary') (Terenin, 2022; Holderrieth et al., 2021b; Azangulov et al., 2023a; Azangulov et al., 2023b).

Neural processes

The limited modelling capacity of Gaussian processes and the difficulty in designing complex, problem-specific kernels motivated the development of *neural processes* (Garnelo et al., 2018), which learn to approximately model a conditional stochastic process directly from data. Neural processes have been extended to model translation invariant (scalar) processes (Gordon et al., 2020) and more generic $\mathrm{E}(n)$-invariant processes (Holderrieth et al., 2021b).

In this chapter we develop two sets of contributions. Firstly, we produce a framework for placing score-based generative models on the spaces of tensor fields on manifolds, and on the spaces of paths on manifolds. We accomplish this by considering the finite marginals of a noising process defined on the infinite dimension

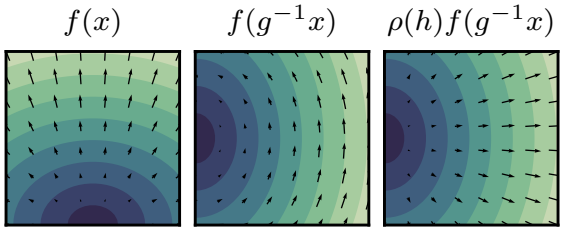


Figure 6.1. Illustration of a vector field $f : \mathbb{R}^2 \to \mathbb{R}^2$ being steered by a group element $g = uh \in \mathbb{E}(2) = \mathbb{R}^2 \rtimes \mathrm{O}(2)$.

space of functions, targeting a Gaussian process as the reference distribution. We investigate a series of methods of parametrising the score function for such a process. We also derive the constraints required on the score function to ensure that the modelled density is invariant, or if we are modelling a conditional distribution, equivariant.

Secondly we design a new series of conditioning methods to sample the posterior distribution of a prior diffusion model, conditioned on a partial observation. This method relies on exploiting a decomposition of the score function and an application of Langevin dynamics. We explore a series of hyperparameter settings for this algorithm, trading off sampling cost with sample quality.

At the end, we empirically evaluate these contributions on a series of experimental tasks.

6.2. Geometric neural diffusion processes

For ease of presentation we begin with our method specialised for tensor fields on Euclidean space. Later we will generalise this to tensor fields on manifolds.

6.2.1. Euclidean tensor fields. Our aim is to model tensor fields on Euclidean space. These are an example of a *signal on a space* (see section 2.1.1). Signal on a space

Recall that a signal on a space consists of an input space X, a set of symmetries of X, G, and an output space, Y. We call tuple (f, ρ) with f as above and ρ a *representation* (see section 2.1.1) of G, a group action of G on Y, a *feature field*. Feature field
Recall the action of G on a signal is given by

$$[g \triangleright f](x) = \rho(g) f(g^{-1}x), \quad x \in X,\ g \in G. \tag{6.1}$$

In our setting X is $\mathbb{R}^d$, Y a real vector space, isomorphic to $\mathbb{R}^{d'}$, and G the *Euclidean group*, the space of rigid symmetries of $\mathbb{R}^d$. ρ is given by the problem being modelled. Euclidean group

The elements $g \in \mathrm{E}(d)$ admit a unique decomposition $g = uh$ where $h \in \mathrm{O}(d)$ is a $d \times d$ orthogonal matrix and $u \in \mathrm{T}(d)$ is a translation which can be identified as an element of $\mathbb{R}^d$; for a vector $x \in \mathbb{R}^d$, and $g = (h, u) \in \mathrm{E}(d)$, $g \cdot x = hx + u$ denotes the action of g on x, with h acting from the left on x by matrix multiplication.

The action of $\mathrm{E}(d)$ on the feature field f given by eq. (6.1) specialises in this case to

$$g \cdot f(x) = (uh) \cdot f(x) \triangleq \rho(h) f\big(h^{-1}(x-u)\big). \tag{6.2}$$

Typical examples of feature fields include scalar fields, $Y = \mathbb{R}$ with $\rho_{\text{triv}}(g) \triangleq 1$, such as temperature fields, and vectors fields, $Y = \mathbb{R}^d$ with $\rho_{\text{Id}}(g) \triangleq h$, such as wind or force fields. For an example of a vector field, see figure 6.1.

6.2.2. Continuous diffusion on function spaces. We construct a diffusion model on functions $f : X \to Y$ by defining a diffusion model for every *finite set of marginals*. Most prior works on infinite-dimensional diffusions consider a noising process on the space of functions (Kerrigan et al., 2023; Pidstrigach et al., 2023; Lim et al., 2023b). In theory, this allows the model to define a consistent distribution over all the finite marginals of the process being modelled. In practice, however, only finite marginals can be modelled on a computer and the score function needs to be approximated, and at this step consistency over the marginals is lost. The only work to stay fully consistent in implementation is Phillips et al. (2022), at the cost of limiting functions that can be modelled to a finite-dimensional subspace.

Finite set of marginals

With this in mind, we eschew the technically laborious process of defining diffusions over the infinite-dimension space and work solely on the finite marginals following Dutordoir et al. (2023). We find that in practice consistency can be well learned from data see, section 6.5, and this allows for more flexible choices of score network architecture and easier training.

For the modelling problem, we assume that there exists an underlying distribution over tensor fields, $f \sim \mu$, for $\mu \in \mathcal{P}(C(X, Y))$, where $\mathcal{P}$ is the space of probability measure on the space of continuous functions $f : X \to Y$. From this we have a *finite set of observations*, observed on a random or static finite index set, that is a dataset

Finite set of observations

$$\Big(\big(x_{i,j}, y_{i,j} = f_i(x_{i,j})\big)_{j=1}^{N_j}\Big)_{i=1}^{N}, \quad x_{i,j} \in X,\ y_{i,j} \in Y,\ f_i \sim \mu \tag{6.3}$$

for a dataset size N with N_j observations per function sample in the dataset. Examples of this sort of data might include wind observations at sparse weather stations, or pixels making up an image.

It should be noted that a measures on general function spaces $f : X \to Y$, not just when $X = \mathbb{R}$, are termed stochastic processes also. We will refer therefore to the distribution over the data functions as a stochastic process, as well as the stochastic processes defining the forward and reverse stochastic differential equations.

Noising process

We denote a stochastic differential equation on the space of functions as $(\boldsymbol{Y}_t)_{t\geq 0}$ For a given data point $(x_i, y_i)_{i=1}^n = (\boldsymbol{x}, \boldsymbol{y})$ of n observations we denote the noising process on the marginals as $(\boldsymbol{Y}_t(\boldsymbol{x}))_{t\geq 0}$ and consider the following forward *noising* process defined by the following multivariate stochastic differential equation

$$\mathrm{d}\boldsymbol{Y}_t(\boldsymbol{x}) = \tfrac{1}{2}\beta_t[\boldsymbol{m}(\boldsymbol{x}) - \boldsymbol{Y}_t(\boldsymbol{x})]\,\mathrm{d}t + \sqrt{\beta_t}\mathbf{K}(\boldsymbol{x},\boldsymbol{x})^{\frac{1}{2}}\,\mathrm{d}\boldsymbol{B}_t, \tag{6.4}$$

where $\boldsymbol{m}(\boldsymbol{x})_i = m(x_i)$ with $m : X \to \mathbb{R}$ a *mean function* and $\mathbf{K}(\boldsymbol{x},\boldsymbol{x})_{i,j} = k(x_i, x_j)$ with $k : X \times X \to \mathbb{R}$ a *kernel*. This will converge with geometric rate to

Mean function
Kernel

$\mathcal{N}(\boldsymbol{m}(\boldsymbol{x}), \mathbf{K}(\boldsymbol{x}, \boldsymbol{x}))$. Using the results of Phillips et al. (2022), this convergence extends to the diffusion on function space $\boldsymbol{Y}_t$, which will converge to a Gaussian process with mean function m and covariance kernel k. We choose to noise towards a Gaussian process (Rasmussen, 2003) as they are the function-space generalisation of a multivariate Gaussian. We hypothesise that by introducing structure into the reference distribution of the diffusion model we can make the sampling process simpler.

In the specific instance where $k(x_i, x_j) = \delta_{x_i}(x_j)$, then the limiting process $\boldsymbol{Y}_\infty$ is simply *white noise*, whilst other choices such as the squared-exponential or Matérn kernel would lead to the associated Gaussian processes. Note that the *white noise* setting is not covered by the existing theory of functional diffusion models, as a Hilbert space and a square integral kernel are required, see Kerrigan et al. (2023) for instance.

White noise

To draw a sample from $p_{t|0}(\boldsymbol{Y}_t(\boldsymbol{x})|\boldsymbol{Y}_0(\boldsymbol{x}))$ we can sample from the multivariate Gaussian

$$\boldsymbol{Y}_t(\boldsymbol{x}) \sim \mathcal{N}\left(\left(1 - e^{-\frac{1}{2}B(t)}\right)\boldsymbol{m}(\boldsymbol{x}) + e^{-\frac{1}{2}B(t)}\boldsymbol{Y}_0(\boldsymbol{x}), \left(1 - e^{-B(t)}\right)\mathbf{K}(\boldsymbol{x}, \boldsymbol{x})\right) \tag{6.5}$$

$$= \mathcal{N}\left(\boldsymbol{m}_{t|0}(\boldsymbol{x}), \sigma^2_{t|0}\mathbf{K}(\boldsymbol{x}, \boldsymbol{x})\right). \tag{6.6}$$

where $B(t) = \int_0^t \beta(s)\,\mathrm{d}s$. This can be quickly sampled simply via

$$\boldsymbol{Y}_t(\boldsymbol{x}) \sim \boldsymbol{m}_{t|0}(\boldsymbol{x}) + \sigma_{t|0}\boldsymbol{K}_{t|0}(\boldsymbol{x}, \boldsymbol{x})^{\frac{1}{2}}\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I}). \tag{6.7}$$

The conditional score of the process is given by

$$\nabla_{\boldsymbol{Y}_t} \log p_t(\boldsymbol{Y}_t(\boldsymbol{x})|\boldsymbol{Y}_0(\boldsymbol{x})) = -\Sigma^{-1}_{t|0}(\boldsymbol{Y}_t(\boldsymbol{x}) - \boldsymbol{m}_{t|0}(\boldsymbol{x})) = -\sigma^{-1}_{t|0}\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x})^{-1/2}\varepsilon, \tag{6.8}$$

allowing for the use of denoising score matching. See appendix F.2.1 for these results.

Denoising process

Using the results presented in section 3.2.3, the time-reversal process $(\boldsymbol{Y}^{\leftarrow}_t)_{t\geq 0}$ also satisfies a stochastic differential equation, given by

$$\mathrm{d}\boldsymbol{Y}^{\leftarrow}_t(\boldsymbol{x}) = \beta_t\left[\tfrac{1}{2}(\boldsymbol{m}(\boldsymbol{x}) - \boldsymbol{Y}_t(\boldsymbol{x})) + \mathbf{K}(\boldsymbol{x}, \boldsymbol{x})\nabla \log p^{\boldsymbol{x}}_{T-t}(\boldsymbol{Y}^{\leftarrow}_t(\boldsymbol{x}))\right]\mathrm{d}t + \sqrt{\beta_t}\mathbf{K}(\boldsymbol{x}, \boldsymbol{x})^{\frac{1}{2}}\,\mathrm{d}\boldsymbol{B}_t,$$

with $p^{\boldsymbol{x}}_t$ the density of $\boldsymbol{Y}_t(\boldsymbol{x})$ with respect to. the Lebesgue measure. In practice, the Stein score $\nabla \log p_{T-t}$ is not tractable and is approximated by a neural network. We then consider the generative stochastic process model defined by first sampling with $\boldsymbol{Y}^{\leftarrow}_0 \sim \mathcal{N}(\boldsymbol{m}(\boldsymbol{x}), \mathbf{K}(\boldsymbol{x}, \boldsymbol{x}))$ and then simulating the reverse diffusion (6.9) (e.g. via Euler-Maruyama discretisation).

Score parametrisation and training loss

As the reverse stochastic differential equation (6.9) involves the preconditioned score $\mathbf{K}(\boldsymbol{x}, \boldsymbol{x})\nabla \log p_t$, we directly approximate it with a neural network $(\boldsymbol{Ks})_\theta : [0, T] \times X^n \times Y^n \to TY^n$.

We learn the preconditioned score $(\boldsymbol{K}\boldsymbol{s})_\theta$ by minimising the following denoising score matching (DSM) loss (Vincent et al., 2010) weighted by $\Lambda(t)$

$$\begin{aligned}\mathcal{L}(\theta;\Lambda(t)) &= \mathbb{E}[\|\boldsymbol{s}_\theta(t,\boldsymbol{Y}_t) - \nabla \log p_t(\boldsymbol{Y}_t|\boldsymbol{Y}_0)\|^2_{\Lambda(t)}] \\ &= \mathbb{E}[\|\sigma_{t|0}\cdot(\boldsymbol{K}\boldsymbol{s})_\theta(t,\boldsymbol{Y}_t) + \boldsymbol{K}(\boldsymbol{x},\boldsymbol{x})^{1/2}\varepsilon\|_2^2], \end{aligned} \tag{6.9}$$

with $\Lambda(t) = \sigma^2_{t|0}\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x})^\top \boldsymbol{K}(\boldsymbol{x},\boldsymbol{x})$ where $\|x\|^2_\Lambda = x^\top \Lambda x$. Note that when targeting a unit-variance white noise, $k(x_i,x_j) = \delta_{x_i}(x_j)$, the loss (6.9) reverts to the denoising score matching loss with weighting $\lambda(t) = 1/\sigma^2_{t|0}$ (Song et al., 2020b).

In appendix F.2.3 we explore several preconditioning terms and associated weighting $\Lambda(t)$. Overall, we found the preconditioned score $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{x})\nabla \log p_t$ parametrisation, in combination with the above loss, to perform best, as shown by the ablation study in appendix F.4.1.

A discussion on model consistency

So far we have defined a generative model over functions via distributions over sets of finite marginals $\boldsymbol{Y}_0(\boldsymbol{x})$. These finite marginals arise from a stochastic process if, as per the Kolmogorov extension theorem (Øksendal, 2003), they satisfy *exchangeability* and *consistency* conditions.

Exchangeability
Consistency

Exchangeability can be satisfied by parametrising the score network such that the score network is equivariant with respect to permutation, i.e. $\mathbf{s}_\theta(t,\sigma\circ x,\sigma\circ y) = \sigma\circ\mathbf{s}_\theta(t,x,y)$ for any $\sigma\in\Sigma_n$.

Additionally, we have that the forward noising processes on the marginals $(\boldsymbol{Y}^{\rightarrow}_t(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$, where $\mathcal{X}$ is the set of all finite index sets trivially consistent for any $t\in\mathbb{R}_+$ since these are the marginals of a stochastic process as per proposition F.1. Consequently, so is the (true) time-reversal marginals $(\boldsymbol{Y}^{\leftarrow}_t(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$.

When approximating the score $\mathbf{s}_\theta \approx \nabla \log p_t$ however, we lose the consistency over the generative process $\boldsymbol{Y}^{\leftarrow}_t(x)$ as the constraint on the score network is non-trivial to satisfy. The consistency constraint is very strong constraint on the model class, and as soon as one goes beyond linearity (of the posterior with respect to the context set), it is non-trivial to enforce without directly parametrising a stochastic process, e.g. as Phillips et al. (2022) do. We therefore avoid this and aim to learn consistency from data. Empirically there seems to be a strong trade-off between satisfying consistency, and the model's ability to fit complex process and scale to large datasets.

6.2.3. Invariant and equivariant geometric neural diffusion processes. So far the model defined does not build in any symmetries desired. In this section we show how we can incorporate such symmetries. In particular, we aim to incorporate the symmetries of tensor fields as described in section 6.2.1.

Invariant stochastic processes

Let (F,Σ_F) be a measurable space, and let G be a group with a group action on F. Let Λ_F be the space of probability measures on F. For $\mu\in\Lambda_F$ we define the action

of G on μ via the measurable sets of Σ_F as

$$[g \cdot \mu](\mathrm{F}) = \mu(\mathrm{g}^{-1} \cdot \mathrm{F}) = \mu\big(\{\mathrm{g}^{-1} \cdot \mathrm{f} \ : \ \mathrm{f} \in \mathrm{F}\}\big) \tag{6.10}$$

for $g \in G, \mathrm{F} \in \Sigma_{\mathrm{F}}$. This action implies

$$[g \cdot \mu](g \cdot \mathrm{F}) = \mu(\mathrm{F}) \tag{6.11}$$

for $g \in G, \mathrm{F} \in \Sigma_{\mathrm{F}}$.

A probability measure μ on F is said to be *G-invariant measure* if

$$g \cdot \mu = \mu \tag{6.12}$$

G-invariant measure

This definition naturally covers stochastic processes of the form we are studying with F as a space of function. In the Gaussian process literature an invariant Gaussian process is often termed as a 'stationary' stochastic process.

Let ν be a measure on sets of finite marginal, that is it a distribution on sets $(x_1, ..., x_n)$ for variable n and randomly generates sets of marginals to evaluate the stochastic process on. Assume this is invariant under the group action $g \cdot (x_1, ..., x_n) = (g \cdot x_1, ..., g \cdot x_n)$.

The invariances of μ and ν together mean that for a given set of observations of the stochastic process, $(\boldsymbol{x}, \boldsymbol{y}) = \big(\{x_i\}_{i=1}^n, \{y_i = f(x_i)\}_{i=1}^n\big)$ for $f \sim \mu$ and $\boldsymbol{x} \sim \nu$, we are as likely to see this observation as we are to see the transformed observations $g \cdot (\boldsymbol{x}, \boldsymbol{y})$. The transformation of observations is defined in the Euclidean tensor field case by

$$g \cdot (\boldsymbol{x}, \boldsymbol{y}) = \big(\{g \cdot x_i\}_{i=1}^n, \{\rho(h) y_i\}_{i=1^n}\big), \tag{6.13}$$

where ρ is a representation of $\mathrm{O}(d) \subset G$. We aim therefore to define a model that preserves this G-invariance.

Invariant diffusion processes

First, we recall such a necessary and sufficient condition for Euclidean Gaussian processes to be invariant.

PROPOSITION 6.1. *We have that a Gaussian process* $\mathrm{GP}(m, k)$ *is G-invariant if and only if its mean m and covariance k are suitably G-equivariant—that is, for all* $x, x' \in \mathcal{X}, g \in G$

$$m(g \cdot x) = \rho(g) m(x) \quad \textit{and} \quad k(g \cdot x, g \cdot x') = \rho(g) k(x, x') \rho(g)^\top. \tag{6.14}$$

Invariant (stationary) Gaussian process (Holderrieth et al., 2021b)

Trivial examples of $\mathrm{E}(n)$-equivariant kernels include diagonal kernels $k = k_0 \mathbf{I}$ with k_0 invariant (Holderrieth et al., 2021b), but see appendix F.4.2 for non-trivial instances introduced by Macêdo and Castro (2010).

Building on proposition 6.1, we then state that our introduced neural diffusion process.

PROPOSITION 6.2. *We denote by* $(\boldsymbol{Y}_t^{\epsilon}(\boldsymbol{x}))_{t \in [0,T]}$ *the process induced by the time-reversal stochastic differential equation* (6.9) *on the marginals* $\boldsymbol{x}$*. The score is approximated by a score network*

$$\mathbf{s}_\theta : [0, T] \times X^n \times Y^n \to TY^n \tag{6.15}$$

Invariant neural diffusion process (Yim et al., 2023)

, and the law of the limiting process is given by

$$\mathcal{L}(Y_0^{\leftarrow}) \sim \mathrm{GP}(m, k)(\boldsymbol{x}) \tag{6.16}$$

, a G-invariant Gaussian process evaluated on the marginals $\boldsymbol{x}$ (proposition 6.1). If we additionally assume that the score network is G-equivariant vector field, i.e. $\mathbf{s}_\theta(t, g \cdot x, \rho(g)y) = \rho(g)\mathbf{s}_\theta(t, x, y)$ for all $x \in X, g \in G$, then for any $t \in [0, T]$ the law of $\left(Y_t^{\leftarrow}(\boldsymbol{x})\right)_{\boldsymbol{x} \in X^n}$ is G-invariant in the sense that

$$g \cdot \mathcal{L}(Y_t^{\leftarrow}(\boldsymbol{x})) = \mathcal{L}(\rho(h)Y_t^{\leftarrow}(g \cdot \boldsymbol{x})) = \mathcal{L}(Y_t^{\leftarrow}(\boldsymbol{x})). \tag{6.17}$$

Proof. This result can be proved in two ways, from the probability flow ordinary differential equation perspective or directly in terms of stochastic differential equation via the Fokker-Planck, see appendix F.3.2. ■

Suitable architectures to produce such equivariant score functions on points include the E(n)-Equivariant Graph Neural Networks of Satorras et al. (2021).

Equivariant posterior maps

Conditional distribution

Amortised conditioning

The invariant model in the previous section defines a model over a stochastic process. Often in generative modelling we wish to model the *conditional distribution* of a stochastic process. That is, given a partial observation of a sample from μ, $C = (\boldsymbol{x}_c, \boldsymbol{y}_c) = \left(\left\{x_{c,i}, y_{c,i} = f(x_{c,i})\right\}\right)_{i=1}^n$, $f \sim \mu$, $\boldsymbol{x}_c \sim \nu$, we want to model the conditional measure of μ given C, $\mu(\cdot|C)$. Explicitly learning the conditional distribution is known as *amortised conditioning*.

G-equivariant

When the prior process μ is G-invariant, the map from the conditioning set to the posterior, $P : C \mapsto \mu(\cdot|C)$, inherits a symmetry from μ. It is *G-equivariant* in the sense that

$$P(g \cdot C) = g \cdot P(C) = g \cdot \mu(\cdot|C) \quad \forall\, C,\ g \in G. \tag{6.18}$$

This is equivalent to saying it satisfies

$$\mu(g \cdot A|g \cdot C) = \mu(A|C) \quad \forall C\ g \in G\ A \in \Sigma_F, \tag{6.19}$$

and so we could also as that the posterior is *invariant* in this sense.

A version of this statement specialised for Euclidean Gaussian processes was first proved in Holderrieth et al. (2021b). We prove a generalisation of that result to measures on tensor fields on general manifolds, theorem F.2.

This implies for a context set $C = (\boldsymbol{x}_c, \boldsymbol{y}_c)$ and a set of observations from the posterior $O = (\boldsymbol{x}_o, \boldsymbol{y}_o)$, $\boldsymbol{x}_o \sim \nu$, we are as likely to see the observations O given we already observed C as we are to see the observation $g \cdot O$ given we already observed $g \cdot C$. Consequently we want to define a conditional generative model that respects this invariance.

Equivariant conditional diffusion processes

In order to construct a diffusion model respecting the conditioning equivariance of the posterior we need to construct a score function of the form

$$s_\theta : [0, T] \times X^{n_o} \times Y^{n_o} \times X^{n_c} \times Y^{n_c} \to TY^{n_o}, \tag{6.20}$$

where n_o is the number of new observations we are modelling the distribution on and n_c is the size of the conditioning set. This score function takes the conditioning observations as an auxiliary input.

In order to encode the observed invariance in likelihood, it is sufficient to apply a minor extension of proposition 6.2. The conditional model will be conditionally equivariant if the reference density is invariant and the conditional score function satisfies

$$s_\theta(t, g \cdot O_t | g \cdot C) = \rho(h) s_\theta(t, O|C) \tag{6.21}$$

for all context sets $C \in X^{n_c} \times Y^{n_c}$, (partially denoised) observation sets $O \in X^{n_o} \times Y^{n_o}$, times $t \in [0, T]$ and $g \in G$.

6.2.4. Extension to tensor fields on manifolds. In the previous section we described the construction of symmetry preserving diffusion models for Euclidean tensor fields. Here we preset how this can be generalised to tensor fields on manifolds.

Significant work has been done on performing convolutions on feature fields on manifolds (a superset of tensor fields on manifolds), core references being Cohen (2021) for the case of homogeneous spaces and Weiler et al. (2023) for more general Riemannian manifolds. We recommend these as excellent mathematical introductions to the topic and build on them to describe how to formulate diffusion models over these spaces.

Tensor fields as sections of bundles

Formally the fields we are interested in modelling are sections σ of *associated tensor bundles of the principle G-bundle* on a Riemannian manifold (M, g) for a given G structure on the manifold. We shall denote such a bundle BM and the space of sections $\Gamma(BM)$. The goal, therefore, is to model *random elements* $f \in \Gamma(BM)$ from this space of sections. For a clear understanding of this definition, please see Weiler et al. (2023, pages 73-95).

Associated tensor bundles of the principle G-bundle

The symmetries we shall want to preserve are the *Riemannian isometries* (see appendix A.8.3) of the manifold, $\mathsf{Isom}_{(M,g)}$. Unlike Euclidean symmetries these may not be global symmetries of the manifold.

Riemannian isometries

Prior work looking at this setting is Hutchinson et al. (2021c) where they construct kernels for tensor fields on manifolds, allowing for the use of Gaussian processes in this setting.

Stochastic processes on spaces of sections

Given we can see sections as maps $\sigma : M \to BM$, where an element in BM is a tuple (m, b), m in the base manifold and b in the typical fibre, alongside the condition

that the composition of the projection $\text{proj}_i : (m, b) \mapsto m$ with the section is the identity, $\text{proj}_i \circ \sigma = \text{Id}$ it is clear we can see distribution over sections as stochastic processes with index set the manifold M, and output space a point in the bundle BM, with the projection condition satisfied. The projection onto finite marginals, i.e. a finite set of points in the manifold, is defined as $\pi_{m_1,...,m_n}(\sigma) = (\sigma(m_1), ..., \sigma(m_n))$.

Noising process

To define a noising process over these marginals, we can use Gaussian processes defined in Hutchinson et al. (2021c) over the tensor fields. The convergence results of Phillips et al. (2022) hold still, and so using these Gaussian processes as noising processes on the marginals also defines a noising process on the whole section.

Reverse process

The results of Cattiaux et al. (2023) are extremely general and continue to hold in this case of stochastic differential equations on the space of sections. Note we don't actually need this to be the case as we work with the reverse process on the marginals only, which are much simpler objects. It is good to know that it is a valid process on full sections though should one want to parametrise a score function on the whole section akin to some other infinite-dimension diffusion models.

Score function

The last thing to do therefore is parametrise the score function on the marginals. If we were trying to parametrise the score function over the *whole* section at once (akin to a number of other works on infinite dimension diffusions), this could present some problems in enforcing the smoothness of the score function. As we only deal with the score function on a finite set of marginals, however, we need not deal with this issue and this presents a distinct advantage in simplicity for our approach. All we need to do is pick a way of numerically representing points on the manifold and pick a basis for the tangent space of each point on the manifold. This lets us represent elements from the tangent space as element of Euclidean space, and therefore also elements from tensor space at each point as elements of Euclidean space as well. This done, we can use the standard Euclidean score-based modelling framework to model these quantities.

Encoding symmetries

A general recipe for encoding symmetries into this generalisation is unlikely to be available. The symmetry acting on sets of observations of sections of tensor bundles on manifolds are more complex than in the Euclidean case. The generalisation of eq. (6.13) is given by

$$g \cdot (\boldsymbol{x}, \boldsymbol{y}) = \left(\{g \cdot x_i\}_{i=1}^n, \{\rho(h(g, x_i)) \cdot y_i\}_{i=1^n}\right), \tag{6.22}$$

where $i \in \text{Isom}_{(M,g)}$ and $h : \text{Isom}(M, g) \times M \to G$ is a function of the global isometry and the specific point x_i.

One case where this simplifies is when there exists an isometric embedding of the manifold into Euclidean space that also preserve the symmetry of the manifold

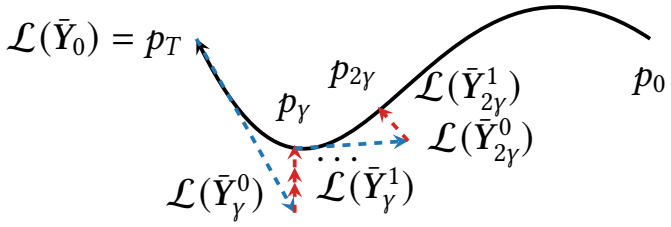


Figure 6.2. Illustration of Langevin corrected conditional sampling. The black line represents the noising process dynamics $(p_t)_{t\in\{0,T\}}$. The time reversal (i.e. predictor) step, is combined with a Langevin corrector step projecting back onto the dynamics.

such that the symmetry in the embedding is a subgroup of the Euclidean group. Examples of this include n-spheres $\mathcal{S}^n$, such as the circle and the sphere. In this case we can use an architecture that is equivariant to Euclidean transformations and project the outputs onto the tangent space of the embedded manifold.

6.2.5. Extension to manifold-valued functions. So far we have defined our generative model for tensor fields $f : \mathcal{M} \to Y$ with $Y \cong \mathbb{R}^d$. We now extend our methodology to model functions of the form $f : X \to \mathcal{M}$ where $\mathcal{M}$ is a Riemannian manifold.

In this case $\mathcal{M}$ does not have the vector space structure necessary to define Gaussian processes. Fortunately we can still target a know distribution independently on each marginal, since this is well-defined, and as such revert to the Riemannian diffusion models introduced in chapter 3 with n independent Langevin noising processes

$$\mathrm{d}Y_t(x_i) = -\tfrac{1}{2}\beta_t \nabla U(Y_t(x_i))\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}B_t^{\mathcal{M}} \tag{6.23}$$

applied to each marginal. Hence, in the limit $t \to \infty$, $Y_t(x)$ has density (assuming it exists) which factors as $\frac{\mathrm{d}p}{\mathrm{dVol}^n_{\mathcal{M}}} \propto \prod_{i=1}^n e^{-U(y_i)}$.

Since the space of this diffusion is the product of finitely many Riemannian manifolds, the time reversal and parametrisation of the score function approaches of chapter 3 also apply, giving us a ready method to define a diffusion model on paths.

6.3. Langevin conditional sampling of diffusion model

Conditional sampling, or posterior sampling, is the problem of sampling from a prior distribution, conditioned on some partial observations, as discussed in

Conditional sampling

section 6.2.3 already. This is a long-standing problem in Bayesian machine learning and Bayesian statistics in general.

Several previous schemes exist for conditional sampling from diffusion models, each with some limitations. These include:

- Replacement sampling (Song et al., 2020b), where the reverse likelihood ordinary differential equation or the reverse stochastic differential equation is evolved but by fixing the conditioning data during the rollout. This method does produce visually coherent sampling in some cases, but is not an exact conditional sampling method.
- SMC-based methods (Trippe et al., 2022), which are an exact method up to the particle filter assumption. These can produce good results but can suffer from the usual SMC methods downsides on highly multimodel data such as particle diversity collapse.
- The RePaint scheme of (Lugmayr et al., 2022). While not originally proposed as an exact sampling scheme, in **??** we show that this method is a specific instantiation of our newly proposed method, and is therefore exact.
- Amortisation methods, e.g. Phillips et al. (2022) and section 6.2.3. While they can be effective, these methods can never perform exact conditional sampling, by definition, and can take significant model capacity to learn approximate conditioning maps.

Our goal in this section is to produce an exact, performant, sampling scheme that does not rely on SMC-based methods. Instead, we base our method on Langevin dynamics. For the initial presentation in the section we consider a distribution over a fixed state space, $\boldsymbol{x} \in X^n$. We will adapt this to the setting of modelling the marginals of a stochastic process later. Consider the setting where we have partially observed $\boldsymbol{x}$, $\boldsymbol{x}^c \in X^{n_c}$, and we wish to sample $\boldsymbol{x}^o \in X^{n_o}$, such that $n_c + n_o = n$. If we have a score function trained over the state space $\boldsymbol{x} = [\boldsymbol{x}^c, \boldsymbol{x}^o]$ we exploit the following score breakdown:

$$\nabla_{\boldsymbol{x}^o} \log p(\boldsymbol{x}^o|\boldsymbol{x}^c) = \nabla_{\boldsymbol{x}^o} \log p([\boldsymbol{x}^o, \boldsymbol{x}^c]) - \nabla_{\boldsymbol{x}^o} \log p(\boldsymbol{x}^c) = \nabla_{\boldsymbol{x}^o} \log p(\boldsymbol{x}). \quad (6.24)$$

This means that if we have access to the score on the joint variables, we have access to the conditional score by simply taking the joint score and ignoring the gradient for the variables we are conditioning on.

In a score-based model we learn the time-dependent score function $s_\theta(t, \boldsymbol{x}) \approx \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$. We could use this to perform Langevin dynamics at $t = \epsilon$, some time very close to 0. Similar to Song and Ermon (2019) however, this produces the twin issues of how to initialise the dynamics, given a random initialisation will start the sampler in a place where the score has been badly learnt, and very slow mixing times for complex distributions, producing inaccurate sampling.

Instead, we follow a scheme of tempered Langevin sampling detailed in algorithm 6.3. Starting at $t = T$ we sample an initialisation of $\boldsymbol{x}^o$ based on the reference distribution. Progressing from $t = T$ towards $t = \epsilon$ we alternate between running a series of inner Langevin corrector steps to sample from the distribution $p_t(\boldsymbol{x}_t^o|\boldsymbol{x}_t^c)$, and a single outer backwards stochastic differential equation step to sample from

Algorithm 6.3 Conditional sampling with Langevin dynamics.

Require: Score network $\mathbf{s}_\theta(t, \boldsymbol{x}, \boldsymbol{y})$, conditioning points $(\boldsymbol{x}^c, \boldsymbol{y}^c)$, query locations $\boldsymbol{x}^o$
$\boldsymbol{x} = [\boldsymbol{x}^c, \boldsymbol{x}^o]$ ▷ Augmented inputs set
$\boldsymbol{y}^o_T \sim \mathcal{N}(\boldsymbol{m}(\boldsymbol{x}^o), \boldsymbol{K}(\boldsymbol{x}^o, \boldsymbol{x}^o))$ ▷ Sample initial noise
for $t \in \{T, T-\gamma, ..., \epsilon\}$ **do**
  $\boldsymbol{y}^c_t \sim p^{\boldsymbol{x}^c}_{t|0}(\cdot|\boldsymbol{y}^c)$ ▷ Sample context noise from forward SDE
  $Z \sim \mathcal{N}(0, \mathbf{I})$ ▷ Sample tangent noise
  $\left[\_, \boldsymbol{y}^o_{t-\gamma}\right] = \left[\boldsymbol{y}^c_t, \boldsymbol{y}^o_t\right] + \gamma\left\{-\frac{1}{2}\left(m(\bar{\boldsymbol{x}}) - \left[\boldsymbol{y}^c_t, \boldsymbol{y}^o_t\right]\right) + \mathrm{K}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})\mathbf{s}_\theta(t, \bar{\boldsymbol{x}}, \left[\boldsymbol{y}^c_t, \boldsymbol{y}^o_t\right])\right\} +$
$\sqrt{\gamma}\mathrm{K}(\boldsymbol{x}, \boldsymbol{x})^{1/2}Z$ ▷ Euler-Maruyama step
  for $l \in \{1, \ldots, L\}$ **do**
    $\tilde{\boldsymbol{y}}^c_{t-\gamma} \sim p^{\boldsymbol{x}^c}_{t-\gamma}(\cdot|\boldsymbol{y}^c_0)$ ▷ Sample context noise from forward SDE
    $Z \sim \mathcal{N}(0, \mathbf{I})$ ▷ Sample tangent noise
    $\left[\_, \tilde{\boldsymbol{y}}^o_{t-\gamma}\right] = \left[\_, \tilde{\boldsymbol{y}}^o_{t-\gamma}\right] + \frac{\gamma}{2}\mathrm{K}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})\mathbf{s}_\theta(t-\gamma, \bar{\boldsymbol{x}}, \left[\boldsymbol{y}^c_{t-\gamma}, \tilde{\boldsymbol{y}}^*_{t-\gamma}\right]) + \sqrt{\gamma}\mathrm{K}(\boldsymbol{x}, \boldsymbol{x})^{1/2}Z$ ▷ Langevin step
  $\boldsymbol{y}^o_{t-\gamma} = \tilde{\boldsymbol{y}}^*_{t-\gamma}$
return $\boldsymbol{y}^o_\epsilon$

$p_t(\boldsymbol{x}^o_{t-\gamma}|\boldsymbol{x}^c_t)$ with a step size γ. At each inner and outer step, we sample a noised version of the conditioning variables $\boldsymbol{x}^c_t$ based on the forward stochastic differential equation applying noise to these context points, $p_{t|0}(\boldsymbol{x}^c_t|\boldsymbol{x}^c)$. For the exactness of this scheme, all that matters is that at the end of the sampling scheme, we are sampling from $p_0(\boldsymbol{x}^o|\boldsymbol{x}^c)$. The rest of the scheme is designed to map from the initial sample at $t = T$ of $\boldsymbol{x}^o$ to a viable sample through *regions where the score has been learnt well.*

We apply this scheme to the problem of sampling from conditional distributions on the marginals of a diffusion model modelling a stochastic process by applying it to a fixed set of marginals at a time. Consider that we have a set of conditioning observations, $(\boldsymbol{x}^c, \boldsymbol{y}^c)$, and a set of index locations we wish to sample conditional observations at, $\boldsymbol{x}^o$. A score-based model on the marginals has learnt the distribution $\nabla_{\boldsymbol{y}^c, \boldsymbol{y}^o} \log p_t^{\boldsymbol{x}^c, \boldsymbol{x}^o}(\boldsymbol{y}^c_t, \boldsymbol{y}^o_t)$. We can therefore compute the conditional score

$$\nabla_{\boldsymbol{y}^o_t} \log p_t^{\boldsymbol{x}^c|\boldsymbol{x}^o}(\boldsymbol{y}^c_t|\boldsymbol{y}^o_t) = \nabla_{\boldsymbol{y}^o_t} \log p_t^{[\boldsymbol{x}^c, \boldsymbol{x}^o]}(\left[\boldsymbol{y}^c_t, \boldsymbol{y}^o_t\right]), \tag{6.25}$$

again by ignoring the gradients of the conditioning variables from the learnt score function. The algorithm for doing this conditioning is laid out in algorithm 6.3.

6.3.1. Noising schemes. Given the noising scheme applied to the context points does not actually play into the theoretical exactness of the scheme, only the practical difficulty of staying near regions of well-learnt score, we could make a series of different choices for how to noise the context set at each step.

The choices that present themselves are

1. The initial scheme of sampling context noise from the forward stochastic differential equation every inner and outer step.

2. Only re-sampling the context noise every outer step, and keeping it fixed to this for each inner step associated with the outer step.

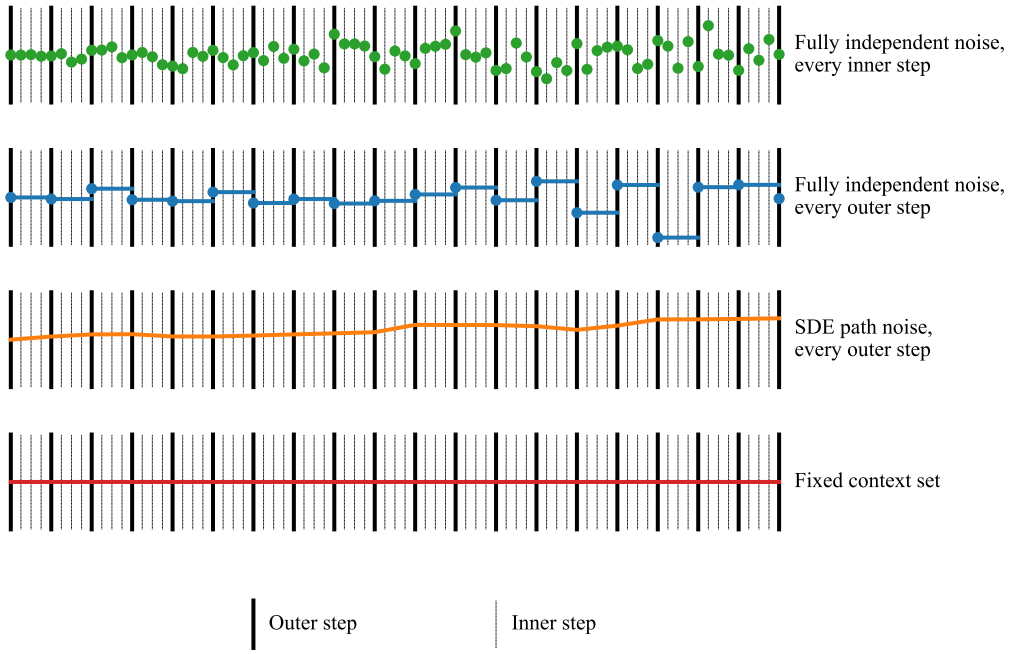


Figure 6.3. Comparison of different context noising schemes for the conditional sampling.

3. Instead of sampling independent marginal noise at each outer step, sampling a single noising trajectory of the context set from the forward SDE and use this as the noise at each time.

4. Perform no noising at all. Effectively the replacement method with added Langevin sampling.

These are illustrated in figure 6.3. The main trade-off of different schemes is the speed at which the noise can be sampled vs sample diversity. In the Euclidean case, we have a closed form for the evolution of the marginal density of the context point under the forward SDE. In this case sampling the noise at a given time is $O(1)$ cost. On the other hand, in some instances such as nosing stochastic differential equations on general manifolds, we have to simulate this noise by discretising the equation. In this case, it is $O(n)$ cost, where n is the number of discretisation steps in the stochastic differential equation. For N outer steps and I inner steps, the complexity of the different noising schemes is compared in table 6.1. Note the conditional sampling scheme other than the noise sampling is $O(NI)$ complexity.

6.3.2. Likelihood evaluation. The likelihood ordinary differential equation formulation for computing likelihoods of the marginals in section 6.2 applies immediately.

For conditional samples using the Langevin sampling method however a similar formulation is not possible. Instead, we rely on the conditional probability rule to evaluate conditional likelihoods, $p_{\boldsymbol{x}^o|\boldsymbol{x}^c}(\boldsymbol{y}^o|\boldsymbol{y}^c) = p_{\boldsymbol{x}^o,\boldsymbol{x}^c}(\boldsymbol{y}^o,\boldsymbol{y}^c)/p_{\boldsymbol{x}^c}(\boldsymbol{y}^c)$. This can be done by solving two probability flow ordinary differential equations on the joint evaluation and context set, and only on the context set.

SCHEME	CLOSED-FORM NOISE	SIMULATED NOISE
Re-sample noise at every inner step	$\mathcal{O}(NI)$	$\mathcal{O}(N^2I^2)$
Re-sample noise at every outer step	$\mathcal{O}(N)$	$\mathcal{O}(N^2)$
Sampling an SDE path on the context	$\mathcal{O}(N)$	$\mathcal{O}(N)$
No noise applied	-	-

Table 6.1. Comparison of complexity of different noise sampling schemes for the context set.

Given that our model does not enforce consistency on the marginals this computation will not be exact, only up to the degree that the model has learnt consistency from the data.

6.4. RELATED WORK

Gaussian processes and the neural process family. One important and powerful framework to construct distributions over functional spaces are Gaussian processes (Rasmussen, 2003). Yet, they are restricted in their modelling capacity and when using exact inference they scale poorly with the number of datapoints. Recently introduced autoregressive Neural Processes (Bruinsma et al., 2023) alleviate this limitation, but they are disadvantaged by the fact that variables early in the autoregressive generation only have simple distributions (typically Gaussian). Finally, (Dupont et al., 2022) model weights of implicit neural representation using diffusion models.

Stationary stochastic processes. The most popular Gaussian process kernels (e.g. squared exponential, Matérn) are stationary, that is, they are translation invariant. This idea can be extended to the entire isometry group of Euclidean spaces (Holderrieth et al., 2021b), allowing for modelling higher order tensor fields, such as wind fields or incompressible fluid velocity (Macêdo and Castro, 2010). Later, Azangulov et al. (2023a) and Azangulov et al. (2023b) extended stationary kernels and Gaussian processes to a large class of non-Euclidean spaces, in particular all compact spaces, and symmetric non-compact spaces. In the context of neural processes, (Gordon et al., 2020) introduced CONVCNP to encode translation equivariance into the predictive process. They do so by embedding the context into a translation equivariant functional representation which is then decoded with a convolutional neural network. Holderrieth et al. (2021b) later extended this idea to construct neural processes that are additionally equivariant with respect to rotations or subgroup thereof.

Spatial structure in diffusion models. A variety of approaches have also been proposed to incorporate spatial correlation in the noising process of finite-dimensional diffusion models leveraging the multiscale structure of data (Jing et al., 2022a; Guth et al., 2022; Ho et al., 2022; Saharia et al., 2022b; Hoogeboom and Salimans, 2023; Rissanen et al., 2023). Our methodology can also be seen as a principled way to modify the forward dynamics in classical denoising diffusion models. Hence, our contribution can be understood in the light of recent advances in generative

modelling on soft and cold denoising diffusion models (Daras et al., 2023; Bansal et al., 2022; Hoogeboom and Salimans, 2023). Several recent works explicitly introduced a covariance matrix in the Gaussian noise, either on a choice of kernel (Biloš et al., 2022), based on Discrete Fourier Transform of images (Voleti et al., 2022), or via empirical second order statistics (squared pairwise distances and the squared radius of gyration) for protein modelling (Ingraham et al., 2022). Alternatively, (Guth et al., 2022) introduced correlation on images leveraging a wavelet basis.

Functional diffusion models. Infinite dimensional diffusion models have been investigated in the Euclidean setting in Kerrigan et al. (2023), Pidstrigach et al. (2023), Lim et al. (2023b), Bond-Taylor and Willcocks (2023), Hagemann et al. (2023), Franzese et al. (2024), Dutordoir et al. (2023), and Phillips et al. (2022). Most of these works are based on an extension of the diffusion models techniques Song et al. (2020b) and Ho et al. (2020) to the infinite-dimensional space, leveraging tools from the Cameron-Martin theory such as the Feldman-Hájek theorem (Kerrigan et al., 2023; Pidstrigach et al., 2023) to define infinite-dimensional Gaussian measures and how they interact. We refer to Da Prato and Zabczyk (2014) for a thorough introduction to Stochastic Differential Equations in infinite dimension. Phillips et al. (2022) consider another approach by defining countable diffusion processes in a basis. All these approaches amount to learn a diffusion model with spatial structure. Note that this induced correlation is necessary for the theory of infinite dimensional stochastic differential equation (Da Prato and Zabczyk, 2014) to be applied but is not necessary to implement diffusion models (Dutordoir et al., 2023). Several approaches have been considered for conditional sampling. Pidstrigach et al. (2023) and Bond-Taylor and Willcocks (2023) modify the reverse diffusion to introduce a guidance term, while Dutordoir et al. (2023) and Kerrigan et al. (2023) use the replacement method. Finally, Phillips et al. (2022) amortise the score function with respect to. the conditioning context.

6.5. Experimental results

Our implementation is built on `jax` (Bradbury et al., 2018), and is publicly available at https://github.com/cambridge-mlg/neural_diffusion_processes. Additional details for each experiment can be found in appendix F.4.

6.5.1. 1D regression over stationary scalar fields. We evaluate GeomNDPs on several synthetic 1D regression datasets. We follow the same experimental setup as Bruinsma et al. (2020) which we detail in appendix F.4.1. In short, it contains Gaussian (Squared Exponential (SE), Matérn($\frac{5}{2}$), Weakly Periodic) and non-Gaussian (Sawtooth and Mixture) sample paths, where Mixture is a combination of the other four datasets with equal weight. Figure 6.4 shows samples for each of these datasets. The Gaussian datasets are corrupted with observation noise with variance $\sigma^2 = 0.05^2$. Table 6.2 reports the average log-likelihood $p(y^*|x^*, C)$ across 4096 test samples, where the context set size is uniformly sampled between 1 and 10 and the target has fixed size of 50. All inputs x^c, x^* are chosen uniformly within their input domain which is $[-2, 2]$ for the training data and 'interpolation' evaluation and $[2, 6]$ for the 'generalisation' evaluation.

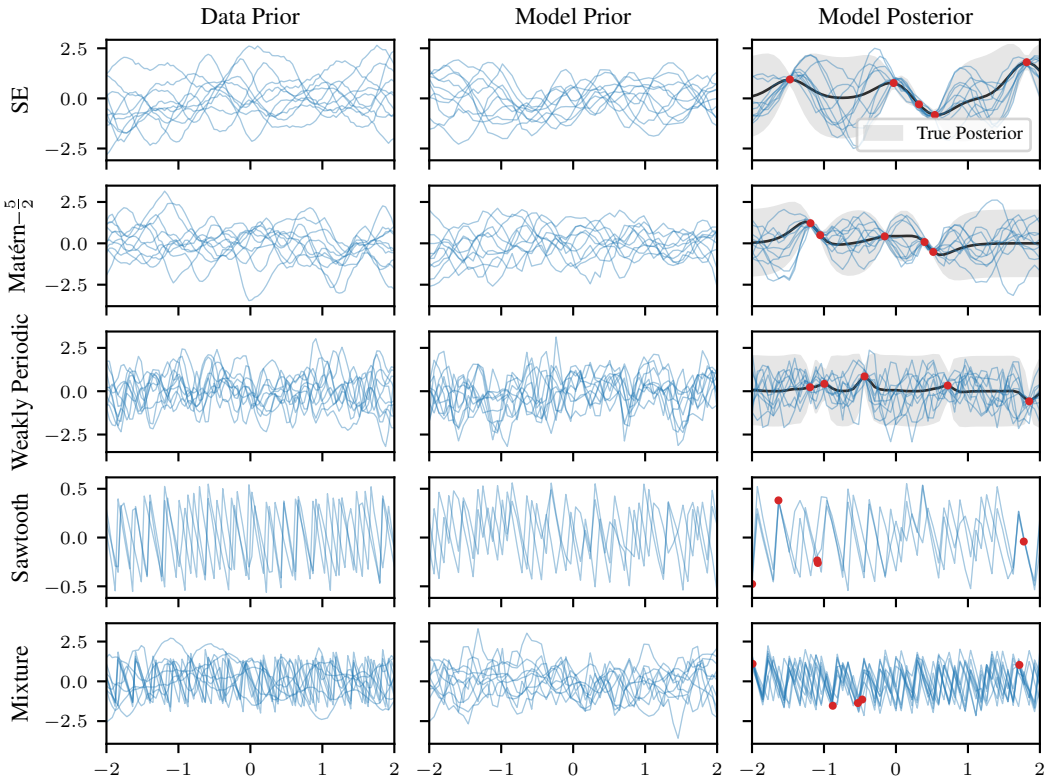


Figure 6.4. Prior and posterior samples (in blue) from the data process and the GeomNDP model, with context points in red and posterior mean in black.

We compare the performance of GeomNDP to a GP with the true hyperparameters (when available), a (convolutional) Gaussian NP (Bruinsma et al., 2020), a convolutional NP (Gordon et al., 2020) and a vanilla attention-based NDP (Dutordoir et al., 2023) which we reformulated in the continuous diffusion process framework to allow for log-likelihood evaluations and thus a fair comparison—denoted NDP*. We enforce translation invariance in the score network for GeomNDP by subtracting the centre of mass from the input x, inducing stationary scalar fields.

On the GP datasets, GNP, ConvNPs and GeomNDP methods are able to fit the conditionals perfectly—matching the log-likelihood of the GP model. GNP's performance degrades on the non-Gaussian datasets as it is restricted by its conditional Gaussian assumption, whilst NDPs methods still performs well as illustrated on figure 6.4. In the bottom rows of table 6.2, we assess the models' ability to generalise outside of the training input range $x \in [-2, 2]$, and evaluate them on a translated grid where context and target points are sampled from $[2, 6]$. Only convolutional Neural Processes (GNP and ConvNP) and T(n)1−GeomNDP are able to model stationary processes and therefore to perform as well as in the interpolation task. The NDP*, on the contrary, drastically fails at this task.

Non white kernels for limiting process. The NDP methods in the above experiment target the white kernel $\mathbb{1}(x = x')$ in the limiting process. In appendix F.4.1, we explore different choices for the limiting kernel, such as SE and periodic kernels with short and long lengthscales, along with several score parametrisations, see appendix F.2.3 for a description of these. We observe that although choosing

Table 6.2. Mean test log-likelihood (TLL) (↑) ± 1 standard error estimated over 4096 test samples are reported. Statistically significant best non-GP model is in **bold**. '*' stands for a TLL below −10. NP baselines from Bruinsma et al. (2020).

		SE	Matérn-$\frac{5}{2}$	Weakly Per.	Sawtooth	Mixture
Interpolate.	GP (optimum)	0.70±0.00	0.31±0.00	−0.32±0.00	-	-
	T(1)−GeomNDP	**0.72**±0.03	**0.32**±0.03	**−0.38**±0.03	**3.39**±0.04	**0.64**±0.08
	NDP*	**0.71**±0.03	**0.30**±0.03	**−0.37**±0.03	**3.39**±0.04	**0.64**±0.08
	GNP	**0.70**±0.01	**0.30**±0.01	−0.47±0.01	0.42±0.01	0.10±0.02
	ConvNP	−0.46±0.01	−0.67±0.01	−1.02±0.01	1.20±0.01	−0.50±0.02
Generalisat.	GP (optimum)	0.70±0.00	0.31±0.00	−0.32±0.00	-	-
	T(1)−GeomNDP	**0.70**±0.02	**0.31**±0.02	**−0.38**±0.03	**3.39**±0.03	**0.62**±0.02
	NDP*	*	*	*	*	*
	GNP	**0.69**±0.01	**0.30**±0.01	−0.47±0.01	0.42±0.01	0.10±0.02
	ConvNP	−0.46±0.01	−0.67±0.01	−1.02±0.01	1.19±0.01	−0.53±0.02

Model	SE	Curl-free	Div-free
GP	$0.56_{\pm 0.00}$	$0.66_{\pm 0.00}$	$0.66_{\pm 0.00}$
NDP*	$0.55_{\pm 0.00}$	$0.62_{\pm 0.01}$	$0.62_{\pm 0.01}$
E(2)-GeomNDP	$\mathbf{0.56_{\pm 0.01}}$	$\mathbf{0.65_{\pm 0.01}}$	$\mathbf{0.66_{\pm 0.01}}$
GP (diag.)	$-1.56_{\pm 0.00}$	$-1.47_{\pm 0.00}$	$-1.47_{\pm 0.00}$
T(2)-ConvCNP	$-1.71_{\pm 0.01}$	$-1.77_{\pm 0.01}$	$-1.76_{\pm 0.00}$
E(2)-SteerCNP	$-1.61_{\pm 0.00}$	$-1.57_{\pm 0.00}$	$-1.57_{\pm 0.01}$

Table 6.3. Quantitative results comparing fitting neural diffusion process models and regular neural process models to samples from GP vector fields. Mean predictive log-likelihood (↑) and confidence interval estimated over 5 random seeds.

such kernels gives a head start to the training, it eventually yields slightly worse performance. We attribute this to the additional complexity of learning a non-diagonal covariance. Finally, across all datasets and limiting kernels, we found the preconditioned score $K\nabla \log p_t$ to result in the best performance.

Conditional sampling ablation. We employ the SE dataset to investigate various configurations of the conditional sampler as we have access to the ground truth conditional distribution through the GP posterior. In figure F.2 we compute the Kullback-Leibler divergence between the samples generated by GeomNDP and the actual conditional distribution across different conditional sampling settings. Our results demonstrate the importance of performing multiple Langevin dynamics steps during the conditional sampling process. Additionally, we observe that the choice of noising scheme for the context values y_c has relatively less impact on the overall outcome.

6.5.2. Regression over Gaussian process vector fields. We now focus our attention to modelling equivariant vector fields. For this, we create datasets using samples from a two-dimensional zero-mean GP with one of the following SE(n)2-equivariant kernels: a diagonal Squared-Exponential (SE) kernel, a zero

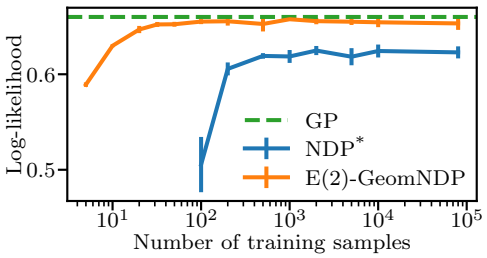


Figure 6.5. Quantitative results for an ablation study fitting models to samples from GP vector fields with varying numbers of training data samples. Mean predictive log-likelihood (↑) and confidence interval estimated over 5 random seeds.

curl (CURL-FREE) kernel and a zero divergence (DIV-FREE) kernel, as described in appendix F.3.1.

We equip our model, GeomNDP, with a $\mathrm{E}(2)$-equivariant score architecture, based on steerable CNNs (Thomas et al., 2018; Weiler et al., 2023). We compare to NDP* with a non-equivariant attention-based network (Dutordoir et al., 2023). We also evaluate two neural processes, a translation equivariant CONVCNP (Gordon et al., 2020) and a $\mathrm{C4} \ltimes \mathbb{R}^2 \subset \mathrm{E}(2)$-equivariant STEERCNP (Holderrieth et al., 2021b). We also report the performance of the data-generating GP, and the same GP but with diagonal posterior covariance GP (DIAG.) . We measure the predictive log-likelihood of the data process samples under the model on a held-out test dataset.

We observe in table 6.3 that the CNPs performance is limited by their diagonal predictive covariance assumption, and as such cannot do better than the GP (DIAG.). We also see that although NDP* is able to fit well GP posteriors, it does not reach the maximum log-likelihood value attained by the data GP, in contrast to its equivariant counterpart $\mathrm{E}(2)$-GEOMNDP . To further explore gains brought by the built-in equivariance, we explore the data-efficiency in figure 6.5 and notice that $\mathrm{E}(2)$-GEOMNDP requires few data samples to fit the data process, since effectively the dimension of the (quotient) state space is dramatically reduced.

6.5.3. Global tropical cyclone trajectory prediction. Finally, we assess our model on a task where the domain of the stochastic process is a non-Euclidean manifold. We model the trajectories of cyclones over the earth, modelled as sample paths of the form $\mathbb{R} \to \mathcal{S}^2$ coming from a stochastic process. The data is drawn from the International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4 (Knapp et al., 2010; Knapp et al., 2018) and preprocessed as per appendix F.4.3, where details on the implementation of the score function, the ODE/SDE solvers used for the sampling, and baseline methods can be found.

Figure 6.6 shows some cyclone trajectories samples from the data process and from a trained GEOMNDP model. We also demonstrate how such trajectories can be interpolated or extrapolated using the conditional sampling method detailed in section 6.3. Such conditional sample paths are shown in figure 6.7. Additionally, we report in table 6.4 the likelihood [1] and MSE for a series of methods. We see that the Gaussian Processes (modelled as $f : \mathbb{R} \to \mathbb{R}^2$, one on latitude/longitude

[1] We only report likelihoods of models defined with respect to the uniform measure on $\mathcal{S}^2$.

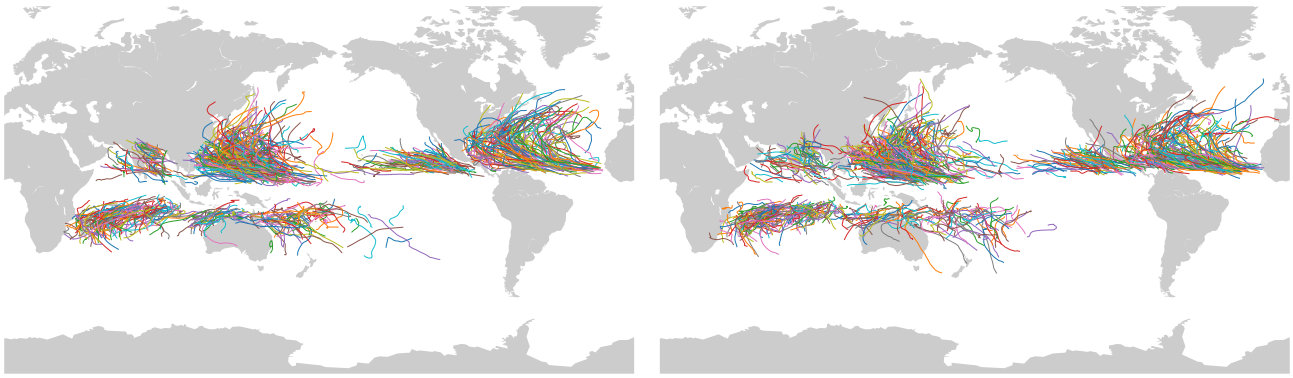


Figure 6.6. *Left:* 1000 samples from the training data. *Right:* 1000 samples from the trained model.

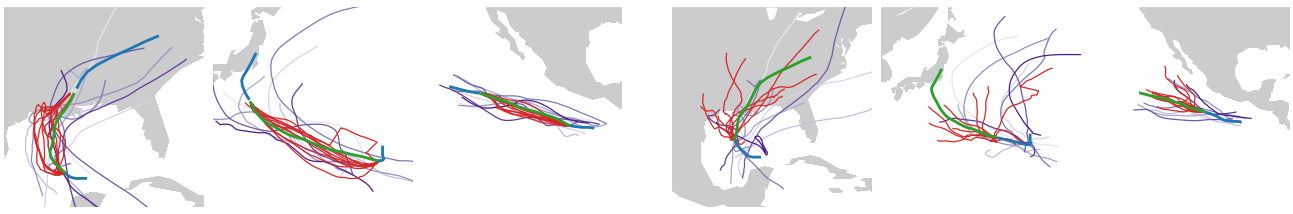
Model	Test data	Interpolation		Extrapolation	
	Likelihood	Likelihood	MSE (km)	Likelihood	MSE (km)
GeomNDP ($\mathbb{R} \to \mathcal{S}^2$)	$\mathbf{802}_{\pm 5}$	$\mathbf{535}_{\pm 4}$	$\mathbf{162}_{\pm 6}$	$\mathbf{536}_{\pm 4}$	$\mathbf{496}_{\pm 14}$
Stereographic GP ($\mathbb{R} \to \mathbb{R}^2/\{0\}$)	$393_{\pm 3}$	$266_{\pm 3}$	$2619_{\pm 13}$	$245_{\pm 2}$	$6587_{\pm 55}$
NDP ($\mathbb{R} \to \mathbb{R}^2$)	-	-	$166_{\pm 22}$	-	$769_{\pm 48}$
GP ($\mathbb{R} \to \mathbb{R}^2$)	-	-	$6852_{\pm 41}$	-	$8138_{\pm 87}$

Table 6.4. Comparative results of different models on the cyclone dataset, comparing test set likelihood, interpolation likelihood and mean squared error, and extrapolation likelihood and mean squared error. Mean and standard deviation are estimated over 5 data splits / random seeds.

coordinates, the other via a stereographic projection, using a diagonal RBF kernel with hyperparameters fitted with maximum likelihood) fail drastically given the high non-Gaussianity of the data. In the interpolation task, the NDP performs as well as the GeomNDP , but the additional geometric structure of modelling the outputs living on the sphere appears to significantly help for extrapolation.

6.6. Discussion

In this chapter, we have extended diffusion models to model invariant stochastic processes over tensor fields. We did so by (a) constructing a continuous noising process over function spaces which correlate input samples with an equivariant kernel, (b) parametrising the score with an equivariant neural network. We have



(a) Interpolation (b) Extrapolation

Figure 6.7. *Top:* Examples of conditional trajectories sampled from the GeomNDP model. *Blue:* Conditioned sections of the trajectory. *Green:* The actual trajectory of the cyclone. *Red:* conditional samples from the model. *Purple:* closest matching trajectories in the dataset to the conditioning data.

empirically demonstrated the ability of our introduced model GEOMNDP to fit complex stochastic processes, and by encoding the symmetry of the problem at hand, we show that it is more data efficient and better able to generalise.

We highlight below some current limitations and important research directions. First, evaluating the model is slow as it relies on costly SDE or ODE solvers. In that respect our model shares this limitation with existing diffusion models. Second, targeting a white noise process appears to over-perform other Gaussian processes. In future work, we would like to investigate the practical influence of different kernels. Third, we wish to apply our method for modelling higher order tensors, such as moment of inertia in classical mechanics (Thomas et al., 2018) or curvature tensor in general relativity. Fourth, strict invariance may sometimes be too strong, we thus suggest softening it by amortising the score network over extra spatial information available from the problem at hand.

7 | Conclusion

7.1. Contributions and applications

This thesis has focused on extending the capabilities of score-based generative models to Riemannian manifolds. Chapter 3 developed methods for modelling on geodesically complete manifolds. Chapters 4 and 5 extended these ideas to geodesically incomplete manifolds, allowing for modelling in the presence of constraints. Chapter 6 looked at modelling data structures that live on manifolds, namely tensor fields and paths on manifolds. This has been achieved by generalising technical results to the manifold setting from the Euclidean setting, proposing practical modelling choices needed to make the manifold setting tractable, and efficient implementations of the algorithms proposed. The main motivation for this has been tackling scientific problems with generative modelling and throughout we have demonstrated practical examples on scientifically relevant tasks.

The work presented in this thesis has already seen a range of applications in scientific modelling problems. The problem of modelling the three-dimension structure of a protein given its amino acid sequence has been studied by Watson et al. (2022), Watson et al. (2023), Wang et al. (2024), and Yim et al. (2023). This involves modelling over products of the group $\mathrm{SE}(3)$. The design of novel proteins, with the additional complexity of modelling the protein backbone residue types has been tackled by Anand and Achim (2022) and Luo et al. (2022). Urain et al. (2022) also consider this setting for the modelling of robotics trajectories. This involves modelling over products of the group $\mathrm{SE}(3)$ as well as a categorical distribution for the amino acid type. Modelling proteins (Wu et al., 2024b) and small molecules (Jing et al., 2022b) by their internal bond rotation angles has also been studied, involving modelling over products of $\mathrm{SO}(2)$. The problem of predicting the interaction between multiple proteins and small molecules has been considered in Corso et al. (2023) and Ketata et al. (2023). Applications in material science have also been considered by Jiao et al. (2024) which involves modelling over Euclidean space with periodic boundary conditions and $\mathrm{E}(3)$ symmetries. Modelling in hyperbolic space has also been considered for modelling tree structures (Wen et al., 2024) and molecules (Wen and Wei, 2023).

7.2. Concurrent work

In the field of machine learning there is never a single group working to tackle a given idea. I would like to briefly touch on other works that tackled similar

problems at similar times to the work presented in this thesis, or made direct theoretical improvements.

Huang et al. (2022) developed methods to tackle a similar setting as presented in chapter 3. The approach taken by the authors differs from chapter 3 in a number of ways. The authors prove a time-reversal result similar to theorem 3.5. In order to define a valid training loss they take a maximum likelihood perspective and generalise the work of Huang et al. (2021) and Song et al. (2021), resulting in an implicit score matching like objective, but not denoising score matching. Finally, in discretising the stochastic differential equations they rely on embedding the manifold into Euclidean space and solving the stochastic differential equation as a Euclidean stochastic differential equation. This relies on expensive nearest point projection operations to keep points on the skin of the manifold, and so can lead to slower and less accurate results than the scheme based on geodesic random walks developed in chapter 3.

Lou et al. (2023) extend the work of chapter 3 and Huang et al. (2022) by providing better tools for heat kernel estimation and sampling on Riemannian manifolds, making these models more tractable.

Lou and Ermon (2023) also worked on the constrained domain setting explored in chapters 4 and 5. The focus of this work is on the hypercube. By exploiting the specific product structure of the hypercube and the tractability of the unit interval they arrive at a tractable approximation to the heat kernel on the hypercube and therefore a denoising score matching objective, and a simulation free forward noising process sampling method. This improves on the implicit score matching objective and the need to simulate the forward noising process. The method is however limited to the hypercube, and domains that can be projected into it, and would not apply to domains with non-zero curvature. By contrast our approach is designed to handle a wider range of settings, such as non-convex polytopes in Euclidean space, or non-Euclidean geometries, where such approximations are not tractable.

Tae (2023) and Liu et al. (2024) also build on the constrained setting of chapters 4 and 5. They explore using mirror maps to project a constrained space into an unconstrained space where unconstrained diffusion modelling is then done. This is similar to the log-barrier method proposed in this thesis, but making use a different map to an unconstrained space. The specific forms of the mirror map can be exploited to allow for a denoising score matching objective and simulation free forward noising process sampling, but at the cost of warping the geometry of the underlying space.

7.3. Future work

7.3.1. Beyond diffusions: flow and bridge matching on manifolds. The paradigm of diffusion modelling starts with samples from an unknown distribution and uses a stochastic differential equation to noise samples from this distribution to a stable reference distribution. By exploiting stochastic differential equations with analytic forward transition densities and sampling, we arrive at efficient score matching training objectives. This has the drawback of limiting us to noising processes

that are aprticularly tractable. When these are not available we have to resort to simulating the forward noising process and either making approximations to the transition density for denoising score matching, or resorting to implicit score matching. This limits the possible maps between target and noise distributions. This is particularly true in the manifold case, as highlighted in chapter 3, where even simple noising processes such as Brownian motion are not analytic.

Flow matching (Lipman et al., 2023; Liu, 2022; Albergo et al., 2023) and bridge matching (Peluchetti, 2023; Liu et al., 2022b; Albergo et al., 2023) take a different perspective. Instead, for samples from two distributions these methods define a deterministic and stochastic map between these points respectively. From these tractable score matching-like objectives can be obtained to define ordinary and stochastic differential equations mapping between the two distributions. The diffusion modelling paradigm can be recovered from the bridge matching paradigm when one of the distributions is a tractable target distribution and can be analytically integrated out of the training objective.

There are two main advantages to these paradigms. Firstly the ability to map two intractable distributions to each other, rather than an intractable one to a tractable one, and secondly a much greater choice in the mapping specified between distributions. For example one can try to match the optimal transport map between distributions, or match the Schrödinger bridge between distributions (De Bortoli et al., 2021; Vargas et al., 2021; Shi et al., 2024; Chen et al., 2022).

These methods potentially provide a way to deal with the additional intractability imposed by modelling on Riemannian manifolds. For example Chen and Lipman (2024) provide a flow matching method that relies only on analytic exponential and logarithm maps on the manifold, rather than needing a tractable noising stochastic differential equation, resulting in significantly better scaling to high dimension problems. Thornton et al. (2022) extends the Schrödinger bridge matching methods to manifolds. There remains ground to explore in this direction for more scalable and efficient generative modelling methods on manifolds.

7.3.2. **Energy-based training.** In this thesis we have explored diffusion modelling methods in the setting where we have a collection of samples $\{x_i\}_{i=1}^N$ from an unknown distribution π. In many scientific problems however the problem setting is that we know a target distribution up to a constant, $q \propto \pi$, without knowing the normalising constant that would allow us to recover π. Classically such problems are sampled with Markov chain Monte Carlo methods (Metropolis et al., 1953) such as Hamiltonian Monte Carlo (Neal, 2011), sequential Monte Carlo (Doucet et al., 2001) techniques, annealed importance sampling (Neal, 2001) and parallel tempering (Geyer, 1991).

Most generative modelling frameworks throw away such an unnormalised target q. It contains significant information however and a number of methods have been developed to make use of this (Noé et al., 2019), most commonly in normalising flow models by mixing these with classic sampling techniques (Midgley et al., 2023; Matthews et al., 2022; Gabrié et al., 2022; Wirnsberger et al., 2022).

For diffusion models it seems natural to be able to incorporate this unnormalised

target into the model as the score at the data is given exactly by $\nabla_x \log \pi(x) = \nabla_x \log q(x)$. This has been exploited in a number of ways.

Zheng et al. (2024) use the known score at the data, $\nabla \log q$, to train the score function at $t = 0$ and then use the continuity equation to enforce path consistency on the rest of the score function. While an exact method, the training objective to enforce consistency both involves the computation of a divergence of the score network, an $O(d^3)$ cost operation in dimension, and is also not a stationary regression objective, eliminating some of the key performance benefits of diffusion models.

Where in section 2.2.1 the forward Kullback-Leibler divergence between the desired diffusion process and the parametrised process, $\mathcal{D}_{KL}\left(P_{\text{data}} \mid P^{\theta}_{\text{model}}\right)$, is used to derive a maximum likelihood objective for diffusion models, Zhang and Chen (2022) and Vargas et al. (2021) utilise the reverse Kullback-Leibler divergence, $\mathcal{D}_{KL}\left(P^{\theta}_{\text{model}} \mid P_{\text{data}}\right)$, to arrive at algorithms that incorporate q and do not require samples from π to compute. The reliance on the reverse Kullback-Leibler divergence however means the algorithms suffer from mode seeking behaviour, and the use of importance sampling presents difficulties in high dimension.

Phillips et al. (2024) develop a particle filtering scheme in which the series of distributions traversed by the particles is defined by a diffusion path, and the score of this path approximated with a simple linear approximation. This approximation, and therefore efficiency of the particle filter, can be improved by training the approximation using samples from the particle filter, bootstrapping the model to generate self-training data.

Akhound-Sadegh et al. (2024) similarly develop a model bootstrapping approach to sampling. They derive an importance sampled loss function incorporating q allowing for arbitrary samples to be used to train the diffusion model.

De Bortoli et al. (2024) introduce a new score matching loss, the target score matching loss, that depends only on q. They show that this loss has lower variance near the start of the noising diffusion, and that denoising score matching has lower variance further into the noising process. By combining these two losses and varying their weights over time they obtain a loss with the best of the properties of denoising and target score matching over time.

This current body of work shows that there is strong promise in combining classic sampling techniques for unnormalised densities with diffusion model ideas. Given that solving many scientific problems requires good unnormalised density sampling further combining the power of diffusion modellings generalisation capabilities with the statistical accuracy of classical sampling techniques seems like an appealing direction for pushing the bounds of science.

Bibliography

M. F. A. R. D. T. (FAIR)†, A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022. Cited on page 1.

N. G. D. C. /. W. D. S. (NGDC/WDS). NCEI/WDS Global Significant Earthquake Database. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ngdc.mgg.hazards:G012153, 2022. doi: doi:10.7289/V5TD9V7K. Cited on page 61.

N. G. D. C. /. W. D. S. (NGDC/WDS). NCEI/WDS Global Significant Volcanic Eruptions Database. https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.ngdc.mgg.hazards:G10147, 2022. doi: 10.7289/V5JW8BSH. Cited on page 61.

R. Abbott, M. S. Albergo, A. Botev, D. Boyda, K. Cranmer, D. C. Hackett, G. Kanwar, A. G. Matthews, S. Racanière, A. Razavi, et al. Normalizing flows for lattice gauge theory in arbitrary space-time dimension. *arXiv preprint arXiv:2305.02402*, 2023. Cited on page 38.

R. Abbott, M. S. Albergo, A. Botev, D. Boyda, K. Cranmer, D. C. Hackett, G. S. Kanwar, A. G. Matthews, S. Racanière, A. Razavi, et al. Sampling QCD field configurations with gauge-equivariant flow models, 2022. Cited on page 38.

R. Abbott, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, G. Kanwar, S. Racanière, D. J. Rezende, F. Romero-López, P. E. Shanahan, et al. Gauge-equivariant flow models for sampling in lattice field theories with pseudofermions. *Physical Review D*, 106(7):074506, 2022. Cited on page 38.

R. Abbott, A. Botev, D. Boyda, D. C. Hackett, G. Kanwar, S. Racanière, D. J. Rezende, F. Romero-López, P. E. Shanahan, and J. M. Urban. Applications of flow models to the generation of correlated lattice QCD ensembles. *Physical Review D*, 109(9):094514, 2024. Cited on page 3.

J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*:1–3, 2024. Cited on pages 1, 25.

P.-A. Absil and J. Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012. Cited on pages 72, 89.

Y. Aflalo, H. Brezis, and R. Kimmel. On the optimality of shape and data representation in the spectral domain. *SIAM Journal on Imaging Sciences*, 8(2):1141–1160, 2015. Cited on pages 22, 220.

Y. Aflalo and R. Kimmel. Spectral multidimensional scaling. *Proceedings of the National Academy of Sciences*, 110(45):18052–18057, 2013. Cited on pages 22, 220.

T. Akhound-Sadegh, J. Rector-Brooks, A. J. Bose, S. Mittal, P. Lemos, C.-H. Liu, M. Sendera, S. Ravanbakhsh, G. Gidel, Y. Bengio, et al. Iterated denoising energy matching for sampling from Boltzmann densities. *arXiv preprint arXiv:2402.06121*, 2024. Cited on page 126.

M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023. Cited on page 125.

E. L. Allgower and K. Georg. *Numerical Continuation Methods: An Introduction*, volume 13. Springer Science & Business Media, 2012. Cited on page 50.

N. Anand and T. Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022. Cited on page 123.

B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. Cited on page 248.

D. Applegate and R. Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 156–163, 1991. Cited on page 95.

K. Atkinson and W. Han. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*, volume 2044. Springer Science & Business Media, 2012. Cited on page 221.

I. Azangulov, A. Smolensky, A. Terenin, and V. Borovitskiy. Stationary Kernels and Gaussian Processes on Lie Groups and their Homogeneous Spaces I: the compact case, 2023. arXiv: 2208.14960 [stat.ME]. Cited on pages 38, 102, 115.

I. Azangulov, A. Smolensky, A. Terenin, and V. Borovitskiy. Stationary Kernels and Gaussian Processes on Lie Groups and their Homogeneous Spaces II: non-compact symmetric spaces, 2023. arXiv: 2301.13088 [stat.ME]. Cited on pages 38, 102, 115.

H. Baatz, J. Granskog, M. Papas, F. Rousselle, and J. Novák. NeRF-Tex: Neural Reflectance Field Textures. In *Computer graphics forum*, volume 41 of number 6, pages 287–301. Wiley Online Library, 2022. Cited on page 24.

D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348. Springer, 2014, pages xx+552. URL: http://dx.doi.org/10.1007/978-3-319-00227-9. Cited on page 269.

V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations: I. Convergence rate of the distribution function. *Probability theory and related fields*, 104:43–60, 1996. Cited on page 250.

A. Bansal, E. Borgnia, H.-M. Chu, J. S. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022. Cited on page 116.

T. Barfoot, J. R. Forbes, and P. T. Furgale. Pose Estimation Using Linearized Rotations and Quaternion Algebra. *Acta Astronautica*, 68(1):101–112, Jan. 2011. ISSN: 0094-5765. DOI: 10.1016/j.actaastro.2010.06.049. Cited on page 282.

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. Cited on page 14.

G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann. Conditional Image Generation with Score-Based Diffusion Models. *arXiv preprint arXiv:2111.13606*, 2021. Cited on page 56.

F. Baudoin. The heat semigroup on a compact Riemannian manifold, 2013. URL: https://fabricebaudoin.blog/2013/08/28/lecture-12-the-heat-semigroup-on-a-compact-riemannian-manifold/. Cited on page 57.

J. A. Bednar, J. Crail, J. Crist-Harif, P. Rudiger, G. Brener, I. Thomas, C. B, J. Mease, J. Signell, M. Liquet, J.-L. Stevens, B. Collins, A. Thorve, thuydotm, S. H. Hansen, esc, kbowen, N. Abdennur, O. Smirnov, maihde, A. Hawley, A. Oriekhov, A. Ahmadia, B. A. B. Jr, C. H. Brandt, C. Tolboom, E. G., E. Welch, J. Bourbeau, and J. J. Schmidt. Holoviz/Datashader, version v0.14.4, Feb. 2023. URL: https://doi.org/10.5281/zenodo.7599872. Cited on page 315.

A. Behboodi, G. Cesa, and T. S. Cohen. A PAC-Bayesian Generalization Bound for Equivariant Networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022. Cited on pages 13, 102.

R. Bellman. Dynamic programming. *Chapter IX, Princeton University Press, Princeton, New Jersey*, 1958. Cited on pages 2, 9.

C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019. Cited on page 1.

J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023. Cited on page 1.

M. Bevis and J.-L. Chatelain. Locating a point on a spherical surface relative to a spherical polygon of arbitrary shape. *Mathematical geology*, 21:811–828, 1989. Cited on pages 99, 315, 316.

R. Bhattacharya. Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *The Annals of Probability*:541–553, 1978. Cited on page 287.

M. Biloš, K. Rasul, A. Schneider, Y. Nevmyvaka, and S. Günnemann. Modeling Temporal Data as Continuous Functions with Process Diffusion, 2022. URL: HTTPS://OPENREVIEW.NET/FORUM?ID=VMJKUYPQ8WR. Cited on page 116.

J.-M. Bismut. Large deviations and the Malliavin calculus. *Birkhauser Prog. Math.*, 45, 1984. Cited on page 58.

S. Bond-Taylor and C. G. Willcocks. ∞-Diff: Infinite Resolution Diffusion with Subsampled Mollified States. In *The Twelfth International Conference on Learning Representations*, 2023. Cited on pages 101, 116.

V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Matern Gaussian Processes on Riemannian Manifolds. In *Advances in Neural Information Processing Systems*, 2020. Cited on page 38.

J. Bose, A. Smofsky, R. Liao, P. Panangaden, and W. Hamilton. Latent variable modelling with hyperbolic normalizing flows. In *International Conference on Machine Learning*, 2020. Cited on page 37.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992. Cited on page 1.

M. Bossy, E. Gobet, and D. Talay. A symmetrized Euler scheme for an efficient approximation of reflected diffusions. *Journal of applied probability*, 41(3):877–889, 2004. Cited on pages 75, 94.

N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. Cited on pages 72, 89.

S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. Cited on pages 69, 70.

D. Boyda, G. Kanwar, S. Racanière, D. J. Rezende, M. S. Albergo, K. Cranmer, D. C. Hackett, and P. E. Shanahan. Sampling using SU (N) gauge equivariant flows. *Physical Review D*, 103(7):074504, 2021. Cited on page 38.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, version 0.2.5, 2018. URL: HTTP://GITHUB.COM/GOOGLE/JAX. Cited on pages 116, 279, 294, 335.

G. Brakenridge. Global active archive of large flood events, 2017. URL: http://floodobservatory.colorado.edu/Archives/index.html. Cited on page 61.

J. Brehmer and K. Cranmer. Flows for simultaneous manifold learning and density estimation. *Advances in neural information processing systems*, 33:442–453, 2020. Cited on pages 37, 40.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges, 2021. arXiv: 2104.13478 [cs.LG]. Cited on pages 3, 15.

M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. Cited on page 3.

T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators, 2024. URL: https://openai.com/research/video-generation-models-as-world-simulators. Cited on page 24.

W. Bruinsma, S. Markou, J. Requeima, A. Y. K. Foong, A. Vaughan, T. Andersson, A. Buonomo, S. Hosking, and R. E. Turner. Autoregressive Conditional Neural Processes. In *International Conference on Learning Representations*, Feb. 2023. Cited on page 115.

W. Bruinsma, J. Requeima, A. Y. K. Foong, J. Gordon, and R. E. Turner. Gaussian Neural Processes. In *3rd Symposium on Advances in Approximate Bayesian Inference*, Feb. 2020. Cited on pages 116–118, 336, 337.

K. Burdzy and Z.-Q. Chen. Discrete approximations to reflected Brownian motion, 2008. Cited on page 95.

K. Burdzy, Z.-Q. Chen, and D. E. Marshall. Traps for reflected Brownian motion. *Mathematische Zeitschrift*, 252:103–132, 2006. Cited on page 74.

K. Burdzy, Z.-Q. Chen, and J. Sylvester. The heat equation and reflected Brownian motion in time-dependent domains. *The Annals of Probability*, 32(1B):775–804, 2004. Cited on pages 74, 288, 289, 291–293.

Y. Canzani. Analysis on Manifolds via the Laplacian, 2013. URL: https://www.math.mcgill.ca/toth/spectral%20geometry.pdf. Cited on page 22.

A. L. Caterini, G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham. Rectangular flows for manifold learning. *Advances in Neural Information Processing Systems*, 34:30228–30241, 2021. Cited on page 37.

P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. Time reversal of diffusion processes under a finite entropy condition. In *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, volume 59 of number 4, pages 1844–1881. Institut Henri Poincaré, 2023. Cited on pages 46, 60, 110, 248, 259, 262, 263, 265, 267, 292.

J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997. Cited on page 227.

I. Chavel. *Eigenvalues in Riemannian Geometry*. Academic press, 1984. Cited on pages 22, 40, 220.

R. T. Chen and Y. Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024. Cited on page 125.

R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018. Cited on pages 2, 26, 36–38.

T. Chen, G.-H. Liu, and E. Theodorou. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representations*, 2022. Cited on page 125.

X. Chen, X.-M. Li, and B. Wu. Logarithmic heat kernel estimates without curvature restrictions. *The Annals of Probability*, 51(2):442–477, 2023. Cited on page 58.

L. Cheng, P. B. Szabó, Z. Schätzle, D. Kooi, J. Köhler, K. J. Giesbertz, F. Noé, J. Hermann, P. Gori-Giorgi, and A. Foster. Highly Accurate Real-space Electron Densities with Neural Networks. *arXiv preprint arXiv:2409.01306*, 2024. Cited on page 24.

X. Cheng, J. Zhang, and S. Sra. Theory and Algorithms for Diffusion Processes on Riemannian Manifolds. *arXiv preprint arXiv:2204.13665*, 2022. Cited on pages 50, 268, 271, 272, 277.

J. Chibane and G. Pons-Moll. Implicit feature networks for texture completion from partial 3d data. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 717–725. Springer, 2020. Cited on page 24.

R. Chitashvili and N. Lazrieva. Strong solutions of stochastic differential equations with boundary conditions. *Stochastics: an international journal of probability and stochastic processes*, 5(4):255–309, 1981. Cited on page 94.

K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. Cited on page 15.

H. Chung, B. Sim, and J. C. Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12413–12422, 2022. Cited on page 56.

S. Cohen, B. Amos, and Y. Lipman. Riemannian convex potential maps. In *International Conference on Machine Learning*, pages 2028–2038. PMLR, 2021. Cited on page 38.

T. Cohen. *Equivariant convolutional networks*. PhD thesis, 2021. Cited on page 109.

T. Cohen and M. Welling. Group Equivariant Convolutional Networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*. PMLR, 2016. Cited on page 15.

T. S. Cohen and M. Welling. Steerable CNNs. In *ICLR*, 2017. Cited on page 15.

G. Corso, B. Jing, R. Barzilay, T. Jaakkola, et al. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. In *International Conference on Learning Representations (ICLR 2023)*, 2023. Cited on page 123.

M. Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression. jl. *arXiv preprint arXiv:2305.01582*, 2023. Cited on page 97.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989. Cited on page 9.

G. Da Prato and J. Zabczyk. *Stochastic equations in infinite dimensions*. Cambridge university press, 2014. Cited on page 116.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):651–676, 2017. ISSN: 1369-7412. DOI: 10.1111/RSSB.12183. URL: HTTPS://DOI.ORG/10.1111/RSSB.12183. Cited on page 334.

G. Daras, M. Delbracio, H. Talebi, A. Dimakis, and P. Milanfar. Soft Diffusion: Score Matching with General Corruptions. *Transactions on Machine Learning Research*, 2023. Cited on page 116.

V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. Cited on pages 331, 332.

V. De Bortoli, M. J. Hutchinson, P. Wirnsberger, and A. Doucet. Target Score Matching. *arXiv preprint arXiv:2402.08667*, 2024. Cited on page 126.

V. De Bortoli, E. Mathieu, M. J. Hutchinson, J. Thornton, Y. W. Teh, and A. Doucet. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 2022. Cited on page 317.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems*, 2021. Cited on pages 125, 340.

V. De Bortoli*, E. Mathieu*, M. J. Hutchinson*, J. Thornton, Y. W. Teh, and A. Doucet. Riemannian score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2022.

M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016. Cited on page 14.

J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022. Cited on page 1.

P. Dhariwal and A. Nichol. Diffusion models beat GAN on Image Synthesis. *arXiv preprint arXiv:2105.05233*, 2021. Cited on pages 39, 101.

R. J. Dormand and J. P. Prince. A Family of Embedded Runge-Kutta Formulae. *Journal of Computational and Applied Mathematics*:19–26, 1980. Cited on page 280.

A. Doucet, N. De Freitas, and N. Gordon. An introduction to sequential Monte Carlo methods. *Sequential Monte Carlo methods in practice*:3–14, 2001. Cited on page 125.

P. G. Dougherty, A. Sahni, and D. Pei. Understanding cell penetration of cyclic peptides. *Chemical Reviews*, 119(17):10241–10287, 2019. Cited on page 83.

E. Dupont, H. Kim, S. A. Eslami, D. J. Rezende, and D. Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. In *International Conference on Machine Learning*, pages 5694–5725. PMLR, 2022. Cited on page 115.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *ArXiv e-prints*, May 2016. arXiv: 1605.01559 [math.ST]. Cited on page 327.

A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017. ISSN: 1050-5164. DOI: 10.1214/16-AAP1238. URL: HTTPS://DOI.ORG/10.1214/16-AAP1238. Cited on page 334.

V. Dutordoir, A. Saul, Z. Ghahramani, and F. Simpson. Neural diffusion processes. In *International Conference on Machine Learning*, pages 8990–9012. PMLR, 2023. Cited on pages 101, 104, 116, 117, 119, 336, 338, 340.

P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017. Cited on page 83.

B. Elesedy. Provably strict generalisation benefit for invariance in kernel methods. *Advances in Neural Information Processing Systems*, 34:17273–17283, 2021. Cited on pages 13, 102.

B. Elesedy and S. Zaidi. Provably strict generalisation benefit for equivariant models. In *International conference on machine learning*, pages 2959–2969. PMLR, 2021. Cited on pages 13, 102.

J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990. Cited on page 15.

EOSDIS. Land, Atmosphere Near real-time Capability for EOS (LANCE) system operated by NASA's Earth Science Data and Information System (ESDIS). "https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data", 2020. Cited on page 61.

L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015. Cited on page 292.

L. Falorsi. Continuous Normalizing Flows on Manifolds. *arXiv preprint arXiv:2104.14959*, Mar. 2021. Cited on page 38.

L. Falorsi, P. de Haan, T. R. Davidson, and P. Forré. Reparameterizing distributions on lie groups. In *International Conference on Artificial Intelligence and Statistics*, pages 3244–3253, 2019. Cited on pages 37, 64, 282.

L. Falorsi and P. Forré. Neural ordinary differential equations on manifolds. *arXiv preprint arXiv:2006.06663*, 2020. Cited on pages 38, 55, 194.

Y. Fang, I. Ohn, V. Gupta, and L. Lin. Intrinsic and extrinsic deep learning on manifolds. *arXiv preprint arXiv:2302.08606*, 2023. Cited on page 24.

A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatain, A. Novikov, F. J. R Ruiz, J. Schrittwieser, G. Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022. Cited on page 1.

H. Federer. *Geometric Measure Theory*. Springer, 2014. Cited on page 255.

W. Feiten, M. Lang, and S. Hirche. Rigid motion estimation using mixtures of projected Gaussians. In *International Conference on Information Fusion*, pages 1465–1472. IEEE, 2013. Cited on page 40.

N. Fishman, L. Klarner, V. De Bortoli, E. Mathieu, and M. J. Hutchinson. Diffusion Models for Constrained Domains. *Transactions on Machine Learning Research*, 2023.

N. Fishman, L. Klarner, E. Mathieu, M. J. Hutchinson, and V. De Bortoli. Metropolis Sampling for Constrained Diffusion Models. In *Advances in Neural Information Processing Systems*, 2023.

H. Föllmer. An entropy approach to the time reversal of diffusion processes. In *Stochastic Differential Systems: Filtering and Control*, pages 156–163. Springer, 1985. Cited on page 248.

G. Franzese, G. Corallo, S. Rossi, M. Heinonen, M. Filippone, and P. Michiardi. Continuous-time functional diffusion processes. *Advances in Neural Information Processing Systems*, 36, 2024. Cited on pages 101, 116.

F. B. Fuchs, D. E. Worrall, V. Fischer, and M. Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. In *NeurIPS*, Nov. 2020. URL: http://arxiv.org/abs/2006.10503. Cited on page 14.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980. Cited on pages 3, 15.

M. Gabrié, G. M. Rotskoff, and E. Vanden-Eijnden. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10):e2109420119, 2022. Cited on page 125.

M. P. Gaffney. A Special Stokes's Theorem for Complete Riemannian Manifolds. *Annals of Mathematics*, 60(1):140–145, 1954. Cited on pages 47, 278.

O.-E. Ganea, X. Huang, C. Bunne, Y. Bian, R. Barzilay, T. S. Jaakkola, and A. Krause. Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking. In *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=GQjaI9mLet. Cited on page 282.

D. García-Zelada and B. Huguet. Brenier–Schródinger problem on compact manifolds with boundary. *Stochastic Analysis and Applications*:1–29, 2021. Cited on page 259.

M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, S. M. A. Eslami, and Y. W. Teh. Neural Processes. *arXiv preprint arXiv:1807.01622*, July 2018. Cited on page 102.

K. Gatmiry and S. S. Vempala. Convergence of the Riemannian Langevin Algorithm. *arXiv preprint arXiv:2204.10818*, 2022. Cited on pages 43, 68, 88.

M. Geiger and T. Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022. Cited on page 338.

M. C. Gemici, D. Rezende, and S. Mohamed. Normalizing flows on riemannian manifolds. *arXiv preprint arXiv:1611.02304*, 2016. Cited on pages 37, 61.

Gemini-Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. Cited on page 1.

C. J. Geyer. Markov chain Monte Carlo maximum likelihood, 1991. Cited on page 125.

J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*. PMLR, 2017. Cited on page 14.

E. Gobet. Euler schemes and half-space approximation for the simulation of diffusion in a domain. *ESAIM: Probability and Statistics*, 5:261–297, 2001. Cited on page 94.

G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013. Cited on page 80.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. Cited on page 2.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. Cited on pages 25, 38.

J. Gordon, W. P. Bruinsma, A. Y. K. Foong, J. Requeima, Y. Dubois, and R. E. Turner. Convolutional Conditional Neural Processes. In *International Conference on Learning Representations*, 2020. Cited on pages 102, 115, 117, 119, 338.

W. Grathwohl, R. T. Chen, J. Bettencourt, and D. Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, page 7, 2019. Cited on pages 26, 38.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(null):723–773, Mar. 2012. ISSN: 1532-4435. Cited on pages 25, 79, 280, 295, 316.

L. Gross. Logarithmic Sobolev inequalities on Lie groups. *Illinois journal of mathematics*, 36(3):447–490, 1992. Cited on page 59.

M. Gunther. Isometric embeddings of Riemannian manifolds, Kyoto, 1990. In *Proc. Intern. Congr. Math.* Pages 1137–1143. Math. Soc. Japan, 1991. Cited on pages 254, 256, 257, 264, 265.

F. Guth, S. Coste, V. De Bortoli, and S. Mallat. Wavelet score-based generative modeling. *Advances in neural information processing systems*, 35:478–491, 2022. Cited on pages 115, 116.

P. Hagemann, L. Ruthotto, G. Steidl, and N. T. Yang. Multilevel Diffusion: Infinite Dimensional Score-Based Diffusion Models for Image Generation. *arXiv preprint arXiv:2303.04772*, 2023. Cited on page 116.

L. Han and L. Rudolph. Inverse Kinematics for a Serial Chain with Joints Under Distance Constraints. In *Robotics: Science and systems*, 2006. Cited on pages 68, 69, 78, 82, 84, 320.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. DOI: 10.1038/s41586-020-2649-2. URL: HTTPS://DOI.ORG/10.1038/S41586-020-2649-2. Cited on page 335.

J. M. Harrison and R. J. Williams. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics: An International Journal of Probability and Stochastic Processes*, 22(2):77–115, 1987. Cited on page 73.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009. Cited on page 13.

U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986. Cited on pages 46, 74, 248, 259–261, 292.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. Cited on page 15.

Y. He. A lower bound for the first eigenvalue in the Laplacian operator on compact Riemannian manifolds. *Journal of Geometry and Physics*, 71:73–84, 2013. Cited on page 259.

B. He*, S. Zaidi*, B. Elesedy*, M. J. Hutchinson*, A. Paleyes, G. Harling, A. Johnson, and Y. W. Teh. Technical Document 3: Effectiveness and Resource Requirements of Test, Trace and Isolate Strategies. *Royal Society DELVE Initiative, Report on Test, Trace, Isolate Systems*, 2020.

L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdóttir, J. Wachowiak, S. M. Keating, V. Vlasov, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. *Nature protocols*, 14(3):639–702, 2019. Cited on page 68.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. Cited on pages 3, 26, 38, 39, 101, 116.

J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. *Journal of Machine Learning Research*, 23:47–1, 2022. Cited on page 115.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. Cited on pages 3, 15.

P. Holderrieth, M. J. Hutchinson, and Y. W. Teh. Equivariant Learning of Stochastic Fields: Gaussian Processes and Steerable Conditional Neural Processes. In *International Conference on Machine Learning*, 2021. Cited on page 38.

P. Holderrieth, M. J. Hutchinson, and Y. W. Teh. Equivariant Learning of Stochastic Fields: Gaussian Processes and Steerable Conditional Neural Processes. In *International Conference on Machine Learning*, 2021. Cited on pages 102, 107, 108, 115, 119, 327, 328, 338.

P. Holderrieth*, M. J. Hutchinson*, and Y. W. Teh. Equivariant Learning of Stochastic Fields: Gaussian Processes and Steerable Conditional Neural Processes. In *International Conference on Machine Learning*, 2021.

E. Hoogeboom, J. Heek, and T. Salimans. Simple Diffusion: End-to-end Diffusion for High Resolution Images, Jan. 2023. DOI: 10.48550/arXiv.2301.11093. Cited on page 30.

E. Hoogeboom and T. Salimans. Blurring Diffusion Models. In *The Eleventh International Conference on Learning Representations*, 2023. Cited on pages 115, 116.

V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006. Cited on page 83.

B. Hou, N. Miolane, B. Khanal, M. C. H. Lee, A. Alansary, S. McDonagh, J. V. Hajnal, D. Rueckert, B. Glocker, and B. Kainz. Computing CNN Loss and Gradients for Pose Estimation with Riemannian Geometry. In A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 756–764, Cham. Springer International Publishing, 2018. ISBN: 978-3-030-00928-1. Cited on page 282.

E. P. Hsu. *Stochastic Analysis on Manifolds*, number 38. American Mathematical Society, 2002. Cited on pages 42, 251, 252, 255–257.

C.-W. Huang, M. Aghajohari, J. Bose, P. Panangaden, and A. C. Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022. Cited on pages 54, 124.

C.-W. Huang, J. H. Lim, and A. C. Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34:22863–22876, 2021. Cited on pages 35, 124.

Z. Huang, C. Wan, T. Probst, and L. Van Gool. Deep learning on lie groups for skeleton-based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6099–6108, 2017. Cited on page 24.

Z. Huang, J. Wu, and L. Van Gool. Building deep networks on grassmann manifolds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32 of number 1, 2018. Cited on page 24.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. DOI: 10.1109/MCSE.2007.55. Cited on page 335.

M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989. Cited on pages 32, 52.

M. J. Hutchinson, M. Reisser, and C. Louizos. Federated Functional Variational Inference. In *Bayesian Deep Learning workshop, Advances in Neural Information Processing Systems*, 2021.

M. J. Hutchinson, A. Terenin, V. Borovitskiy, S. Takao, Y. Whye Teh, and M. P. Deisenroth. Vector-valued Gaussian Processes on Riemannian Manifolds via Gauge Equivariant Projected Kernels. In *Advances in Neural Information Processing Systems*, 2021.

M. J. Hutchinson, A. Terenin, V. Borovitskiy, S. Takao, Y. Whye Teh, and M. P. Deisenroth. Vector-valued Gaussian Processes on Riemannian Manifolds via Gauge Equivariant Projected Kernels. In *Advances in Neural Information Processing Systems*, volume 34, pages 17160–17169, 2021. Cited on pages 38, 109, 110.

M. J. Hutchinson*, C. L. Lan*, S. Zaidi*, E. Dupont, Y. W. Teh, and H. Kim. LieTransformer: Equivariant Self-Attention for Lie Groups. In *International Conference on Machine Learning*, 2021.

A. Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 2005. Cited on pages 31, 39.

N. Ikeda and S. Watanabe. *Stochastic Differential Equations and Diffusion Processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second edition, 1989, pages xvi+555. ISBN: 0-444-87378-3. Cited on page 257.

N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014. Cited on page 287.

J. Ingraham, M. Baranov, Z. Costello, V. Frappier, A. Ismail, S. Tie, W. Wang, V. Xue, F. Obermeyer, A. Beam, et al. Illuminating protein space with a programmable generative model. *bioRxiv*:2022–12, 2022. Cited on page 116.

V. E. Ismailov. A three layer neural network can represent any multivariate function. *Journal of Mathematical Analysis and Applications*, 523(1):127096, 2023. Cited on page 9.

N. Jaquier, L. Rozo, D. G. Caldwell, and S. Calinon. Geometry-aware manipulability learning, tracking, and transfer. *The International Journal of Robotics Research*, 40(2-3):624–650, 2021. Cited on pages 79, 80, 102.

R. Jiao, W. Huang, P. Lin, J. Han, P. Chen, Y. Lu, and Y. Liu. Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems*, 36, 2024. Cited on page 123.

B. Jing, G. Corso, R. Berlinghieri, and T. Jaakkola. Subspace diffusion generative models. In *European Conference on Computer Vision*, pages 274–289. Springer, 2022. Cited on page 115.

B. Jing, G. Corso, J. Chang, R. Barzilay, and T. Jaakkola. Torsional diffusion for molecular conformer generation. *Advances in Neural Information Processing Systems*, 35:24240–24253, 2022. Cited on page 123.

P. W. Jones, M. Maggioni, and R. Schul. Manifold parametrizations by eigenfunctions of the Laplacian and Heat Kernels. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6):1803–1808, 2008. ISSN: 0027-8424. URL: HTTPS://WWW.JSTOR.ORG/STABLE/25451369. Cited on page 58.

M. I. Jordan. Serial order: A parallel distributed processing approach. In *Advances in psychology*. Volume 121, pages 471–495. Elsevier, 1997. Cited on page 15.

E. Jørgensen. The central limit problem for geodesic random walks. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(1-2):1–64, 1975. Cited on pages 40, 49, 258.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. Cited on page 3.

O. Kallenberg. *Foundations of Modern Probability*. Springer International Publishing, 1997. Cited on pages 223, 243.

G. Kallianpur. *The General Filtering Problem and the Stochastic Equation of the Optimal Filter (Part I)*. In *Stochastic Filtering Theory*. Springer New York, New York, NY, 1980, pages 192–224. ISBN: 978-1-4757-6592-2. DOI: 10.1007/978-1-4757-6592-2_8. URL: HTTPS://DOI.ORG/10.1007/978-1-4757-6592-2_8. Cited on page 232.

W. Kang and K. Ramanan. On the submartingale problem for reflected diffusions in domains with piecewise smooth boundaries, 2017. Cited on pages 92, 93, 289, 300, 313.

R. Kannan and H. Narayanan. Random Walks on Polytopes and an Affine Interior Point Method for Linear Programming. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, pages 561–570, Bethesda, MD, USA. Association for Computing Machinery, 2009. ISBN: 9781605585062. DOI: 10.1145/1536414.1536491. URL: HTTPS://DOI.ORG/10.1145/1536414.1536491. Cited on pages 68, 78, 88.

L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422, 1960. Cited on page 25.

G. Kanwar, M. S. Albergo, D. Boyda, K. Cranmer, D. C. Hackett, S. Racaniere, D. J. Rezende, and P. E. Shanahan. Equivariant flow-based sampling for lattice gauge theory. *Physical Review Letters*, 125(12):121601, 2020. Cited on page 38.

A. Karpatne, I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar. Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8), 2018. Cited on page 40.

T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. Cited on pages 30, 37, 326.

T. Karras, M. Aittala, J. Lehtinen, J. Hellsten, T. Aila, and S. Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24174–24184, 2024. Cited on pages 30, 35.

I. Katsman, A. Lou, D. Lim, Q. Jiang, S. N. Lim, and C. M. De Sa. Equivariant manifold flows. *Advances in Neural Information Processing Systems*, 34:10600–10612, 2021. Cited on page 38.

B. Kawar, G. Vaksman, and M. Elad. SNIPS: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021. Cited on page 56.

B. Kawar, G. Vaksman, and M. Elad. Stochastic image denoising by sampling from the posterior distribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1866–1875, 2021. Cited on page 56.

G. Kerrigan, J. Ley, and P. Smyth. Diffusion Generative Models in Infinite Dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 9538–9563. PMLR, 2023. Cited on pages 101, 104, 105, 116.

M. A. Ketata, C. Laue, R. Mammadov, H. Stärk, M. Wu, G. Corso, C. Marquet, R. Barzilay, and T. S. Jaakkola. Diffdock-pp: Rigid protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889*, 2023. Cited on page 123.

R. Ketzner, V. Ravindra, and M. Bramble. A robust, fast, and accurate algorithm for point in spherical polygon classification with applications in geoscience and remote sensing. *Computers & Geosciences*, 167:105185, 2022. Cited on pages 99, 315, 316.

P. Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021. Cited on page 37.

D. P. Kingma. Adam: A method for stochastic optimization. In 2014. Cited on pages 280, 294, 339.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. Cited on pages 2, 25, 38.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. Cited on page 14.

L. Klein, A. Foong, T. Fjelde, B. Mlodozeniec, M. Brockschmidt, S. Nowozin, F. Noé, and R. Tomioka. Timewarp: Transferable acceleration of molecular dynamics by learning time-coarsened dynamics. *Advances in Neural Information Processing Systems*, 36, 2024. Cited on page 24.

A. Klimovskaia, D. Lopez-Paz, L. Bottou, and M. Nickel. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature communications*, 11(1):1–9, 2020. Cited on page 40.

P. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2011. ISBN: 9783540540625. URL: HTTPS://BOOKS.GOOGLE.FR/BOOKS?ID=BCVTSSOM1CMC. Cited on page 254.

H. J. Knapp Kenneth R. Diamond, J. P. Kossin, M. C. Kruk, and C. J. I. Schreck. International Best Track Archive for Climate Stewardship (IBTrACS) Project, Version 4. Technical report, NOAA National Centers for Environmental Information, 2018. DOI: HTTPS://DOI.ORG/10.25921/82TY-9E16. Cited on page 119.

K. R. Knapp, M. C. Kruk, D. H. Levinson, H. J. Diamond, and C. J. Neumann. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying Tropical Cyclone Data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010. DOI: HTTPS://DOI.ORG/10.1175/2009BAMS2755.1. Cited on page 119.

L. Koestler, D. Grittner, M. Moeller, D. Cremers, and Z. Lähner. Intrinsic neural fields: Learning functions on manifolds. In *European Conference on Computer Vision*, pages 622–639. Springer, 2022. Cited on page 24.

J. Köhler, L. Klein, and F. Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pages 5361–5370. PMLR, 2020. Cited on pages 56, 57, 328.

I. Kolár, P. W. Michor, and J. Slovák. *Natural operations in differential geometry*. Springer Science & Business Media, 2013. Cited on page 15.

Y. Kook, Y.-T. Lee, R. Shen, and S. Vempala. Sampling with Riemannian Hamiltonian Monte Carlo in a constrained space. *Advances in Neural Information Processing Systems*, 35:31684–31696, 2022. Cited on pages 68, 78, 88.

R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 384(6693):eadl2528, 2024. Cited on page 25.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. Cited on page 1.

A. Krogh and J. Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991. Cited on page 9.

S. Kullback. *Information Theory and Statistics*. Dover Publications, Inc., Mineola, NY, 1997, pages xvi+399. ISBN: 0-486-69684-7. Reprint of the second (1968) edition. Cited on page 269.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. Cited on page 25.

T. G. Kurtz, É. Pardoux, and P. Protter. Stratonovich stochastic differential equations driven by general semimartingales. In *Annales de l'IHP Probabilités et statistiques*, volume 31 of number 2, pages 351–377, 1995. Cited on page 254.

K. Kuwada. Convergence of time-inhomogeneous geodesic random walks and its application to coupling methods. *The Annals of Probability*, 40(5):1945–1979, 2012. Cited on pages 50, 51.

T. J. Lane. Protein structure prediction has reached the single-structure frontier. *Nature Methods*:1–4, 2023. Cited on page 81.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*:1302–1338, 2000. Cited on page 303.

Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989. Cited on pages 3, 10, 15.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, volume 86 of number 11, pages 2278–2324, 1998. URL: HTTP://CITESEERX.IST.PSU.EDU/VIEWDOC/SUMMARY?DOI=10.1.1.42.7665. Cited on page 338.

J. Lee. *Introduction to Topological Manifolds*, volume 202. Springer Science & Business Media, 2010. Cited on pages 15, 252.

J. M. Lee. *Introduction to Riemannian manifolds*. Springer, 2018. Cited on pages 58, 252, 253, 256, 261, 267.

J. M. Lee. *Introduction to Smooth Manifolds*. Springer, 2013. Cited on pages 15, 55, 94, 173, 184, 202–204, 252, 300, 302.

J. M. Lee. *Riemannian Manifolds: An Introduction to Curvature*, volume 176. Springer Science & Business Media, 2006. Cited on pages 15, 252, 292.

S.-g. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S. Yoon, and T.-Y. Liu. PriorGrad: Improving Conditional Denoising Diffusion Models with Data-Dependent Adaptive Prior. In *International Conference on Learning Representations*, 2022. Cited on page 56.

Y. T. Lee and S. S. Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1115–1121, 2018. Cited on pages 68, 88.

Y. T. Lee and S. S. Vempala. Geodesic Walks in Polytopes. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 927–940, Montreal Canada. ACM, June 2017. ISBN: 978-1-4503-4528-6. DOI: 10.1145/3055399.3055416. Cited on pages 68, 70, 71, 73, 78, 88, 286.

G. Leobacher and A. Steinicke. Existence, uniqueness and regularity of the projection onto differentiable manifolds. *Annals of Global Analysis and Geometry*, 60(3):559–587, 2021. Cited on page 254.

C. Léonard. Girsanov theory under a finite entropy condition. In *Séminaire de Probabilités XLIV*, pages 429–465. Springer, 2012. Cited on page 265.

C. Léonard. Some properties of path measures. *Séminaire de Probabilités XLVI*:207–230, 2014. Cited on page 35.

C. Léonard, S. Rœlly, J.-C. Zambrini, et al. Reciprocal processes: a measure-theoretical point of view. *Probability Surveys*, 11:237–269, 2014. Cited on page 263.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993. Cited on page 9.

P. Li. Large time behavior of the heat equation on complete manifolds with nonnegative Ricci curvature. *Annals of Mathematics*, 124(1):1–21, 1986. Cited on page 259.

J. H. Lim, N. B. Kovachki, R. Baptista, C. Beckham, K. Azizzadenesheli, J. Kossaifi, V. Voleti, J. Song, K. Kreis, J. Kautz, C. Pal, A. Vahdat, and A. Anandkumar. Score-based Diffusion Models in Function Space, 2023. Cited on page 101.

J. H. Lim, N. B. Kovachki, R. Baptista, C. Beckham, K. Azizzadenesheli, J. Kossaifi, V. Voleti, J. Song, K. Kreis, J. Kautz, C. Pal, A. Vahdat, and A. Anandkumar. Score-based Diffusion Models in Function Space, 2023. Cited on pages 104, 116.

P.-L. Lions and A.-S. Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on pure and applied Mathematics*, 37(4):511–537, 1984. Cited on pages 73, 288, 289.

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023. Cited on page 125.

R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes. I*, volume 5 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, expanded edition, 2001, pages xvi+427. ISBN: 3-540-63929-2. General theory, Translated from the 1974 Russian original by A. B. Aries, Stochastic Modelling and Applied Probability. Cited on page 266.

G.-H. Liu, T. Chen, E. Theodorou, and M. Tao. Mirror diffusion models for constrained and watermarked generation. *Advances in Neural Information Processing Systems*, 36, 2024. Cited on pages 89, 91, 124.

Q. Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022. Cited on page 125.

S. Liu, T. Kanamori, and D. J. Williams. Estimating density models with truncation boundaries using score matching. *Journal of Machine Learning Research*, 23(186):1–38, 2022. Cited on pages 77, 317.

X. Liu, L. Wu, M. Ye, et al. Let us Build Bridges: Understanding and Extending Diffusion Generative Models. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022. Cited on page 125.

Y. Liu. Discretization of a class of reflected diffusion processes. *Mathematics and computers in simulation*, 38(1-3):103–108, 1995. Cited on page 94.

Y. Liu. *Numerical approaches to stochastic differential equations with boundary conditions*. Purdue University, 1993. Cited on page 94.

J. Loper. Uniform Ergodicity for Brownian Motion in a Bounded Convex Set. *Journal of Theoretical Probability*, 33(1):22–35, 2020. Cited on page 74.

A. Lou and S. Ermon. Reflected diffusion models. In *International Conference on Machine Learning*, pages 22675–22701. PMLR, 2023. Cited on pages 90, 91, 124.

A. Lou, M. Xu, A. Farris, and S. Ermon. Scaling Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 36:80291–80305, 2023. Cited on page 124.

G. Louppe, J. Hermans, and K. Cranmer. Adversarial variational optimization of non-differentiable simulators. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1438–1447. PMLR, 2019. Cited on page 24.

L. Lovász and S. Vempala. Hit-and-Run from a Corner. *SIAM Journal on Computing*, 35(4):985–1005, 2006. DOI: 10.1137/S009753970544727X. Cited on pages 68, 78.

L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007. Cited on page 95.

A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. Cited on pages 112, 330–333.

Y. M. Lui. Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6-7):380–388, 2012. Cited on page 40.

J. M. Lukens, K. J. Law, A. Jasra, and P. Lougovski. A practical and efficient approach for Bayesian quantum state estimation. *New Journal of Physics*, 22(6):063038, 2020. Cited on pages 68, 69, 84.

S. Luo, Y. Su, X. Peng, S. Wang, J. Peng, and J. Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022. Cited on page 123.

C. Lyle, M. van der Wilk, M. Kwiatkowska, Y. Gal, and B. Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020. Cited on pages 13, 102.

I. Macêdo and R. Castro. *Learning divergence-free and curl-free vector fields with matrix-valued kernels*. Pré-Publicações / A.: Pré-publicações. IMPA, 2010. Cited on pages 107, 115, 328.

V. Maiorov and A. Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1-3):81–91, 1999. Cited on page 9.

D. Maoutsa, S. Reich, and M. Opper. Interacting particle solutions of Fokker–Planck equations through gradient–log–density estimation. *Entropy*, 22(8):802, 2020. Cited on page 36.

K. V. Mardia, C. C. Taylor, and G. K. Subramaniam. Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, 63(2):505–512, 2007. Cited on page 37.

E. Mathieu, C. L. Lan, C. J. Maddison, R. Tomioka, and Y. W. Teh. Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders. *arXiv preprint arXiv:1901.06033*, 2019. Cited on page 60.

E. Mathieu and M. Nickel. Riemannian Continuous Normalizing Flows. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc., 2020. Cited on pages 38, 40, 48, 61.

E. Mathieu*, V. Dutordoir*, M. J. Hutchinson*, V. De Bortoli, Y. W. Teh, and R. Turner. Geometric neural diffusion processes. In *Advances in Neural Information Processing Systems*, 2024.

A. Matthews, M. Arbel, D. J. Rezende, and A. Doucet. Continual repeated annealed flow transport Monte Carlo. In *International Conference on Machine Learning*, pages 15196–15219. PMLR, 2022. Cited on page 125.

W. McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010. Cited on page 335.

T. A. Mellan, H. H. Hoeltgebaum, S. Mishra, C. Whittaker, R. P. Schnekenberg, A. Gandy, H. J. T. Unwin, M. A. Vollmer, H. Coupland, I. Hawryluk, et al. Subnational analysis of the COVID-19 epidemic in Brazil, 2020.

A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. Cited on page 3.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953. Cited on page 125.

S. P. Meyn and R. L. Tweedie. Stability of Markovian processes II: Continuous-time processes and sampled chains. *Advances in Applied Probability*, 25(3):487–517, 1993. Cited on page 287.

L. I. Midgley, V. Stimper, G. N. Simm, B. Schölkopf, and J. M. Hernández-Lobato. Flow Annealed Importance Sampling Bootstrap. In *The Eleventh International Conference on Learning Representations*, 2023. Cited on page 125.

A. Millet, D. Nualart, and M. Sanz. Integration by parts and time reversal for diffusion processes. *The Annals of Probability*:208–238, 1989. Cited on page 248.

N. Miolane, N. Guigui, A. L. Brigant, J. Mathe, B. Hou, Y. Thanwerdas, S. Heyder, O. Peltre, N. Koep, H. Zaatiti, H. Hajri, Y. Cabanes, T. Gerald, P. Chauchat, C. Shewmake, D. Brooks, B. Kainz, C. Donnat, S. Holmes, and X. Pennec. Geomstats: A Python Package for Riemannian Geometry in Machine Learning. *Journal of Machine Learning Research*, 21(223):1–9, 2020. Cited on pages 279, 294.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. Cited on page 1.

M. Monod, A. Blenkinsop, X. Xi, D. Hebert, S. Bershan, V. C. Bradley, Y. Chen, H. Coupland, S. Filippi, J. Ish-Horowicz, and others (14th/29). Age groups that sustain resurging COVID-19 epidemics in the United States. *Science*, 2021.

F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5115–5124, 2017. Cited on page 14.

B. J. Morris. Improved bounds for sampling contingency tables. *Random Structures & Algorithms*, 21(2):135–146, 2002. Cited on pages 68, 84.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997. Cited on page 25.

J. Nash. The imbedding problem for Riemannian manifolds. *Annals of mathematics*, 63(1):20–63, 1956. Cited on page 23.

Natural Earth. Natural Earth, 2023. URL: HTTPS://WWW.NATURALEARTHDATA.COM/. Cited on pages 98, 315.

R. M. Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001. Cited on page 125.

R. M. Neal. MCMC Using Hamiltonian Dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011. Cited on page 125.

E. Nelson. *Dynamical Theories of Brownian Motion*. en. Princeton University Press, Feb. 1967. Cited on pages 31, 248.

Y. Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018. Cited on page 70.

Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994. Cited on pages 70, 89.

A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. Cited on pages 30, 48.

M. Noble, V. De Bortoli, and A. Durmus. Unbiased constrained sampling with self-concordant barrier Hamiltonian Monte Carlo. *Advances in Neural Information Processing Systems*, 36:32672–32719, 2023. Cited on pages 68, 73, 78, 88.

F. Noé, S. Olsson, J. Köhler, and H. Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019. Cited on pages 25, 125.

D. Nualart. *The Malliavin calculus and related topics*, volume 1995. Springer, 2006. Cited on page 287.

M. Oechsle, L. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019. Cited on page 24.

M. O. of the United Kingdom. *Cartopy: A Cartographic Python Library with a Matplotlib Interface*. Exeter, Devon, 2015. URL: HTTPS://SCITOOLS.ORG.UK/CARTOPY. Cited on pages 98, 315.

B. Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003. Cited on pages 35, 106, 236, 237, 240, 244.

OpenAI et al. GPT-4 Technical Report, 2024. arXiv: 2303.08774 [cs.CL]. URL: HTTPS://ARXIV.ORG/ABS/2303.08774. Cited on pages 1, 24.

B. Pacchiarotti, C. Costantini, and F. Sartoretto. Numerical approximation for functionals of reflecting diffusion processes. *SIAM Journal on Applied Mathematics*, 58(1):73–102, 1998. Cited on pages 75, 94.

M. Paganini, L. de Oliveira, and B. Nachman. Accelerating science with generative adversarial networks: an application to 3D particle showers in multilayer calorimeters. *Physical review letters*, 120(4):042003, 2018. Cited on page 24.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019. Cited on page 328.

E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962. Cited on page 24.

D. Peel, W. J. Whiten, and G. J. McLachlan. Fitting mixtures of Kent distributions to aid in joint set identification. *Journal of the American Statistical Association*, 96(453):56–63, 2001. Cited on pages 37, 40, 60, 61, 280.

S. Peluchetti. Non-denoising forward-time diffusions. *arXiv preprint arXiv:2312.14589*, 2023. Cited on page 125.

X. Pennec. Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006. Cited on page 43.

F. Petit. Time Reversal and Reflected Diffusions. *Stochastic Processes and their Applications*, 69(1):25–53, July 1997. ISSN: 03044149. DOI: 10.1016/S0304-4149(97)00035-5. Cited on pages 68, 74, 78, 88, 292.

R. Pettersson. Approximations for stochastic differential equations with reflecting convex boundaries. *Stochastic processes and their applications*, 59(2):295–308, 1995. Cited on page 94.

R. Pettersson. Penalization schemes for reflecting stochastic differential equations. *Bernoulli*:403–414, 1997. Cited on page 94.

D. Pfau, S. Axelrod, H. Sutterud, I. von Glehn, and J. S. Spencer. Accurate computation of quantum excited states with neural networks. *Science*, 385(6711):eadn0137, 2024. Cited on page 24.

D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Res.*, 2:033429, 3, 2020. Cited on page 24.

A. Phillips, H.-D. Dau, M. J. Hutchinson, V. De Bortoli, G. Deligiannidis, and A. Doucet. Particle Denoising Diffusion Sampler. In *Forty-first International Conference on Machine Learning*, 2024. Cited on page 126.

A. Phillips, T. Seror, M. J. Hutchinson, V. De Bortoli, A. Doucet, and E. Mathieu. Spectral Diffusion Processes. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022. Cited on pages 104–106, 110, 112, 116, 323.

A. Phillips*, T. Seror*, M. J. Hutchinson, V. De Bortoli, A. Doucet, and E. Mathieu. Spectral Diffusion Processes. In *Advances in Neural Information Processing Systems*, 2022.

J. Pidstrigach, Y. Marzouk, S. Reich, and S. Wang. Infinite-Dimensional Diffusion Models for Function Spaces, Feb. 2023. URL: HTTP://ARXIV.ORG/ABS/2302.10130. Cited on pages 101, 104, 116.

A. Pilipenko. *An introduction to stochastic differential equations with reflection*, volume 1. Universitätsverlag Potsdam, 2014. Cited on page 94.

S. Prokudin, P. Gehler, and S. Nowozin. Deep Directional Statistics: Pose Estimation with Uncertainty Quantification. In *European Conference on Computer Vision (ECCV)*, Oct. 2018. Cited on page 282.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. Cited on page 102.

K. Ramanan. Reflected diffusions defined via the extended Skorokhod map, 2006. Cited on pages 300, 313.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-Shot Text-to-Image Generation. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021. URL: HTTPS://PROCEEDINGS.MLR.PRESS/V139/RAMESH21A.HTML. Cited on page 24.

C. E. Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003. Cited on pages 102, 105, 115.

A. Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961. Cited on page 25.

D. Revuz and M. Yor. *Continuous Martingales and Brownian Motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999, pages xiv+602. ISBN: 3-540-64325-7. DOI: 10.1007/978-3-662-06400-9. URL: HTTPS://DOI.ORG/10.1007/978-3-662-06400-9. Cited on page 254.

D. Revuz and M. Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013. Cited on pages 73, 293, 294.

D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538. PMLR, 2015. Cited on pages 2, 26, 38.

D. J. Rezende and S. Racanière. Implicit riemannian concave potential maps. *arXiv preprint arXiv:2110.01288*, 2021. Cited on page 38.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014. Cited on pages 2, 25, 38.

D. J. Rezende, G. Papamakarios, S. Racanière, M. S. Albergo, G. Kanwar, P. E. Shanahan, and K. Cranmer. Normalizing flows on tori and spheres. *arXiv preprint arXiv:2002.02428*, 2020. Cited on page 38.

S. Rissanen, M. Heinonen, and A. Solin. Generative Modelling with Inverse Heat Dissipation. In *The Eleventh International Conference on Learning Representations*, 2023. Cited on page 115.

G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*:341–363, 1996. Cited on pages 42, 246, 335.

M. Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956. URL: HTTPS://DOI.ORG/10.1214/AOMS/1177728190. Cited on page 24.

P. Rotkiewicz and J. Skolnick. Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of computational chemistry*, 29(9):1460–1465, 2008. Cited on page 83.

D. M. Roy, C. Kemp, V. Mansinghka, and J. B Tenenbaum. Learning annotated hierarchies from relational data, 2007. Cited on page 40.

N. Rozen, A. Grover, M. Nickel, and Y. Lipman. Moser Flow: Divergence-based Generative Modeling on Manifolds. *Advances in Neural Information Processing Systems*, 2021. Cited on pages 24, 38, 40, 60–62, 64, 271, 282.

P. Rudiger, J. A. Bednar, J.-L. Stevens, M. Liquet, C. B, S. H. Hansen, B. Little, Andrew, J. Signell, M. Hedley, C. Bosley, R. Hattersley, G. Brener, ea42gh, kbowen, A. Hilboll, I. Lustig, J. deWerd, N. Hand, R. Signell, J. Bampton, pmav99, scaine1, and zassa. holoviz/geoviews: Version 1.9.6, version v1.9.6, Jan. 2023. DOI: 10.5281/ZENODO.7543863. URL: HTTPS://DOI.ORG/10.5281/ZENODO.7543863. Cited on page 315.

W. Rudin. *Real and Complex Analysis*. McGraw-Hill, Inc., 3rd edition, 1987. Cited on page 213.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986. *Biometrika*, 71:599–607, 1986. Cited on pages 3, 15, 25.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. Cited on page 1.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35, 2022. Cited on pages 1, 24.

C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022. Cited on page 115.

L. Saloff-Coste. Precise estimates on the rate at which certain diffusions tend to equilibrium. *Mathematische Zeitschrift*, 217(1):641–677, 1994. Cited on pages 58, 59, 74, 221.

F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017. Cited on page 65.

S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 1st edition, 2019. DOI: 10.1017/9781108186735. Cited on page 324.

V. G. Satorras, E. Hoogeboom, and M. Welling. E(n) Equivariant Graph Neural Networks, 2021. arXiv: 2102.09844 [cs.LG]. Cited on pages 14, 108.

F. Schuller. Lectures on the Geometric Anatomy of Theoretical Physics, 2013. URL: HTTPS://MATHSWITHPHYSICS.BLOGSPOT.COM/2016/07/LECTURES-ON-GEOMETRIC-ANATOMY-OF.HTML. Cited on pages 15, 192.

R. Senanayake and F. Ramos. Directional grid maps: modeling multimodal angular uncertainty in dynamic environments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3241–3248. IEEE, 2018. Cited on page 40.

J. Serre. *Linear Representations of Finite Groups*. Collection Méthodes. Mathématiques. Springer-Verlag, 1977. ISBN: 9783540901907. Cited on page 12.

M. V. Shapovalov and R. L. Dunbrack Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011. Cited on page 40.

Y. Shi, V. De Bortoli, A. Campbell, and A. Doucet. Diffusion Schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36, 2024. Cited on page 125.

H. Shima and K. Yagi. Geometry of Hessian manifolds. *Differential geometry and its applications*, 7(3):277–290, 1997. Cited on page 71.

M. Shkolnikov and I. Karatzas. Time-Reversal of Reflected Brownian Motions in the Orthant, July 2013. URL: HTTP://ARXIV.ORG/ABS/1307.4422. Cited on pages 68, 78, 88.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016. Cited on page 1.

A. Sinha, J. Song, C. Meng, and S. Ermon. D2C: Diffusion-Denoising Models for Few-shot Conditional Generation. *arXiv preprint arXiv:2106.06819*, 2021. Cited on page 56.

A. V. Skorokhod. Stochastic equations for diffusion processes in a bounded region. *Theory of Probability & Its Applications*, 6(3):264–274, 1961. Cited on pages 73, 292.

L. Słomiński. On approximation of solutions of multidimensional SDE's with reflecting boundary conditions. *Stochastic processes and their Applications*, 50(2):197–219, 1994. Cited on pages 90, 94.

R. L. Smith. Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions. *Operations Research*, 32(6):1296–1308, Dec. 1984. DOI: 10.1287/opre.32.6.1296. URL: https://ideas.repec.org/a/inm/oropre/v32y1984i6p1296-1308.html. Cited on pages 68, 78.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*. PMLR, 2015. Cited on pages 3, 26, 38.

Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021. Cited on pages 35, 124.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019. Cited on pages 3, 26, 37, 39, 101, 112.

Y. Song and S. Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, 2020. Cited on page 48.

Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, 2020. Cited on page 32.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2020. Cited on pages 26, 33, 36, 38, 39, 44, 48, 50, 51, 55, 101, 106, 112, 116, 248, 279, 294, 327.

A. Sperduti. Encoding labeled graphs by labeling RAAM. *Advances in Neural Information Processing Systems*, 6, 1993. Cited on page 3.

M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1):41–78, 2005. Cited on page 40.

R. S. Strichartz. Analysis of the Laplacian on the complete Riemannian manifold. *Journal of functional analysis*, 52(1):48–79, 1983. Cited on pages 22, 220.

D. W. Stroock and S. S. Varadhan. Diffusion processes with boundary conditions. *Communications on Pure and Applied Mathematics*, 24(2):147–225, 1971. Cited on pages 88, 91, 93, 300, 313, 314.

D. W. Stroock and S. S. Varadhan. *Multidimensional diffusion processes*. Springer, 2007. Cited on pages 333, 334.

Y. Sun, N. Flammarion, and M. Fazel. Escaping from saddle points on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 32, 2019. Cited on page 275.

D. Sustretov. Hessian metrics with distribution coefficients on a 2-sphere, 2022. arXiv: 2212.10640 [math.AG]. URL: HTTPS://ARXIV.ORG/ABS/2212.10640. Cited on page 72.

R. Sutton. The Bitter Lesson. 2019. URL: HTTP://WWW.INCOMPLETEIDEAS.NET/INCIDEAS/BITTERLESSON.HTML (visited on 06/15/2024). Cited on page 1.

E. G. Tabak and C. V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013. Cited on pages 2, 26, 38.

E. G. Tabak and E. Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010. Cited on pages 2, 26, 38.

J. Tae. Mirror Diffusion Models. *arXiv preprint arXiv:2308.06342*, 2023. Cited on pages 89, 124.

T. Tao. *An Introduction to Measure Theory*. American Mathematical Society, 2011. Cited on page 229.

T. Tao. *Analysis, Volume I*. Springer, 4th edition, 2022. Cited on page 162.

G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. Cited on page 24.

Y. W. Teh, A. Bhoopchand*, P. Diggle*, B. Elesedy*, B. He*, M. J. Hutchinson*, U. Paquet*, J. Read*, N. Tomasev*, and S. Zaidi*. Efficient Bayesian Inference of Instantaneous Re-production Numbers at Fine Spatial Scales, with an Application to Mapping and Nowcasting the Covid-19 Epidemic in British Local Authorities. *Joural of the Royal Statistical Society: Series A*, 2021.

A. Terenin. *Gaussian Processes and Statistical Decision-making in Non-Euclidean Spaces*. PhD thesis, Imperial College London, 2022. Cited on pages 102, 223.

I. Thiele, N. Swainston, R. Fleming, A. Hoppe, S. Sahoo, M. Aurich, H. Haraldsdottir, M. Mo, O. Rolfsson, M. Stobbe, S. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A. Chavali, P. Dobson, W. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novère, N. Malys, A. Mazein, J. Papin, N. Price, E. Selkov, M. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. Westerhoff, D. Kell, P. Mendes, and B. Palsson. A community-driven

global reconstruction of human metabolism. English. *Nature Biotechnology*, Mar. 2013. ISSN: 1087-0156. DOI: 10.1038/NBT.2488. Cited on pages 68, 84.

N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3D point clouds. *arXiv preprint arXiv:1802.08219*, May 2018. URL: HTTP://ARXIV.ORG/ABS/1802.08219. Cited on pages 119, 121, 338.

K. S. Thorne, C. W. Misner, and J. A. Wheeler. *Gravitation*. Freeman San Francisco, 2000. Cited on page 15.

J. Thornton, M. J. Hutchinson, E. Mathieu, V. De Bortoli, Y. W. Teh, and A. Doucet. Riemannian Diffusion Schrödinger Bridge. In 2022. Cited on page 125.

M. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR, Aug. 2009. URL: HTTPS://PROCEEDINGS.MLR.PRESS/V5/TITSIAS09A.HTML. Cited on page 102.

B. L. Trippe, J. Yim, D. Tischer, T. Broderick, D. Baker, R. Barzilay, and T. Jaakkola. Diffusion Probabilistic Modeling of Protein Backbones in 3D for the Motif-Scaffolding Problem, June 2022. DOI: 10.48550/ARXIV.2206.04119. Cited on pages 81, 112.

A. B. Tsybakov. Nonparametric estimators. *Introduction to Nonparametric Estimation*:1–76, 2009. Cited on page 9.

H. J. T. Unwin, S. Mishra, V. C. Bradley, A. Gandy, T. A. Mellan, H. Coupland, J. Ish-Horowicz, M. A. Vollmer, C. Whittaker, S. L. Filippi, et al. State-level tracking of COVID-19 in the United States. *Nature communications*, 2020.

J. Urain, N. Funk, G. Chalvatzaki, and J. Peters. SE (3)-DiffusionFields: Learning cost functions for joint grasp and motion optimization through diffusion. *arXiv preprint arXiv:2209.03855*, 2022. Cited on page 123.

H. Urakawa. Convergence rates to equilibrium of the heat kernels on compact Riemannian manifolds. *Indiana University Mathematics Journal*:259–288, 2006. Cited on page 259.

G. Van Rossum and F. L. Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995. Cited on page 335.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000. Cited on page 13.

G. Vardi and O. Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pages 4224–4258. PMLR, 2021. Cited on page 9.

F. Vargas, P. Thodoroff, A. Lamacraft, and N. Lawrence. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021. Cited on pages 125, 126.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. Cited on pages 3, 14, 338.

P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017. Cited on page 14.

P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. Cited on pages 32, 39.

T. Vincent, L. Risser, and P. Ciuciu. Spatially Adaptive Mixture Modeling for Analysis of fMRI Time Series. *IEEE Trans. Med. Imag.*, 29(4):1059–1074, Apr. 2010. ISSN: 0278-0062. DOI: 10.1109/TMI.2010.2042064. Cited on page 106.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019. Cited on page 1.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. DOI: 10.1038/s41592-019-0686-2. Cited on page 335.

V. Voleti, C. Pal, and A. M. Oberman. Score-Based Denoising Diffusion with Non-Isotropic Gaussian Noise Models, 2022. URL: HTTPS://OPENREVIEW.NET/FORUM?ID=IGC8CJKCB0Q. Cited on page 116.

M. A. Vollmer, S. Mishra, H. J. T. Unwin, A. Gandy, T. A. Mellan, V. Bradley, H. Zhu, H. Coupland, I. Hawryluk, M. J. Hutchinson, et al. Report 20: Using mobility to estimate the transmission intensity of COVID-19 in Italy: a subnational analysis with future scenarios. *MedRxiv*, 2020.

A. Waibel, T. Hanazawa, and G. Hinton. Phoneme Recognition Using Time-Delay Neural Networks. *IEEE transactions on acoustics, speech, and signal processing*, 31(3), 1989. Cited on pages 3, 15.

Y. Wang, L. Wang, Y. Shen, Y. Wang, H. Yuan, Y. Wu, and Q. Gu. Protein conformation generation via force-guided se (3) diffusion models. *arXiv preprint arXiv:2403.14088*, 2024. Cited on page 123.

J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. Cited on page 123.

J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, N. Hanikel, S. J. Pellock, A. Courbet, W. Sheffler, J. Wang, P. Venkatesh, I. Sappington, S. V. Torres, A. Lauko, V. D. Bortoli, E. Mathieu, R. Barzilay, T. S. Jaakkola, F. DiMaio, M. Baek, and D. Baker. Broadly Applicable and Accurate Protein Design by Integrating Structure Prediction Networks and Diffusion Generative Models, Dec. 2022. DOI: 10.1101/2022.12.09.519842. Cited on pages 81, 123.

M. Weiler, P. Forré, E. Verlinde, and M. Welling. *Equivariant and Coordinate Independent Convolutional Networks. A Gauge Field Theory of Neural Networks*. 2023. Cited on pages 24, 109, 119, 329, 338.

M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. S. Cohen. 3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data. In *Advances in Neural Information Processing Systems*, volume 31, 2018. Cited on page 338.

E. Weinan, T. Li, and E. Vanden-Eijnden. *Applied Stochastic Analysis*. en. American Mathematical Soc., Sept. 2021. Cited on page 242.

J. Welty and M. Jeffries. Combined wildfire datasets for the United States and certain territories, 1878-2019: US Geological Survey data release, 2020. Cited on pages 98, 315.

L. Wen, X. Tang, M. Ouyang, X. Shen, J. Yang, D. Zhu, M. Chen, and X. Wei. Hyperbolic Graph Diffusion Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38 of number 14, pages 15823–15831, 2024. Cited on page 123.

L. Wen and X. Wei. Hyperbolic Graph Diffusion Model for Molecule Generation. *arXiv preprint arXiv:2306.07618*, 2023. Cited on page 123.

H. Whitney. Differentiable manifolds. *Annals of Mathematics*:645–680, 1936. Cited on page 186.

M. Wiatrak, S. V. Albrecht, and A. Nystrom. Stabilizing generative adversarial networks: A survey. *arXiv preprint arXiv:1910.00927*, 2019. Cited on page 25.

C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2 of number 3. MIT press Cambridge, MA, 2006. Cited on page 1.

R. J. Williams. Reflected Brownian Motion with Skew Symmetric Data in a Polyhedral Domain. *Probability Theory and Related Fields*, 75(4):459–485, Aug. 1987. ISSN: 0178-8051, 1432-2064. DOI: 10.1007/BF00320328. Cited on pages 68, 78, 88.

J. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. Deisenroth. Efficiently sampling functions from Gaussian process posteriors. In *ICML*, pages 10292–10302. PMLR, 2020. Cited on page 102.

P. Wirnsberger, G. Papamakarios, B. Ibarz, S. Racaniere, A. J. Ballard, A. Pritzel, and C. Blundell. Normalizing flows for atomic solids. *Machine Learning: Science and Technology*, 3(2):025009, 2022. Cited on page 125.

F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR, 2019. Cited on page 14.

K. E. Wu, K. K. Yang, R. van den Berg, S. Alamdari, J. Y. Zou, A. X. Lu, and A. P. Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1):1059, 2024. Cited on page 81.

K. E. Wu, K. K. Yang, R. van den Berg, S. Alamdari, J. Y. Zou, A. X. Lu, and A. P. Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1):1059, 2024. Cited on page 123.

O. Yadan. Hydra - A framework for elegantly configuring complex applications. Github, 2019. URL: HTTPS://GITHUB.COM/FACEBOOKRESEARCH/HYDRA. Cited on page 335.

W. Yifan, L. Rahmann, and O. Sorkine-hornung. Geometry-Consistent Neural Shape Representation with Implicit Displacement Fields. In *International Conference on Learning Representations*, 2022. Cited on page 24.

J. Yim, B. L. Trippe, V. De Bortoli, E. Mathieu, A. Doucet, R. Barzilay, and T. Jaakkola. SE (3) diffusion model with application to protein backbone generation. In *International Conference on Machine Learning*, pages 40001–40039. PMLR, 2023. Cited on pages 107, 123, 329.

T. Yoshikawa. Manipulability of robotic mechanisms. *The international journal of Robotics Research*, 4(2):3–9, 1985. Cited on pages 78, 79.

M. Zaheer, S. Kottur, S. Ravanbhakhsh, B. Póczos, R. Salakhutdinov, and A. J. Smola. Deep Sets. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 3394–3404, Long Beach, California, USA. Curran Associates Inc., 2017. ISBN: 9781510860964. Cited on page 14.

J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D.-Y. Yeung. Gaan: Gated attention networks for learning on large and spatiotemporal graphs. *arXiv preprint arXiv:1803.07294*, 2018. Cited on page 14.

J. Zhang, G. Zhu, R. W. Heath Jr, and K. Huang. Grassmannian learning: Embedding geometry awareness in shallow and deep learning. *arXiv preprint arXiv:1808.02229*, 2018. Cited on page 24.

Q. Zhang and Y. Chen. Path Integral Sampler: A Stochastic Control Approach For Sampling. In *International Conference on Learning Representations*, 2022. Cited on page 126.

S. Zheng, J. He, C. Liu, Y. Shi, Z. Lu, W. Feng, F. Ju, J. Wang, J. Zhu, Y. Min, et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*:1–10, 2024. Cited on pages 24, 126.

A | Manifolds

In this chapter, we provide background material on manifolds. Manifolds are a natural generalisation of Euclidean space that allow us to describe spaces that are not flat in the usual sense. Many types of real-world data naturally exist on non-flat manifolds. Incorporating the geometry of the manifolds that data lives on provides a more accurate and insightful representation of the underlying data, and allows machine learning algorithms to perform better.

First appendix A.1 contains an introduction to topological spaces and contains definitions and concepts useful for the rest of this section. Appendix A.2 introduces the concept of topological manifolds, abstract spaces that locally resemble Euclidean space. This concept allows us to generalize the familiar notions of continuity and smoothness to curved spaces.

Appendix A.3 explores smooth manifolds, which equip topological manifolds with a differentiable structure, paving the way for the application of calculus and differential geometry. After this we explore this differential structure. Appendix A.4.3 reviews linear algebra and tensors, in preparation for the next section. Appendix A.5 introduces tangent, cotangent, and tensor spaces at each point on the manifold, key objects in the study of differentiation on manifolds at a point. Appendix A.6 introduces tangent, cotangent, and tensor bundles, structures that attach elements of tangent, cotangent, and tensor spaces to each point on the manifold. Appendix A.7, explores Lie groups, symmetry groups that are also themselves manifolds.

Finally, in appendix A.8, we delve into Riemannian manifolds and metrics, which introduce a notion of distance and curvature, enabling us to explore the intrinsic geometry of the underlying data. Appendix A.9 studies how to do integration on manifolds. We build up the notions of differential forms, the types of objects that we can integrate on manifolds, and orientations. Appendix A.10 introduces connections, a way of differentiating tensors, how they lead to geodesics, straight paths on manifolds, and parallel transport, a method for comparing tangent vectors at different points on a manifold.

A.1. Topological Spaces

Intuitively, the study of topology is the study of spaces that we identify as the same if we can "continuously" deform one into another - as if the space is made of rubber that can be stretched and deformed. Another way to think about it is the weakening of the structure of a *metric space* that allows us to extend certain

Metric space

tools to new spaces. It allows us to discuss whether points are "near" each other without needing to being able to say how far apart exactly they are.

Metric space

DEFINITION A.1. *A metric space (X, d) is a space X with a two argument function $d : X \times X \to \mathbb{R}_{\geq 0}$ such that d satisfies:*

1. *$d(x, x) = 0$ for $x \in X$.*
2. *If $x \neq y$ for $x, y \in X$ then $d(x, y) > 0$ (Positivity)*
3. *$d(x, y) = d(y, x)$ for $x, y \in X$ (Symmetry)*
4. *$d(x, z) \leq d(x, y) + d(y, z)$ for $x, y, z \in X$ (Triangle inequality)*

Distance function

This encodes in its structure our intuitive notion of a *distance function* in d between points in the space X. On top of this we can create two notions.

Open sets

First is that of *open sets*.

One motivation behind the modern definition of a topology comes from the "open set criterion", the observation that continuous functions between metric spaces (defined in the (ϵ, δ) sense, Tao (2022, chapter 9)) can be defined by knowing only the open sets of the metric spaces and throwing out the metrics that generated the open sets in the first place.

Topology
Open sets

DEFINITION A.2. *A topological space $(X, \mathcal{O}_X)$ is a set X along with a collection of subsets $\mathcal{O}_X$ of X, called the* OPEN SETS *of X that satisfy the following:*

- *X and $\varnothing$ are elements of $\mathcal{O}_X$.*
- *$\mathcal{O}_X$ is closed under* FINITE *intersections.*
- *$\mathcal{O}_X$ is closed under* ARBITRARY *unions.*

Neighbourhood

A point in a topological space can be said to be "near" another is it is in a *neighbourhood* of the point. A neighbourhood of $x \in X$ is simply an open set containing x. The neighbourhood of a set $S \subset X$ is an open set containing S.

Closed sets

This definition of a topology takes open sets as the primary element. Relatedly of importance are *closed sets*. A set in a topology is closed if its compliment in X, $X \backslash S$ is open. Note despite the name, these properties are not opposite. Sets can be neither, just closed, just open, or both. In all topologies X and $\varnothing$ are both open and closed. A topology can just as well be defined by its closed sets, but by convention we specify the open ones.

On any given space we can define many topologies, although only some are useful. We say that a topology is *coarser* than another if it is a subset of the other topology. Conversely, a topology that contains another topology is said to be *finer* than that topology.

Interior

The *interior* of a set S, $\operatorname{Int} S$ is the union of all open sets in $\mathcal{O}_X$ contained in S. The interior of any set is open.

Exterior

The *exterior* of a set S, $\operatorname{Ext} S$, is the union of all open sets in $\mathcal{O}_X$ contained in $X \backslash S$. The exterior of any set is open.

Closure

The *closure* of a set S, $\overline{S}$, is the intersection of all closed set in $\mathcal{O}_X$ containing S. The closure of any set is closed.

The *boundary* of a set S, ∂S, is the set of all points not in Int S or Ext S. The boundary of any set is a closed set. Boundary

A point $p \in S$ is said to be *isolated* in S if p has a neighbourhood in $\mathcal{O}_X$ such that $U \cap S = \{p\}$. Isolated

A point $p \in X$ (not S) is said to be a *limit point* of S if every neighbourhood of p contains a point in S that is not p. Topologically closed sets contain all such limit points within themselves. Limit point

A set S is said to be *dense* in another set T if $\overline{S} = T$. A classic example is that the rational numbers $\mathbb{Q}$ are dense in the reals $\mathbb{R}$. Dense

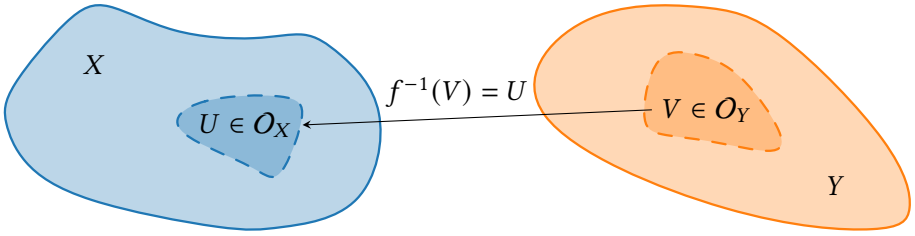
Figure A.1. From left to right: S, Int S, Ext S, $\overline{S}$, ∂S.

A topological space is *disconnected* if there exists two disjoint open sets whose union is the whole space. The converse is a *connected* space. Disconnected Connected

A.1.1. Continuity & Convergence. The two most important notions in topology are those of *continuous functions* and *sequential convergence*.

Continuous functions between two topological spaces preserve in one direction only the structure of the topology of the two spaces.

DEFINITION A.3. *A function $f : X \to Y$ between two topological spaces $(X, \mathcal{O}_X)$ and $(Y, \mathcal{O}_Y)$ is* CONTINUOUS *if for every open set $V \in \mathcal{O}_Y$, its preimage is open in X, $f^{-1}(V) \in \mathcal{O}_X$.* Topological continuity



Compositions of continuous functions are continuous, and so are their restrictions to open sets. Continuity can be proved by proving continuity just for a basis on $\mathcal{O}_Y$.

If a bijective function and its inverse are both continuous, then that function is said to be a *homeomorphism*, as it preserve the topological structure of the two spaces. If there exists a homeomorphism between X and Y then we say that X and Y are *homeomorphic*, denoted $X \cong_{\text{top}} Y$ (isomorphic in the topological sense). Homeomorphism Homeomorphic

A weaker but important notion is that of a *local homeomorphism*. A function f is a local homeomorphism if for every point $x \in X$ has a neighbourhood U such that the restriction of f to U, $f|_U$, is a homeomorphism from U to $f(U)$. If there Local homeomorphism

exists a local homeomorphism between X and Y then we say they are *locally homeomorphic*.

Locally homeomorphic

Convergence in the topological sense implies a sequence becomes arbitrarily close to a point in the sense of neighbourhoods of the point. The point $x \in X$ is said to be the *limit of a sequence* $\{x_i\}_{i=0}^{\infty}$ if for every neighbourhood U of x, there is and $N_U \in \mathbb{N}$ such that for all $i > N_U$, $x_i \in U$. Alternatively we might say that $\{x_i\}_{i=0}^{\infty}$ *converges* to x, denoted $\{x_i\}_{i=0}^{\infty} x$ or even $x_i \to x$ for short. A set S is said

Limit of a sequence

Converges

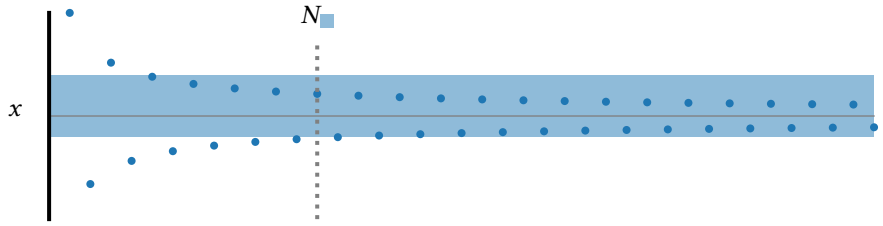


Figure A.2. The sequence $\{\bullet_i\}$ converges to the point x. We see that on the neighbourhood ■, for $i > N_{■}$, $\bullet_i \in$ ■

to be *sequentially closed* if it contains the limits of all sequences in S. This notion is *different* to that of *topologically closed* introduced earlier, and is different to the notion of containing all of its *topological limit points*. Topological closure implies sequential closure, but not the converse.

Sequentially closed

A useful combination of these notions is the convergence of sequences under continuous maps. If a sequence $x_i \to x$, and we have $f : X \to Y$ continuous, then $f(x_i) \to f(x)$.

A.1.2. Goldilocks topologies. The definition of a topology along with the notions we have introduced can sometimes lead to paradoxical results. On one end if we endow X with the topology with the least sets, the *trivial topology*, $\mathcal{O}_X = \{X, \varnothing\}$, then every sequence converges to every point in X. By contrast if we endow X with the *discrete topology*, the topology that defines *every* subset of X as open, then it is *totally disconnected*, i.e. that every point is completely separated from every other point. In order to avoid these pathologies in "nice" spaces, we introduce a some more notions so that our topologies have the right amount of open sets to be useful to us.

Trivial topology

Discrete topology

A *Hausdorff space* ensures that we have *enough* open sets in our topology. A space is said to be Hausdorff if for any two points $x_1, x_2 \in X$, we can find two open sets U_1, U_2 such that $x_1 \in U_1$, $x_2 \in U_2$, but $U_1 \cap U_2 = \varnothing$. Intuitively this says if I have any two points, I can always separate them into two distinct neighbourhoods. Being a Hausdorff space implies two very useful properties:

Hausdorff space

1. Every 1-point set $\{x\}$ is closed.
2. If a sequence $\{x_i\}$ in X converges to a limit x, then this limit is *unique*.

On a finite set, only the discrete topology is Hausdorff. On infinite sets, many might be. The Hausdorff property is not the only possible "separability" criteria, it exists in a scale of strictness, but it is the most useful to us.

Countability

Secondly we have the notion of *countability*. This ensures we don't have *too many* open sets in our topology. To introduce this, we first introduce the notion of a *basis* of a topology.

Basis

A topology can often be constructed by specifying some special set of open sets, and "completing" this in a suitable manner. A *basis* $\mathcal{B}$ of X is a collection of subsets such that their union is X, and for all $B_1, B_2 \in \mathcal{B}$, there is a $B_3 \in \mathcal{B}$ such that $B_3 \subset B_1 \cap B_2$.

A basis can be completed into a topology by adding to it all possible unions of the sets in the basis. This topology is said to be *generated* by the basis. Bases are useful as they give us something *simpler* from which we can reconstruct the whole topology.

Neighbourhood basis

In some sense opposite, we have the notion of a *neighbourhood basis*. If we consider *all* the neighbourhoods of a point or set, $\mathcal{N}(x)$, then a neighbourhood basis $\mathcal{B}$ is a subset of $cN(x)$ such that for any neighbourhood $U \in \mathcal{N}(x)$ there is an element in the neighbourhood basis $V \in \mathcal{B}$ such that $V \subset U$. This says that for any neighbourhood we can find a set in the neighbourhood basis that is contained within it. This is useful as it allows us to study properties only "very close" to the point, forgetting about the rest of the topology.

In the discrete topology, the set $\{\{x\} : x \in X\}$ is a basis and the set $\{x\}$ is a neighbourhood basis of x as all points are separated from each other.

First-countable

A space is *first-countable* if every point has a countable neighbourhood basis. First countability implications regarding *sequential closure* and *topological closure*. Previously we had that topological closure $\implies$ sequential closure, but topological closure $\not\Longleftarrow$ sequential closure. First countability results in topological closure $\iff$ sequential closure, i.e. the topological and sequential notions of closure become identical. Every metric space is first countable, but not necessarily the converse.

Second-countable

A space is *second-countable* if its topology has a countable basis. Second countability implies first countability, as it must contain a countable neighbourhood basis for each point.

Cover
Open cover
Subcover

This condition is often used to reduce the number of sets on needs to "cover" a space X. A collection of subsets $\mathcal{U}$ is called a *cover* of X if for every point in X, there is a $U \in \mathcal{U}$ such that $x \in U$. An *open cover* is a cover where all the subsets are open. A *subcover* is a subset of a cover that remains still a cover. In a second countable space, every open cover of the space has a *countable* subcover.

Compact

If we can further reduce every open cover to having a *finite* sub-cover, then we say the space is *compact*. This generalise the notion of a set being closed and bounded in metrics spaces.

A.1.3. Constructing topologies. Many import spaces can be constructed out of other spaces. When doing this we need to define the topology on these new spaces. 4 examples of particular importance are: *subspaces*, *products*, *disjoint unions* and *quotients*. We will choose canonical ways to define these new topologies that preserve sensible properties on the new spaces.

Subspaces

Subset topology

If we have $S \subseteq X$, then the *subset topology* is the topology $\mathcal{O}_S = \{U \cap S \ : \ U \in \mathcal{O}_X\} = S \cap \mathcal{O}_X$, i.e. the intersection of the topology of X with S. We define the *inclusion map* $\iota_S : S \hookrightarrow X$ as the restriction of the identity of X to S, $\mathrm{id}|_S$.

Inclusion map

Topological embedding

A continuous, *injective* map $f : X \to Y$ is called a *topological embedding* if it is a homeomorphism onto its image in Y, $f(X) \subseteq Y$, endowed with the subspace topology.

Characteristic property of the subset topology

This definition is motivated by the *characteristic property of the subset topology*: if Y is a topological space, $f : Y \to S$ is continuous if and only if the composition

$$Y \xrightarrow{f} S \xhookrightarrow{\iota_S} X \tag{A.1}$$

is continuous. The subset topology is the unique topology satisfying this property.

This induces a number of properties with respect to $\mathcal{O}_X$ and the subset topology on S, $S \cap \mathcal{O}_X$:

1. The inclusion map is a topological embedding.
2. The closed subsets of S are the intersection of the closed subsets of X with S.
3. If $f : X \to Y$ is continuous, then $f|_S : S \to Y$ is continuous.
4. If $\mathcal{B}$ is a basis of X, then $\mathcal{B}_S = \{B \cap S \ : \ B \in \mathcal{B}\}$ is a basis of S.
5. If we have $T \subset S \subset X$ then the topology $T \cap (S \cap \mathcal{O}_X)$ agrees with $T \cap \mathcal{O}_X$.
6. If X is Hausdorff, then S is Hausdorff.
7. If X is first countable, then S is first countable.
8. If X is second countable, then S is second countable.

Continuity is local

The converse of 3. allows us to show that *continuity is local*. If we have $f : X \to Y$, and for every point $x \in X$ we have a neighbourhood U such that $f|_U$ is continuous, then f is continuous.

Glue continues maps

Alternatively, we can *glue continues maps* together. If we have a) $B_1, ..., B_n$, a finite collection of closed sets or b) $\{B_i\}_{i \in A}$, any collection of open sets such that $\cup_{a \in A} B_a = X$, and a continuous function for each B, $f_i : B_i \to Y$, and these functions agree on intersections, $f_i(B_i \cap B_j) = f_j(B_i \cap B_j)$, then there is a unique continuous function $f : X \to Y$ such that $f|_{B_i} = f_i$. An important corollary of this is that to check continuity it is sufficient to check continuity on a basis of X.

Product spaces

Now consider the product of a finite collection of spaces $X_1, ..., X_n$. Define a basis by

$$\mathcal{B} = \left\{U_1 \times ... \times U_n \ : \ U_i \in \mathcal{O}_{X_i}, i = 1, .., n\right\}, \tag{A.2}$$

the Cartesian product of each combination of sets from each topology. The topology generated by this basis is the *product topology*.

Product topology

Characteristic property of the product topology

The *characteristic property of the product topology* is that, for any topological space Y, $f : B \to X_1 \times ... X_n$ is continuous if and only if the composition with each projection function $\pi_i : x_1, ..., x_n \mapsto x_i$, $f_i = \pi_i \circ f$ is continuous. The product topology is the unique topology for which this holds.

This induces a number of properties with respect to product topology on $X_1 \times ... \times X_n$:

1. Each projection map π_i is continuous.
2. For continuous functions $f_i : X_i \to Y_i$, the *product map* $f_1 \times ... \times f_n : X_1 \times .. \times X_n \to Y_1 \times ... \times Y_n$, $f(x_1, ..., x_n) = (x_1, ..., x_n)$ is continuous. Product map
3. If we have subspaces $S_i \subseteq X_i$, then the product topology $S_1 \cap O_{X_1} \times ... \times S_n \cap O_{X_n}$ coincides with the subspace topology $(S_1 \times ... \times S_n) \cap O_{X_1 \times ... \times X_n}$.
4. For any i, and picking $a_j \in X_j$, $i \neq j$, then the map $x \mapsto (a_1, ..., x, ..., a_n)$ it a topological embedding into the product space $X_1 \times ... \times X_n$.
5. If $\mathcal{B}_i$ is a basis for each X_i, then $\mathcal{B}\{B_1 \times ... \times B_n \ : \ B_i \in \mathcal{B}_i\}$ is a basis of $X_1 \times ... \times X_n$.
6. Every finite product of Hausdorff spaces is Hausdorff.
7. Every finite product of first countable spaces is first countable.
8. Every finite product of second countable spaces is second countable.

For products of an infinite collection of spaces, a piece of subtly creeps in. We want all the properties above to continue to hold. But, as it turns out the basis we defined earlier is not a good choice. On infinite spaces this is known as the *box topology*, where we take the products of all the open sets of each space. This topology fails in a number of ways. The box topology fails to preserve compactness through products. For example, functions we expect to be continuous fail to be continuous, and topological convergence becomes a significantly strict condition under the box topology. Box topology

Instead, we introduce the concept of *cylinder sets*. For collection of spaces, each with a collection of subsets, $(X_a, \mathcal{S}_a)_{a \in A}$, we take the Cartesian product of these to give $X = \prod_{a \in A}$. We get a canonical projection for each a, $\pi_a : X_a \to X$, that elements of X_a to its component in X. A *cylinder set* is then a set Cylinder sets

$$\bigcap_{i=1}^{n} \pi_{a_i}^{-1}(S_{a_i}), \qquad S_{a_i} \in \mathcal{S}_{a_i}, a_i \in A. \tag{A.3}$$

Note this is a *finite* intersection. Alternatively we can see the cylinder sets as

$$\left\{U_1 \times U_2 \times ... \ : \ U_i \in O_{X_i}, i = 1, 2, ..., U_i \neq X_i \text{ for finitely many } i\right\}, \tag{A.4}$$

products of subsets of each X_a where only *finitely many* of the subsets are not the whole space X_a.

We then define the *product topology on infinite products* to be the one generated by the basis of cylinder sets on the product of $(X_a, O_{X_a})_{a \in A}$. This topology is the one compatible with the previously stated conditions, and is the default topology on infinite products. In the product topology, topological convergence coincides with the notion of *pointwise convergence*, i.e. a sequence of functions converge if and only if each of the component functions f_i converge. Product topology on infinite products. Pointwise convergence

On finite products, we can see that these two definitions coincide, and so we can use the cylinder sets to define the topology on finite products too.

Disjoint unions

An alternative way to join sets together from the Cartesian product is the *disjoint union*. For a collection of sets $\{X_a\}_{a \in A}$, the disjoint union is given by appending a Disjoint union

to each element of X_a.

$$\coprod_{a\in A} X_a = \{(x,a) \ : \ a \in A, x \in X_a\}. \tag{A.5}$$

Disjoint union topology

We can define a canonical inclusion map $\iota_a : X_a \to \coprod_{a\in A} X_a$ for each a by $\iota_a : x \mapsto (x,a)$. We define the open sets of the *disjoint union topology* to be the set that when intersected with each X_a, are open in $\mathcal{O}_{X_a}$. Note this is defined for infinite index sets A.

Characteristic property of the disjoint union topology

The *characteristic property of the disjoint union topology* is that for a topological space Y, a function $f : \coprod_{a\in A} X_a \to Y$ is continuous if and only if $f \circ \iota_a \to Y$ is continuous for each $a \in A$. The disjoint union topology is the unique topology satisfying this property.

This induces a number of properties with respect to disjoint union topology on $\coprod_{a\in A} X_a$:

1. A subset of $\coprod_{a\in A} X_a$ is closed if and only if its intersection with each X_a is closed.
2. Each injection $\iota_a : X_a \to \coprod_{a\in A} X_a$ is a topological embedding.
3. Every disjoint union of Hausdorff spaces is Hausdorff.
4. Every disjoint union of first countable spaces is first countable.
5. Every disjoint union of second countable spaces is second countable.

Quotient spaces

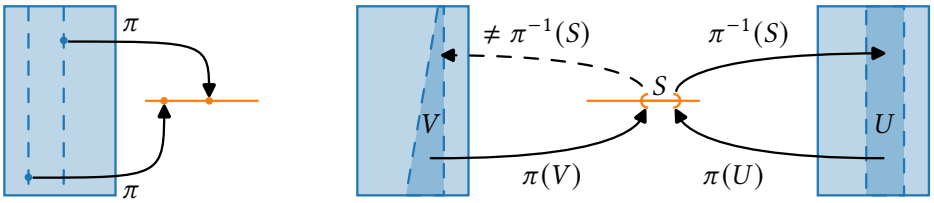


Figure A.3. *Left:* The quotient map π maps fibres of blue into points in orange. *Right:* The set U is a saturated set in blue, as it is the preimage of S. V is not.

Quotient map
Quotient topology

Finally, we discuss creating quotient spaces. If we have a topological space X, a set Y, and a continuous surjective map $\pi : X \to Y$, called a *quotient map*, then we define the open sets of the *quotient topology* $\mathcal{O}_Y$ to be the sets $U \subseteq Y$ such the preimage $\pi^{-1}(U)$ is open in $\mathcal{O}_X$. Denote this topology $\mathcal{O}_X \setminus \pi$.

Equivalence relation

The most common quotient spaces are induced by an *equivalence relation*, $\sim$. This is a truth operation on two elements on X such that it is reflexive ($x \sim x$ for all $x \in X$), transitive($x \sim y$ and $y \sim z \implies x \sim z$) , and symmetric ($x \sim y \iff y \sim x$). We

Equivalence class

denote the *equivalence class* of x, $[x]$, as all the elements y of X such that $x \sim y$. The set of all equivalence classes partitions X into disjoint sets. If we denote the set of all equivalence classes on X as $X\setminus\sim$, then there is a natural quotient map $\pi : X \to X\setminus\sim$, $\pi : x \mapsto [x]$.

Fibre
Saturated set

A *fibre* of a quotient map $\pi : X \to Y$ is the preimage of a point in Y, $\pi^{-1}(y) \subset X$. A *saturated set* is a union of fibres. Alternatively, it is a subset $U \subseteq X$ such that $U = \pi^{-1}(\pi(U))$.

The *characteristic property of the quotient topology* is that, for any quotient map $\pi : X \to Y$, for any topological space B, a map $f : Y \to B$ is continuous if and only if the composition $f \circ \pi : X \to B$ is continuous. The quotient topology is the unique topology satisfying this condition. Characteristic property of the quotient topology

This induces a number of properties with respect to quotient topology on Y:

1. $K \subseteq Y$ is closed if and only if $\pi^{-1}(K)$ is closed in X.
2. If π is injective (and therefore bijective), it is a homeomorphism.
3. If $U \subseteq X$ is a saturated open or closed set, then $\pi|_U : U \to \pi(U)$ is a quotient map.
4. Any composition of quotient maps is a quotient map.

Importantly, quotient topologies do not necessarily play nicely with products, subspaces or disjoint unions, and do not necessarily preserve the properties of Hausdorff, first or second countability, these have to be checked.

Quotients have two important properties that will be important later.

Passing to the quotient. If $\pi : X \to Y$ is a quotient map, B a topological space, and we have a continuous function $f : X \to B$ such that it is *constant* on the fibres of π, i.e. that $\pi(p) = \pi(q) \implies f(p) = f(q)$, then there is a unique continuous map $\tilde{f} : Y \to B$ such that $f = \tilde{f} \circ \pi$. Passing to the quotient

Homeomorphic quotient spaces can be detected by analysing their quotient maps. If for two quotient maps $\pi_1 : X \to Y_1$ and $\pi_2 : X \to Y_2$ we have that $\pi_1(p) = \pi_1(q) \iff \pi_2(q) = \pi_2(q)$, then Y_1 and Y_2 are homeomorphic, and there exists a homeomorphism $\phi : Y_1 \to Y_2$ such that $\pi_2 = \phi \circ \pi_2$. Homeomorphic quotient spaces

A.2. Topological Manifolds

Definition A.4. *An n-dimension topological manifold is a Hausdorff and second-countable topological space for which every point has a neighbourhood homeomorphic to an open subset of $\mathbb{R}^n$.* Topological manifold

This definition tells us that these new curved spaces we are interested in should *locally* look like Euclidean space, but that they do not have to globally. The Hausdorff and second-countable conditions are conditions on the topology of the manifold that neatly prescribe the space to be "nice" enough, and prevent certain pathological spaces being classed as manifolds. It should be noted that some definitions of topological manifolds, second countability, and therefore the implication of paracompactness is omitted. Examples that fit this reduced definition include the long line, manifolds in general relativity near black holes, or various infinite dimension manifolds such as the manifold of smooth functions. We are not interested in any spaces like this, and so stick with keeping second countability in the definition, but other sources may differ.

This last condition is also stated as being *locally Euclidean of dimension n*. Note that the requirement of second countability is sometimes replaced with *paracompactness*. In locally Euclidean Hausdorff spaces, these two notions imply each other and so this is a moot point. Paracompactness

This means therefore a manifold X is in fact a tuple $(X, \mathcal{O}_X)$, containing a set along with a suitable topology.

Euclidean space of dimension n is clearly a topological manifold of dimension n, when endowed with the topology induced by the usual Euclidean metric.

We may also be interested in manifolds that have *edges*, aptly names manifolds with boundary.

Topological manifold with boundary

DEFINITION A.5. *An n-dimension TOPOLOGICAL MANIFOLD WITH BOUNDARY is a HAUSDORFF and SECOND-COUNTABLE topological space for which every point has a neighbourhood homeomorphic to an open subset of the n-dimension* $\mathbb{H}^n$*,*

Upper half-space

where $\mathbb{H}^n$ is the *upper half-space*, $\mathbb{H}^n = \{(x_1, ..., x_n) \in \mathbb{R}^n \ : \ x_n \geq 0\}$. We call any point that has a neighbourhood homeomorphic to the upper half-space a *boundary point*, and denote the collection of these the *boundary*, denoted ∂X. Points that are not boundary points are call *interior points*.

Boundary point
Interior points

Interior of a manifold with boundary
Boundary of a manifold with boundary

The *interior of a manifold with boundary*, $\operatorname{Int} X$, is an n-dimension topological manifold *without boundary*. The *boundary of a manifold with boundary*, ∂X, is an $n-1$-dimension topological manifold *without boundary*, and typically properties of ∂X are *induced* by corresponding properties on $\operatorname{Int} X$.

The required structure of a manifold (with boundary) implies additional properties: all topological manifolds (with boundary) are locally path-connected, locally compact, and paracompact.

Using the previously outlined ways of producing topological spaces from other topological spaces, it is clear we can create new manifolds automatically via taking subspaces of manifolds, taking products of manifolds, or by taking direct products of manifolds. We can also produce new manifolds via quotients, however we must check that this quotient preserves the Hausdorff and second countable properties manually.

A.2.1. Bundles and Sections. The operations just discussed can create many types of new manifold. However, there exist a very useful class of manifold that we can study that while *locally* look like the direct product of two manifolds, *globally* they do not. The prototypical example of this is the Möbius strip.

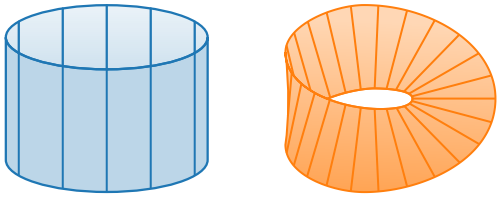


Figure A.4. The Both the Möbius strip and the cylinder look a bit like the direct product of the circle and the unit interval. The cylinder is exactly this, but the Möbius strip has an additional "twist" in it.

We can make this notion more precise with the study of *bundles*.

DEFINITION A.6. *A* BUNDLE *is a triple* (E, π, B) *consisting of a topological space* E *the* TOTAL SPACE*,* $p : E \to B$ *the* PROJECTION*, a continuous surjective map, and a topological space* B *the* BASE SPACE*.* Bundle

The *fibre* at a point $p \in B$, F_p, is given by the preimage of the point under the projection, $F_p = \pi^{-1}(p)$. In the other direction, the projection map will send all points in the fibre F_p to p. Fibre

Intuitively, we can think about this bundle as having "glued" all the fibres to each point on the manifold, and, in addition, we have specified thought E how fibres next to each other are to be glued together. This is glueing of the fibres to each other is how we procure the difference between the Möbius strip and the cylinder.

The simplest example of a bundle is the direct product, $(B \times F, \pi : (b, f) \mapsto b, B)$.

This is also the simplest example of a *fibre bundle*, which is a restriction of the idea of a bundle, where every fibre is homeomorphic to some canonical fibre.

DEFINITION A.7. *A* FIBRE BUNDLE *is a quadruple* (E, π, B, F) *where the first three arguments are a bundle such that* Fibre Bundle

$$\pi^{-1}(p) = F_p \underset{\text{top}}{\cong} F \; \forall \; p \in B \tag{A.6}$$

for an additional topological space F*.*

The Möbius strip and the cylinder are both examples of fibre bundles. The have the same base space, S^1, and canonical fibre, $[0, 1]$, but different total space and projection.

One concept that will be very important to us is the notion of a *section* of a bundle.

DEFINITION A.8. *A* SECTION *of a bundle* (E, π, B) *is a continuous map* $\sigma : B \to E$ *such that the composition of the section with the projection is the identity, i.e. that* Section

$$\pi \circ \sigma = \mathrm{id}_B \tag{A.7}$$

Intuitively, a section defines a map that assigns to every point in $p \in B$ a point in the fibre F_p. In the sense of fibre bundles, we can see this as a way of defining functions from the base space into the typical fibre, but where the sense of continuity for this function can be different to that of a typical continuous function from $B \to F$, which would correspond to the product bundle of B and F.

A.2.2. Charts and atlases. While we can view manifolds as their whole, it is useful to look at them only as a patch at a time. We can formalise this notion via charts, where we map some part of the manifold onto a subset of R^n.

We call any open subset of X along with the homeomorphism $\phi : U \to \underset{\text{open}}{\subset} \mathbb{R}^n$ a *chart* of X. A collection of charts $\{(U_a, \phi_a) \; : \; a \in A\}$ such that $\bigcup_{a \in A} U_a = X$ an *atlas* of X, denoted $\mathcal{A}_X$. For any two charts $(U, \phi_U), (V, \phi_V)$ in an atlas where $U \cap V \neq \varnothing$ we can define the *transition map* between them as the map $\phi_{UV} : \phi_U(U \cap V) \to \phi_V(U \cap V)$ as the composition $\phi_{UV} = \phi_V \circ \phi_U^{-1}$. An atlas is called a *maximal atlas* if it is not a subset of some larger atlas. Chart, Atlas, Transition map, Maximal atlas

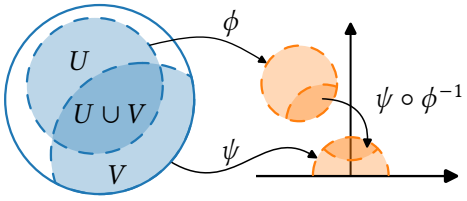


Figure A.5. Two charts of the atlas $\mathcal{A}_X$ of X, the unit disk, a manifold with boundary. (U, ϕ_U), a chart mapping into a subset of $\mathbb{R}^n$, and (V, ϕ_V), a chart mapping into a subset of the upper half-space $\mathbb{H}^2$.

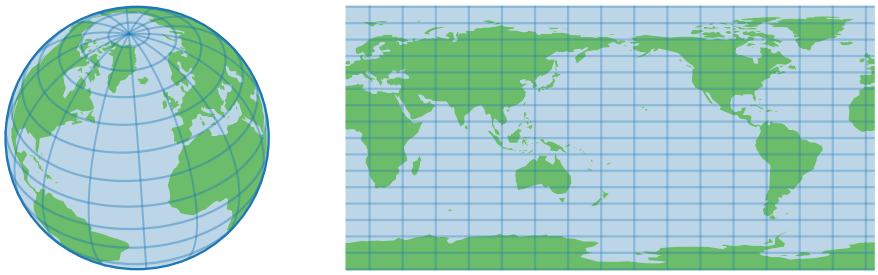


Figure A.6. The projection of the earth onto a flat map is the prototypical example of a chart. While the Mercator projection it looks like the chart covers the whole globe, but mathematically it is in fact missing the poles and a line connecting them as we need the set to be an open one.

While not required for the definition of a manifold, the concept of charts and atlases is key to the study of additional structure we can impose on a manifold. In addition, from a computational perspective they are incredibly useful as they give us one way of representing points on a manifold as a series of real numbers.

A.2.3. Partitions of unity. A very common strategy, both when defining properties and proving theorems, on manifolds, is to break apart the manifold into an *open cover* (see appendix A.1.2), define or prove things locally, and then using for example the *gluing lemma* (see appendix A.1.3), patch the result back together. We run into issues when using any cover however, as at any given point we try to glue together an uncountable number of functions. A *partition of unity* is the type of structure we need to reduce this problem.

Partition of unity

DEFINITION A.9. *Let X be a topological manifold, and $\mathcal{U} = \{U_a\}_{a \in A}$ be an open cover of X. A* PARTITION OF UNITY, *$\{\psi_a\}_{a \in A}$ is a collection of continuous functions $\psi_a : X \to [0, 1]$ such that*

1. *$\operatorname{supp} \psi_a \in \mathcal{U}$, each function is non-zero only on some set in the cover,*
2. *For a point $p \in X$, only finitely many $psi_a(p)$ are non-zero,*
3. *$\sum_{a \in A} \psi(p) = 1$ for all $p \in X$.*

The existence of partitions of unity

THEOREM A.10. *Let X be a topological manifold, and $\mathcal{U} = \{U_a\}_{a \in A}$ be an open cover of X. Then there exists a partition of unity subordinate to $\mathcal{U}$.*

A.3. Differentiable and Smooth Manifolds

To add additional structure to our manifolds, we might require that the atlas we specify has some additional properties. For any property, ■, we say that an atlas is ■-compatible if for all pairs of charts in the atlas the transition map between charts has the particular property. We call a manifold equipped with an atlas that is ■-compatible a ❄-manifold.

By definition, all atlases are C^0, and so all manifolds are C^0-manifolds. In addition, we could require our manifold to be a *(C^k)-differentiable manifolds*, where all the transition maps are C^k, i.e. they have up to k differentials. Of most interest to us will be *smooth manifolds*, manifolds for which the transition maps are infinity differentiable, i.e. they are in C^∞.

(C^k)-differentiable manifolds
Smooth manifolds

A *smooth structure* on X is a *maximal smooth atlas*, and so a *smooth manifold (with boundary)* is a triple $(X, \mathcal{O}_X, \mathcal{A}_X)$ with the right properties. Every coordinate chart of a smooth manifold is itself smooth. The smooth structure of a smooth manifold with boundary can be determined entirely by a smooth structure on the interior $\operatorname{Int} X$ by continuously extending the charts of the interior to cover the boundary also.

Smooth structure
Smooth manifold (with boundary)

Showing a manifold is smooth is not always simple from the definition. The following is useful as it lets us create a smooth manifold with suitable charts:

Smooth manifold chart lemma

Lemma A.11. *Suppose X is a set, and we have a subset of collections $\{U\}_a$ along with maps $\phi_a : U \to \mathbb{R}^n$. If the following are satisfied:*

1. *for each a, ϕ_a is a bijection between U_a and an OPEN subset of $\mathbb{R}^n$ with the usual topology.*
2. *For each a, b, the sets $\phi_a(U_a \cap U_b)$ and $\phi_b(U_a \cap U_b)$ are open in $\mathbb{R}^n$ with the usual topology.*
3. *For all a, b where $U_a \cap U_b \neq \varnothing$, the map $\phi_a^{-1} \circ \phi_b : \phi_a(U_a \cap U_b) \to \phi_b(U_a \cap U_b)$ is smooth.*
4. *Countably many U_a cover X.*
5. *When $p, q \in X$, $p \neq q$, there exists some U_a containing p and q, or there exist disjoint U_a, U_b with $p \in U_a$, $q \in U_n$.*

Then X has a unique smooth manifold structure such that each (U_a, ϕ_a) is a smooth chart.

Proof. (Lee, 2013, p.22) ■

The topology of the structure is induced by the topology on $\mathbb{R}^n$ and the (U_a, ϕ_a) (1-3).
The basis for the topology on X is given by $\mathcal{B} = \left\{\phi_a^{-1}(S) \ : \ S \text{ is open in } \phi_a(U_a) \text{ for all } a\right\}$.
4. ensures second countability, and 5. Hausdorff.

Smooth functions on manifolds

We can define *smooth functions on manifolds*. $f : X \to \mathbb{R}^k$ is a smooth function if for every point $p \in X$ and chart $(U, \phi) \in \mathcal{A}_X$ the function $f \circ \phi^{-1} : \phi(U) \to \mathbb{R}^k$ is smooth. Typically, we denote the space of smooth function $f : X \to \mathbb{R}^k$ as $C^\infty(X, \mathbb{R}^k)$, and $f : X \to \mathbb{R}$ $C^\infty(X)$. Like continuity, smoothness is a local property,

and it suffices to check that for a sub-atlas of the atlas on a smooth manifold that a function is smooth on the restriction to each chart to show it is globally smooth.

Smooth functions between manifolds

We can define *smooth functions between manifolds* as well. $f : X \to Y$, where X, Y are smooth manifolds, is smooth if for each pair of charts $(\phi, U) \in \mathcal{A}_X, (\psi, V) \in \mathcal{A}_Y$ the composition $\psi \circ f \circ \phi^{-1} : \phi(U) \to \psi(V)$ is smooth. We denote the space of smooth functions between X and Y as $\mathbb{C}^\infty(X, Y)$. A *diffeomorphism* between manifolds is a smooth bijection with smooth manifolds. A pair of manifolds for which a diffeomorphism exists are said to be *diffeomorphic*, $X \cong_{\text{diff}} Y$.

Diffeomorphism

Diffeomorphic

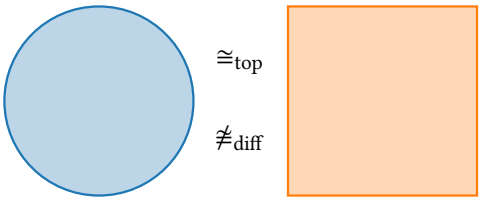


Figure A.7. Diffeomorphism is a stronger condition that homeomorphic. The disk is homeomorphic to the square, but not diffeomorphic, as one runs into smoothness difficulties with the sharp corners.

A.4. Tensors

Having introduced differentiable structure to manifolds, we are ready to start talking about derivatives on manifolds. Derivatives naturally form a vector space in multivariate calculus. Therefore, we briefly recap vector spaces, and discuss the notion of tensors. Tensors will be central to the study of deep structure on manifolds later.

A.4.1. Algebraic structures. Briefly, we recall the definitions of a few algebraic structures.

Group

Definition A.12. *A* group *is a set G and an operation $\cdot : G \times G \to G$ such that*

1. *For $a, b \in G$, $a \cdot b \in G$.*
2. *The operation is* associative*, for $a, b, c \in G$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.*
3. *There exists an* identity element *$e \in G$ such that for any $a \in G$, $e \cdot a = a \cdot e = a$.*
4. *For every element $a \in G$, there is another element $a^{-1} \in G$ such that $a \cdot a^{-1} = e = a^{-1} \cdot a$, its* inverse*.*

Abelian group

Definition A.13. *An* Abelian group *is a group $(G, \cdot)$ where in addition $\cdot$ is* commutative*, i.e. for $a, b \in G$, $a \cdot b = b \cdot a$.*

Field

Definition A.14. *A* field *is a set F equipped with two maps, $+ : F \times F \to F$ and $* : F \times F \to F$ such that*

- *$(F, +)$ is an Abelian group, with identity element 0.*
- *$(F \backslash \{0\}, *)$ is an Abelian group with identity element 1.*

In addition, the operations DISTRIBUTE *in the following way, for* $a, b, c \in F$,

$$(a + b) * c = a * c + b * c \tag{A.8}$$

The prototypical example of a field is the real numbers, R, equipped with the usual addition and division, excluding division by zero. The complex, $\mathbb{C}$, and rational, Q numbers are also fields.

Also of use the weaker definition of a *ring*.

DEFINITION A.15. *A* RING *is a set* F *equipped with two maps,* $+ : F \times F \to F$ *and* $* : F \times F \to F$ *such that* Ring

- $(F, +)$ *is an Abelian group, with identity element* 0.
- *For* $(F \setminus \{0\}, *)$, *the* $*$ *operator only satisfies associativity, for* $a, b, c \in F \setminus \{0\}$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.

This definition is useful for when discussing spaces of functions, for example the square integrable functions on some domain D, $L^2(D)$. In this case, if we define the addition operation in the usual pointwise way, the identity function (the constant zero function) is not the only function we need to exclude to make division by a function sensible. We would need to exclude any function with a zero at any point. For this same reason, defining the inverse of a function under multiplication is not sensible, as the inverse of the points in a function that are zero are undefined. If we were to restrict ourselves to strictly positive L^2 functions, then we could form a field from these. As it stands however this is not so useful, so we satisfy ourselves with a ring.

A.4.2. Vector Spaces.

DEFINITION A.16. *A* VECTOR SPACE *over a field* F *is a set* V *equipped with a* VECTOR ADDITION *operation* $\oplus : V \times V \to V$ *and a* SCALAR MULTIPLICATION *operation* $\odot : F \times V \to V$ *such that* Vector space over a field

- $(V, \oplus)$ *is an Abelian group.*
- $\odot$ *is an* ACTION *on* $(V, \oplus)$ *such that*
  1. *Scalar multiplication distributes over vector addition, for* $\lambda \in K, v, w \in V$, $\lambda \odot (v \oplus w) = (\lambda \odot v) \oplus (\lambda \odot w)$.
  2. *Field addition distributes over vector addition, for* $\lambda, \mu \in K, v \in V$, $(\lambda + \mu) \odot v = (\lambda \odot v) \oplus (\mu \odot v)$.
  3. *Field multiplication is compatible with scalar multiplication, for* $\lambda, \mu \in K$, $v \in V$, $(\lambda * \mu) \odot v = \lambda \odot (\mu \odot v)$.
  4. *The field identity is preserved, for* $v \in V$, $1 \odot v = v$.

When discussing different vector spaces over the same field, properly one should denote their vector addition and scalar multiplication operations with different notation, e.g. $\oplus, \boxplus$, to differentiate these operations from each other and the field operations. Typically, however this is suppressed and the same symbol is used for field multiplication and addition, as well as the vector counterparts on different

vector spaces as one can infer the correct operation from context. We will use this convention.

Topological vector space

A *topological vector space* is a vector space over a topological field, equipped with a topology so that vector addition and scalar multiplication are continuous. Most commonly we deal with vector spaces over the *real numbers* $\mathbb{R}$, but sometimes also over the *complex numbers* $\mathbb{C}$. From now on we will only discuss real vector spaces.

Linear subspace
Linear combination
Span
Linearly dependent
Linearly independent

A subset of a vector space that is closed under the vector operation is called a *linear subspace*, distinct from a topological subspace. If we have a finite set of vectors $S = \{v_1, ..., v_n\}, v_i \in V$, a *linear combination* of them is denoted by $\sum_{i=1}^{n} a_i v_i$ with $a_i \in \mathbb{R}$. The *span* of a set of vectors, denoted $\text{span}(S)$ is the set of all linear combinations of the vectors, and is a linear subspace. If $V = \text{span}(S)$ then S *spans* V. A set S is said to be *linearly dependent* is there exist $a_i \in \mathbb{R}$ such that the linear combination is $0 \in V$. Otherwise, they are *linearly independent*.

Linear map
Linear functional
Image
Kernel

A *linear map* between vector spaces V and W are maps $T : V \to W$ such that they preserve the structure of vector addition and scalar multiplication. I.e. $T(av+bw) = aT(v) + bT(w)$. A map $T : V \to \mathbb{R}$ is called a *linear functional*. The *image* of T, $\text{im}\, T$ is the set $\text{im}\, T = \{w = T(v) \ : \ v \in V\}$. The *kernel* of T, $\ker T$ is the preimage of 0 under T, $\ker T = \{v \ : \ T(v) = 0, v \in V\}$.

Vector isomorphism

If V, W are vector spaces over the same field, a bijective linear map $T : V \to W$ then T is called a *vector isomorphism*, T has an inverse, and V, W are said to be (vector) isomorphic, $V \cong_{\text{vec}} W$.

Homeomorphisms

The space of all linear maps from a vector space V to a vector space W over the same field is denoted as $\text{Hom}(V, W)$, the *homeomorphisms* from V to W. This space of homeomorphisms can in fact be itself turned into a vector space by defining vector addition

$$\boxplus : \text{Hom}(V, W) \times \text{Hom}(V, W) \to \text{Hom}(V, W) \tag{A.9}$$

$$(f \boxplus g)(v) \mapsto f(v) + g(v) \tag{A.10}$$

and scalar multiplication

$$\boxdot : K \times \text{Hom}(V, W) \to \text{Hom}(V, W) \tag{A.11}$$

$$(\lambda \boxdot f)(v) = \lambda f(v) \tag{A.12}$$

Endomorphism
Automorphism

An *endomorphism* of V is a linear map onto itself, and the space of these $\text{End}(V) = \text{Hom}(V, V)$. An *automorphism* of V is a linear isomorphism onto itself, and the space of these is $\text{Aut}(X) = \{f \in \text{End}\, V \ : \ f \text{ is an isomorphism}\}$.

Covector
Dual space

For any vector space V, we can define a *covector* as a linear map $\omega : V \to \mathbb{R}$. We call the space of covectors the *dual space* of V, denoted V^*, and it is equal to $\text{Hom}(V, K)$, where V is a vector space over the field K.

Dual map

For any linear map $A : V \to W$, we can define its *dual map*, *transpose*, or *adjoint* $A^* : W^* \to V^*$ as the map

$$(A^*\omega)(v) = \omega(Av), \quad \omega \in W^*, v \in v. \tag{A.13}$$

This adjoint distributes over composition, but reverse terms, $(A \circ B)^* = B^* \circ A^*$, and preserves the identity, $(\text{id}_V)^* = \text{id}_{V^*}$.

The *double dual* of V, $V^{**} = (V^*)^*$, the dual of the dual, for finite dimension vector spaces is isomorphic to the original space V. There is a natural map $\xi : V \to V^{**}$ as Double dual

$$\xi(v)(\omega) = \omega(v) \quad , \omega \in V^*, v \in V. \tag{A.14}$$

Because we can unambiguously identify V with V^{**} for finite dimension vector spaces, we typically only talk about V and its dual V^*. Sometimes one may see the inner product between a vector and a covector written as $\langle \omega, v \rangle$ or $\langle v, \omega \rangle$. The second denotes the action of $\xi(v)$ on ω, but the result of both are identical.

A.4.3. Tensors. Covectors defined linear maps over vector spaces. Tensors allow us to define *multilinear maps*. For vector spaces $V_1, ..., V_k, W$, a multilinear map is a map $f : V_1 \times ... \times V_k \to W$ if for any i, Multilinear maps

$$f(v_1, ..., av_i + bv_i', ..., v_k) = af(v_1, ..., v_i, ..., v_k) + bf(v_1, ..., v_i', ..., v_k). \tag{A.15}$$

Tensors are multilinear maps over copies of the same vectors space and its dual. We can define the following spaces of multilinear maps

Map	Notation	"Archaic" name
$f : \underbrace{V \times ... \times V}_{k \text{ copies}} \to \mathbb{R}$	$T^k(V^*) = T^{(0,k)}(V)$	covariant k-tensor on V
$f : \underbrace{V^* \times ... \times V^*}_{k \text{ copies}} \to \mathbb{R}$	$T^k(V) = T^{(k,0)}(V)$	contravariant k-tensor on V
$f : \underbrace{V^* \times ... \times V^*}_{k \text{ copies}} \times \underbrace{V \times ... \times V}_{l \text{ copies}} \to \mathbb{R}$	$T^{(k,l)}(V)$	mixed (k, l)-tensor on V

The "archaic" name refers to how the basis coefficients transform with the basis vectors. *Covariant* tensor basis coefficients transform *with* them, and *contravariant opposite* to them. In the modern setting where we typically work basis-free these make less sense. By convention, and for sensible reasons, the a tensor in $T^{(0,0)}$ is just a scalar.

The *tensor product* is an operation $\otimes : T^{(k,l)}(V) \times T^{(p,q)}(V) \to T^{(k+p,l+q)}(V)$. It is defined by its evaluation Tensor product

$$(F \otimes G)(\omega^1, ..., \omega^k, \omega^{k+1}, ..., \omega^{k+p}, v_1, ..., v_l, v_{l+1}, ..., v_{l+q}) \tag{A.16}$$
$$= F(\omega^1, ..., \omega^k, v_1, ..., v_l) G(\omega^{k+1}, ..., \omega^{k+p}, v_{l+1}, ..., v_{l+q}). \tag{A.17}$$

This is an associative but *not* a commutative operation.

Some immediately interesting tensor spaces arise

- $T_1^0 V = \text{Hom}(V) = V^*$
- $T_1^1 V = V \otimes V^*$. This is in fact isomorphic to $\text{End}(V^*)$. To see this, note that for a fixed $\omega \in V^*$, and $T \in T_1^1 V$, $T(\cdot, \omega) : V \to F$, and since this is a linear map, this must be an element of V^*.

A.4.4. Bases. The concepts introduced so far are abstract, and have no easy way to be represented in a computer or on paper and manipulated numerically. In order to do this, we will need to express vectors in a *basis*.

Hamel basis

A *Hamel basis* for V is a linearly independent subset $S \subseteq V$ such that $\text{span}(S) = V$. Every $v \in V$ has a *unique* expression as a linear combination of the basis. The choice of basis is non-unique, although often there is a sensible choice. All bases of a vector space are of the same cardinality, called the dimension of the space, $\dim V$. If the basis is finite of size n, the space is said to be finite-dimension or $n-$dimension. If it is infinite, infinite-dimension. If we can ascribe an order to a basis (if it is finite or countable) then we call it an *ordered basis*, and we can identify the elements $v \in V$ of the vector space with the ordered set of field coefficients $(a_i 1, ...)$ such that $\sum_i a^i v_i = v$. This is called the *basis representation*. All bases of the same vector space are of the same dimension.

Ordered basis

Basis representation

For a finite dimension vector space then it is common to express a vector as an ordered list of numbers, with each element of the list being a basis coefficient of the vector expressed in a basis,

$$v = \sum_{i=1}^{n} v^i e_i = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} \tag{A.18}$$

Standard dual basis

For a finite dimension vector space, if we have a basis on V, $e_1, ..., e_n$, then we can define the *standard dual basis* $\epsilon^1, ..., \epsilon^n$ on V^* by $\epsilon^i(e_j) = \delta^i_j$, showing it is also n-dimension. Note the use of upper indices, inline with Einstein summation notation. We write basis vectors of the dual space with upper indices so that they appear once upper and once lower when computing the interaction of a vector and a covector.

$$\omega(v) = (\omega_i \epsilon^i)(v^j e_j) = \omega_i v^j \epsilon^i(e_j) = \omega_i v^j \delta^j_i = \omega_i v^i \tag{A.19}$$

Basis on a tensor space

We can build a *basis on a tensor space* then out of the basis on V and V^*, giving basis elements of the form

$$e_{i_1} \otimes ... \otimes e_{i_k} \otimes \epsilon^{j_1}, \otimes ... \otimes \epsilon^{j_l}, \qquad i_a \in [1, ..., k],\ j_b \in [1, ..., l] \tag{A.20}$$

with the obvious action on basis vectors

$$e_{i_1} \otimes ... \otimes e_{i_k} \otimes \epsilon^{j_1}, \otimes ... \otimes \epsilon^{j_l}\left(\epsilon^{s_1}, ..., \epsilon^{s_k}, e_{t_1}, ..., e_{t_l}\right) = \delta^{s_1}_{i_1} ... \delta^{s_k}_{i_k} \delta^{j_1}_{t_1} ... \delta^{j_l}_{t_l}. \tag{A.21}$$

This makes a $T^{(k,l)}V$ an n^{k+l}-dimension vector space. For a tensor $T \in T^p_q V$, we can express this in basis coefficients

$$T = \underbrace{\sum_{a_1=1}^{n} ... \sum_{b_q=1}^{n}}_{p+q \text{ sums}} T^{a_1,...a_p}_{b_1,...,b_q} e_{a_1} \otimes ... \otimes e_{a_p} \otimes \epsilon^{b_1} \otimes ... \otimes \epsilon^{b_q} \tag{A.22}$$

Components of the map

For a linear map f between finite dimension vectors spaces, V, W, each with a basis $e^V_1, ..., e^V_n, e^W_1, ..., e^W_m$, the *components of the map* can be expressed by the coefficients $f^i_j = f(e^V_i, e^W_j)$. For a vector $v \in V$ expressed in the basis, $v = \sum_{i=1}^n v^i e^V_i$, its image under the linear map f, $f(v)$, can then be expressed in coordinates in the basis on W as $w = f(v) = \sum_{j=1}^m (\sum_{i=1}^n f^j_i v^i) e^W_j = \sum_{j=1}^m w^j e^W_j$.

A *change of basis* of a vector can be performed between two different bases $e_1, ..., e_n$ and $\tilde{e}_1, ..., \tilde{e}_n$. Each basis vector $\tilde{e}_j$ can be expressed as a linear combination of the first basis, $\tilde{e}_j = \lambda^i_j e_i$. Any vector expressed in the first basis, $v = v^i e_i$ can then be re-expressed in the second basis via $v = \tilde{v}^j \tilde{e}_j$ where $\tilde{v}_j = \lambda^j_i v^i$. As such we can see a change of basis as an automorphism of V. Change of basis

We can classify bases into two equivalence classes, called an *orientation*. To do so, we pick a basis to be the reference. We called another basis *positively oriented* if the determinant of the transition matrix λ^j_i is positive, and *negatively oriented* if this is negative. Orientation Positively oriented Negatively oriented

A.4.5. Einstein summation convention. Einstein summation convention is a notation method for reducing the notation overhead when dealing with vector spaces and linear maps. It consists of suppressing explicit summation signs and replacing them with the implicit assumption that repeated indices in an expression are summed over. Indices can appear in subscript or superscript, and can only appear in each position at most once in an expression. For example, we can write

$$v = \sum_{i=1}^{n} v^i e_i = v^i e_i \tag{A.23}$$

for the expression of a vector in a basis, and

$$T = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} T^{ij}_k f^k e_i \otimes e_j = T^{ij}_k f^k e_i \otimes e_j \tag{A.24}$$

for the more complex construction of a $(2, 0)-$tensor.

Indices that are summed over are called *dummy indices*. They appear in pairs and it is not important what letter we choose to label them with. Indices that appear once, and are therefore not summed, are called *free indices*, and must appear on both sides of an expression. For example Dummy indices Free indices

$$T^{ij}_k = T^{imj}_{km} \tag{A.25}$$

represents a trace-like summation.

There is an additional convention to which indices are superscript and which are subscript.

- Basis vectors of a vector space are subscripted.
- Basis vectors of dual vector spaces are superscripted.
- Other placements are derived from the above two rules.

As an example of the last rule, coefficients of a vector in a basis must be in superscript so that they are correctly summed with the basis vectors, and the opposite for covector coefficients.

When working with Einstein summation convention and maps, care should be taken to only apply the notation to *linear* maps, or take additional care. For example, for the bilinear map $\phi : V \times W \to \mathbb{R}$ we can write this as

$$\phi(v, w) = \phi(v^i e_i, w^j \tilde{e}_j) = v^i w^j \phi(e_i, \tilde{e}_j). \tag{A.26}$$

We can then represent the map ϕ via the set of coefficients $\phi_{ij} = \phi(e_i, \tilde{e}_j)$, and write

$$\phi(v, w) = v^i w^j \phi_{ij}. \tag{A.27}$$

This is only true however if ϕ is bilinear. This makes Einstein summation convention very suitable in linear algebra, and by extension the study of tensor fields in geometry, as linear maps are omnipresent.

A.4.6. Symmetric tensors. For the space of tensors of one type only ($T^k(V^*) = T^{(0,k)}(V)$ or $T^k(V) = T^{(k,0)}(V)$) we can define two special types of tensor. We will focus here on $T^k(V^*)$ tensors, but the same applies identically to $T^k(V)$ tensors.

Symmetric tensor

A *symmetric tensor* is one such that if you take any set of vectors $v_1, ..., v_k$ and permute a pair, then the evaluation of the tensor remains the same. I.e.

$$F(v_1, v_i, ..., v_j ...., v_k) = F(v_1, v_j, ..., v_i, ..., v_k). \tag{A.28}$$

This implies *any* permutation of arguments by $\sigma \in S_k$, the group of permutation of k elements, leave it unchanged. We denote this *linear* subspace of $T^k(V^*)$ as $\Sigma^k(V^*)$. The tensor product does *not* preserve symmetry. We can *symmetrise a tensor*, denoted $\operatorname{Sym} F$ by

Symmetrise a tensor

$$(\operatorname{Sym} F)(v_1, ..., v_k) = \frac{1}{k!} \sum_{\sigma \in S^k} F\big(v_{\sigma(1)}, ..., v_{\sigma(k)}\big). \tag{A.29}$$

Symmetric product

We can then define the *symmetric product* of $F \in \Sigma^k(V^*), G \in \Sigma^l(V^*)$ as

$$FG = \operatorname{Sym}(F \otimes G), \tag{A.30}$$

$FG \in T^{k+l}(V^*)$. This is a *bilinear, associative, commutative* operation. On 2 1-tensors this gives

$$\omega\eta = \frac{1}{2}(\omega \otimes \eta + \eta \otimes \omega). \tag{A.31}$$

All 0- and 1-tensors are symmetric trivially.

Alternating tensor

A.4.7. Alternating tensors. An *alternating tensor* is one such that if you take any set of vectors $v_1, ..., v_k$ and permute a pair, then the evaluation of the tensor flips sign. I.e.

$$F(v_1, v_i, ..., v_j ...., v_k) = -F(v_1, v_j, ..., v_i, ..., v_k). \tag{A.32}$$

This implies that *even* permutations of S_k (made up of an even number of pair flips) leaves the sign unchanged and that *odd* permutations flip i,

$$F(v_{\sigma(1)}, ..., v_{\sigma(k)}) = \operatorname{sgn}(\sigma) F(v_1, ..., v_k). \tag{A.33}$$

We denote this *linear* subspace of $T^k(V^*)$ as $\Lambda^k(V^*)$. The tensor product does *not* preserve alternation. We can define the *alternation of a tensor*, denoted $\operatorname{Alt} F$ by

Alternation of a tensor

$$(\operatorname{Alt} F)(v_1, ..., v_k) = \frac{1}{k!} \sum_{\sigma \in S^k} \operatorname{sgn}(\sigma) F\big(v_{\sigma(1)}, ..., v_{\sigma(k)}\big). \tag{A.34}$$

For any tensor, this returns us an alternating tensor.

Let $I = (i_1, ..., i_k)$ define a *multi-index*. Let $I_\sigma = \left(i_{\sigma(1)}, ..., i_{\sigma(n)}\right)$. Assume we have a basis for V^*, $\epsilon^1, ..., \epsilon^n$. We can define an *elementary alternating tensor* by its action on k vectors, $v_1, ..., v_k$

Multi-index
Elementary alternating tensor

$$\epsilon^I(v_1, ..., v_k) = \det \begin{bmatrix} \epsilon^{i_1}(v_1) & \dots & \epsilon^{i_1}(v_1) \\ \vdots & \ddots & \vdots \\ \epsilon^{i_k}(v_1) & \dots & \epsilon^{i_k}(v_k) \end{bmatrix} \tag{A.35}$$

where det is the usual determinant. If we express the vectors $v_1, ..., v_k$ in the dual basis to $\epsilon^1, ..., \epsilon^n$, $e_1, ..., e_n$, then this is

$$\epsilon^I(v_1, ..., v_k) = \det \begin{bmatrix} v_1^{i^1} & \dots & v_k^{i^1} \\ \vdots & \ddots & \vdots \\ v_1^{i^k} & \dots & v_k^{i^k} \end{bmatrix}. \tag{A.36}$$

Clearly this is an alternating tensor, as swapping any two columns in the determinant results in the change in a sign, and it is zero if any two vectors are co-linear.

Evaluating an elementary alternating tensor on basis elements $e_{j_1}, ...e_{j_k}$ gives

$$\epsilon^I(e_1, ..., e_k) = \begin{bmatrix} \delta_{j_1}^{i^1} & \dots & \delta_{j_k}^{i^1} \\ \vdots & \ddots & \vdots \\ \delta_{j_1}^{i^k} & \dots & \delta_{j_k}^{i^k} \end{bmatrix} = \delta_J^I, \tag{A.37}$$

where $J = (j_1, ..., j_k)$.

By picking a subset of the elementary alternating tensors we can produce a basis on the space of alternating tensors. We need to exclude any that are zero (I with repeated indices) and any I that is a permutation of another element of the basis set. For the space alternating tensors $\Lambda^k(V^*)$ over an n-dimension vector space V^*, we can therefore pick the basis

$$\mathbb{I} = \left\{\epsilon^I \ : \ I \text{ is a strictly increasing multiindex of size } k\right\}. \tag{A.38}$$

The *wedge product* of two alternating tensors, denoted by $\wedge$, maps two alternating tensors to an alternating tensor of higher dimension,

Wedge product

$$\wedge : \Lambda^k(V^*) \times \Lambda^l(V^*) \to \Lambda^{k+l}(V^*). \tag{A.39}$$

For $F \in \Lambda^k(V^*), G \in \Lambda^l(V^*)$, is is defined as

$$F \wedge G = \frac{(k+l)!}{k!l!} \operatorname{Alt}(F \otimes G), \tag{A.40}$$

$F \wedge G \in \Lambda^{k+l}(V^*)$. This is a *bilinear, associative, anti-commutative* ($F \wedge G = (-1)^{kl} G \wedge F$) operation. All 0- and 1-tensors are alternating trivially.

The *interior product* is an operation $\lrcorner : \Lambda^k(V^*) \times V \to \Lambda^{k-1}(V^*)$ defined by

Interior product

$$(v \lrcorner \omega)(w_1, ..., w_{k-1}) = \omega(v, w_1, ..., w_{k-1}). \tag{A.41}$$

Volume form

Finally, we define the notion of a *volume form*. For an n-dimension vector space V, a volume form is a tensor $\omega \in \Lambda^n V$. For a series of vectors $v_1, ..., v_n$, the *volume* they span is given by

$$\text{vol}(v_1, ..., v_n) = \omega(v_1, ..., v_n) \tag{A.42}$$

If the set of vectors are not linearly independent, this volume spanned will be zero. Intuitively this makes sense as then the vectors span a hyperplane in V, which should have zero volume in the whole of V.

Volume form defines an orientation

Using a particular choice of a *volume form defines an orientation* of an ordered basis. If the volume form ω applied to an ordered basis $e_1, ..., e_n$ is positive then the basis is positively oriented, otherwise it is negatively oriented.

A.5. Tangent, cotangent, and tensor spaces

A.5.1. Tangent spaces and derivations. One of the key advantages of studying smooth manifolds is that it allows us to bring the notions of calculus to the study of manifolds. The most basic concept is that of the *tangent vector*.

We can begin looking at tangent vectors on manifolds by first looking at tangent vectors on $\mathbb{R}^n$. Since all we have defined on smooth manifolds so far is smooth functions and smooth charts, we will need a definition that works in terms of these.

Geometric tangent space

Consider the set $\mathbb{R}^n_p = \{p\} \times \mathbb{R}^n$. We will call this the *geometric tangent space* at p. We call an element of this, (p, v), $v \in \mathbb{R}^n$, usually written at $v|_p$ or v_p, a *geometric tangent vector*. This is clearly a vector space with operations

Geometric tangent vector

$$v_p + w_p = (v + w)_a \qquad c(v_p) = (cv)_p. \tag{A.43}$$

The subscript p is to allow us to differentiate between tangent spaces at different points p and q. Given the canonical basis of $\mathbb{R}^n$, $e_1, ..., e_n = (1, ..., 0), ..., (0, ..., 1)$, we can put a natural basis on $\mathbb{R}^n_p$ as $e_1|_p = (p, e_1), ..., e_n|_p = (p, e_n)$ via the *canonical isomorphism* $\iota_p : \mathbb{R}^n \to \{p\} \times \mathbb{R}^n$, $\iota_p : v \mapsto (p, v)$.

Canonical isomorphism

Directional derivative

For a smooth function $f \in C^\infty(\mathbb{R}^n)$ we can define the *directional derivative* $D_{v_a} : C^\infty(\mathbb{R}^n) \to \mathbb{R}$ as

$$D_{v_a} f = D_v f(a) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} f(a + tv). \tag{A.44}$$

This operation obeys the product rule

$$D_{v_a}(fg) = f(a) D_{v_a} f + g(a) D_{v_a} f. \tag{A.45}$$

For the basis vectors $e_i|_p$,

$$D_{e_i|_p} f = \frac{\partial f}{\partial x_i}(p), \tag{A.46}$$

simply the partial derivative with respect to the ith input. So for $v_p \in \mathbb{R}^n_p$ expressed as $v = \sum_{i=1}^n v_i \, e_i|_p$, we have that

$$D_{v_p} f = \sum_{i=1}^n v_i \frac{\partial f}{\partial x_i}(p) \tag{A.47}$$

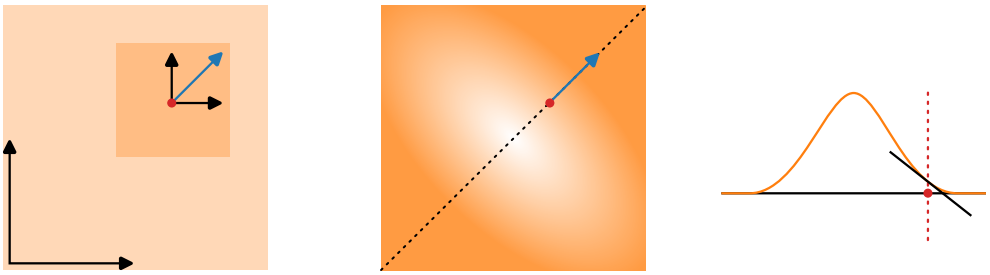


Figure A.8. Directional derivative.

Tangent vectors as algebraic structures

With this intuitive definition in mind, let us take another look at the space of smooth functions, $C^\infty(\mathbb{R}^n)$. We recall a few more algebraic definitions.

DEFINITION A.17. *An ALGEBRA A over a FIELD F is a set A with 3 operations,* Algebra over a field

$$+ : A \times A \to A$$

$$\cdot : F \times A \to A$$

$$\bullet : A \times A \to A$$

where

- $(A, +, \cdot)$ *forms a VECTOR SPACE over F*
- *and $\bullet$ is BI-LINEAR product with respect to the field F.*

Additionally, this can be *associative* if the product is associate, *unital* if the product has an identity element, and *commutative* if the product commutes.

We can make the space of smooth functions $C^\infty(\mathbb{R}^n)$ an associative, unital, commutative algebra over the field $\mathbb{R}$ by setting

$+ : C^\infty(\mathbb{R}^n) \times C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n)$ to be pointwise addition of functions.

$\cdot : \mathbb{R} \times C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n)$ to be pointwise multiplication by a scalar.

$\bullet : C^\infty(\mathbb{R}^n) \times C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n)$ to be pointwise multiplication of functions.

Often for this particular algebra, we will suppress the notation of $f \bullet g$ to simply fg for $f, g \in C^\infty(\mathbb{R}^n)$.

DEFINITION A.18. *For an algebra A, a DERIVATION is a linear map $D : A \to A$ satisfying the product rule over the algebra product, for $f, g \in A$,* Derivation on an algebra

$$D(f \bullet g) = D(f) \bullet g + f \bullet D(g) \tag{A.48}$$

We denote the space of derivations as $\operatorname{Der} A$.

For our algebra of smooth functions $C^\infty(\mathbb{R}^n)$, we can restrict these derivations to a point $p \in \mathbb{R}^n$

Derivation on $\mathbb{C}^\infty(\mathbb{R}^n)$ at $p \in \mathbb{R}^n$

DEFINITION A.19. *For the algebra $C^\infty(\mathbb{R}^n)$ and a point $p \in \mathbb{R}^n$, a DERIVATION AT p is a linear map $D : C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n)$ satisfying a reduced product rule over the algebra product, for $f, g \in C^\infty(\mathbb{R}^n)$, at a point $p \in X$,*

$$D(f \bullet g)(p) = (D(f) \bullet g)(p) + (f \bullet D(g))(p) \tag{A.49}$$

We denote the space of derivations as $\mathrm{Der}_p\, C^\infty(\mathbb{R}^n)$. This space of derivations at a point from a *vector space* over the *field* $\mathbb{R}$ by defining vector addition of derivations by

$$[(D+\tilde{D})f](p) = [Df](p)+[\tilde{D}f](p), \quad D, \tilde{D} \in \mathrm{Der}_p\, C^\infty(\mathbb{R}^n),\ f \in C^\infty(\mathbb{R}^n),\ p \in \mathbb{R}^n \tag{A.50}$$

and scalar multiplication as

$$[(\lambda D)f](p) = \lambda[Df](p), \quad D \in \mathrm{Der}_p\, C^\infty(\mathbb{R}^n),\ f \in C^\infty(\mathbb{R}^n),\ p \in \mathbb{R}^n,\ \lambda \in \mathbb{R}. \tag{A.51}$$

We can then prove that our directional derivative and the derivations at p are isomorphic.

$\mathrm{Der}_p\, C^\infty(\mathbb{R}^n) \cong_{\mathrm{vec}} \mathbb{R}^n_p$

PROPOSITION A.20.

1. *For each $v_p \in \mathbb{R}^n_p$, D_{v_p} is a derivation.*
2. *The map $v_p \mapsto D_{v_p}$ is an isomorphism from $\mathbb{R}^n_p$ onto $\mathrm{Der}_p\, C^\infty(\mathbb{R}^n)$.*

Proof. (Lee, 2013, proposition 3.2) ∎

This therefore connects the intuitive idea of directional derivatives that we are used to when working with Euclidean space with the more abstract definition of the vector space of derivations on the algebra of smooth functions at a point. We call the space of derivations at a point the *tangent space*, and denote it $T_p\mathbb{R}^n = \mathrm{Der}_p\, C^\infty(\mathbb{R}^n)$.

Tangent space

Tangent space at a point on a manifold

Deviation on $C^\infty(U)$ at $p \in U$

This then give us a natural way to define the *tangent space at a point on a manifold*. For a manifold and an open set of the manifold $U \subset X$, we say that a *deviation on $C^\infty(U)$ at $p \in U$* is a linear map $D : C^\infty(U) \to R$ satisfying

$$D(fg) = D(f)g(p) + f(p)D(g), \tag{A.52}$$

and we denote this space of derivations as $\mathrm{Der}_p\, C^\infty(U)$.

We define the tangents space at a point on the manifold $T_pX = \mathrm{Der}_p\, C^\infty(U)$. We can prove that the choice of chart is independent of the final object, and therefore is not notated. For an n-dimension manifold, the tangent space is n-dimension vector space, isomorphic (in the vector space sense) to $\mathbb{R}^n$ using the same vector operations as for Euclidean space.

Intuitively, this is the space of all possible directions one could move in from p.

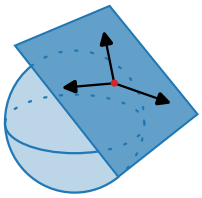


Figure A.9. The tangent space of the point $\bullet \in \mathcal{S}_2$ with example tangent vectors.

A.5.2. The differential. If we have a smooth function between manifolds $F : X \to Y$, we can also define a natural map between their tangent spaces at a point $p \in X$ and $F(p) \in Y$. We call this the *differential* of F at p, denoted $\mathrm{d}F : T_pX \to T_{F(p)}Y$, and it is defined by its action on a tangent vector $v \in T_pX$ and a smooth function on Y, $f \in C^\infty(Y)$. Differential

$$\big[\mathrm{d}F_p(v)\big](f) = v(f \circ F)(p), \quad v \in T_pM,\ f \in C^\infty(Y). \tag{A.53}$$

This is a *linear* map that distributes over compositions of maps between manifolds, $\mathrm{d}_p(F \circ G) = \mathrm{d}F_{G(p)} \circ \mathrm{d}G_p$, and preserves the identity map, $\mathrm{d}(\mathrm{id}_X)_p = \mathrm{id}_{T_pX}$. Importantly if F is a *diffeomorphism*, then $\mathrm{d}F$ is an *isomorphism* between tangent spaces and $(\mathrm{d}F)^{-1} = \mathrm{d}\big(F^{-1}\big)$.

It is worth thinking about what the differential means intuitively. It says that if we have a smooth function f on Y, we can take a derivation of this with respect to tangent vector $v \in T_pX$ by first transporting the function f from Y to X via F, and then take the derivation of this new function with v.

Sometimes the differential is referred to as the *pushforward* of a tangent vector by F, and is denoted Pushforward

$$(F_*)_p : T_pX \to T_pY \tag{A.54}$$

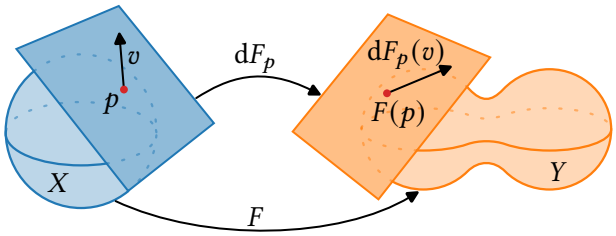


Figure A.10. A smooth map between smooth manifolds $F : X \to Y$ induces a natural map between the tangent spaces of p and $F(p)$, the differential $\mathrm{d}F_p$.

A.5.3. Submersions, immersions, and embeddings. The definition of the differential allows us to classify smooth maps between manifolds. If we have a map $F : X \to Y$, at any point we can define the *rank of a map* F at a point $p \in X$ to be the rank of the linear map $\mathrm{d}F_p$. This is the dimension of the image of the differential, $\dim \operatorname{im} \mathrm{d}F_p \subseteq T_{F(p)}Y$. The rank of a linear map is bounded by the minimum of the dimensions of its domain and codomain, so $\operatorname{rank} F_p \leq \min\{\dim X, \dim Y\}$. If F has the same rank everywhere, it is said to be of constant rank, $\operatorname{rank} F$. Rank of a map

A *smooth submersion* is a smooth map $F : X \to Y$ with $\operatorname{rank} F = \dim Y$. The simplest example of this is the Euclidean projection onto a subset of coordinates, Smooth submersion

Smooth immersion

$F : \mathbb{R}^{m+n} \to \mathbb{R}^m$, $F : (x_1, ..., x_{m+n}) \mapsto (x_1, ..., x_m)$. A *smooth immersion* is a smooth map $F : X \to Y$ with $\operatorname{rank} F = \dim X$. The simplest example of this is the Euclidean inclusion of coordinates, $F : \mathbb{R}^m \to \mathbb{R}^{m+n}$, $F : (x_1, ..., x_m) \mapsto (x_1, ..., x_{m+n})$. Of

Smooth embeddings

most importance to us will be *smooth embeddings*. These are smooth immersions $F : X \to Y$ that are also *topological embeddings*(appendix A.1.3), an immersion that is a homeomorphism when inheriting the subspace topology in its image $F(X) \subseteq Y$.

Smooth embeddings are particularly useful when we can embed the manifold into Euclidean space. This is useful from a technical perspective as it allows us to transfer tools from the simplified setting of Euclidean space onto the surface of the manifold easily, and from a computational point of view as it allows us to represent points on the surface of the manifold in a global way. Instead of having to pick a collection of charts to cover the manifold and represent, we can pick a global (but slightly redundant) coordinate system. In addition, from a deep learning perspective, embeddings make it much easier to define smooth maps between manifolds. Again if we were to use charts, we would need to pick more than one to cover most manifolds, and as such we would need to ensure that the function we define (usually a learnable function) smoothly agrees on the overlap of the charts. With a pair of embeddings for the input and output manifolds into Euclidean space, we can instead define a single smooth function from Euclidean space to Euclidean space and restrict its inputs and outputs to the embedded manifolds.

A question then remains, can we always embed a smooth manifold smoothly into Euclidean space? The answer is yes.

Whitney Embedding Theorem

THEOREM A.21. *Every smooth n dimension smooth manifold (with or without boundary) admits a proper smooth embedding into* $\mathbb{R}^{2n+1}$.

Proof. Whitney (1936, Theorem 1) ∎

Note the *proper* part of this definition additionally restricts the preimage of compact sets under the embedding to also be compact.

A.5.4. Local coordinates. The concepts of tangent vectors and differentials just introduced are very abstract. To understand them better, and make them computationally useful, we can introduce the concept of *coordinate vectors* to produce a basis on tangent spaces.

Coordinate vectors

For the special case of $\mathbb{R}^n$, we already know how to define a basis on the tangent space at $p \in \mathbb{R}^n$, $T_p\mathbb{R}^n$. We simply take the partial differentials with respect to each input at the given point, $\left.\frac{\partial}{\partial x_1}\right|_p, ..., \left.\frac{\partial}{\partial x_n}\right|_p$. Any derivation can then be expressed as

$$v = \sum_{i=1}^{n} v^i \left.\frac{\partial}{\partial x_i}\right|_p \qquad v \in T_p\mathbb{R}^n,\ v_i \in \mathbb{R}. \tag{A.55}$$

Their action on a smooth function $f \in C^\infty(\mathbb{R}^n)$ is simply

$$\left.\frac{\partial}{\partial x_i}\right|_p (f) = \frac{\partial f}{\partial x_i}(p). \tag{A.56}$$

Using this basis on $T_p\mathbb{R}^n$, we can place a basis on a tangent space of a manifold T_pX at point using a smooth chart $(U, \phi) \in \mathcal{A}_X$ containing that point. Remember that a smooth chart $\phi : U \to \mathbb{R}^n$ is a diffeomorphism onto its image. Therefore, the differential $\mathrm{d}\phi_p : T_pX \to T_{\phi(p)}\mathbb{R}^n$ is an isomorphism. We can define a basis on T_pX then as the *preimage* of the standard basis on $T_{\phi(p)}\mathbb{R}^n$ under ϕ. Let us then define the basis vector

$$\left.\frac{\partial}{\partial\phi_i}\right|_p = (\mathrm{d}\phi_p)^{-1}\left(\left.\frac{\partial}{\partial x_i}\right|_{\phi(p)}\right) = d(\phi^{-1})_{\phi(p)}\left(\left.\frac{\partial}{\partial x_i}\right|_{\phi(p)}\right). \tag{A.57}$$

Often, $\left.\frac{\partial}{\partial\phi_i}\right|_p$ is denoted as $\left.\frac{\partial}{\partial x_i}\right|_p$, obscuring the dependence on ϕ and potentially confusing it with the basis vector of $T_{\phi(p)}\mathbb{R}^n$ that it is related to, $\left.\frac{\partial}{\partial x_i}\right|_{\phi(p)}$. For further simplicity, the notation is often shortened further to

$$\left.\frac{\partial}{\partial\phi_i}\right|_p = \partial\phi_i|_p = \partial x_i|_p = \partial_i|_p \tag{A.58}$$

when the chart and local coordinates are implied. The action of one of these basis vectors on a function $f \in C^\infty(X)$ is then

$$\left.\frac{\partial}{\partial\phi_i}\right|_p f = \left.\frac{\partial}{\partial x_i}\right|_{\phi(p)}(f \circ \phi^{-1}). \tag{A.59}$$

We can then express any tangent vector $v \in T_pX$ as

$$v = \sum_{i=1}^{n} v_i \left.\frac{\partial}{\partial\phi_i}\right|_p \qquad v \in T_pX,\ v_i \in \mathbb{R} \tag{A.60}$$

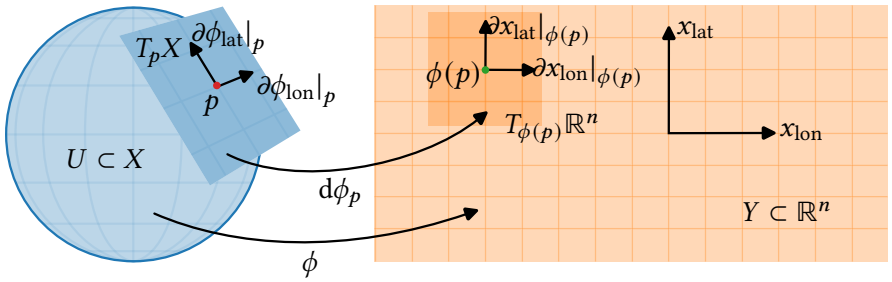


Figure A.11. Using a chart on the sphere that maps from points on the sphere to their latitude and longitude, we can induce a basis on the tangent space on the sphere in the latitude and longitude direction via the differential of the chart.

Differential in local coordinates

We can also study how the *differential in local coordinates* looks. First again let us consider the special case of $F : U \subseteq \mathbb{R}^n \to V \subseteq \mathbb{R}^m$, with the standard basis for the tangent spaces $\left.\frac{\partial}{\partial x_1}\right|_p, \ldots, \left.\frac{\partial}{\partial x_n}\right|_p$ and $\left.\frac{\partial}{\partial y_1}\right|_q, \ldots, \left.\frac{\partial}{\partial y_m}\right|_q$. Applying the definition of

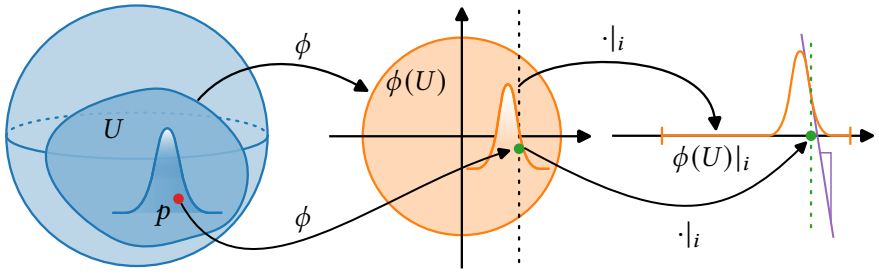


Figure A.12. Taking the differential of f, a bump function, evaluated at p in the i'th direction. Push the function f onto the chart U through ϕ to give the function $f \circ \phi^{-1} : \mathbb{R}^k \to \mathbb{R}$. Then restrict this function to the i'th component to give $f \circ \phi^{-1}\big|_i : \mathbb{R} \to \mathbb{R}$. Finally, evaluate the derivative of this as usual, in purple here at the i'th coordinate of $\phi(p)$.

the differential and the chain rule to a function $f \in C^\infty(\mathbb{R}^m)$, we get

$$\begin{aligned} \mathrm{d}F_p\left(\left.\frac{\partial}{\partial x_i}\right|_p\right) f &= \left.\frac{\partial}{\partial x_i}\right|_p (f \circ F) \\ &= \sum_{j=1}^{m} \frac{\partial f}{\partial y_j}(F(p)) \frac{\partial F_j}{\partial x_i}(p) \\ &= \left(\sum_{j=1}^{m} \frac{\partial F_j}{\partial x_i}(p) \left.\frac{\partial}{\partial y_j}\right|_{F(p)}\right) f. \end{aligned} \tag{A.61}$$

Therefore,

$$\mathrm{d}F_p\left(\left.\frac{\partial}{\partial x_i}\right|_p\right) = \sum_{j=1}^{m} \frac{\partial F_j}{\partial x_i}(p) \left.\frac{\partial}{\partial y_j}\right|_{F(p)}. \tag{A.62}$$

If we have $v = \sum_{i=1}^{n} v_i \left.\frac{\partial}{\partial x_i}\right|_p$, then its image under the differential, $w = \mathrm{d}F(v)$ in basis coefficients is

$$\begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = \begin{bmatrix} \frac{\partial F_1}{\partial x_1}(p) & \cdots & \frac{\partial F_1}{\partial x_n}(p) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1}(p) & \cdots & \frac{\partial F_m}{\partial x_n}(p) \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}. \tag{A.63}$$

Jacobian matrix

So we see in local coordinates the differential is associated with a matrix with entries $\frac{\partial F_j}{\partial x_i}$. This is exactly the *Jacobian matrix* of F. The differential is therefore the coordinate-free definition of the Jacobian.

In order to apply this to the differential of $F : X \to Y$, we again look at the behaviour with respect to charts. Pick two charts $(U, \phi) \in \mathcal{A}_X$, $p \in U$, and $(V, \psi) \in \mathcal{A}_Y$, $F(p) \in V$, and write that $\tilde{F} = \psi \circ F \circ \phi^{-1}$, so $F = \psi^{-1} \circ \tilde{F} \circ \phi$. We know that $\mathrm{d}\tilde{F}_{\phi(p)}$ in coordinates is the matrix described above, $\frac{\partial \tilde{F}_j}{\partial x_i}$. We can then

write

$$\mathrm{d}F_p\left(\left.\frac{\partial}{\partial\phi_i}\right|_p\right) = \mathrm{d}\left(\psi^{-1}\circ\tilde{F}\circ\phi\right)_p\left(\left.\frac{\partial}{\partial\phi_i}\right|_p\right) = \mathrm{d}\psi^{-1}{}_{F(p)}\circ\mathrm{d}\tilde{F}_{\phi(p)}\circ\mathrm{d}\phi_p\left(\left.\frac{\partial}{\partial\phi_i}\right|_p\right) \tag{A.64}$$

$$= \mathrm{d}\psi^{-1}{}_{F(p)}\circ\mathrm{d}\tilde{F}_{\phi(p)}\left(\left.\frac{\partial}{\partial x_i}\right|_{\phi(p)}\right) \tag{A.65}$$

$$= \mathrm{d}\psi^{-1}{}_{F(p)}\left(\sum_{j=1}^{m}\frac{\partial\tilde{F}_j}{\partial x_i}\left.\frac{\partial}{\partial y_j}\right|_{\psi(p)}\right) \tag{A.66}$$

$$= \sum_{j=1}^{m}\frac{\partial\tilde{F}_j}{\partial x_i}\left.\frac{\partial}{\partial\psi_j}\right|_p. \tag{A.67}$$

So again we see that the differential is associated in local coordinates with the Jacobian matrix $\frac{\partial\tilde{F}_j}{\partial x_i}$. The rank of this Jacobian matrix in the usual matrix sense is equal to the more abstract rank of the differential, $\dim\operatorname{im}\mathrm{d}F_p \subseteq T_{F(p)}N$.

One particular example of the differential in coordinates is in the effect on tangent vectors under a change in chart, called a *change in coordinate*. If we have p in two charts (U,ϕ), (V,ψ) and a vector v at p, in the two different chart bases, $v = \sum_{i=1}^n a_i\left.\frac{\partial}{\partial\phi_i}\right|_p = \sum_{i=1}^n b_i\left.\frac{\partial}{\partial\psi_i}\right|_p$, then the basis coefficients are related by $b_j = \sum_{i=1}^n\frac{\partial\psi\circ\phi_j^{-1}}{\partial x_i}a_i$. As a small abuse of notation I will write $\frac{\partial\psi\circ\phi_j^{-1}}{\partial x_i} = \frac{\partial\psi_j}{\partial\phi_i}$. Since we know this differential is an isomorphism, so must its Jacobian, i.e. is it full rank, so $\frac{\partial\psi_j}{\partial\phi_i}\in\mathrm{GL}(n)$.

Change in coordinate

A.5.5. Cotangent and tensor spaces. On a manifold, we denote the *cotangent space* at p, T_p^*X to be the dual of the tangent space

Cotangent space

$$T_p^*X = \left(T_pX\right)^*. \tag{A.68}$$

Elements of this are called *cotangent vectors*.

Cotangent vectors

With the tangent space and cotangent space defined at a point p, we can define also define the *tensor space* $(T_s^r)_pX$ as

Tensor space

$$(T_s^r)_pX = T_s^r(T_pX) = \underbrace{T_pX\otimes\ldots\otimes T_pX}_{r\text{ copies}}\otimes\underbrace{T_pX^*\otimes\ldots\otimes T_pX^*}_{s\text{ copies}} \tag{A.69}$$

Given a chart (U,ϕ), and the induced basis on the tangent space, $\left\{\partial\phi_i|_p\right\}$, we can define a basis on the cotangent space, $\left\{\mathrm{d}\phi^i\big|_p\right\}$ by $\mathrm{d}\phi^i\big|_p(\partial\phi_j\big|_p) = \delta^i_j$.

Previously, we saw how under a change of coordinates from (U,ϕ) to (V,ψ) the coefficients of a tangent vector $v\in T_pX$ transformed under the Jacobian of the differential. The coefficients of a cotangent vector also transform under the

Jacobian, but in the opposite way. For $v \in T_pX$, $v = v^i \partial\phi_i|_p = \tilde{v}^i \partial\psi_i|_p$ and $\omega \in T_p^*X$, $\omega = \omega_i \mathrm{d}\phi_i|_p = \tilde{\omega}_i \mathrm{d}\psi_i|_p$ we get the transforms

$$\underbrace{\tilde{v}^j = \frac{\partial\psi^j}{\partial\phi^i}(p)v^i}_{\text{Tangent vector transformation}} \qquad \underbrace{\omega_i = \frac{\partial\psi^j}{\partial\phi^i}(p)\tilde{\omega}_j}_{\text{Cotangent vector transformation}} \quad . \tag{A.70}$$

This transform definition is equivalent to saying that $\tilde{\omega}_j = \left(\frac{\partial\psi^j}{\partial\phi^i}(p)\right)^{-1}\omega_i$. Plugging this into the definition of the cotangent vectors and working it through in coordinates, we see that this leave the action of the cotangent vector on the tangent vector in coordinates invariant to the choice of coordinates, as we would expect.

Pullback

Previously we saw the definition of the *pushforward* of a tangent vector along a diffeomorphism $F : X \to Y$ from T_pX to T_pY. We can instead *pullback* a covector from T_pY^* to T_pX^* via the pullback $(F^*)_p : T_pY^* \to T_pX^*$, defined by its action on a vector $v \in T_pX$,

$$((F^*)_p(\omega))(v) = \omega((F_*)_p(v).) \tag{A.71}$$

A.6. Tensor Bundles

A.6.1. Tangent bundles and vector fields. We have just considered the tangent space of a point p on a manifold X. What if we want to consider the tangent space at every point on a manifold simultaneously? This leads us to the notion of the tangent *bundle*.

Tangent bundle

We construct the *tangent bundle* by taking the set *disjoint union* of the tangent spaces,

$$TX = \coprod_{p\in X} T_pX = \{(x, v) \ : \ x \in X, v \in T_xX\}. \tag{A.72}$$

Canonical projection

Along with this we have a *canonical projection* $\pi : TX \to X$, defined by $\pi : (x, v) \mapsto x$. The topology we place on this is *not* however the disjoint union topology. The disjoint union topology would leave each tangent space topologically disconnected from each other.

Natural topology

Instead, we give it a *natural topology* that renders the tangent bundle itself also a manifold of dimension $2n$.

The tangent bundle is a smooth manifold

Proposition A.22. *For any smooth n-dimension manifold X, the tangent bundle is a $2n$-dimension smooth manifold with a natural topology and smooth structure. With respect to this structure, the projection $\pi : TX \to X$ is smooth.*

Proof. To do this, we construct charts that satisfy the smooth chart lemma (lemma A.11).

We construct our charts on TX out of the atlas of the base manifold $\mathcal{A}_X$.

Take any chart (U, ϕ) in $\mathcal{A}_X$. We construct a new chart $\Phi : TU \to \mathbb{R}^{2n}$ ($TU = \coprod_{p\in U} T_pX$) by taking the coordinate functions $\phi_i : U \to \mathbb{R}$ and the induced basis on the tangent space T_pX, $\partial\phi_i|_p$. Defining

$$\Phi(p, v^i \partial\phi_i|_p) = (\phi_1(p), ..., \phi_n(p), v_1, ..., v_n). \tag{A.73}$$

This is a bijection as its inverse is

$$\Phi^{-1}(x_1, ..., x_n, v_1, ..., v_n) = \left(\phi^{-1}(x_1, ..., x_n), v^i \phi_i|_{\phi^{-1}(x_1,...,x_n)}\right). \tag{A.74}$$

For Ψ constructed similarly to Φ, the transition function $\Psi \circ \Phi^{-1}$ is given by

$$\Psi \circ \Phi^{-1}(x_1, ..., x_n, v_1, ..., v_n) = \left(\left(\psi \circ \phi^{-1}\right)(x_1, ..., x_n), \sum_{i=1}^{n} \frac{\partial \psi_1}{\partial \phi_i} v_i, ..., \sum_{i=1}^{n} \frac{\partial \psi_n}{\partial \phi_i} v_i\right) \tag{A.75}$$

which is smooth.

To satisfy the smooth chart lemma:

1. For any chart (TU, Φ), this is a bijection onto $\phi(U) \times \mathbb{R}^n \underset{\text{open}}{\subseteq} \mathbb{R}^n \times \mathbb{R}^n$, an open set of $\mathbb{R}^{2n}$.
2. Similarly, for (TU, Φ), (TV, Ψ), $\Phi(TU \cap TV)$ and $\Psi(TU \cap TV)$ are open subsets of $\mathbb{R}^n$.
3. The transition maps $\Psi \circ \Phi^{-1}$ are smooth as shown.
4. We can pick a countable cover of TX by taking a countable cover of X and use those charts to construct our charts on TX.
5. If we have $(p, v_1), (p, v_2)$ and a set $p \in U$, then $(p, v_1), (p, v_2) \in TU$. If we have $(p, v), (q, w)$, $p \neq q$, then from the charts of X we have U, V such that they are disjoint and $p \in U$, $q \in V$. Therefore $(p, v) \in TU$, and $(q, w) \in TV$, but TU, TV are disjoint.

To show that π is smooth, note that $\pi : (x, v) \mapsto x$. In charts (U, ϕ), (TU, Φ) this is simply a restriction to the first half of the coordinates, with is smooth. ∎

This object we have constructed is an example of a fibre bundle introduced earlier (definition A.7), as the triple (TX, π, X).

The sections of the tangent bundle are called *vector fields*. In the Euclidean setting, this amounts to continuous map from $\mathbb{R}^n \to \mathbb{R}^n$ (or open subsets of), that can be visualised as an "arrow" at each point. In the sense of the direction derivative discussed earlier, this akin to attaching a directional derivative to each point. A *vector field on a smooth manifold* is then a section of the tangent bundle instead.

Vector fields

Vector field on a smooth manifold

A *smooth vector field* is one that is a smooth function between X and TX. A *rough vector field* is one that is not.

Smooth vector field

Rough vector field

We will denote the space of vector fields as $\Gamma(TX)$. Sometimes it is written as $\mathfrak{X}(X)$.

Algebraic structure of the tangent bundle

In the previous section we saw that tangent spaces *at a point* can be viewed as a specific algebraic structure, namely as the *vector space* of *derivations at a point* over the *algebra* of smooth functions $(C^\infty(X), +, \cdot, \bullet)$. Recall that we had:

- $+ : C^\infty(X) \times C^\infty(X) \to C^\infty(X)$ to be pointwise addition of functions.
- $\cdot : \mathbb{R} \times C^\infty(X) \to C^\infty(X)$ to be pointwise multiplication by a scalar.

- $\bullet : C^\infty(X) \times C^\infty(X) \to C^\infty(X)$ to be pointwise multiplication of functions.

Let us then define:

- $\oplus : \Gamma(TX) \times \Gamma(TX) \to \Gamma(TX)$, $(\sigma \oplus \tau)(p) \mapsto \sigma(p) + \tau(p)$, pointwise vector addition of vector fields.
- $\odot : C^\infty(X) \times \Gamma(TX) \to \Gamma(TX)$, $(f \odot \sigma)(p) \mapsto f(p)\sigma(p)$, pointwise scaling of a vector field by a function.

We can construct a *ring* (definition A.15) from the smooth functions on a manifold as $(C^\infty(X), +, \bullet)$. The triple $(\Gamma(TX), \oplus, \odot)$ then satisfies the conditions of a *vector space* with respect to this *ring*. Vector spaces over rings are very different to vector spaces over fields, and are often called *modules*.

Modules

It is possible to define $\Gamma(TX)$ as a vector space over the field $\mathbb{R}$ by using *global* scaling of a vector field by a constant, but the *basis* of this vector space necessarily becomes uncountably infinite and difficult to work with. By contrast $\Gamma(TX)$ as a vector space over the ring $C^\infty(X)$ by using *local* scaling of a vector field by a function is more natural and pleasant to work with. See (Schuller, 2013).

As a summary:
T_pX is a *vector space* over the *field* $\mathbb{R}$ of *derivations at a point* of the *algebra* $C^\infty(X)$.
$\Gamma(TX)$ is a *module* over the *ring* $C^\infty(X)$ of *derivations* of the *algebra* $C^\infty(X)$.

This algebraic description of the space of vector fields in fact is an alternative way to define the space, in lieu of the previously presented construction.

Lie brackets

One additionally useful method of combining vector fields is the *Lie bracket*. For two vector fields $V, W \in \Gamma(TX)$ the *Lie bracket* is defined by its action on a function $f \in C^\infty(X)$,

Lie bracket

$$[V, W]f = V(Wf) - W(Vf) \tag{A.76}$$

The Lie bracket of two smooth vector fields is again a smooth vector field.

In addition to being a product, the Lie bracket is

- *antisymmetric*, $[V, V] = 0$

Antisymmetric

- and satisfies the *Jacobi identity*, $[V, [W, X]] + [W, [X, V]] + [X, [V, W]] = 0$.

Jacobi identity

More general than this application to vector fields, any product which is also antisymmetric and satisfies the Jacobi identity is said to be a *Lie bracket*. An algebra whose product is a Lie bracket is termed a *Lie algebra*.

Lie bracket
Lie algebra

As it stands, $(\Gamma(TX), \oplus, \odot)$ is *not* a vector space over a field, it is a vector space over a ring. As such, the combination of the tangent bundle and the Lie bracket is *not* a Lie algebra. It is important to point this out as later we will meet a class of manifold called *Lie groups* for which a subset of vector fields *do* form a *Lie algebra* under the Lie bracket.

Local bases for vector fields

We can write a vector field using its *component functions* at a given point using the basis vectors induced by a chart (U, ϕ), Component functions

$$V_p = \sum_{i=1}^{n} V_i(p) \left.\frac{\partial}{\partial \phi_i}\right|_p . \tag{A.77}$$

This defines the component function $V_i(p) : U \to \mathbb{R}$. A vector field is smooth if and only if its component functions in each chart are smooth.

A chart induces *coordinate vector fields* on $U \subseteq X$ by the assignment $p \mapsto \left.\frac{\partial}{\partial \phi_i}\right|_p$. Coordinate vector fields

These are denoted as $\frac{\partial}{\partial \phi_i} = \partial \phi_i = \partial_i$ where the rest of the information is implicit. Together, the coordinate vector fields $\partial\phi_1, \ldots \partial\phi_n$ produce a *smooth local frame* called a *coordinate frame*. A *smooth local frame* a set of smooth (restricted to an open set U) vector fields that at each point span the tangent space. We can similarly define *continuous local frames*, with the vector fields being only continuous. Coordinate frame, Smooth local frame, Continuous local frames

If $U = X$, this is a global frame, and we can have a global choice of basis for the tangent space, although these do not exist for most manifolds. For example,

THEOREM A.23. *There exists no non-vanishing smooth vector field on even dimension n-spheres.* Hairy ball, or Hedgehog theorem

Since there are no non-vanishing smooth vector fields, none of these can be a basis field globally, and so there is no smooth global basis.

For some specific manifolds there exists a set of d vector fields that span the tangent bundle. Such manifolds are called *parallelisable* Parallelisable

The lack of existence of a global basis for $\Gamma(TX)$ initially makes it difficult to see how to parametrise smooth vector fields on manifolds in a computer, as patching together local chart while maintaining smoothness without a basis is tricky.

The failure of the existence of a smooth global basis is exactly due to $\Gamma(TX)$ being a module rather than a vector space. The solution to this lies in the study of modules.

DEFINITION A.24. *Let $(M, \oplus, \odot)$ be a module over the ring $(R, +, \cdot)$. A GENERATOR of the module is a subset $G \subset M$ such that for any element $m \in M$ there exists coefficients $(r_g)_{g \in G}$, $r_g \in R$ such that* Generator of a module

$$m = \sum_{g \in G} r_g g. \tag{A.78}$$

If there exists a generator of M that is finite we say that it is FINITELY GENERATED *Finitely generated*

The generator of a module is similar to the basis of a vector space, but we give up injectivity between coefficients in the ring and elements of the module. That is, there may be more than one set of coefficients in the ring that lead to the same element in the module.

Fortunately for working with vector fields on manifolds, we have the following result.

Modules of vector fields on smooth manifolds are finitely generated

THEOREM A.25. *Let X be a second countable smooth manifold. Then the module of smooth vector fields $\Gamma(TX)$ over smooth functions $C^\infty(X)$ is a finitely generated module.*

Proof. Falorsi and Forré (2020, see). The proof is an application of the fact that we can for a second countable manifold arrive finite size local trivialization of the manifold. We can use these charts to create local bases for the tangent bundle, and a *partition of unity* (see definition A.9) subordinate to the trivialisation to smoothly glue together local vector fields represented in these local bases into a global vector field. ∎

This gives us a strategy to parametrise smooth vector fields on manifolds.

1. Pick a finite generator of $\Gamma(TX)$, $(V_i)_{i=1}^m$
2. Parametrise smooth functions $f^i : X \to \mathbb{R}$.
3. Define the vector field as
$$V = f^i V_i \tag{A.79}$$

Using this result therefore, we will always be able to pick a finite set of fields from which we can parametrise the space of vector fields.

Tangent and non-tangent vector fields

Supposed we have X, and a submanifold $Y \subset X$, immersed or embedded, with inclusion map $\iota : Y \to X$. The differential of the inclusion map $\mathrm{d}\iota : \Gamma(TY) \to \Gamma(TX)$ identifies $\Gamma(TY)$ with a linear subspace of $\Gamma(TX)$. We can then project any vector field $V \in \Gamma(TX)$, restricted to the surface of Y, into $\Gamma(TY)$ by projecting onto this linear subspace. We say that a vector field is $V \in \Gamma(TX)$ *tangent to a submanifold* if this projection leave the vector field unchanged, and that is it *normal to a submanifold* if the vector field is zero under this projection.

Tangent to a submanifold

Inward-pointing vector field

For manifolds with boundary, X, we say that an *inward-pointing vector field* is a vector field $V \in \Gamma(TX)$ if at every point on its restriction to the boundary it is normal to the boundary and the tangent vectors point *into* the interior. We define the converse for *outward-pointing vector fields*.

Outward-pointing vector fields

A.6.2. Tensor bundles and fields. The tangent bundle is one of the most fundamental bundles we deal with in geometry. We can however introduce a wider range of bundles by bundling arbitrary tensors over the manifold (appendix A.4.3). All of these bundles will be examples of *vector bundles*, which are fibre bundles where the typical fibre is isomorphic to $\mathbb{R}^n$ with the usual vector structure.

In constructing these bundles, it is very handy to have the following lemma.

Vector bundle chart lemma

LEMMA A.26. *Let X be a smooth manifold (with boundary), and suppose that for each $p \in X$ we have E_p, a k-dimension vector space. Let $E = \coprod_{p\in X} E_p$, and $\pi : E \to X$ be the map such that $\pi : E_p \mapsto p$. If we have the following:*

1. *an open cover of X, $\{U_a\}_{a\in A}$.*
2. *for each $a \in A$, a bijection $\Phi : \pi^{-1}(U_a) \to U_a \times \mathbb{R}^k$.*

3. *for each $a, b \in A$ with $U_a \cap U_b \neq \emptyset$, a smooth map $\tau_{ab} U_a \cap U_b \to \mathrm{GL}(k, \mathbb{R})$ such that the map $\Phi_a \circ \Phi_b^{-1} : (U_a \cap U_b) \times \mathbb{R}^k \to (U_a \cap U_b) \times \mathbb{R}^k$ is of the form*

$$\Phi_a \circ \Phi_b^{-1}(p, v) = (p, \tau_{ab}(p)v) \tag{A.80}$$

Then E has a unique topology and smooth structure making it into a smooth manifold with or without boundary and a smooth rank-k vector bundle over M; with π as projection and $\{(U_a, \Phi_a)\}$ as smooth local trivializations.

Using this we could have constructed the target bundle by creating the transition maps as we did, and applying this lemma.

Identically to the tangent bundle then, we can build the *cotangent bundle*, denoted by T^*X, from cotangent spaces. This too is a $2n$-dimension manifold and a vector bundle. We build the natural coordinates again out of the basis induced by charts, in this case from the dual basis of the tangent basis. Sections of this are called *covector fields*, denoted $\omega \in \Gamma(T^*X)$. The action of a covector field $\omega \in \Gamma T^*X$ on a vector field $V \in \Gamma(TX)$ is defined by

Cotangent bundle

Covector fields

$$\omega(V) : X \to \mathbb{R}, \qquad \omega(V)(p) = \omega_p(V_p). \tag{A.81}$$

One central example of a covector field is the *gradient of a smooth function.* Remember that a smooth function $f \in C^\infty(X)$ is a smooth function $f : X \to \mathbb{R}$. The *differential* of this is then a smooth function $\mathrm{d}f : \Gamma(TX) \to \mathbb{R}$, and therefore exactly a covector field. We in particular term the differential of a smooth function as its gradient. This is in contrast to normal multivariate calculus where the gradient of a function is a vector field.

Gradient of a smooth function

More generally we can bundle tensor spaces at every point to give *tensor bundles*, denoted by $T_q^p TX$. We gain construct this using the vector bundle chart lemma, and use a basis everywhere from the basis of the tangent bundle and cotangent bundle, the same as in appendix A.4.4.

Tensor bundles

There are a series of special cases.
The *tangent bundle* is denoted $TX = T_0^1 TX$
The *cotangent bundle* is denoted $T^*X = T_1^0 TX$
The *bundle of covariant k-tensors* is denoted $T^k T^*X = T_k^0 TX$.
The *bundle of contravariant k-tensors* is denoted $T^k TX = T_0^k TX$.
The *bundle of symmetric k-tensors* is denoted $\Sigma^k T^*X = \mathrm{Sym}\, T_k^0 TX$.
The *bundle of alternating k-tensors* is denoted $\Lambda^k T^*X = \mathrm{Alt}\, T_k^0 TX$.

A *tensor field* is then a *smooth section* of a tensor bundle. The space of smooth sections of a tensor bundle is denoted $\Gamma\left(T_q^p T\right)$. $\Gamma\left(T_0^0 TX\right)$ is simply the space of smooth functions $C^\infty(X)$.

Tensor field

Bases for sections of tensor bundles

For tensor bundles we can again define smooth local *frames*. If we have a smooth local frame for the tangent bundle $e_1, ..., e_n$, then we can define the local *dual coframe* as the vector fields $\epsilon^1, ..., \epsilon^n$ such that $\epsilon^i(e_j) = \delta^i_j$. If we have a chart (ϕ, U) and therefore tangent frames $\partial\phi_i$, then the dual frames are the differentials of the

Dual coframe

chart, $\mathrm{d}\phi_i$. From this we can construct smooth local frames on any tensor bundle in terms of a local frame of the tangent bundle. For a $T^p_q TX$ bundle, the smooth local frames are given by

$$e_{i_1} \otimes ... \otimes e_{i_p} \otimes \epsilon^{j_1} \otimes ... \otimes \epsilon^{j_q}. \tag{A.82}$$

For a given chart (U, ϕ), a tensor field F can be expressed locally as

$$F = F^{i_1,...,i_p}_{j_1,...,j_q} \partial\phi_{i_1} \otimes ... \otimes \partial\phi_{i_p} \otimes \mathrm{d}\phi^{j_1} \otimes ... \otimes \mathrm{d}\phi^{j_q} \tag{A.83}$$

where the $F^{i_1,...,i_p}_{j_1,...,j_q} : U \to \mathbb{R}$ are smooth functions.

Smooth multilinear maps over $C^\infty(X)$

One of the most important interpretations of tensor fields is as *smooth multilinear maps over* $C^\infty(X)$. For a $T^p_q TX$ tensor field $F \in \Gamma\left(T^p_q TX\right)$, this can be seen as the map

$$F : \underbrace{T^*X \times ... \times T^*m}_{p \text{ covector fields}} \times \underbrace{TX \times ... \times Tm}_{p \text{ vector fields}} \to C^\infty(X). \tag{A.84}$$

This map is smooth, and multilinear. In fact, all smooth multilinear maps of this form are induced by tensor fields (**??**Lemma 12.24 extends naturally to mixed tensor fields]lee2013smooth).

Pullbacks of tensor fields

Previously we say how using the differential we could pushforward and pullback tangent and cotangent vectors between the tangent spaces of two manifolds related by a smooth function $F : X \to Y$. For generic tensor fields we can't define the pushforward, but we can define, the *pullback of a covariant tensor field* pointwise. For $A \in \Gamma(T^kT^*Y)$, the *pointwise pullback* of $\mathrm{d}F(A) \in \Gamma(T^kT^*X)$ is given by

Pullback of a covariant tensor field

$$[\mathrm{d}F^*(A)]_p(v_1, ..., v_n) = A_p(\mathrm{d}F_p(v_1), ..., \mathrm{d}F_p(v_k)) \tag{A.85}$$

for $v_1, ..., v_k \in \Gamma(TX)$.

A.6.3. Lie derivatives. Now we have defined tensor fields, a question remains - can we take derivatives of these? In Euclidean space, if we restrict ourselves to vector fields, there is a natural notion of the *direction derivative* of the vector field. For a vector field $W \in \Gamma TX$ and a vector $V \in \mathbb{R}^n$, the *directional derivative of a vector field* is given by

Directional derivative of a vector field

$$[D_v W](p) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} W_{p+tv} = \lim_{t\to 0} \frac{W_{p+tv} - W_p}{t} \tag{A.86}$$

for $D_v : \Gamma(TX) \to \Gamma(TX)$ and this can be show to be equal to

$$[D_v W](p) = D_v W^i(p)\, \partial x^i\big|_p \tag{A.87}$$

where on the right-hand side, D_v is acting as the derivative of the *scalar* component functions of the vector field, $W^i \in C^\infty(\mathbb{R}^n)$.

A few things hint that this will not immediately generalise to manifold settings. Firstly, the sum on the right-hand side violates the einsum notation convention,

indicating we are not respecting geometric rules. Secondly, we are taking the difference between W_{p+tv} and W_p, elements of two *different* tangent spaces, an operation that is not possible. We get away with this in Euclidean space as we can canonically identify all tangents spaces with one another, but this does not work on manifolds.

To define this kind of derivative on manifolds, we need a few more definitions.

Integral curves

A *curve on a manifold* is a smooth function from an open subset of $\mathbb{R}$ to the manifold, $\gamma : (a, b) \subset \mathbb{R} \to X$. The derivative of this curve is an element of the tangent space of the manifold, $\left[\frac{\mathrm{d}}{\mathrm{d}t}\ \gamma\right](t) \in T_{\gamma(t)}X$. We call the point $\gamma(0) \in X$ the *starting point of a curve.*

Curve on a manifold

Starting point of a curve

An *integral curve of a vector field* $V \in \Gamma(TX)$ is curve for which its derivative at any point is equal to the vector field at that point,

Integral curve of a vector field

$$\frac{\mathrm{d}}{\mathrm{d}t}\ \gamma(t) = V_{\gamma(t)}. \tag{A.88}$$

Viewed in any chart, this defines an *ordinary differential equation* for the coordinates of the curve. Under the right conditions, ordinary differential equation have smooth, unique solutions.

THEOREM A.27. *For any vector field $V \in \Gamma(TX)$ and any point $p \in X$, there exists an integral curve $\gamma : (-\epsilon, \epsilon) \to X$ of V such that $\gamma(0) = p$, for $\varepsilon > 0$.*

Existence of integral curves

Using the definition of integral curves we can relate vector fields on two different manifolds. For a smooth map between manifolds $F : X \to Y$ and two vector fields $V \in \Gamma(TX)$ and $W \in \Gamma(TY)$, we say they are *F-related vector fields* if and only if for any integral curve γ of V, then $F \circ \gamma$ is an integral curve of W.

F-related vector fields

We can shift the time value of an integral curve as we please.

LEMMA A.28. *For an integral curve $\gamma : (a, b) \to X$ of $V \in \Gamma(TX)$ and $s \in \mathbb{R}$, the function $\hat{\gamma} : (a + s, b + s) \to X$ given by $\hat{\gamma}(t) = \gamma(s + t)$ is also an integral curve of V.*

Translation lemma of integral curves

Flows

Theorem A.27 does not guarantee that the integral curve for a vector field will exist for all time, only on $(-\epsilon, \epsilon)$. For now, let us assume that $V \in \Gamma(TX)$ is such that the integral curves of it *do* exist for all times. We define the *flow of a vector field* V as the map $\theta_t : X \to X$ as the map that sends each point on the manifold to $\gamma_p(t)$ where $\gamma_p(0) = p$, the unique integral curve that starts at p.

Flow of a vector field

Using lemma A.28 we see that the composition $\theta_t(p) \circ \theta_s(p) = \theta_{t+s}(p)$ combining this with the fact that $\theta_0(p) = \mathrm{id}_X(p) = p$, we see that this forms a group with the additive structure of the real numbers, $(\mathbb{R}, +)$. Consequently, we see that the map $\theta : \mathbb{R} \times X \to X$ also has additive group structure, since $\theta(s, \theta(t, p)) = \theta(s + t, p)$ and $\theta(0, \cdot) = \mathrm{id}_X$.

This motivates the definition of a *global flow*. A global flow is a smooth map $\theta : \mathbb{R} \times X \to X$ which under composition of the second argument has the additive

Global flow

group structure of the reals in the first. I.e.

$$\theta : \mathbb{R} \times M \to M, \quad \theta(s, \theta(t,p)) = \theta(s+t,p), \quad \theta(0,\cdot) = \mathrm{id}_X \tag{A.89}$$

Infinitesimal generator of a global flow

The *infinitesimal generator of a global flow* is defined as a (not necessarily smooth) vector field $V \in \Gamma(TX)$ such that

$$V_p = \left.\frac{\mathrm{d}}{\mathrm{d}t}\,\theta(t,p)\right|_{t=0}. \tag{A.90}$$

We can show that all global flows are generated by smooth vector fields.

Global flows are generated by smooth vectro fields

Proposition A.29. *Let $\theta : \mathbb{R} \times X \to X$ be a global flow on X. Then the generator of θ is a smooth vector field $V \in \Gamma(TX)$ and each curve $\theta(\cdot, p) : \mathbb{R} \to X$ is an integral curve of V.*

Local flow
Complete vector field

The converse of this is not true. Not every smooth vector field generates a global flow. A *local flow* $\theta : \mathcal{D} \to M$ is the flow generated by a smooth vector field $V \in \Gamma(TX)$ that is only defined on a subset $\mathcal{D} \subset \mathbb{R} \times X$. A *complete vector field* is a vector field that generates a global flow.

Lie derivatives of vector fields

With this definition of a flow, we can produce a coherent definition of a derivative of a vector field. Instead of differentiating with respect to a single tangent vector, we will differentiate with respect to *another vector field.*

Lie derivative of a vector field

Definition A.30. *For vector fields $V, W \in \Gamma(X)$, the Lie derivative of W with respect to V is given by*

$$(\mathcal{L}_V W)_p = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mathrm{d}(\theta_{-t})_{\theta_t(p)}(W_{\theta_t^V(p)}) \tag{A.91}$$

$$= \lim_{t\to 0} \frac{\mathrm{d}(\theta_{-t})_{\theta_t(p)}\left(W_{\theta_t(p)}\right) - W_p}{t} \tag{A.92}$$

where this limit exists.

where θ_t is the flow induced by V.

Unpacking this a little, this says we take a point $p \in X$ and evolve it under the flow of the vector field V for time t. At this new point we evaluate W. In order to bring this point into the tangent space of p, we map this under the differential of the flow for time $-t$. We then compare this tangent vector to the value of W at p. The limit of this operation as $t \to 0$ defines the derivative.

Existence of the Lie derivative for vector fields

Lemma A.31. *Let X be a smooth manifold, and $V, W \in \Gamma(TX)$ be smooth vector fields. Then the Lie derivative $(\mathcal{L}_V W)_p$ exists for all p, and $\mathcal{L}_V W$ is a smooth vector field.*

Computing such a derivative however seems a tricky task. Fortunately however, this derivative is exactly given by the *Lie bracket*, giving this a geometric interpretation.

The Lie derivite of vector fields is the Lie bracket

Theorem A.32. *Let X be a smooth manifold and $V, W \in \Gamma TX$. Then*

$$\mathcal{L}_V W = [V, W]. \tag{A.93}$$

This gives the Lie derivative therefore all the structure of a *Lie algebra.*

Lie derivatives of tensor fields

With the Lie derivative of vector fields defined, we can define the Lie derivative for general tensor fields.

Lie derivative of a tensor field

DEFINITION A.33. *For vector field $V \in \Gamma(X)$ and $A \in \Gamma(T^k T^* X)$ tensor field on X, the LIE DERIVATIVE of A with respect to V is given by*

$$(\mathcal{L}_V A)_p = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \mathrm{d}(\theta_t)_p^*(A_{\theta_t(p)}) \qquad = \lim_{t\to 0} \frac{\mathrm{d}(\theta_t)_p^*\left(A_{\theta_t(p)}\right) - A_p}{t} \tag{A.94}$$

$$= \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (\theta_t^* A)(p) \qquad = \lim_{t\to 0} \frac{(\theta_t^* A)(p) - A_p}{t} \tag{A.95}$$

where this limit exists.

This definition is almost identical, except we have used the pullback of a covariant tensor field.

Existence of the Lie derivative for tensor fields

LEMMA A.34. *Let X be a smooth manifold, $V \in \Gamma(TX)$ a smooth vector field, and $A \in \Gamma(T^k T^* X)$ a smooth tensor field. Then the Lie derivative $(\mathcal{L}_V A)_p$ exists for all p, and $\mathcal{L}_V A \in \Gamma(T^k T^* X)$ is a smooth tensor field.*

This definition of the Lie derivative is compatible with a number of operations:

- For $f \in C^\infty(X)$,
$$\mathcal{L}_V f = Vf, \tag{A.96}$$
since $C^\infty(X) = \Gamma\left(T^0 T^* X\right)$.
- For $A \in \Gamma(T^k T^* X)$, $B \in \Gamma(T^l T^* X)$,
$$\mathcal{L}_V(A \otimes B) = (\mathcal{L}_V(A)) \otimes B + A \otimes (\mathcal{L}_V B) \tag{A.97}$$
- For smooth vector fields $X_1, ..., X_n \in \Gamma(TX)$,
$$\mathcal{L}_V(A(X_1, ..., X_n)) = \mathcal{L}_V(A)(X_1, ..., X_n) + A(\mathcal{L}_V X_1, ..., X_1) + ... + A(X_1, ..., \mathcal{L}_V X_n).$$

A.7. LIE GROUPS

Next we introduce *Lie groups*. These are interesting as manifolds in their own right, but are incredibly important in the study of manifolds more generally. Firstly, they act as *global* symmetry groups. For example the group SE(3) is the symmetry group of the sphere $\mathcal{S}_2$. They also act as *local* symmetry groups on manifolds. We have already seen how elements of the group GL(n) acts as change of basis transforms under a change in chart for vectors expressed in a basis.

Lie group

A *Lie group* G is a smooth manifold that is also an algebraic, and in this case also topological, group. That is, they come equipped with an operator $\cdot : G \times G \to G$, an identity element $e \in G$ such that $e \cdot g = g \cdot e = g$ for all $g \in G$, and an inverse operator $i : G \to G$ such that $g \cdot i(g) = g \cdot g^{-1} = e$, and that all these are continuous maps. Typically, the notation of $\cdot$ is suppressed.

Subgroup

A *subgroup* of a Lie group is a subset $H \subset G$ of the group that is itself a group under $\cdot$, and that has a topology and smooth structure that make it an *immersed submanifold* of the parent group.

The most important example of a Lie group for our purposes is the group $\mathrm{GL}(v, \mathbb{R})$ and its subgroups. There do exist Lie group that are subgroups of $\mathrm{GL}(v, \mathbb{C})$, but we will not consider these. This is the group of $n \times n$ invertible matrices with real entries, and this forms a group under matrix multiplication. As we have seen, this provides the space of possible change of basis transforms of a basis for a tangent space under a change in chart.

Important subgroups of $\mathrm{GL}(n)$ include

- $\mathrm{GL}(n)^+$, the subgroup of $\mathrm{GL}(n)$ with positive determinant.
- $\mathrm{O}(n)$, the subgroup of $\mathrm{GL}(n)$ with determinant of ± 1.
- $\mathrm{SO}(n)$, the subgroup of $\mathrm{GL}(n)$ with determinant 1.

A.7.1. Homogeneous spaces. For a subgroup H of a group G, the set of elements $gH = \{g \cdot h \ : \ h \in H\}$ is called the *left coset* of g. The left cosets of G by H partition G into equivalence classes of G, if two elements g_1, g_2 have the same left coset $g_1 H = g_2 H$.

Left coset

Homogeneous space

The resulting *quotient space*, $X = G/H$ is called a *homogeneous space*. The action of G on X, $\cdot : G \times X \to X$ is defined by mapping the original group action to the quotient space. For a point $x \in X$ and a representative element $g_x \in G$, $g \cdot x = g g_x H = g_y H = y$. This action is independent of the choice of g_x made, and so is well-defined. This action is a *transitive action* as using it the group G maps every point in X to every other point in X.

Transitive action

A common example of this is that the n-sphere is a homogeneous space of $\mathrm{O}(n)$, $\mathcal{S}_n \sim \mathrm{O}(n)n + 1/\mathrm{O}(n)$.

A.7.2. Group actions and equivariant maps. While the action of a group on a homogeneous space is a global symmetry, manifolds can have symmetries that don't map every point to every other point. A symmetry, or a *G-action*, is a map $\cdot : G \times X \to X$ such that

G-action

$$g_1 \cdot (g_2 \cdot x) = (g_1 \cdot g_2) \cdot x \quad \text{for all } x \in X \ g_1, g_2 \in G \tag{A.98}$$

and

$$e \cdot x = x \quad \text{for all } x \in X \text{ and } e \text{ is the identity element of } G. \tag{A.99}$$

If there is only one orbit, the action is transitive and the space a homogeneous space. Such a symmetry partitions the manifold into a set of *orbits*,

Orbits

$$G \cdot x = \{g \cdot x \ : \ g \in G\}. \tag{A.100}$$

Stabiliser

For each point, we can define the *stabiliser* as

$$G_x = \{g \cdot x = x \ : \ g \in G\}, \tag{A.101}$$

the set of elements that leave x in place. An action is said to be a *free action* if the stabiliser of each point is the identity element only, i.e. that no non-identity group element leaves a point stationary. Free action

Suppose that we have two manifolds X, Y, and a group G, and each manifold has a left G-action. A map $F : X \to Y$ is said to be *equivariant* with respect to G if Equivariant

$$F(g \cdot x) = g \cdot F(x) \quad \text{for all } x \in X,\ g \in G. \tag{A.102}$$

Sometimes it is said that the function F *intertwines* the action of G on X and Y. Intertwines

A.7.3. Semi-direct products. In addition to the typical direct product of group actions, there is a construction called the *semi-direct product* of two Lie groups. For Lie groups H and N, supposed we have a left action of H on N, $\theta : H \times N \to N$. If for all $h \in H$ $\theta_h : N \to N$ is an isomorphism on N, then we can define a joint group action of H and N together by Semi-direct product

$$(h, n) \cdot (h', n') = (h \cdot h', n \cdot \theta_h n). \tag{A.103}$$

Using this action we define a new group $G = H \ltimes_\theta N$. Often this is written as $H \ltimes N$ where the action θ is implied. As a manifold this has the direct product structure $H \times N$.

The most important example of this the *Euclidean group*, $\mathrm{E}(n)$. This is given by the semi-direct product of $\mathrm{T}(n)$, $\mathbb{R}^n$ considered as a Lie group of translation actions, and $\mathrm{O}(n)$, to give $\mathrm{E}(n) = \mathrm{O}(n) \ltimes \mathbb{R}^n$. The action of $\mathrm{O}(n)$ on $\mathrm{T}(n)$ is given by the usual rotation of vectors. $\mathrm{E}(n)$ action on $\mathbb{R}^n$ by $(R, t) \cdot x = Rx + t$ for $R \in \mathrm{O}(n),\ t \in \mathrm{T}(n),\ x \in \mathbb{R}^n$. Euclidean group

A.7.4. Representations. Many Lie groups can be realised as subgroups of $\mathrm{GL}(n)$. We can extend the analysis of groups in general by considering *(linear) representations* of a group, which is any *group homeomorphism* $\rho : G \to \mathrm{GL}(n)$, i.e. that (linear) representations

$$\rho(g)\rho(g') = \rho(g \cdot g') \tag{A.104}$$

where the left hand side is matrix multiplication. It is also possible to have complex representations, but again we will not consider these.

A *faithful* representation is one that is an injection into $\mathrm{GL}(n)$. Many representations are not injections however, and will map multiple group elements of G to the same element of $\mathrm{GL}(n)$. Faithful

The theory of representations if vast and spans harmonic analysis, differential equations, number theory, and has major applications in differential geometry also.

A.8. Riemannian Metrics

We now turn our study to *Riemannian* manifolds. So far we have defined manifolds with notions of topology, smoothness, and differentiation. However, we are yet to introduce geometry, in simple terms the study of lengths and angles. In metric spaces such ideas are studied through the direct definition of a metric. On manifolds

however the more natural object to study these concepts through is a *Riemannian metric*.

For the rest of this section, we assume that X is a smooth manifold with or without boundary.

A.8.1. Riemannian metrics. The most important concept to transfer from Euclidean space is the notion of lengths of vectors, and inner products between vectors.

Riemannian metric

DEFINITION A.35. *A RIEMANNIAN METRIC g is a smooth section of the symmetric covariant 2-tensor field on X. That is,*

$$g \in \Sigma^2 T^* X \tag{A.105}$$

Riemannian manifold

DEFINITION A.36. *A RIEMANNIAN MANIFOLD (with or without boundary) (X, g) is a smooth manifold (with or without boundary) X along with a Riemannian metric $g \in \Sigma^2 TX$.*

Distance functions

It should be noted a Riemannian metric is *not* the same as a metric in the metric space sense and we will distinguish these by calling them *distance functions*.

By looking at the metric restricted to a particular point on the manifold, $g_p : T_pX \times T_pX \to \mathbb{R}$, we see that this defines an inner product on the tangent space, as the metric is a symmetric tensor and therefore $g_p(v, v) \geq 0$ for $v \in T_pX$.

We can express the metric in the coordinates of a chart (U, ϕ) as

$$g = g_{ij}\, \mathrm{d}\phi^i \otimes \mathrm{d}\phi^j = g_{ij}\, \mathrm{d}\phi^i\, \mathrm{d}\phi^j \tag{A.106}$$

The simplest metric is the *Euclidean metric*,

$$g = \delta_{ij}\, \mathrm{d}\phi^i\, \mathrm{d}\phi^j = (\mathrm{d}\phi^i)^2 \tag{A.107}$$

and from this we recover all the usual notations of Euclidean space. In particular note that the inner product defined for two vectors $v, w \in T_pX$ is given by

$$g_p(v, w) = \delta_{ij} v^i w^j = \sum_{i=1}^{n} v^i w^i, \tag{A.108}$$

the usual dot product. Note how the last expression violates the index convention of einsum notation, necessitating the use of explicit sums, revealing how the usual definition of the dot product leave some inaccuracy in geometric terms. Often an alternative notation is used, $g_p(v, w) = \langle v, w \rangle_g$.

The metric naturally extends to a map on vector fields,

$$g : \Gamma(TX) \times \Gamma(TX) \to C^\infty(X), \quad V, W \in \Gamma(TX). \tag{A.109}$$

Every smooth manifold admits a metric (Lee, 2013, proposition 13.3). Every smooth manifold admits a huge variety of metrics in fact, and as such there is no canonical choice of Riemannian metric.

We can define the *product metric* on a product of Riemannian manifolds by applying each metric to the components of the tangent vectors that cam from each manifold in the product, and so define *product Riemannian manifolds*.

Product metric

Product Riemannian manifolds

From the definition of a Riemannian metric, we can then define the *length* of a tangent vector as

Length

$$\|v\|_g = \langle v, v\rangle_g^{1/2} = g_p(v,v)^{1/2}, \tag{A.110}$$

called the *Riemannian norm* and the *angle between tangent vectors* as

Riemannian norm

Angle between tangent vectors

$$\cos\theta = \frac{\langle v, w\rangle}{|v|_g |w|_g}. \tag{A.111}$$

We can say that two tangent vectors are *orthonormal* if $\langle v, w\rangle = 0$. An *orthonormal frame* on a manifold is a frame for which at each point on the manifold, the basis on the tangent space consists of vectors orthonormal to each other. The existence of smooth *locally orthonormal frames* is guaranteed (Lee, 2013, corollary 13.8), but the existence of smooth *globally orthonormal frames* is not guaranteed, and in fact only exist for a very small class of manifolds, such as Euclidean space.

Orthonormal

Orthonormal frame

Locally orthonormal frames

Globally orthonormal frames

A.8.2. Tangent-cotangent isomorphism. The Riemannian metric on a manifold allows us to specify an isomorphism between the tangent and cotangent spaces of a manifold.

Define the map

$$g^\downarrow : TX \to T^*X, \quad [g^\downarrow(v)(w)] = g_p(v,w), \quad v, w \in T_pX \tag{A.112}$$

which maps elements of the tangent space to elements of the cotangent space. Using this then we define a bundle homeomorphism between TX and T^*X, and so also create an isomorphism between $\Gamma(TX)$ and $\Gamma(T^*X)$ via

$$g^\downarrow : \Gamma(TX) \to \Gamma(T^*X), \quad [g^\downarrow(V)](W)_p = g_p(V,W), \quad V, W \in \Gamma(TX) \tag{A.113}$$

In a chart (U, ϕ) where $V = V^i \partial\phi_i$, $W = W^i \partial\phi_i$, we have that

$$[g^\downarrow(V)](W) = g_{ij} V^i W^j \tag{A.114}$$

and so in charts

$$g^\downarrow(V) = g_{ij} V^i \,\mathrm{d}\phi^j = V_j \,\mathrm{d}\phi^j \tag{A.115}$$

where $V_j = g_{ij} V^i$. This is the motivation for the notation $g^\downarrow$, as this operation is said to *lower an index* of the coefficients of the vector field in charts. Since this operation is unique on a manifold, we will use as shorthand

Lower an index

$$V^\downarrow = g^\downarrow(V). \tag{A.116}$$

The inverse of this operation is called *raising and index*. We define

Raising and index

$$g^\uparrow = (g^\downarrow)^{-1} : \Gamma(T^*X) \to \Gamma(TX). \tag{A.117}$$

In a chart (U, ϕ) for a covector field $\omega \in \Gamma(T^*X)$ this is given by

$$g^\uparrow(\omega) = g^{ij} \omega_j \partial\phi_i = \omega^i \partial\phi_i \tag{A.118}$$

where $\omega^i = g^{ij}\omega_j$, and g^{ij} are the coefficients of the matrix inverse of g_{ij}, such that $g^{ij}g_{jk} = \delta^i_k$, and we have the same shorthand that

$$\omega^\uparrow = g^\uparrow(\omega). \tag{A.119}$$

Gradient of a smooth function on a Riemannian manifold

The most important use of this isomorphism is to define the *gradient of a smooth function on a Riemannian manifold* as

$$\operatorname{grad}_g f = (\mathrm{d}f)^\uparrow, \quad f \in C^\infty(X). \tag{A.120}$$

This makes the gradient the unique *vector field* (not *covector field*) satisfying

$$g\Big(\operatorname{grad}_g f, X\Big) = Xf, \quad f \in C^\infty(X),\ X \in \Gamma(TX), \tag{A.121}$$

and in a chart (U, ϕ) it is given by

$$\operatorname{grad}_g f = g^{ij}\frac{\partial f}{\partial \phi_i}\partial\phi_i \tag{A.122}$$

A.8.3. Pullback metrics and submanifolds. If we have a smooth map $F : X \to Y$ and a Riemannian metric on Y, then the *pullback of the Riemannian metric* F^*g is a Riemannian metric on X if and only if F is a smooth immersion (Lee, 2013, proposition 13.9).

Pullback of the Riemannian metric

Riemannian isometry
Isometric
Locally isometric
Riemannian geometry
Flat

If we have two Riemannian manifolds (X, g) and $(\tilde{X}, \tilde{g})$, and a diffeomorphism $F : X \to \tilde{X}$, then F is a *Riemannian isometry* if $F^*\tilde{g} = g$. If there exists such an isometry between two manifolds, we say they are *isometric*. If this property only holds locally for two manifolds, but not globally, then they are said to be *locally isometric*. The study of properties that are invariant under global or local Riemannian isometries is called *Riemannian geometry*. One such property of a manifold is if it is *flat*. A manifold is flat if it is locally isometric to Euclidean space equipped with the Euclidean metric.

In particular, when the isometry is a map between the same manifold, $F : X \to X$, this is an *automorphism* of the manifold, a self-isometry. We denote this set of isomorphisms $\operatorname{Isom}_g(X)$. These are the intrinsic isometries of the manifold. For example for *homogeneous spaces* (see appendix A.7.1) of Lie groups, this will be the original Lie group. For most manifolds it will not be a global isometry group.

Induced Riemannian metric

The most useful case for pullback metrics is for inducing metrics on submanifolds. If we have a Riemannian manifold (X, g) and a submanifold $S \subset X$, immersed or embedded, with or without a boundary, then this manifold automatically inherits a metric, ι^*g, the *induced Riemannian metric*, where $\iota : S \to X$ is the inclusion map.

Nash embedding theorem

Previously we saw that the *Whitney embedding theorem* (see theorem A.21) allowed us to always find an embedding of a smooth n-dimension manifold in to at most $2n + 1$-dimension Euclidean space, while preserving the smooth structure of the manifold. A cornerstone result of Riemannian geometry says that we can do one step better with Riemannian manifolds, and preserve the metric as well.

THEOREM A.37. *Let (X, g) be an n-dimension Riemannian manifold. Then there exists a C^k isometric embedding of X into $\mathbb{R}^m$, $f \in C^k(X, \mathbb{R}^m)$, such that $g = f^*g_e$, where g_e is the Euclidean metric on $\mathbb{R}^m$, and $m \geq n(n+1)(3n+11)/2$.*

The bound on the dimension of Euclidean space needed to embed a manifold into is not strict. For example, the n-torus can be embedded into $2n$ Euclidean space, and the n-sphere into $n + 1$ Euclidean space.

This theorem is incredibly useful. On the theoretical side, it allows us at any time to transfer the manifold into Euclidean space as an embedding. We can then (with care) use mathematical tools that apply to Euclidean space and use them for analysis on manifolds.

On the practical side, this gives us a good way to represent Riemannian manifolds in a computer. We can simply pick an isometric embedding of the manifold. Particularly, this means that if we are careful, computational geometric operations will be compatible. For example, if we have a function $f : X \to \mathbb{R}$ and compose this with the inverse inclusion to get a function $\iota^{-1} \circ f : \iota(X) \subset \mathbb{R}^m \to \mathbb{R}$, we can represent this in a computer, and we can then also apply operations such as automatic differentiation to the function. As long as we project the resulting tangent vectors back into the tangent space of X immersed in the tangent space of $\mathbb{R}^m$, then this will give us exactly the function $\iota^{-1} \circ \operatorname{grad}_g f$. I.e.

$$\operatorname{proj}_{TX \subset T\mathbb{R}^m} \operatorname{grad}(\iota^{-1} \circ f) = \iota^{-1} \circ \operatorname{grad}_g f \tag{A.123}$$

Where grad is the usual Euclidean gradient, equal to $\operatorname{grad}_{g_e}$, the Riemannian gradient with Euclidean metric. This also applies to more complex operators, such at the divergence and Laplace-Beltrami operators.

Riemannian distance function

Using the metric, we can also define a sense of the length of a given curve on a manifold, and from this, a notion of a distance between two points on a manifold. Let a continuous function $\gamma : I \to X$ be a *curve on* X where $I \subset \mathbb{R}$ is an interval. A *smooth curve* is a one that is smooth as a function. A *regular curve* is a curve for which the velocity $\gamma'(t) \neq 0$ everywhere, so that it has no "kinks". An *admissible curve* is a curve that can be split into a partition of regular curves. A *reparametrisation of a curve* is a map $\varphi : I' \to I$ such that $\tilde{\gamma} = \gamma \circ \varphi$ is also a curve.

Curve on X
Smooth curve
Regular curve
Admissible curve
Reparametrisation of a curve

We then define the *length of a curve* for any admissible curve $\gamma : [a, b] \to X$ as

Length of a curve

$$L_g(\gamma) = \int_a^b \|\gamma'(t)\|_g \, \mathrm{d}t. \tag{A.124}$$

This definition obeys a few natural properties; it is additive if we split a curve into sections, it is independent of any reparametrisation of $[a, b]$, and it is invariant under isometries.

We define the *speed of a curve* as $\|\gamma'(t)\|_g$, and a *unit-speed curve* as one such that its speed is one everywhere. Any regular curve can be reparametrised to have unit speed.

Speed of a curve
Unit-speed curve

We have the important result

Proposition A.38. *If X is a connected smooth manifold, then any two points on X can be connected by an admissible curve.*

Points can be connected by curves

Riemannian distance function

Using this, we can define the *Riemannian distance function* as

$$d_g(p,q) = \inf_{\gamma \text{ is an admissible curve}} L_g(\gamma), \quad p,q \in X. \tag{A.125}$$

This definition is also invariant under isomorphisms.

This distance function defines a metric in the sense of metric spaces, and so equipping a smooth manifold with a Riemannian metric turns it into a metric space as well. Bringing us full circle, the *topology generated by this metric* is exactly the manifold topology we chose initially in constructing the topological manifold.

A.9. Differential Forms, orientations, and integration

In this section we will study a particular subset of tensor fields, namely section of the alternating tensor bundles $\Lambda^K T^*X = \operatorname{Alt} T^0_k TX$. These are typically called *differential k-forms*, or k-forms, and the space of them is denoted $\Omega^k(X) = \Gamma\left(\Lambda^K T^*X\right)$. The integer k is called the *degree* of the k-form.

Differential k-forms
Degree

While this may seem odd, it turns out that differential forms are exactly the type of objects to which we can apply a coherent theory of integration to on smooth manifolds, and also differentiation, but in a different sense to that of Lie derivatives. This notion is called the *exterior derivative* and it is a generalisation of the notion of the gradient of a smooth function. This sense of derivative generalises the notions of divergence, cross products and curl.

In appendix A.4.7 we saw two ways of mapping alternating tensors to alternating tensors of different degrees, the *wedge product* and the *interior product*. Both generalise immediately to differential forms by applying them pointwise. For $\omega \in \Omega^k(X)$, $\eta \in \Omega^l(X)$, the *wedge product* $\wedge : \Omega^k(X) \times \Omega^l(X) \to \Omega^{k+l}(X)$ is defined as

Wedge product

$$(\omega \wedge \eta)(p) = \omega(p) \wedge \eta(p), \tag{A.126}$$

as long as $k + l \leq n$. For $\omega \in \Omega^k(X)$ and $V \in \Gamma(TX)$ the *interior product* $\lrcorner : \Omega^k(X) \times \Gamma(TX) \to \Omega^{k-1}(X)$ is defined as

Interior product

$$(\omega \lrcorner V)(p) = \omega(p) \lrcorner V(p), \tag{A.127}$$

as long as $1 \leq k \leq n$.

We can express a differential form in a chart via basis functions. 1-forms are exactly the space of covector fields on a smooth manifold. For the chart (U, ϕ) the basis of 1-forms is given by the derivatives of the coordinate functions, $\mathrm{d}\phi_1, \ldots, \mathrm{d}\phi_n$. For a k-form, we can produce the *elementary k-forms* similar to generating the basis for alternating tensors (see appendix A.4.7) by taking wedge products of $\mathrm{d}\phi_1, \ldots, \mathrm{d}\phi_n$. We can then write the k-form in coordinates as

Elementary k-forms

$$\omega = \sum_{I \in \mathbb{I}} \omega_I \, \mathrm{d}\phi^I, \tag{A.128}$$

remembering that $\mathbb{I}$ is the set of strictly increasing multi-indexes, and $\mathrm{d}\phi^I = \mathrm{d}\phi^{i_i} \wedge \ldots \wedge \mathrm{d}\phi^{i_k}$, and ω_I are smooth functions.

We can apply naturally the pullback of tensor fields to differential forms. One result that gives a hint as to how we will define coordinate free integration is by looking at the pullback of top forms in charts. A *top form* is an n-form on an n-dimension smooth manifold. The basis if a top form at every point is one dimension and so a top form in a chart (U, ϕ) can be written as

Top form

$$\omega = f \, \mathrm{d}\phi^i \wedge ... \wedge \mathrm{d}\phi^n \tag{A.129}$$

for a smooth function $f \in C^\infty(X)$.

Pullback of top-degree forms

PROPOSITION A.39. *For a pair of n-dimension smooth manifold X, Y, a smooth map $F : X \to Y$, two charts $(U, \phi) \in \mathcal{A}_X$ and $(V, \phi) \in \mathcal{A}_Y$, and a continuous real-valued function $f : V \cap F(U) \to \mathbb{R}$, then*

$$F^*(f \, \mathrm{d}\psi^1 \wedge ... \wedge \mathrm{d}\psi^n) = (\det DF)(f \circ F) \, \mathrm{d}\phi^1 \wedge ... \wedge \phi^n \tag{A.130}$$

where $\det DF$ *is the* JACOBIAN MATRIX *of F in the two charts (appendix A.5.4).*

In particular, if we look at two charts of the same manifold, and set $F = \mathrm{id}_X$,

COROLLARY A.40. *For an n-dimension smooth manifold X, two charts $(U, \phi), (V, \phi) \in \mathcal{A}_X$, then*

$$\mathrm{d}\psi^1 \wedge ... \wedge \mathrm{d}\psi^n = \det\left(\frac{\partial \psi^j}{\partial \phi^i}\right) \mathrm{d}\phi^1 \wedge ... \wedge \phi^n \tag{A.131}$$

A.9.1. Exterior derivatives. Exterior derivatives are a type of differential operator that we can apply to k-forms to turn then into $k+1$-forms. We have already met the start point for exterior derivative in the form of the gradient of a smooth function. A smooth function is a 0-form, the *differential* (see appendix A.6.2) maps it to a covector field, all of which are 1-forms.

More generally, we can build on of this to extend the differential to higher forms. The *exterior derivative* of a k-form is defined by its action in a chart. For $\omega \in \Omega^l(X)$ in a chart (U, ϕ),

Exterior derivative

$$\mathrm{d}\omega = \sum_{I \in \mathbb{I}} \mathrm{d}\omega_I \wedge \mathrm{d}\phi^I \tag{A.132}$$

$$= \sum_{I \in \mathbb{I}} \sum_{i=1}^{n} \frac{\partial \omega_I}{\partial \phi^i} \mathrm{d}\phi^i \wedge \mathrm{d}\phi^I \tag{A.133}$$

Where $\mathrm{d}\omega^I$ is the differential of the I coordinate function, and the second line is its expansion into basis covector fields. The exterior derivative

- Is linear over $\mathbb{R}$.
- For $\omega \in \Omega^k(X)$, $\eta \in \Omega^l(X)$,

$$\mathrm{d}(\omega \wedge \eta) = \mathrm{d}\omega \wedge \eta + (-1)^k \omega \wedge \mathrm{d}\eta. \tag{A.134}$$

- $d \circ \mathrm{d} = 0$.
- For a smooth map $F : X \to Y$ and $\omega \in \Omega^k(X)$,

$$\mathrm{d}F^*(\omega) = F^*(\mathrm{d}\omega). \tag{A.135}$$

Closed k-form
Exact form

We call a k-form ω for which $\mathrm{d}\omega = 0$ a *closed k-form*, and if there is a $(k-1)$-form η such that $\omega = \mathrm{d}\eta$ an *exact form*. Since $d \circ \mathrm{d} = 0$, all exact forms are closed, but not necessarily the converse.

A.9.2. Orientations. Recall that the definition of an *orientation of a vector space* (see appendix A.4.4) involves picking a particular basis to label as "positively oriented" and defining the orientation of all other bases relative to this, and that we can equivalently define this by choice of a *volume form* (see appendix A.4.7).

We want to apply this definition to the space of vector fields of smooth manifolds. Here, instead of an isolated vector space, we have a smooth continuous collection of vector spaces. Remembering that we commonly *do not have a smooth global basis for vector fields* (see theorem A.23), we make this definition by patching together

Local orientation

orientations of local bases on the tangent space. A *local orientation* is defined by a *continuous local frame* (see appendix A.6.1), given by the orientation at each point defined by the local frame.

Orientation of a smooth manifold

DEFINITION A.41. *An ORIENTATION ON A SMOOTH MANIFOLD, O_X is a continuous choice of orientation of each tangent space T_pX such that for every neighbourhood on X the orientation is defined by a continuous local frame.*

The existence of an orientation of a smooth manifold however is far from guaranteed, to the extent that we classify manifolds as orientable or non-orientable.

Orientable manifold
Oriented manifold

An *orientable manifold* is a manifold for which an orientation exists. An *oriented manifold* is a smooth manifold, along with a particular choice of orientation.

Relative to an orientation O_X, a chart of a smooth manifold (U, ϕ) is said to be positively oriented if the *coordinate frame* (see appendix A.6.1) of the chart is

Consistently oriented

positively oriented. An atlas of X is said to be *consistently oriented* if the *determinant of the Jacobian* transition map between two charts, $\det \frac{\partial \phi^i}{\partial \phi^j}$ is positive everywhere. The implication of this is that the *coordinate frames* (see appendix A.6.1) of all that charts will have the same orientation. This in fact lets us have an alternative definition of an orientation.

Orienation determinied by an atlas

PROPOSITION A.42. *For a smooth manifold X and a CONSISTENTLY ORIENTED ATLAS $\mathcal{A}_X$, there exists a unique orientation on X, O_X, such that every chart in $\mathcal{A}_X$ is positively oriented.*

Similar to a vector space, we can also define an orientation by an n-form.

Volume form

DEFINITION A.43. *Let X be an n-dimension smooth manifold. A VOLUME FORM $\omega \in \Omega^n(X)$ is a NON-VANISHING n-form on X.*

The *non-vanishing* part of this definition is key, it means our volume form is not allowed to assign zero volume to any point on the manifold.

Orientation determined by an n-form

PROPOSITION A.44. *Let X be an n dimension smooth manifold. Any non-vanishing n-form $\mu \in \Omega^n(X)$ determines an orientation on X by setting any ordered basis for a tangent space, $e_1, ..., e_n \in T_pX$, to be positively oriented if*

$$\mu(e_1, ..., e_n) > 0. \tag{A.136}$$

If X is orientable, then such a form exists.

For a pair of oriented manifolds, X, Y, a diffeomorphism $F : X \to Y$ is *orientation-preserving* if the pullback of a positively oriented volume forms $F^*\omega$ is also positively oriented, or *orientation-reversing* they are negatively oriented. Note we can flip which maps preserve and reverse orientation reversing the orientation on one of X and Y.

Orientation-preserving
Orientation-reversing

One important property of manifolds *with boundary* X is that the orientation on the boundary ∂X is entirely determined by the orientation on $\operatorname{Int} M$.

Definition A.45. *Let X be an n-dimension manifold with boundary with non-vanishing volume form $\omega \in \Omega^n(X)$, and let $N \in \Gamma(TX)$ be an* outward-pointing vector field *(see appendix A.6.1). Then the $(n-1)$-form $\iota^*_{\partial X}(N \lrcorner \omega)$ is a volume form on ∂X, and induces the same orientation for any choice of N, where $\iota^*_{\partial X} : \Gamma(TX) \to \Gamma(T\partial X)$ is the pullback of a differential form from X to ∂X.*

Induced orientation

Riemannian volume form

The choice of volume form on an orientable manifold is not unique. We can however, make a unique choice on Riemannian orientable manifolds. On Riemannian manifolds, we are always able to pick smooth, *orthonormal* local frames on the manifold.

Proposition A.46. *Ley X be an n-dimension orientable Riemannian manifold (with or without boundary). Then there exists a smooth n-form, $\mathrm{d}\operatorname{vol}_g \in \Omega^n(X)$, such that for any local, oriented, orthonormal frame $e_1, ..., e_n$,*

$$\mathrm{d}\operatorname{vol}_g(e_1, ..., e_n) = 1. \tag{A.137}$$

Riemannian volume form

This is called the Riemannian volume form

In a chart (U, ϕ), the Riemannian volume form has the form

$$\mathrm{d}\operatorname{vol}_g = \sqrt{\det(g_{ij})}\, \mathrm{d}\phi^1 \wedge ... \wedge \mathrm{d}\phi^n, \tag{A.138}$$

a very handy expression for being able to compute with. Our choice of notation $\mathrm{d}\operatorname{vol}_g$ will be motivated shortly, but it does *not* mean that the Riemannian volume form is the exterior derivative of an $(n-1)$-form.

A.9.3. Integration on manifolds. Finally, we arrive at how to define integration on manifolds. The objects that we are going to integrate are *volume forms* (see definition A.43), and we will do so by relating integration on a smooth manifold to integration in a chart.

Definition A.47. *Let X be a smooth oriented manifold and $\omega \in \Omega^k(X)$ with* compact support, *i.e. it is non-zero only on a compact set. Let (U, ϕ) be a smooth chart of X such that $\operatorname{supp}\omega \subset U$. Then the integral of ω is defined as*

Integration of a compactly supported n-form

$$\int_U \omega \overset{1}{=} \int_{\phi(U)} (\phi^{-1})^*\omega \overset{2}{=} \int_{\phi(U)} f_\omega\, \mathrm{d}\phi^1 \wedge ... \wedge \mathrm{d}\phi^n \overset{3}{=} \int_{\phi(U)} f_\omega\, \mathrm{d}\phi^1 ... \mathrm{d}\phi^n \tag{A.139}$$

Lets unpack this a little. Step 1 pulls back the n-form onto Euclidean space. Step 2 rewrites the n-form by its expression in coordinates in the chart, using the canonical

n-form on Euclidean space given by $\mathrm{d}x^1 \wedge \ldots \wedge \mathrm{d}x^n$, which is a volume form. Step 3 equates this integration by the Euclidean volume form to multiple Lebesgue integration. Using the orientable property of the manifold, it can then be shown this integration is independent of the choice of chart.

To then integrate generic n-forms, we split the form into a series of compactly supported forms.

Integration of compactly supported n-forms

PROPOSITION A.48. *Let X be an oriented manifold, and $\omega \in \Omega^n(X)$ a COMPACTLY SUPPORTED n-form. Let $\{(U_a, \phi_a)\}_{a=1}^k$ be a finite set of charts covering the manifold, and $\{(U_a, \phi_a), \rho_a\}_{a=1}^k$ a PARTITION OF UNITY (see definition A.9) subordinate to this cover. Then the integral of ω is given by*

$$\int_M \omega = \sum_{a=1}^{k} \int_{U_a} (\phi_a^{-1})^* (\rho_a \omega) \tag{A.140}$$

and this does not depend on the choice of chart or partition of unity

This integral obeys a number of properties

1. It is *linear*, for $\omega, \eta \in \Omega^n(X)$, $a, b \in \mathbb{R}$,
$$\int_X a\omega + b\eta = a \int_X \omega + b \int_X \eta \tag{A.141}$$
2. A *reverse in orientation* causes a reverse in integral sign. Denote $-X$ as X with the reverse orientation. Then
$$\int_X \omega = -\int_{-X} \omega \tag{A.142}$$
3. Positive volume forms have *positive integral*. If ω is a positively oriented volume form, then $\int_X \omega > 0$.
4. It is *diffeomorphism invariant*. For oriented manifolds X, Y, an n-form $\omega \in \Omega^n(Y)$, and an orientation preserving diffeomorphism $F : X \to Y$, $\int_Y \omega = \int_X F^*\omega$. If F is orientation-reversing, then $\int_Y \omega = -\int_X F^*\omega$.

Integrable forms on manifolds

volume forms define measures

The definition of integration in proposition A.48 motivates our choice of notation for the Riemannian volume form by $\mathrm{d}\,\mathrm{vol}_g$. This, or any other volume form ω, defines a *measure in the measure-theoretic sense* (see appendix B.1) of the *Borel σ-algebra* (see appendix B.1) of the open sets on the manifold. On a chart (U, ϕ), for any open set $V \subseteq U$, the local measure μ_U is defined by

$$\mu_{\omega,U}(V) = \int_V \omega = \int_{\phi(V)} f_\omega \, \mathrm{d}\phi^1 \wedge \ldots \wedge \phi^n \tag{A.143}$$

We then patch together the local measures into a global measure via a partition of unity $((U_a, \phi_a))$ and hence the d in $\mathrm{d}\,\mathrm{vol}_g$ denotes that this is integration against a measure.

Proposition A.48 also only allows us to define the integral of *compactly supported* n-forms. This is somewhat limiting, and not strictly necessary. Taking a hint from the measure theoretic measure a volume form denotes, we can extend the types of forms that are integrable.

For a given volume form on a manifold ω, we can write any n-form η as $\eta = f\omega$ for $f \in C^\infty(X)$. Let μ_ω be the measure induced by ω. We say η is L^p integrable with respect to μ_ω if

$$\left(\int_X |f|^p \mu_\omega\right)^{\frac{1}{p}} < \infty \tag{A.144}$$

Stoke's theorem

The most central result of integration on manifolds is *Stoke's theorem.*

THEOREM A.49. *Let X be an orientable smooth n-dimension manifold with boundary, and ω a compactly supported $n-1$-form on X. Then* Stoke's theorem

$$\int_X \mathrm{d}\omega = \int_{\partial X} \omega \tag{A.145}$$

where ∂X has the STOKE'S ORIENTATION (see definition A.45), and ω on the right-hand side is understood as $\imath^*_{\partial X}\omega$, the pullback on ω onto the boundary.

This is a very generic theorem that generalises the fundamental theory of calculus, and a number of other classical results.

For example, is we set $X = [a, b]$ and pick a function $f \in \mathbb{C}^\infty([a, b])$, then Stoke's theorem states

$$\int_{[a,b]} \mathrm{d}f = \int_{\partial[a,b]} f = \int_{\{\{a\},\{b\}\}} f = f(b) - f(a) \tag{A.146}$$

which is exactly the fundamental theorem of calculus.

Integration on Riemannian manifolds

Adding the structure of a Riemannian metric to an orientable manifold allows us to extend the results of integration further. The primary reason for this is the canonical choice we can make for the volume form we use for integration, namely the *Riemannian volume form* (see proposition A.46). Using this canonical choice, we can produce a canonical correspondence $* : C^\infty(X) \to \Omega^n(X)$ between smooth functions $f \in C^\infty(X)$ on the manifold and n-forms on the manifold, namely

$$\star f = f \,\mathrm{d}\,\mathrm{vol}_g\,. \tag{A.147}$$

We see then that the integral of a smooth function on a Riemannian manifold in a chart (U, ϕ) is simply given by

$$\int_U f \,\mathrm{d}\,\mathrm{vol}_g = \int_{\phi(U)} f\sqrt{g_{ij}}\,\mathrm{d}\phi^1 \wedge \ldots \wedge \mathrm{d}\phi^n. \tag{A.148}$$

Divergence on a Riemannian manifold

We use this correspondence to define the *divergence on a Riemannian manifold* $\operatorname{div} : \Gamma(TX) \to C^\infty(X)$ by

$$\operatorname{div} X = \star^{-1} \mathrm{d}\big(X \lrcorner \mathrm{d}\operatorname{vol}_g\big) \tag{A.149}$$

and using Stoke's theorem we arrive at a generalisation of the divergence theorem,

Divergence theorem

THEOREM A.50. *Let (X, g) be an oriented Riemannian manifold with boundary, and $V \in \Gamma(TX)$ with compact support. Then*

$$\int_X (\operatorname{div} V)\, \mathrm{d}\operatorname{vol}_g = \int_{\partial X} \langle V, N\rangle_g \, \mathrm{d}\operatorname{vol}_{\tilde{g}} \tag{A.150}$$

where N is the outward-pointing normal vector field on ∂X, $\tilde{g}$ is the induced metric on ∂X and X on the right-hand side is restricted to ∂X.

The divergence has a very geometric interpretation; it tells us how much a particular piece of space contracts or expands under a vector field. If D is a *compact, regular domain* of X, that is, a compact submanifold of X of the same dimension as X, then the rate of change of the volume of D under the *flow induced by* X (see appendix A.6.3) is exactly the integral of the divergence of X over D,

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \operatorname{vol}\Big(\theta_t^X(D)\Big) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} \int_{\theta_t^X(D)} \mathrm{d}\operatorname{vol}_g = \int_D (\operatorname{div} X)\, \mathrm{d}\operatorname{vol}_g \tag{A.151}$$

Hodge star

There are a number of other ways of defining the divergence on a manifold. We can generalise the $\star$ operator, called the *Hodge star* operator, to map any k-form to an $(n-k)$-form. For any $\omega, \eta \in \Lambda^k T^*X$,

$$\omega \wedge \star\eta = \langle \omega, \eta\rangle_g \, \mathrm{d}\operatorname{vol}_g \,. \tag{A.152}$$

We can then define the divergence as

$$\operatorname{div}(X) = \star^{-1} \mathrm{d} \star X. \tag{A.153}$$

We can also define it via the *Lie derivative* as

$$\frac{1}{2} \operatorname{tr}_g(\mathcal{L}_X g) \tag{A.154}$$

and also via the *Levi-Civita connection* (see appendix A.10.7). These three are equivalent definitions.

In a particular chart (ϕ, U) the Riemannian divergence can be expressed as

$$\operatorname{div} V = \frac{1}{\sqrt{\det g}} \partial\phi_i \Big(\sqrt{\det g}\; V_i\Big) \tag{A.155}$$

where V_i are the components of V in the chart ϕ.

A.9.4. Measures induced by volume forms. In order to work with measure theory on manifolds (appendix B.1), and therefore deal with probability formally, we need to define a base measure on a Riemannian manifold. We can construct this on a Riemannian manifold using the Reisz representation theorem and the integration defined by the volume form.

THEOREM A.51. *Let X be a locally compact Hausdorff space, and let Λ be a positive linear functional on $C(X)$. Then there exists a σ-algebra, Σ_X, on X containing all open sets on X, and there exists a unique positive measure μ on Σ_X which represents Λ in the sense that:* Riesz-Markov-Kakutani representaiton theorem

1. *$\Lambda f = \int_X f \, \mathrm{d}\mu$ for every $f \in C(X)$, in the Lebesgue integration sense.*
2. *$\mu(K) < \infty$ for every compact set $K \subset X$.*
3. *For every $A_x \in \Sigma_X$, we have that*
$$\mu(A_x) = \inf\{\mu(V) \ : \ E \subset V, \ V \text{ open}\} \tag{A.156}$$
4. *The relation*
$$\mu(E) = \sup\{\mu(K) \ : \ K \subset E, \ K \text{ compact}\} \tag{A.157}$$
holds for every open set E, and for every $E \subset \Sigma_X$ with $\mu(E) < \infty$.
5. *If $E \in \Sigma_X$, $A \subset E$, and $\mu(E) = 0$, then $A \in \Sigma_X$.*

Proof. Rudin (1987, Theorem 2.14) ∎

To apply this, we simply let $\Lambda : f \mapsto \int_M f \, \mathrm{d}\,\mathrm{vol}_g$, and the result follows. This allows us to place a base measure on a Riemannian manifold against which we can define densities, with a σ-algebra compatible with the topology of the manifold.

This is an example of a *Radon measure*, defined by properties 2, 3, 4 and 5. Radon measure

We therefore define the *Riemannian volume measure* to be the measure induced by the Riemannian volume form by theorem A.51. Riemannian volume measure

A.10. CONNECTIONS, GEODESICS, AND PARALLEL TRANSPORT

Previously we saw in the *Lie derivative* (see definition A.33) one way of differentiating tensor fields with respect to a vector field on a manifold. Intuitively, the Lie derivative measures how a tensor field changes as one moves along the flow defined by a vector field. We are now going to introduce a different way of differentiating tensors with respect to a vector field, namely via *connections*. Intuitively, this is a generalisation of the notion of a *directional derivative* (see appendix A.5.1) to tensor fields. The Lie derivate fails to be a directional derivative, most obviously because it is not linear with respect to multiplication by a smooth function,

$$\mathcal{L}_{fX} \neq f\mathcal{L}_X, \quad f \in C^\infty(X), \ X \in \Gamma(TX) \tag{A.158}$$

To recover a notion of directional derivative then, we define a new type of object with all the explicit properties we want it to have.

DEFINITION A.52. *Let X be a smooth manifold. Let $T^k_l X$ be a tensor bundle over X. Then a connection is a map*

$$\nabla : \Gamma(TX) \times \Gamma\left(T^k_l X\right) \to \Gamma\left(T^k_l X\right), \tag{A.159}$$

usually denoted

$$\nabla_X A, \quad X \in \Gamma(TX),\ A \in \Gamma\left(T^k_l X\right) \tag{A.160}$$

satisfying

1. *∇ is LINEAR in the first argument over THE MODULE $(\Gamma(TX), \oplus, \odot)$ (see appendix A.6.1),*
$$\nabla_{f_1 V + f_2 W} A = f_1 \nabla_V A + f_2 \nabla_W A \tag{A.161}$$
for $f_1, f_2 \in C^\infty(X),\ V, W \in \Gamma(TX),\ A \in \Gamma\left(T^k_l X\right)$
2. *∇ is LINEAR in the second argument over THE FIELD $\mathbb{R}$,*
$$\nabla_V (aA + bB) = a\nabla_V A + b\nabla_V B \tag{A.162}$$
for $a, b \in \mathbb{R},\ V \in \Gamma(TX),\ A, B \in \Gamma\left(T^k_l X\right)$.
3. *∇ satisfies the product rule*
$$\nabla_X (fA) = f\nabla_X A + (Xf)A \tag{A.163}$$
for $f \in C^\infty(X),\ V \in \Gamma(TX),\ A \in \Gamma\left(T^k_l X\right)$.

Importantly, the choice of connection on a smooth manifold is *not unique*. Different choices of connection will lead to different properties. We can however prove that a connection exists on any smooth manifold.

A.10.1. Connections on the tangent bundle and tensor bundles. Since there are many different tensor bundles, it would seem that we need to pick a connection for each one. Fortunately, if we make a choice of connection on the tangent bundle, we can extend this choice to any tensor bundle.

If we have a connection on the tangent bundle $\nabla : \Gamma(TX) \times \Gamma(TX) \to \Gamma(TX)$, then we can express it in a chart (U, ϕ) on coordinate fields by

$$\nabla_{\partial\phi_i} \partial\phi_j = \Gamma^k_{ij} \partial\phi_k. \tag{A.164}$$

Coordinate expression of a connection

For two vector fields $V, W \in \Gamma(TX)$, using the properties of the connection, then we have in the chart that the *coordinate expression of a connection* is

$$\nabla_V W = \left(V\left(W^k\right) + V^i W^j \Gamma^k_{ij}\right)\partial\phi_k. \tag{A.165}$$

To extend this to any tensor bundles we note a few things.

1. On smooth functions, the action of any connection is given by the usual derivative, $\nabla_V f = Vf$ for $f \in C^\infty(X)$, $V \in \Gamma(TX)$.

2. ∇ obeys a product rule over tensor products,

$$\nabla_V(A \otimes B) = A \otimes (\nabla_V B) + (\nabla_V A) \otimes B \tag{A.166}$$

3. For $V, W \in \Gamma(TX)$ and $\omega \in \Gamma(T^*X)$,

$$\nabla_V \omega(W) = [\nabla_V \omega](W) + \omega(\nabla_V W) \tag{A.167}$$

Using these we can clearly extend the connection on the tangent bundle to any tensor bundle, using 4. to define the connection on covector fields, and 2. to build up the operation to any tensor field from there.

In addition, ∇ commutes with the trace operation,

$$\operatorname{tr}(\nabla_V A) = \nabla_V(\operatorname{tr} A) \tag{A.168}$$

A.10.2. Covariant derivatives. We can view a connection in a slightly alternative manner by not passing the vector field we want to differentiate with respect to. In this case we see it as a map

$$\nabla : \Gamma\left(T^k_l X\right) \to \Gamma\left(T^k_{l+1} X\right) \tag{A.169}$$

by partially evaluating the connection on the tensor field only, leaving the vector field un-inserted. This form is called the *covariant derivative*.

Covariant derivative

In this way, the connection can be seen as a generalisation (in the sense of the type of function it is) of the *exterior derivative*, as it will also map $(0, l)$-tensors to $(0, l+1)$-tensors. In this case, note then that for a smooth function $f \in C^\infty(X)$,

$$\nabla f = \mathrm{d}f. \tag{A.170}$$

It is possible to choose a connection that agrees with the exterior derivative. The action of a connection on 0-forms will always agree, and typically its action on 1-forms as well. On higher order forms however it will usually not be made to agree, and will not map alternating tensors to alternating tensors, as we run into geometric issues when generalising to non-forms.

For a vector field X and covector field ω, the *coordinate expression of the covariant derivative* in a chart (U, ϕ) is given by

Coordinate expression of the covariant derivative

$$\nabla V = \left[\partial\phi_j(V^i) + V^k \Gamma^i_{jk}\right] \partial\phi_i \otimes \mathrm{d}\phi^j \qquad \nabla\omega = \left[\partial\phi_j(\omega^i) - \omega^k \Gamma^i_{jk}\right] \mathrm{d}\phi^i \otimes \mathrm{d}\phi^j$$

The covariant derivative gives us a way to define the *Riemannian gradient of a tensor field*. We define it as

Riemannian gradient of a tensor field

$$\operatorname{grad}_g : \Gamma\left(T^p_q\right) \to \Gamma\left(T^{p+1}_q\right) \quad \operatorname{grad}_g : T \mapsto (\nabla T)^\uparrow \tag{A.171}$$

, i.e. the raising of an index of the covariant derivative of the tensor field. We can also use it to define the *divergence of a tensor field* as

Divergence of a tensor field

$$\operatorname{div}_g : \Gamma T^p_q \to \Gamma T^{p-1}_q \quad \operatorname{div}_g : T \mapsto \operatorname{tr}(\nabla T) \tag{A.172}$$

i.e. the trace of the covariant derivative of the tensor field.

Second covariant derivative

Having defined the covariant derivative, we can also define the *second covariant derivative* as

$$\nabla^2 A = \Delta A = \nabla(\nabla A) \tag{A.173}$$

for a tensor field A. This can be evaluated on two vector fields V, W by inserting them into the last 2 slots of the resulting tensor. It should be noted though that

$$\nabla^2_{V,W} A = [\nabla^2 A](..., V, W) \neq \nabla_V(\nabla_W A) \tag{A.174}$$

instead

$$\nabla^2_{V,W} A = \nabla_V(\nabla_W A) - \nabla_{\nabla_V W} A. \tag{A.175}$$

Hessian of a smooth function

The most common example of the second covariant derivative is the second covariant derivative of a scalar function, which is a $(0,2)$-tensor field, and is the *hessian of a smooth function*, encoding how it fast the gradient of a function varies in a particular direction.

A.10.3. Vector fields on curves. The most important application of connections is that they tell us how to transport tensors, and in particular vectors, along a path on a manifold in a coherent way. To do so, let us consider a vector field on a curve.

Smooth vector field on a curve
Extendable

Let $\gamma : I \to X$ be a smooth curve on the manifold, and let a *smooth vector field on a curve* be a smooth map $V : I \to TX$ such that $V(t) \in T_{\gamma(t)}X$. Such a vector field is said to be *extendable* if there exists a smooth vector field $W \in \Gamma(TX)$ such that $V(t) = W_{\gamma(t)}$. This definition is necessary as if our path crosses itself, we may assign two different vectors to the same location. We say that an *extension of an extendable vector field on a curve* is any vector field $\tilde{V}$ such that $V(t) = \tilde{V}_{\gamma(t)}$. We can similarly define *extendable smooth tensor fields on a curve.*

Using this concept, we can define the *covariant derivative along a curve.*

Covariant derivative along a curve

DEFINITION A.53. *Let X be a smooth manifold with a connection ∇. Let $\gamma : I \to X$ be a smooth curve on X, and let V be a smooth extendable vector field on γ. Let $\tilde{V}$ be an extension of V. The* COVARAINT DERIVATIVE OF V ON γ *is given by*

$$[D_\gamma V](t) = [\nabla_{\gamma'} \tilde{V}](\gamma(t)) \tag{A.176}$$

where γ' is the DERIVATIVE OF THE CURVE *(see appendix A.6.3)*

This definition is independent of the choice of extension.

A.10.4. Geodesics. Using the definition of the covariant derivative on a curve, we can define the *acceleration of a curve* simply as

Acceleration of a curve

$$D_\gamma \gamma', \tag{A.177}$$

i.e. the covariant derivative of the velocity of the curve.

Geodesic

A *geodesic* on a manifold is very simply then any curve γ for which the acceleration of γ is zero everywhere. That is, that a geodesic is a curve that follows a 'straight' line over the manifold with respect to the connection specified.

Importantly, geodesics exist, and are unique.

THEOREM A.54. *Let X be a smooth manifold with a connection ∇. For every point $p \in X$, $w \in T_pX$, and $t_0 \in \mathbb{R}$ there exists an open interval $I \subset \mathbb{R}$ containing t_0 and a geodesic $\gamma : I \to X$ satisfying $\gamma(t_0) = p$ and $\gamma'(t_0) = w$. I is not unique, but for another geodesic $\bar{\gamma}$ with domain $\bar{I}$, they agree on $I \cap \bar{I}$.* Existance and uniqueness of geodesics

A.10.5. The exponential and logarithm maps. With geodesics in mind, let us denote $\gamma : TX \times \mathbb{R} \to X$ as the map $(p, v), t \mapsto \gamma_v(t)$, where γ_v is the unique geodesic satisfying $\gamma_v(0) = p$ and $\gamma_v'(0) = v$. This assignment obeys the *rescaling lemma*, that is Rescaling lemma

$$\gamma_{cv}(t) = \gamma_v(ct), \quad c, t \in \mathbb{R},\ v \in TX. \tag{A.178}$$

This allows us to say a few things: Firstly the map $v \mapsto \gamma_v$ creates a map between TX and the space of geodesics. Secondly it tells us that straight lines through the origin in the tangent space map to geodesics. Lastly, it means we can restrict ourselves to evaluating γ_v at $t = 1$ and still arrive at any point we could have before, by application of the rescaling lemma.

If we define then the domain

$$\mathcal{E} = (v \in TM \ :\ \text{the domain of } \gamma_v \text{ contains } [0, 1]) \tag{A.179}$$

then we can define a map

$$\exp : TX \to X, \quad \exp(v) = \gamma_v(1) \tag{A.180}$$

as *the exponential map*. The exponential map

The domain $\mathcal{E}$ is by no means guaranteed to cover the whole of TX. Manifolds for which $\mathcal{E} = TX$ are said to be *geodesically complete*. The *Hopf-Rinow theorem* is a cornerstone result that states for any connected Riemannian manifold, geodesic completeness and completeness in the metric sense (i.e. that all Cauchy sequences converge) are equivalent. Geodesically complete Hopf-Rinow theorem

The inverse of the exponential map is *the logarithm map*, defined by The logarithm map

$$\log : X \times X \to TX, \quad \exp(\log_p(q)) = q,\ \log_p(q) \in T_pX \tag{A.181}$$

and provides a map from points on the manifold q to the tangent space T_pX, giving the tangent vector at p that under the exponential map sends p to q.

A.10.6. Parallel transport. Lastly, we can use a connection to define a what it means for a vector field to be parallel along a curve, and consequently a way of transporting vector from one tangent space to another.

For a vector field V on a curve γ, we say that V is *parallel along* γ if Parallel along γ

$$D_\gamma V = 0. \tag{A.182}$$

In this way, we can define a geodesic as a curve for which the velocity is parallel along the curve. Since a connection can be extended to work on any tensor field, this notion of parallel naturally extends to any tensor field.

To map tensors along curves in a parallel fashion, we need a way to extend a tensor at a point to a tensor field along a curve. This can be done uniquely by the following theorem.

Uniqueness of parallel tensor fields

THEOREM A.55. *Let X be a smooth manifold and ∇ a connection on X. Given a smooth curve $\gamma : I \to M$, $t_0 \in I$, and $v_0 \in T^k_l T_{\gamma(t_0)} X$, there exists a unique parallel tensor field A along γ such that A is parallel along γ with respect to ∇ such that $V(\gamma(t_0)) = v$.*

Parallel transport of a tensor along a curve

We then define the *parallel transport of a tensor along a curve* as

$$P^{\gamma}_{t_0,t_1} : T^k_l T_{\gamma(t_0)} \to T^k_l T_{\gamma(t_0)}, \quad v \mapsto V(\gamma(t_1)). \tag{A.183}$$

Parallel basis along a curve

This map provides an *isometry* between tangent spaces. We can use this map to create a *parallel basis along a curve* by parallel transporting a the basis vectors of the tangent space at one point along the curve, to produce parallel basis fields.

Flat connection

It should be noted that this parallel transport is strongly dependent on the path taken. A connection for which parallel transport is path independent is called a *flat connection*. The existence of a flat connection on a smooth manifold is equivalent to the tangent bundle being *parallelisable* (see appendix A.6.1), as we can define non-vanishing coordinate fields on the manifold by parallel transporting a frame at a given point around the manifold.

Parallel transport in coordinates

A.10.7. The Levi-Civita connection. On Riemannian manifold, with the additional structure of the Riemannian metric, we can require a connection to be compatible with the metric on the manifold in the following way:

Metric compatible connection

DEFINITION A.56. *Let (X, g) be a Riemannian manifold. A connection $\nabla : \Gamma(TX) \times \Gamma(TX) \to \Gamma(TX)$ on the tangent bundle is* COMPATIBLE WITH THE METRIC *if for any vector fields $U, V, W \in \Gamma(TX)$*

$$\nabla_U g(V, W) = g(\nabla_U V, W) + g(V, \nabla_U W) \tag{A.184}$$

There are a few other properties that are equivalent to this definition:

1. $\nabla g = 0$, i.e. the metric is parallel under the connection.
2. The coefficients of the connection in a chart (U, ϕ) satisfy
$$\Gamma^l_{ki} g_{lj} + \Gamma^l_{kj} g_{il} = \partial \phi_i(g_{ij}). \tag{A.185}$$
Christoffel symbols — These are called the *Christoffel symbols*.
3. Parallel transport of a vector along a curve preserves their Riemannian norm.
4. Parallel transport of a pair of vectors along a curve preserves their Riemannian inner product.
5. Parallel transport of an orthonormal frame along a curve remains orthonormal.

Torsion tensor

Unfortunately, this does not uniquely specify a connection. There remains a degree of freedom in the choice of connection, described by its *torsion*. The *torsion tensor* of a connection is given by

$$\tau(V, W) = \nabla_V W - \nabla_W V + [V, W]. \tag{A.186}$$

Torsion free
Symmetric

If a connection has zero torsion for all vector fields, it is said to be *torsion free* or *symmetric*.

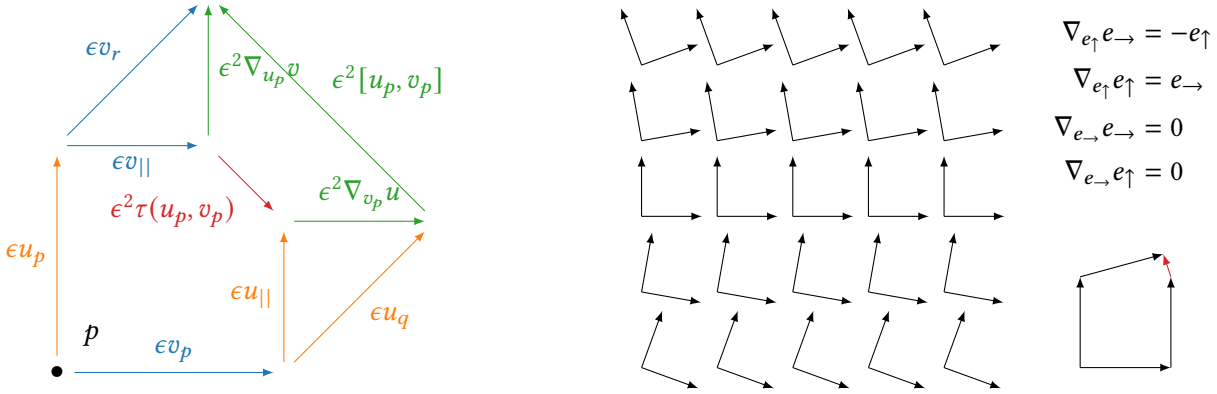


Figure A.13. *Left:* We can understand torsion intuitivly, but not precisely, via the above diagram. For vectors v_p and u_p, we extend these to smooth vector fields on an infinitesimal neighbourhood. If we take an infinitesimal step in each direction, $r = \exp_p \epsilon v_p$ and $q = \exp_p \epsilon u_p$, and parallel transport the step vectors along the opposite step vectors, $\epsilon v_{||} = P_{\epsilon u_p}(\epsilon v_p)$ and $\epsilon u_{||} = P_{\epsilon v_p}(\epsilon u_p)$, then, ignoring the issues of adding vectors in different tangent spaces, the torsion, $\epsilon^2 \tau(v_p, u_p)$, is the lack of agreement between $\epsilon v_p + \epsilon u_{||}$ and $\epsilon u_p + \epsilon v_{||}$. *Right:* An example a connection with torsion on two dimension Euclidean space. On the left we see a basis frame transported around the plane, twisting from point to point. On the right, the basis representation of the connection, and a diagram showing the non-zero torsion in red.

THEOREM A.57. *Let (X, g) be a Riemannian manifold. Then there exists a unique connection ∇ that is both* METRIC-COMPATIBLE *(see definition A.56) and* TORSION-FREE *(see appendix A.10.7).*

Fundimental theorem of Riemannian Geometry

We call this unique connection the *Levi-Civita connection*, and when ambiguous we denote it ∇_g, however when we refer to a connection without specifying it, we always refer to the Levi-Civita connection as it is the canonical choice on a manifold. In addition, when we refer *geodesics* (see appendix A.10.4), the *exponential* (see appendix A.10.5) and *logarithm* (see appendix A.10.5) maps and *parallel transport* (see appendix A.10.6) on a Riemannian manifold, they are with respect to the Levi-Civita connection unless otherwise specified.

Levi-Civita connection

A.10.8. The Laplace-Beltrami operator. One particular operator that is of great use and is well studied is the *Laplace-Beltrami operator*. This operator is the coordinate free generalisation of the *Laplace operator*, a second order differential operator, given by

Laplace-Beltrami operator

$$\Delta : C^\infty(\mathbb{R}^n) \to C^\infty(\mathbb{R}^n) \qquad \Delta f = \nabla \cdot (\nabla f) = \nabla^2 f = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial {x_i}^2}. \tag{A.187}$$

On Riemannian manifolds we can define this in two ways,

$$\Delta : C^\infty(X) \to C^\infty(X) \qquad \Delta f = \operatorname{div}_g(\operatorname{grad}_g f) = \operatorname{tr}(\nabla^2 f), \tag{A.188}$$

either the *Riemannian divergence* (see appendix A.9.3) of the *Riemannian gradient* (see appendix A.8.2) of a smooth function, or as the *trace* of the *second covariant derivative* (see appendix A.10.2) of the *Levi-Civita connection* (see appendix A.10.7).

In local coordinates, we can write this operator as

$$\Delta = \frac{1}{\sqrt{\det g}} \sum_{i=1}^{n} \partial\phi_i \left(g^{ij} \sqrt{\det g} \partial\phi_j \right) \tag{A.189}$$

This can also be generalised to tensor bundles using the generalisations of the Riemannian gradient and divergence to tensor fields specified earlier.

The operator self-adjoint in the sense that for $f \in C^\infty(X)$

$$\langle \nabla f, \nabla f \rangle_g = \langle \Delta f, f \rangle_g = \langle f, \Delta f \rangle_g. \tag{A.190}$$

This can be extended further to L^2-integrable functions on the manifold, and also to L^2-integrable vector fields on the manifold where $-\Delta$ is a self-adjoint unbounded positive-definite operator (Strichartz, 1983, theorem 2.4).

A.10.9. Eigenspectrum of the Laplace-Beltrami operator. Since the space of L^2-integrable functions are Hilbert spaces, we can leverage spectral theory to decompose functions in the Hilbert space over eigenvalues and eigenfunctions of $-\Delta$. In particular for compact manifolds, the spectrum of $-\Delta$ is countable.

Strun-Liouville decomposition

THEOREM A.58. *Let (X, g) be a compact manifold. Then there exists an orthonormal basis of $L^2(X)$, $f_n \in L^2(X)$, $n \in \mathbb{Z}^+$, such that*

$$-\Delta f_n = \lambda_n \tag{A.191}$$

and that the basis is ordered in the sense that $0 \leq \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$, and $\lambda_n \to \infty$ as $n \to \infty$.

Proof. Chavel (1984) ∎

This means we can write any function $f \in L^2(X)$ on a compact manifold as the infinite sum $f = \sum_{i=1}^{\infty} \langle f, f_n \rangle_g f_n$. This basis expansion can be truncated to a finite length $f_N \approx \sum_{i=1}^{N} \langle f, f_n \rangle_g f_n$. The error of this basis truncation can be bounded.

Bounded error of the Sturm-Liouville decomposition

THEOREM A.59. *Let (X, g) be a compact Riemannian manifold, and $(f_n, \lambda_n)_{n=0}^{\infty}$ be the eigen-spectrum of $-\Delta_g$. Let $f \in L^2(X)$. Let f_N be the truncated expansion of f. Then*

$$\|f - f_N\|_g^2 \leq \frac{\left\|\nabla_g f\right\|^2}{\lambda_{N+1}}. \tag{A.192}$$

Proof. (Aflalo and Kimmel, 2013) ∎

It can also be shown that this approximation is optimal in the sense that no other basis will obtain a better error bound than the truncated Sturm-Liouville approximation (Aflalo et al., 2015).

Since they will be of use to us, we give example of these eigen-spectrum on the d-dimension torus and the d-dimensions sphere.

The d-torus

Let $\{b_i\}_{i=1}^d$ be a basis of $\mathbb{R}^d$. We consider the associated lattice on $\mathbb{R}^d$, i.e. $\Gamma = \left\{\sum_{i=1}^d \alpha_i b_i \ : \ \{\alpha_i\}_{i=1}^d \in \mathbb{Z}^d\right\}$. Finally, the associated d-dimensional torus is defined as the quotient $\mathbb{T}_\Gamma = \mathbb{R}^d/\Gamma$.

Denote $\mathrm{B} = (b_1, \ldots, b_d) \in \mathbb{R}^{d\times d}$. Let $\{\bar{b}_i\}_{i=1}^d \in (\mathbb{R}^d)^d$ such that $(\mathrm{B}^{-1})^\top = (\bar{b}_1, \ldots, \bar{b}_d)$. We define $\Gamma^\star = \left\{\sum_{i=1}^d \alpha_i \bar{b}_i \ : \ \{\alpha_i\}_{i=1}^d \in \mathbb{Z}^d\right\}$ as the dual lattice. Note that for any $x \in \Gamma$ and $y \in \Gamma^\star$ we have that $\langle x, y\rangle \in \mathbb{Z}$ and that if $\{b_i\}_{i=1}^d$ is an orthonormal basis then $\Gamma = \Gamma^\star$.

As a result, the torus $\mathbb{R}^d/\Gamma$ is a (flat) compact Riemannian manifold. The set of eigenvalues and functions of the Laplace–Beltrami operator are associated with the dual lattice Γ^*. For each element of the dual lattice, $y \in \Gamma^*$ the eigen-value is given by

$$\lambda_y = -4\pi^2 \|y\|^2. \tag{A.193}$$

For each element of the dual lattice there are two eigenfunction, given by

$$\phi_{1,\lambda}(x) = \sin(2\pi\langle x, y\rangle) \tag{A.194}$$

and

$$\phi_{2,\lambda}(x) = \cos(2\pi\langle x, y\rangle). \tag{A.195}$$

The d-sphere

Next, we investigate the case of the d-dimensional sphere (see Saloff-Coste, 1994). The set of eigenvalues of the Laplace–Beltrami operator is given by

$$\{-k(k+d-1) \ : \ k \in \mathbb{N}\}. \tag{A.196}$$

Note that $\lambda_k = k(k+d-1)$ has multiplicity $d_k = (k+d-2)!/\{(d-1)!k\}(2k+d-1)$, i.e. that it has this many eignefunctions associated with it.

The eigenfunctions of the Laplace–Beltrami operator are known as the spherical harmonics and can be defined in terms of Legendre polynomials. When investigating the heat kernel on the d-dimensional sphere, we are interested in the product

$$(x, y) \mapsto \sum_{\phi\in\Phi_n} \phi(x)\phi(y) \tag{A.197}$$

, where Φ_n is the set of eigenfunctions associated with the eigenvalue λ_n for $n \in \mathbb{N}$. This function can be described using the Gegenbauer polynomials (see Atkinson and Han, 2012, Theorem 2.9). More precisely, we have that for any $n \in \mathbb{N}$ and $x, y \in \mathbb{S}^d$

$$G_n(x, y) = \sum_{\phi\in\Phi_n} \phi(x)\phi(y) \tag{A.198}$$

$$= n!\Gamma((d-1)/2) \sum_{k=0}^{\lfloor n/2\rfloor} \frac{(-1)^k(1-\langle x, y\rangle^2)\langle x, y\rangle^{n-2k}}{4^k k!(n-2k)!\Gamma(k+(d-1)/2)}, \tag{A.199}$$

where here $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ is given for any $v > 0$ by $\Gamma(v) = \int_0^{+\infty} t^{v-1}\mathrm{e}^{-t}\mathrm{d}t$. In the special case where $d = 1$, then the heat kernel coincide with the wrapped Gaussian density and can be easily evaluated.

B | Stochastic Differential Equations

Stochastic differential equations are the second core technical area used in this thesis. They form the mathematical foundation for dealing processes that evolve randomly in time through some state space. The core mathematics of diffusion models are built on stochastic differential equations, and so having an understanding of the underlying mathematical tools is helpful for making advances.

In this chapter I aim to give a gentle introduction to the formalisation of stochastic differential equations. First I provide a review of measure theory basics needed to fully formalise stochastic processes. I then provide an introduction to stochastic processes, then stochastic differential equations, including definitions of stochastic integrals and tools for working with them. These include the Kolmogorov equations, Langevin dynamics, connections with ordinary differential equations and methods to approximately sample stochastic differential equations. Finally, we discuss how to define, and tools to work with, stochastic differential equations on manifolds.

B.1. A review of measure theory based probability

It is necessary when talking about stochastic processes to work with probability under the *measure theory* approach. Without this, there is no clear way to discuss probability on infinite dimension spaces, such as the space of continuos functions on interval $[0, 1]$. There are a number of other motivations for studying measure theory, particularly concerning technical difficulties that arise with the usual definitions of probability that one encounters.

Measure theory

The presentation is based on Kallenberg (1997), and takes inspiration from Terenin (2022).

A *σ-algebra*, much like a topology, is a collection of sets obeying a set of rules. The definition even looks very similar. But this is where the similarity ends.

Definition B.1. *A measurable space (X, Σ_X) is a set X along with a collection of subsets Σ_X of X, called a sigma/σ-algebra, that satisfy the following:*

Measurable space
Sigma/σ-algebra

- *$X \in \Sigma_X$, and is called the universal set.*
- *Σ_X is closed under compliments.*

- Σ_X *is closed under* COUNTABLE *unions.*

From this it also follows that a σ-algebra is closed under countable intersections, and that $\varnothing \in \Sigma_X$.

The definition of a σ-algebra means that it contains within itself the limits of any sequence of sets contained in it, which allows us to define some sensible concepts of convergence of random variables, to be discussed later.

Borel σ-algebras

Of particular relevance are *Borel σ-algebras.* For a topological space $(X, \mathcal{O}_X)$, the Borel σ-algebra, $\mathcal{B}((X, \mathcal{O}_X))$ is the smallest sigma algebra containing the topology $\mathcal{O}_X$. On a Riemannian manifold this is given by the *Riemannian volume form* (see proposition A.46)

Measure

A *measure* on a measurable space is a function $\mu : \Sigma_X \to \mathbb{R}$ that is also non-negative, is zero on the empty set, and is countably additive. This is that for disjoint sets $\mathbb{E}_1, \ldots \in \Sigma_X$, $\mu(\bigcup_i E_i) = \sum_i \mu(E_i)$. With this in mind we can interpret the σ-algebra as the set of sets we can measure. The restriction we place on this set of measurable sets is exactly designed to avoid technical issues that can appear. For example, on the real numbers $\mathbb{R}$ for a particular subset $[a, b]$ we would make this simply $|b - a|$. In this way, we are measuring the length of this set. A *measure space* is a measurable space along with a metric, (X, Σ_X, μ).

Measure space

Products of measurable spaces

Product σ-algebra

We can take *products of measurable spaces.* For a countable sequence of spaces $(X_i, \Sigma_{X_i})_{i=1}^{\infty}$ the *product σ-algebra* $\bigotimes_i \Sigma_{X_i}$ is given by the smallest σ-algebra containing the *cylinder sets* (eq. (A.4)) of the Σ_{X_i}. Note the similarity to the definition of the product topology. Indeed, for *Hausdorff* (appendix A.1.2) (and other types of separability) the product σ-algebra and the product topology interact nicely.

$$\mathcal{B}\left(\prod_i (X_i, \mathcal{O}_{X_i})\right) = \bigotimes_i \mathcal{B}((X, \mathcal{O}_X)) \tag{B.1}$$

Measurable function

Composition

A function $f : X \to Y$ between two measurable spaces, (X, Σ_X), (Y, Σ_Y), $f : X \to Y$ is said to be a *measurable function*, denoted Σ_X/Σ_Y-measurable, if for any $U \in \Sigma_Y$, $f^{-1}(U) \in \Sigma_X$. This says if I can measure a set in Y, then I can measure its preimage in X. Notice again the similarity to this definition to the definition of continuity. Indeed, for any pair of topological spaces with the Borel σ-algebra, $(X, \mathcal{O}_X, \mathcal{B}(\mathcal{O}_X))$, $(Y, \mathcal{O}_Y, \mathcal{B}(\mathcal{O}_Y))$, any continuous function $f : X \to Y$ is $\mathcal{B}(\mathcal{O}_X)/\mathcal{B}(\mathcal{O}_Y)$-measurable. Like continuous functions, the *composition* of measurable functions is again measurable.

Measurable functions on product spaces

For *measurable functions on product spaces* the function $f : X \to Y \times Y'$ is $\Sigma_X/\Sigma_Y \otimes \Sigma_{Y'}$-measurable if each component function is measurable. The converse is *not* true. $f : X \times X' \to Y$ is not necessarily measurable if $f(\cdot, x') : X \to Y$ and $f(x, \cdot) : X' \to Y$ are measurable.

Probability measure

Probability space

A *probability measure* is a measure $\mu : X \to \mathbb{R}$ on a measurable space (X, Σ_X) such that $\pi_X(X) = 1$. A *probability space*, $(\Omega, \mathcal{F}, \mathbb{P})$ in standard notation, is a measurable space along with a probability measure. In this way, we can think about $\mathcal{F}$ as the set of sets of events $A_\omega \subset \Omega$ that we can ask questions about probability about. We are precluded from asking questions about sets that are not in $\mathcal{F}$ as it creates technical issues. In this sense a probability measure describes the uncertainty

we have about the events $\omega \in \Omega$. Typically, when dealing with probabilities, the space Ω is very abstract, and we simply choose one with "enough elements" and a σ-algebra with enough sets in be able to achieve what we want to. We can see Ω as a bag of random numbers and $\mathbb{P}$ as a random generator function from this bag, not dissimilar to a random number generator influencing events in a computer simulation.

Random variable

We define a *random variable* as a measurable function $\mathsf{X} : \Omega \to X$ from a probability space into a measurable space. A random variable can then be thought as mapping abstract randomly generated events $\omega \in \Omega$ into real observations in X. The *distribution of a random variable* is given by $\pi_{\mathsf{X}} = \mathsf{X}^* \mathbb{P}$, where this denotes the *pushforward* of $\mathbb{P}$ through X, defined by $(\mathsf{X}^* \mathbb{P})(A_x) = \mathbb{P}(\mathsf{X}^{-1}(A_x))$ for $A_x \in \Sigma_X$. The probability of a collection of events A_x happening in X is then the probability of their preimage happening in Ω. The fact that X is measurable guarantees this is well defined.

Distribution of a random variable

The existence of random variables

To expand on the point of being able to choose Ω to include enough events to be interesting, it is not guaranteed that for a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and distribution π_{X} that there exists a random variable $\mathsf{X} : \Omega \to X$ such that $\pi_{\mathsf{X}} = \mathsf{X}^* \mathbb{P}$. However, given a π_{X} and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ it is possible to find a *bigger* probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ along with a measurable surjection $i : \Omega' \to \Omega$ such that $i_* \mathbb{P}' = \mathbb{P}$ and a random variable $\mathsf{X} :' \Omega \to X$ exists with distribution π_{X}.

Expectation

The *expectation* of a random variable is only defined is the space of the random variable has enough structure, namely vector space structure. For real valued random variables, $\mathsf{X} : \Omega \to \mathbb{R}$, $\mathbb{E}[\mathsf{X}]$ can be thought of as the weighted average value of X under the values of X generated by $\mathbb{P}$. For a suitable notion of integration, namely Lebesgue integration, this is

$$\mathbb{E}[\mathsf{X}] = \int_X x \, \mathrm{d}\pi_{\mathsf{X}}(x) = \int_\Omega \mathsf{X}(\omega) \, \mathrm{d}\, \mathbb{P}(\omega) \tag{B.2}$$

where this integration happens against the measure π_{X}.

Manifold-valued random variables

For *manifold-valued random variables*, $\mathsf{X} : \Omega \to M$, however this structure is rarely presents, and so the expectation is undefined. In order to study such random variables we often look at their behaviour under composition with measurable test functions $f : M \to \mathbb{R}$ as we can then take the expectation of the random variable $f \circ \mathsf{X} : \Omega \to \mathbb{R}$. This definition of expectation applies clearly to real vector-valued random variables as well.

Covariance

The *covariance* of a pair of random variables, if it exists, is then given by $\mathrm{Cov}(\mathsf{X}, \mathsf{X}') = \mathbb{E}[\mathsf{X}\mathsf{X}'] - \mathbb{E}[\mathsf{X}]E[\mathsf{X}']$. This definition also extends to real vector-valued random variables as well by applying it component-wise.

Equal surely

For a pair of random variables, X, X' on a measurable space (X, Σ_X) relative to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we can define equality in a number of ways. We say that they are *equal surely*, denoted $\mathsf{X} = \mathsf{X}'$, if $\mathsf{X}(\omega) = \mathsf{X}'(\omega)$ for all $\omega \in \Omega$, i.e. they are equal as functions. As it turns out this definition is generally too strict.

Equal almost-surely

We say that they are *equal almost-surely*, denoted $\mathsf{X} = \mathsf{X}'$ a.s., if $\mathbb{P}(\mathsf{X} = \mathsf{X}') = 1$. This means that they agree as functions *except* on sets of probability zero. Almost-sure equality is implied by sure equality.

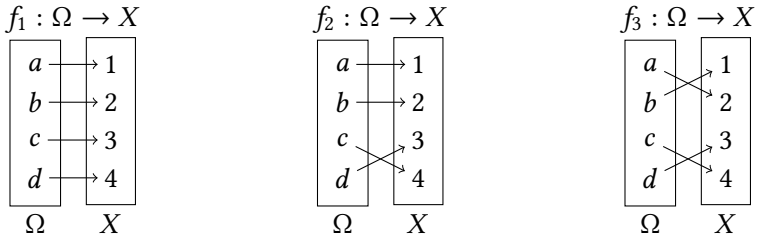


Figure B.1. Consider random variables f_1, f_2, f_3, with $\Omega = \{a, b, c, d\}$ and $X = \{1, 2, 3, 4\}$. Since this is a finite set, we can make the σ-algebra $\mathcal{F}$ the power set (the set of all subsets) of Ω (this does not work on uncountable spaces like $\mathbb{R}$), and the same for Σ_X. Let $\mathbb{P}(\{a\}) = \mathbb{P}(\{b\}) = 0.5$ and $\mathbb{P}(\{c\}) = \mathbb{P}(\{d\}) = 0$. Clearly then f_1, f_2, f_3 are equal to themselves surely. f_1 and f_2 are equal almost surely as the only sets they don't agree on are $\{c\}, \{d\}$, and these are measure 0 under $\mathbb{P}$. f_1, f_2 are equal to f_3 in distribution only. The measures $\pi_{f_1}, \pi_{f_2}, \pi_{f_3}$ all assign the same probabilities to sets of X ($\pi_f(\{1\}) = \pi_f(\{2\}) = 0.5$, $\pi_f(\{1, 2\}) = 1$, $\pi_f(\{3\}) = \pi_f(\{4\}) = 0.0$, etc), but the are not the same as functions on sets of non-zero measure, ($f1, 2(\{a\}) \neq f_3(\{a\})$, $f_{1,2}(\{2\}) \neq f_3(\{2\})$)

Equal in distribution

Weaker than this, we say that two random variables are *equal in distribution*, denoted $\mathsf{X} \overset{d}{=} \mathsf{X}'$, if the measures $\pi_{\mathsf{X}} = \mathsf{X}_* \mathbb{P}$ and $\pi_{\mathsf{X}'} = \mathsf{X}'_* \mathbb{P}$ are equal. In other words this implies that $\mathsf{X}(A) = \mathsf{X}'(A)\ \forall A \in \Sigma_X$. This means that they assign the same probabilities to events in X, but these events may *not* have been generated by the same random events in Ω. Equality in distribution is implied by almost-sure equality.

These notions of equality also lead to definitions of limits of sequences of random variables. Let us consider a random variable X and a sequence of random variables $(X_n)_{n=1}^{\infty}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Sure convergence
Pointwise convergence

Sure convergence or *pointwise convergence* implies that X_n converges as a function to X, that is

$$\lim_{n\to\infty} X_n(\omega) = X(\omega)\ \forall\ \omega \in \Omega, \quad \text{and is denoted by} \quad X_n \to X. \tag{B.3}$$

Almost sure convergence

As before, this notion is often too strong, and so we define *almost sure convergence* as

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1, \quad \text{and is denoted by} \quad X_n \overset{\text{a.s.}}{\to} X \tag{B.4}$$

which means that X_n need not converge to X on sets of measure zero. Almost-sure convergence is implied by sure convergence.

Convergence in probability

If X has metric space structure we can define *convergence in probability* as, for all ϵ

$$\lim_{n\to\infty} \mathbb{P}(d(X, X_n) \geq \epsilon) = 0, \quad \text{and is denoted by} \quad X_n \overset{p}{\to} X. \tag{B.5}$$

Convergence in probability is implied by almost-sure convergence.

Convergence in distribution

We define *convergence in distribution* as

$$\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A), \quad \text{and is denoted by} \quad X_n \overset{L^p}{\to} X. \tag{B.6}$$

Convergence in distribution is implied by convergence in probability.

Finally, if we have structure that allows for expectations we can define *convergence in L^p-norm*, if $\mathbb{E}\left[\|X_n\|_p^p\right]$ and $\mathbb{E}\left[\|X\|_p^p\right]$ exist, as

Convergence in L^p-norm

$$\lim_{n\to\infty} \mathbb{E}\left[\|X_n - X\|_p^p\right] = 0, \quad \text{and is denoted by} \quad X_n \xrightarrow{\text{a.s.}} X. \tag{B.7}$$

Convergence in L^p for $p \geq 1$ implies convergence in probability, and is usually easier to work with directly.

Interactions between random variables

We will often need to work with the interaction between one or more random variables.

Consider a joint random variable $\mathsf{Z} : \Omega \to X \times Y$, a $\mathcal{F}/\Sigma_X \otimes \Sigma_Y$ measurable function, with distribution $\pi_{\mathsf{Z}} : X \times Y \to \mathbb{R}$.

The *marginal random variable* $\mathsf{X} : \Omega \to X$ is given by

Marginal random variable

$$\text{proj}_X \circ \mathsf{Z} : \Omega \to X, \quad \text{where } \text{proj}_X : X \times Y \to X,\ (x, y) \mapsto x, \tag{B.8}$$

where the proj_X function projects onto the X coordinate. We can create from the *marginal distribution* $\pi_{\mathsf{X}} : X \to \mathbb{R}$, given by

Marginal distribution

$$\pi_{\mathsf{X}}(A_x) = \pi_{\mathsf{Z}}(A_x \times Y),\ A_x \in \Sigma_X. \tag{B.9}$$

A *probability kernel* is a function $\pi_{\mathsf{X}|\mathsf{Y}} : \Sigma_X \times Y \to \mathbb{R}$ is a function that 1) $\pi_{\mathsf{X}|\mathsf{X}}(\cdot \mid y) : X \to \mathbb{R}$ is a probability measure and 2) $\pi_{\mathsf{X}|\mathsf{X}}(A_x \mid \cdot) : Y \to \mathbb{R}$ is $\Sigma_Y/\mathcal{B}(\mathbb{R})$-measurable. These are useful for a few reasons, defining stochastic processes, but also conditional distributions.

Probability kernel

We define the *conditional distribution*, $\pi_{\mathsf{X}|\,y} : X \times Y \to \mathbb{R}$, where it exists, as the unique probability kernel satisfying

Conditional distribution

$$\pi_{\mathsf{X},\mathsf{X}}(A_x \times A_y) = \int_{A_y} \pi_{\mathsf{X}|\mathsf{X}}(A_x|y)\, \mathrm{d}\pi_{\mathsf{X}}(y) \tag{B.10}$$

where $\pi_{\mathsf{X},\mathsf{X}}$ is a joint distribution and π_{Y} is the marginal distribution of y.

This can also be viewed from the perspective of *disintegration of a measure* (Chang and Pollard, 1997).

Disintegration of a measure

Densities

When we have two measures on the same measurable space (X, Σ_X), we can sometimes relate them to one another through a simple function. For measures μ, ν, if for every $A \in \Sigma_X$, if $\nu(A) = 0$ implies that $\mu(A) = 0$, we say that μ is *absolutely continuous* with respect to ν. This is often denoted $\mu << \nu$. If this is the case, then there exists a $\Sigma_X/\mathcal{B}(\mathbb{R})$-measurable function f such that

Absolutely continuous

$$\mu(A + x) = \int_{A_x} f(x)\, \mathrm{d}\nu(x). \tag{B.11}$$

Radon-Nikodym derivative

Such an f if called the *Radon-Nikodym derivative* of μ with respect to ν, and is often denoted as $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$. The existence of this is not guaranteed, however. Note it is also possibly that both $\nu << \mu$ and $\mu << \nu$ at the same time, so that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ both exist, and $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}^{-1}$, but again this is not guaranteed.

Distribution with respect a measure

The existence of Radon-Nikodym derivates lets us connect the measure theory formalisation back to the more comment density based formulation. We say that a distribution $\pi : X \to \mathbb{R}$ has a *distribution with respect a measure* $\lambda : X \to \mathbb{R}$ if its Radon-Nikodym derivative exists, i.e. $\pi << \lambda$.

If we have a joint distribution $\pi_{\mathsf{X},\mathsf{Y}} : X \times Y \to \mathbb{R}$ that has a density with respect to product measure $\lambda : X \times Y \to \mathbb{R}$, i.e. $\lambda(A_x \times A_y) = \lambda_{\mathsf{X}}(A_x)\lambda_{\mathsf{Y}}(A_y)$, with $f_{\mathsf{X},\mathsf{Y}} = \frac{\mathrm{d}\pi_{\mathsf{X},\mathsf{Y}}}{\mathrm{d}\lambda}$, then

- The marginal distribution π_{X} has a density f_{X} with respect to the measure $\lambda(\cdot \times Y) = \lambda_{\mathsf{X}}(\cdot)$, and is given by $f_{\mathsf{X}}(x) = \int_Y f_{\mathsf{X},\mathsf{Y}}(x, y)\, \mathrm{d}\lambda(x, y)$ and the same for π_{Y}.
- The density of the conditional distribution $\pi_{\mathsf{X}|\mathsf{Y}}$ is given by $f_{\mathsf{X}|\mathsf{Y}} = \frac{f_{\mathsf{X},\mathsf{Y}}}{f_{\mathsf{Y}}}$, and the same for $\pi_{\mathsf{Y}|\mathsf{X}}$.

Bayes rule

This gives us exactly the usual *Bayes rule*, as we can write from this $f_{\mathsf{X}|\mathsf{Y}} = \frac{f_{\mathsf{X},\mathsf{Y}}}{f_{\mathsf{Y}}} = \frac{f_{\mathsf{Y}|\mathsf{X}} f_{\mathsf{X}}}{f_{\mathsf{Y}}}$

B.2. Stochastic processes

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, (S, Σ_S) be a measurable space, called the *observation space*, and (I, Σ_I) a measurable, called the *index set*.

Stochastic process

A *stochastic process* is a function $X : \Omega \times I \to S$ that 1) $X(\cdot, i) : \Omega \to S$ is a random variable for all $i \in I$ and 2) $X(\omega, \cdot) : I \to S$ is measurable for all $\omega \in \Omega$. The index set and observation space can be any measurable space that we like, restricting our choices avoids many technical difficulties, however. We will limit ourselves to the case of Polish spaces as they poses many nice properties.

Observation space

Typically, the *observation space* is Euclidean space of some dimension, $S = \mathbb{R}^n$. We will begin with this, and then later study the case where $S = X$, a Riemannian manifold. The stochastic process can even take values in more abstract spaces, such as the tangent space at a point on a manifold.

Index set

In the simple case, the *index set* could be a finite collection of indices, $1, ..., n$, or a countable number of points $1, 2, ...$ with the discrete σ-algebra. Random variables over $\mathbb{R}$, and multivariate distributions fall into these camps. More interesting is when the index set is uncountable, and comes with some additional structural information, such as a topology or a metric. The index set I is often chosen to be time and can be $[0, \infty)$ or $[0, T]$, but can also be other spaces such as Riemannian manifolds or other metric spaces. For now, we will use time, $t \in [0, \infty)$ or $t \in [0, T]$ as our index set. This index set is both a metric space, and so has a topology, and

Totally ordered

is also *totally ordered*, such that for any two elements in the set we have access to boolean comparisons $<, >, \leq, \geq$.

By fixing the index argument to a specific value a stochastic process can be thought of as an indexed collection of random variables $\{X_i : \Omega \to S = X(\cdot, i) \ : \ i \in I\}$. By fixing the ω argument, we can think of the *sample paths* of X as the functions $X(\omega, \cdot) : I \to S$. From this point of view we can see a stochastic process as a *function-valued* random variable.

Sample paths

The *law of a stochastic process* is the *distribution*, defined as $\mu = \mathbb{P} \circ X^{-1} = X_* \mathbb{P}$, where the pushforward is in the Ω argument. This can be thought of as a distribution on the space of all possible functions $f : I \to S$. The *finite-dimension marginals* of a stochastic process are the pushforwards the law onto a finite index set by the projection onto a finite set of coordinate. This can be thought of as a distribution on space of values functions $f : I \to S$ can assume a some finite subset of I. For $\mathcal{I}(i_1, ..., i_k) \subset I$, $\mu_{i_1,...,i_k} = (\text{proj}_{\mathcal{I}})_* \mu$. This is equivalent to saying $\mu_{i_1,...,i_k}(F_1 \times ... \times F_k) = \mathbb{P}[\{\omega \ : \ X(\omega, i_1) \in F_1, ..., X(\omega, i_1) \in F_n\}]$. When the index set for a stochastic process is uncountable, the law will not have a density with respect to any measure. The finite dimension marginals may well have densities however with respect to the products of some reference measure on the state space. For the typical example of functions $f : \mathbb{R} \to \mathbb{R}^d$, products of the Lebesgue measure. This makes the finite dimension marginals significantly easier to work with as we can manipulate these densities.

Law of a stochastic process

Finite-dimension marginals

One of the most important theorems in the study of stochastic processes is the *Kolmogorov extension theorem*. Opposite to the projection from law of the process to law of the marginals, it tells us that if we can write down all sets of finite-dimension marginals, and they agree properly, then we can conclude the existence of a stochastic process. There exist many statements of this theorem, often with specific index sets or observation space. A more generic statement for topological spaces is

THEOREM B.2. *Let $\left(\left(X_a, \Sigma_{X_a}\right), \mathcal{O}_{X_a}\right)_{a \in A}$ be an arbitrary collection of measurable spaces $\left(X_a, \Sigma_{X_a}\right)$ each equipped with topology $\mathcal{O}_{X_a}$. For each finite $B \subset A$, let μ_B be an inner regular probability measure on the sigma algebra $\Sigma_{X_B} = \prod_{a \in B} \Sigma_{X_a}$ with the product topology $\mathcal{O}_{X_B} = \prod_{a \in B} \mathcal{O}_{X_a}$. Let $C \subset B$. If for all C, B, the measures are compatible in the sense that*

Kolmanagorov Extentsion Theorem

$$(\pi_{C \leftarrow B})_* \mu_B = \mu_C, \tag{B.12}$$

then there exists a unique probability measure μ_A on Σ_{X_A} such that

$$(\pi_B)_* \mu_A = \mu_B \tag{B.13}$$

for all finite $B \subset A$.

Proof. Tao (2011, theorem 2.4.3) ∎

This is very useful as it lets us specify the finite marginals of a stochastic process we wish to construct, and then prove that the process does exist. We can then manipulate these marginals to prove properties of the stochastic process. A common example of this in machine learning is the study of *Gaussian processes*, where the finite dimension marginals are multivariate Gaussian distributions.

Gaussian processes

The above generally does not uniquely specify a unique stochastic process though. There is a lot of "wiggle room" in the choice of X_t that still matches the right

Kolmogorov continuity theorem

marginals and law. The *Kolmogorov continuity theorem* states that we can choose X_t, under certain conditions, that is *continuous* almost surely, i.e. for any set of non-zero measure, $A_\omega \in \mathcal{F}$, the functions $X(\omega, \cdot) : I \to S,\ \omega \in A_\omega$ are topologically continuous. This means we can restrict the law of the process to be a measure on the space of continuous functions. This is extremely useful as this is again a Polish space with its useful properties. One other property that continuity brings is that the stochastic processes will now be measurable $\mathcal{F} \otimes \Sigma_i / \Sigma_X$, not just individually measurable in each argument.

Filtration

An important class of stochastic process if that of *martingales*. In order to define this, we need the concept of a *filtration*. A *filtration* on a sigma algebra Σ is a collection of sigma algebras indexed by a totally ordered set $\{\Sigma_i\}_{i \in I}$, such that for all $0 \leq i < j$,

$$\Sigma_i \subset \Sigma_j \subset \Sigma. \tag{B.14}$$

Adapted to a filtration

We say that a stochastic process $(X_t)_{t \geq 0}$ on a measurable space (X, Σ_X) with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is *adapted to a filtration* $(\mathcal{F}_t)_{t \geq 0}$ if for all t, $X_t : \Omega \to X$ is an $\mathcal{F}_t / \Sigma_X$-measurable function. Being adapted to a filtration says that a random variable at time t is only dependent on events in the "history" (times $< t$) of the filtration, and not its "future" (times $> t$). For any stochastic process we can give

Natural filtration

its *natural filtration*, which is made of us at each time t the smallest σ-algebra that makes all past events of the stochastic process measurable. A stochastic process is always adapted to its natural filtration, and we say one stochastic process is adapted to another if it is adapted to the natural filtration of the other process.

Martingale

Further, if X_t is a adapted to the filtration $\mathcal{F}_t$, it is a *martingale* if

1. $\mathbb{E}[\|X_t\|_2] < \infty$ for all t.
2. $\mathbb{E}[X_s \mid \mathcal{F}_t] = X_t$ for all $s \geq t$.

These conditions effectively say that the distribution of the process 1. that it is integrable, and 2. that the best guess for the process into the future is the last observed value. The properties of martingales make them both a practically reasonable case to study, and technically easier to manage than general stochastic processes.

Semimartingale
Local martingale

We can relax the definition of a martingale a little to arrive at the definition of a *semimartingale* This is a stochastic process that can be decomposed into two more stochastic processes, $W_t = M_t + A_t$, where M_t is a *local martingale*, a relaxed, localised, version of the martingale property , and A_t is a stochastic process with bounded total variation. Again, the technical properties of semimartingales make them attractive to study.

Markov processes

One final class of stochastic processes of interest are *Markov processes*. For a stochastic processes $(X_t)_{t \geq 0}$ on a measurable space (X, Σ_X) with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ adapted to the filtration $(\mathcal{F}_t)_{t \geq 0}$, it is Markov if

$$\mathbb{E}[f(X_t)]\mathcal{F}_s = \mathbb{E}[f(X_t)]\sigma X_s \quad \text{for all} \quad f : X \to \mathbb{R} \text{ bounded and measurable} \tag{B.15}$$

where $\sigma(X_s)$ is the smallest σ-algebra that makes X_s measurable. Interpreting this, it means that the expectation of a function of the process at time t, given the path

the process takes up to time s, encoded in $\mathcal{F}_s$, depends *only* on the value of the process at time s, encoded by $\sigma(X_s)$. In essence, the process is *memoryless*.

The most important process that is both a martingale and Markov studied is that of the Brownian motion. The *Brownian motion on* $\mathbb{R}^n$ can be defined via its finite dimension marginals. Let

Brownian motion on $\mathbb{R}^n$

$$p(t, \boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{1}{2\pi t^{n/2}} e^{-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2}{2t}} \quad \text{for } \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n,\ t > 0. \tag{B.16}$$

For $0 \leq t_1 \leq \ldots \leq t_k$ we define the finite dimension marginal $\nu_{t_1,\ldots,t_k}$ by picking some $\boldsymbol{x}$ to start the Brownian motion at and setting

$$\nu_{t_1,\ldots,t_k}(F_1, \ldots, F_k) = \int_{F_1,\ldots,F_k} p(t_1, \boldsymbol{x}, \boldsymbol{x}_1) p(t_2 - t_1, \boldsymbol{x}_1, \boldsymbol{x}_2) \ldots p(t_k - t_{k-1}, \boldsymbol{x}_{k-1}, \boldsymbol{x}_k)\, \mathrm{d}\boldsymbol{x}_1 \ldots \mathrm{d}\boldsymbol{x}_n. \tag{B.17}$$

We use the convention that $p(0, \boldsymbol{x}_1, \boldsymbol{x}_2) = \delta_{\boldsymbol{x}_1}(\boldsymbol{x}_2)$, and that $\mathrm{d}\boldsymbol{x}_1 \ldots \mathrm{d}\boldsymbol{x}_n$ is the Lebesgue measure on $\mathbb{R}^{n^k}$.

This satisfies the conditions of Kolmonogorov's extension theorem, giving us a stochastic process, B_t, with these marginals, and with law we denote $\mathbb{P}^{\boldsymbol{x}}$ such that

$$\mathbb{P}^{\boldsymbol{x}}(B_{t_1} \in F_1, \ldots, B_{t_k} \in F_k) = \nu_{t_1,\ldots,t_k}(F_1, \ldots, F_k). \tag{B.18}$$

There remains however the previously mentioned ambiguity, there exist multiple $B_t, \mathbb{P}^{\boldsymbol{x}}$ that satisfy this. Using the Kolmogorov continuity theorem we pick the Brownian motion that has continuous paths.

B.3. Stochastic differential equations

Differential equations are a wide class of equations that relate an unknown function f to its derivatives.

Differential equations

Ordinary differential equations (ODEs) are equations of a function of a single variable, $f(x)$, and its derivatives, $\frac{\mathrm{d}^n f}{\mathrm{d}x^n}$.

Ordinary differential equations (ODEs)

Partial differential equations (PDEs) are equations of a function of multiple variables, $f(x_1, \ldots x_n)$, and its various partial derivatives $\frac{\partial^{i_1+\ldots+i_n} f}{\partial x^{1\,i_1} \ldots \partial x_n{}^{i_n}}$.

Partial differential equations (PDEs)

Linear differential equation are differential equations that are linear combinations of derivatives, with coefficients that are differentiable functions $x_1, \ldots, x_n$.

Linear differential equation

The *order of a differential equation* is the highest order derivative that appears in the equation.

Order of a differential equation

Finding solutions to differential equations is often difficult or impossible analytically, although for a larger class of equations solutions can be found via numerical integration.

One of the most common settings for differential equations is when one of the input to is function is time and its output some state space, $f : \mathbb{R} \to X$, making f a path through the space X. For now, we will consider the state space $X = \mathbb{R}^d$.

One of the more simple differential equations can be written as

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} = b(t, f(t)), \tag{B.19}$$

with $f : \mathbb{R} \to \mathbb{R}^n$ and $b : \mathbb{R} \times \mathbb{R}^n$. This is a *first order, non-linear* differential equation. Coupled with an initial point ($f(t_0) = f_0$) and conditions on b, that it is continuous in t and Lipschitz-continuous in X, this constitutes an *initial value problem*.

Initial value problem

By the *Picard–Lindelöf theorem*, also known as the *Cauchy–Lipschitz theorem*, this has a unique solution, defined by the integral equation

Picard–Lindelöf theorem

$$f(t) = f(t_0) + \int_{t_0}^{t} b(s, f(t))\,\mathrm{d}s \tag{B.20}$$

for times $[t_0 - \epsilon, t_0 + \epsilon]$. More strongly, if b is bounded, i.e. that $\|b\| < K$ for some finite K, then the solutions are guaranteed to exist for all times.

If we are lucky, the integral may be analytic. More likely we will need to use numerical solvers to approximate the solution to this equation. In a small abuse of notation, this ODE is typically written as

$$\mathrm{d}f(t) = b(t, f(t))\,\mathrm{d}t. \tag{B.21}$$

From now on we will replace $f(t)$ with X_t to denote the path through a space X over time.

Stochastic differential equations

The study of *stochastic differential equations* aims to make rigorous the modification of differential equations with the addition of a "noise" term. I.e. how can we make the equation

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = b(t, X_t) + \sigma(t, X_t) \cdot \text{“noise”} \tag{B.22}$$

rigorous, and how can we integrate this to give us solutions so that the integration

$$X_t = X_0 + \int_0^t b(s, X_s)\,\mathrm{d}s + \int_0^t \sigma(s, X_s) \cdot \text{”noise”}\,\mathrm{d}s \tag{B.23}$$

makes sense.

A stochastic process seems the natural notion to use for the "noise" as it is a time-indexed random variable, so let us denote this $(W_t)_{t \geq 0}$. A natural set of conditions for this might be that

- $W_{t_1} \perp\!\!\!\perp W_{t_2}$ for $t_1 \neq t_2$.
- W_t is *stationary*, i.e. the distribution of the marginals $W_{t+t_1}, ..., W_{t+t_n}$ does not depend on t.
- $\mathbb{E}[W_t] = 0$.

However, it turns out there is no stochastic process that satisfies the first two conditions and has continuous paths. If we add another condition, that the process have unit variance, $\mathbb{E}[W_t] = 1$, then the function $(\omega, t) \to W_t(\omega)$ cannot even be measurable with respect to joint sigma algebra $\mathcal{F} \otimes \Sigma_I$ to Σ_S (Kallianpur, 1980, p.

180). It is *possible* to construct such a *white noise process* satisfying these conditions, however it is defined on a very different, larger space of *generalized functions*, a space where notions such as the Dirac delta can be made rigorous. White noise process

In order to come up with a coherent way of integrating a noise term, and to avoid the difficulty of the white noise process described above, we generalise a specific theory of integration of deterministic integrals, namely the *Riemann-Stieltjes integral*, to stochastic integrals. Briefly, we recap the construction of this.

B.4. The Riemann-Stieltjes integral

The Riemann-Stieltjes integral is one of a number of useful theories of integration. It is a generalisation of the Riemannian integral that fixes a number of issues. While not as general in some senses as Lebesgue or Lebesgue-Stieltjes integration, the analogy of Riemann sums will be useful when trying to define stochastic integrals.

The *Riemann-Stieltjes integral* of a function $f : [a, b] \subset \mathbb{R} \to \mathbb{R}$ with respect to another function $g : \mathbb{R} \to \mathbb{R}$ is denoted by Riemann-Stieltjes integral

$$\int_a^b f(x)\,\mathrm{d}g(x). \tag{B.24}$$

Given a *partition* $P = \{x_0 = a, x_1, ..., x_n = b\}$, $x_i < x_{i=1}$, of the interval $[a, b]$, we approximate the integral with a summation by Partition

$$S(P, f, g) = \sum_{i=0}^{n-1} f(c_i)[g(x_{i+1}) - g(x_i)], \tag{B.25}$$

where $c_i \in [x_i, x_{i+1}]$. The *norm of a partition* (or mesh) is the length of the largest subinterval, $\max_i x_{i+1} - x_i$. To define the integral, define a sequence of partitions such that $\lim_{n\to\infty} \|P_n\| \to 0$. The the Riemann-Stieltjes integral is defined as Norm of a partition

$$\int_a^b f(x)\,\mathrm{d}g(x) = \lim_{n\to\infty} S(P_n, f, g), \tag{B.26}$$

where this converges for all choices of $c_i \in [x_i, x_{i+1}]$ and for all sequences of partitions with norm converging to 0.

The class of integrands f and integrators g that are integrable is not simple and a number of different conditions exist under which f is integrable by g. One major class is when g is of bounded variation (on $[a, b]$).

The *total variation* of a function $f : I \to \mathbb{R}$ on $[a, b] \subset I$ is given by Total variation

$$TV_a^b(f) = \sup_{P\in\mathcal{P}} \sum_{i=1}^{|P|-1} |f(x_{i+1}) - f(x_i)| \tag{B.27}$$

where $\mathcal{P}$ is the set of all finite size partitions of $[a, b]$. A function is said to be of *bounded variation* if its total variation is bounded. If g is of bounded variation, then all continuous functions f are integrable. Bounded variation

B.5. The Ito and Stratonovich integrals

To define a stochastic integral, we follow a similar process to the Riemann-Stieltjes integral. The Riemann-Stieltjes integral is not immediately applicable however, for example because the types of object we will want to integrate against, such as Brownian motion, are *not* of bounded variation, and are not continuous.

First let us focus on the one dimension setting, where we have a drift function $b : \mathbb{R} \times [0, T] \to \mathbb{R}$ and a noise functions $\sigma : \mathbb{R} \times [0, T] \to \mathbb{R}$. In this case we are studying a stochastic process with index set $[0, T]$ and state space $\mathbb{R}$.

First we need to choose something to be our noise which we can integrate against. Consider a discretisation of the stochastic process X_t as

$$X_{k+1} - X_k = b(t_k, X_k)\Delta t_k + \sigma(t_k, X_k)\text{“noise”}\Delta t_k \tag{B.28}$$

on a partition $0 = t_0 < \ldots < t_n = t$, $X_k = X_{t_k}$, $W_k = W_{t_k}$, $\Delta t_k = t_{k+1} - t_k$. Let us replace “noise”$\Delta t_k$ with $\Delta V_k = V_{t_{k+1}} - V_{t_k}$, the difference between some time points of a stochastic process $(V_t)_{t\geq 0}$, as this seems another reasonable choice for the noise.

If we review our previous conditions for the “noise” term (stationary, independent, mean 0) and instead apply them to the *increments* ΔV_k, it turns out that exactly one stochastic process that satisfies these conditions if we make another requirement, that of the increments being *Gaussian*, and that is Brownian motion.

Thus, we approximate our stochastic differential equation by

$$X_k = X_0 + \sum_{j=0}^{k-1} b(t_j, X_j)\Delta t_j + \sum_{j=0}^{k-1} \sigma(t_j, X_j)\Delta B_j, \tag{B.29}$$

giving rise to and integral of the form

$$X_t = X_0 + \int_0^t b(s, X_x)\,\mathrm{d}s + \text{“}\int_0^t \sigma(s, X_s)\,\mathrm{d}B_s\text{”}. \tag{B.30}$$

The first term is the usual Riemann-Stieltjes integral, but the second is something we will need to see find a way to define this integral.

For a sequence of partitions P_n such that $\lim_{n\to\infty}\|P_n\| = 0$, define

$$S(P_n, f, (B_t)_{t\geq 0}, \omega) = \sum_{i=0}^{n-1} f(c_i, \omega)\left[B_{t_{i+1}} - B_{t_i}\right](\omega), \tag{B.31}$$

again with $c_i \in [t_i, t_{i+1}]$, and then define the integral as

$$\int_0^t f(s, \omega)\,\mathrm{d}B_s(\omega) = \lim_{n\to\infty} S(P_n, f, (B_t)_{t\geq 0}, \omega), \tag{B.32}$$

where the limit converges in L^2*-norm* (see appendix B.1). This limit is independent of the choice of P_n.

As it has been alluded to, this definition is not well behaved with respect to the choice of c_i, as, although we can choose the paths of our Brownian motion to

be continuous, it is nowhere differentiable, and the total variation of the paths is almost surely infinite, making it very poorly behaved as a Riemann-Stieltjes integral.

As an example of this, choose $f(t,\omega) = B_t(\omega)$, and choose $c_i = t_i + \alpha(t_{i+1} - t_i)$, where α interpolates between choosing the start and end of the intervals in the discretisation. Then

$$\mathbb{E}\left[S(P_n, f, (B_t)_{t\geq 0}, \omega)\right] = \mathbb{E}\left[\sum_{i=0}^{n-1} f(c_i,\omega)\left[B_{t_{i+1}} - B_{t_i}\right](\omega)\right] \tag{B.33}$$

$$= \sum_{i=0}^{n-1} \mathbb{E}\left[B_{t_i+\alpha(t_{i+1}-t_i)}(\omega)\left[B_{t_{i+1}} - B_{t_i}\right](\omega)\right] \tag{B.34}$$

$$= \sum_{i=0}^{n-1} \mathbb{E}\left[\left[B_{t_i+\alpha(t_{i+1}-t_i)} - B_{t_i}\right](\omega)^2\right] \tag{B.35}$$

$$= \alpha t. \tag{B.36}$$

Seeing as this is independent of the discretisation, we see that the choice of α affects greatly the value the integral converges to. Nonetheless, we can develop a coherent theory of integration for stochastic integrals by fixing an α and working through the consequences. There is much debate in the literature as to whether this choice of α is philosophically important, or if it is simply a modelling assumption. While theory exists for $\alpha \in [0,1]$, the two choices that prove most useful are *The Itô integral*, $\alpha = 0$, usually denoted by

The Itô integral

$$\int_a^b f(t,\omega)\,\mathrm{d}B_t(\omega), \tag{B.37}$$

along with the abuse of notation differential equation form

$$\mathrm{d}X_t = \sigma(t, X_t)\,\mathrm{d}B_t \tag{B.38}$$

and *the Stratonovich integral*, $\alpha = \frac{1}{2}$, usually denoted

The Stratonovich integral

$$\int_a^b f(t,\omega) \circ \mathrm{d}B_t(\omega), \tag{B.39}$$

along with the abuse of notation differential equation form

$$\mathrm{d}X_t = \sigma(t,\omega) \circ \mathrm{d}B_t. \tag{B.40}$$

For $\alpha \in [0,1]$ I will use $\circ_\alpha$.

As mentioned earlier, we have introduced these integrals for integrating against Brownian motion, however we can be more general than this. It can be shown that these integrals can be defined (i.e. the Riemann-Stieltjes sums converge in probability on L^2 in the limit) for any integrator $(W_t)_{t\geq 0}$ that is a *semimartingale* (see appendix B.2).

Against a semimartingale $(W_t)_{t\geq 0}$ we can integrate any stochastic process $(H_t)_{t\geq 0}$ that has left-continuous, locally bounded paths that is *adapted* (see appendix B.2) to $(W_t)_{t\geq 0}$.

It should be noted that the time function $W_t = t$ is a semimartingale, and so the drift term and its integrals can also be defined in terms of an Itô or Stratonovich integral. We can therefore write any stochastic differential equation as a collection of semimartingales $(Z_t^i)_{i=1}^n$, coefficient functions, $(c_i)_{i=1}^n$, and a start point X_0 with defining equation

$$X_t = X_0 + \int_0^t \sum_{i=1}^n c_i(t, X_t) \circ_\alpha \mathrm{d}Z_t^i. \tag{B.41}$$

The previously presented form with a drift and diffusion term are given by $Z = (t, B_t)$ and $c = (b(t, X_t), \sigma(t, X_t))$.

Given this, we can also rewrite our stochastic process X_t as (X_t, t), add an additional coefficient function c_{n+1}, constant in the time index, and add another integrator $\mathrm{d}Z_t^{n+1} = \mathrm{d}t$. This allows us to incorporate the time variable into the state space variable, and therefore our coefficient functions can depend only on the state-space variable, as time is explicitly included in this. While it seems like trickery, much of the study of stochastic differential equations is made easier with *time-homogenous* coefficient functions, those that do not have an explicit time dependence separate for a state space dependence. Using this trick we can easily map between the notational convenience of an explicit time dependence, and the technical convenience of time-homogenous coefficients.

Time-homogenous

B.5.1. Advantages to different stochastic integrals. How then should we choose the value of α to use? Both choices of α have advantages and disadvantages. The Itô integral primarily is attractive as function evaluations, $f(c_i, \omega)$, are independent of the Brownian motion increments, $\left[B_{t_{i+1}} - B_{t_i}\right]$. Firstly, this is attractive in definition as the process doesn't "look ahead" to evaluate the integrand. Secondly, if we look at *Itô processes*, stochastic processes that can be written as stochastic differential equations of the form

Itô processes

$$X_t = X_0 + \int_0^t b(s, \omega)\, \mathrm{d}s + \int_0^t \sigma(s, \omega)\, \mathrm{d}B_t, \tag{B.42}$$

and the fact that $\mathbb{E}\left[\int_0^t f(t, \omega)\, \mathrm{d}B_t\right] = 0$ we arrive at the *martingale representation theorem*, which gives the connection that

Martingale representation theorem

$$X_t \text{ is a martingale} \iff X_t \text{ is an Itô process} \tag{B.43}$$

(Øksendal, 2003, theorem 4.3.4). This property is very useful mathematically and makes proving statements about Itô integrals much easier.

On the other hand, the Itô integral behaves clunkily with a change of variable.

Itô's formula *or* Integration by parts

THEOREM B.3. *For an Itô process* $(X_t)_{t\geq 0}$

$$X_t = X_0 + \int_0^t b(s, \omega)\, \mathrm{d}s + \int_0^t \sigma(s, \omega)\, \mathrm{d}B_t \tag{B.44}$$

and a twice differentiable function $g : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ *(i.e.* $g \in C^2(\mathbb{R} \times [0, \infty))$*)*

1. $(Y_t)_{t\geq 0}$, $Y_t = g(X_t, t)$ *is also a Itô process and*

2. *they are related by the stochastic differential equation*

$$\begin{aligned} \mathrm{d}Y_t =& \frac{\partial g}{\partial t}(t, X_t)\,\mathrm{d}t + \frac{\partial g}{\partial x}(t, X_t)\,\mathrm{d}X_t + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(t, X_t)\cdot(\mathrm{d}X_t)^2 && \text{(B.45)} \\ =& \left[\frac{\partial g}{\partial t}(t, X_t) + b(t, X_t)\frac{\partial g}{\partial x}(t, X_t) + \frac{\sigma(t, X_t)^2}{2}\frac{\partial^2 g}{\partial x^2}(t, X_t)\right]\mathrm{d}t \\ & + \sigma(t, X_t)\frac{\partial g}{\partial x}(t, X_t)\,\mathrm{d}B_t && \text{(B.46)} \end{aligned}$$

Proof. Øksendal (2003, theorem 4.1.5). A sketch is as follows:

1. Approximate the integral of Y_t via elementary sums, as per appendix B.5

$$g(t, X_t) = g(0, X_0) + \sum_i \Delta g(t_i, X_i) \tag{B.47}$$

2. Apply a Taylor expansion to this

$$=g(0, X_0) + \sum_j \frac{\partial g}{\partial t}\Delta t_i + \frac{\partial g}{\partial x}\Delta X_i + \frac{1}{2}\frac{\partial^2 g}{\partial t^2}(\Delta t_i)^2 + \frac{1}{2}\,\frac{\partial^2 g}{\partial x \partial t}\Delta B_i \Delta t_i + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}(\Delta B_i)^2 \tag{B.48}$$

   plus higher order terms.

3. We then prove that $\sum_i a_i(\Delta t_i)^2 \to 0$ and $\sum_i a_i \Delta t_i \Delta B_i \to 0$, but that remarkably $\sum_i a_i(\Delta B_i)^2 \to \int_0^t a(s)\,\mathrm{d}s$.

4. We can then expand the above sum in terms of Δt_i and ΔB_i, and take the limit. Collecting the various terms results in the two above formulas.

■

The term-complexity of this change of variable formula can make working with Itô integrals unwieldy.

By contrast, the Stratonovich integral, even against Brownian motion, is in general *not* a martingale. They also "look into the future" in the sense that the evaluation of the integrand in the discretised intervals is statistically dependent on the value of the jump of the integrator. It instead has a different intuitive interpretation. If we define a series of stochastic processes, $\left(B_t^\tau\right)_{t\geq 0}$, that converge to the white noise process in the limit $\tau \to 0$, for example a Gaussian process with kernel $\Sigma_{ij} = \exp\left[\frac{-(t_i - t_j)^2}{\tau^2}\right]$, and we define a series of stochastic differential equations as

$$\frac{\mathrm{d}X_t^\tau}{\mathrm{d}t} = b(t, X_t^\tau) + \sigma(t, X_t^\tau)\,\frac{\mathrm{d}B_t^\tau}{\mathrm{d}t}\,, \tag{B.49}$$

then the solutions to this series of differential equations converges to the solution to the Stratonovich differential equation

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t) \circ \mathrm{d}B_t. \tag{B.50}$$

Change of variable formula for Stratonovich integrals

The *change of variable formula for Stratonovich integrals* is much nicer to work with. It is given by the formula

$$\begin{aligned} \mathrm{d}Y_t &= \frac{\partial g}{\partial t}(t, X_t)\,\mathrm{d}t + \frac{\partial g}{\partial x}(t, X_t) \circ \mathrm{d}X_t && \text{(B.51)} \\ &= \left[\frac{\partial g}{\partial t}(t, X_t) + b(t, X_t)\frac{\partial g}{\partial x}(t, X_t)\right]\mathrm{d}t + \sigma(t, X_t)\frac{\partial g}{\partial x}(t, X_t) \circ \mathrm{d}B_t. && \text{(B.52)} \end{aligned}$$

More generically, for any α, the change of variable formula is given by

$$\begin{aligned} \mathrm{d}Y_t =& \frac{\partial g}{\partial t}(t, X_t)\,\mathrm{d}t + \frac{\partial g}{\partial x}(t, X_t) \circ_\alpha \mathrm{d}X_t + \left(\frac{1}{2} - \alpha\right)\frac{\partial^2 g}{\partial x^2}(t, X_t) \circ_\alpha (\mathrm{d}X_t)^2 && \text{(B.53)} \\ =& \left[\frac{\partial g}{\partial t}(t, X_t) + b(t, X_t)\frac{\partial g}{\partial x}(t, X_t) + \left(\frac{1}{2} - \alpha\right)\sigma(t, X_t)^2\frac{\partial^2 g}{\partial x^2}(t, X_t)\right]\mathrm{d}t \\ &+ \sigma(t, X_t)\frac{\partial g}{\partial x}(t, X_t) \circ_\alpha \mathrm{d}B_t && \text{(B.54)} \end{aligned}$$

B.5.2. Connections between different stochastic integrals. From a technical and philosophical point of view, it is useful that there is a connection between integrals defined with $\alpha \in [0, 1]$. For a stochastic differential equation of the form

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t) \circ_\alpha \mathrm{d}B_t, \tag{B.55}$$

Modified Itô equation

with solution $(X_t)_{t\geq 0}$ then X_t is also the solution to a *modified Itô equation*

$$\mathrm{d}X_t = \left[b(t, X_t) + \alpha\frac{\partial \sigma}{\partial x}(t, X_t)\sigma(s, X_t)\right]\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}B_t. \tag{B.56}$$

Using this connection we can map between using any α we desire, particularly between the Itô and Stratonovich forms for the technical properties.

Additive noise

One very useful observation here is that when the noise is spatially constant, known as *additive noise*, the solutions for all $\alpha \in [0, 1]$ coincide.

B.5.3. Multi-dimension stochastic differential equations. The mathematics we have developed extends to multi-dimension settings as well.

Now consider that we are studying a stochastic process with again index set $[0, T]$, but state space $\mathbb{R}^d$. We have a drift function $\boldsymbol{b} : \mathbb{R}^d \times [0, T] \to \mathbb{R}^d$ and a noise function $\Sigma : \mathbb{R}^d \times [0, T] \to \mathbb{R}^{d\times d}$, where this function is matrix positive semi-definite everywhere. Define a vector of Brownian motions $\mathbf{B}(\omega, t) = (B_1(\omega, t), ..., B_d(\omega, t))$. We can the define a vector of stochastic processes as

$$\begin{aligned} \mathrm{d}\boldsymbol{X}_{t,1} &= \boldsymbol{b}_1(t, \boldsymbol{X}_t)\,\mathrm{d}t + \Sigma_{1,1}(t, \boldsymbol{X}_t) \circ_\alpha \mathrm{d}B_{1,t} + ...\Sigma_{1,d}(t, \boldsymbol{X}_t) \circ_\alpha \mathrm{d}B_{d,t} \\ &\vdots \\ \mathrm{d}\boldsymbol{X}_{t,d} &= \boldsymbol{b}_d(t, \boldsymbol{X}_t)\,\mathrm{d}t + \Sigma_{d,1}(t, \boldsymbol{X}_t) \circ_\alpha \mathrm{d}B_{1,t} + ...\Sigma_{d,d}(t, \boldsymbol{X}_t) \circ_\alpha \mathrm{d}B_{d,t}, \end{aligned} \tag{B.57}$$

or, in matrix notation,

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \Sigma(t, \boldsymbol{X}_t) \circ_\alpha \mathrm{d}\mathbf{B}_t, \tag{B.58}$$

where the stochastic integral of the matrix coefficient Σ against the vector of Brownian motions $\boldsymbol{B}_t$ can be understood as the column-wise sum $\sum_{i=1}^{d} \Sigma_i(t, \boldsymbol{X}_t) \circ_\alpha \mathrm{d}\mathbf{B}_t^i$, in the manner of eq. (B.41).

The change of variable formula can also be extended to the multi-dimension and any α setting.

THEOREM B.4. *Let $(\boldsymbol{X}_t)\_t \geq 0$ multivariate stochastic process defined by the stochastic differential equation* Multivariate change of variable

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \Sigma(t, \boldsymbol{X}_t) \circ_\alpha \mathrm{d}\mathbf{B}_t \tag{B.59}$$

, and assume that it has a solution and that it is unique.

For a coordinate-wise twice differentiable function $g : \mathbb{R}^d \times [0, T] \to \mathbb{R}$, the stochastic process $(\boldsymbol{Y}_t)_{t\geq 0}$, $\boldsymbol{Y}_t = \boldsymbol{g}(t, \boldsymbol{X}_t)$ is given by

$$\begin{aligned}
\mathrm{d}\boldsymbol{Y}_t &= \frac{\partial g}{\partial t}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \nabla g(t, \boldsymbol{X}_t) \cdot \mathrm{d}\boldsymbol{X}_t + \left(\frac{1}{2} - \alpha\right)\nabla\nabla^\top g(t, \boldsymbol{X}_t) : \mathrm{d}\boldsymbol{X}_t\,\mathrm{d}\boldsymbol{X}_t^\top \\
&= \left[\frac{\partial g}{\partial t}(t, \boldsymbol{X}_t) + \nabla g(t, \boldsymbol{X}_t) \cdot \boldsymbol{b}(t, \boldsymbol{X}_t) + \left(\frac{1}{2} - \alpha\right)\nabla\nabla^\top g(t, \boldsymbol{X}_t) : \frac{1}{2}\Sigma(t, \boldsymbol{X}_t)\Sigma(t, \boldsymbol{X}_t)^\top\right]\mathrm{d}t \\
&\quad + \nabla g(t, \boldsymbol{X}_t) \cdot \Sigma(t, \boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{B}_t
\end{aligned}$$

Where $\cdot$ is the dot product, $:$ is the matrix dot product, $A : B \mapsto \sum_{i,j} A_{i,j}, B_{i,j}$, ∇ is the gradient operator, $(\nabla f)_i = \frac{\partial f}{\partial x_i}$, and $\nabla\nabla^T$ is the Hessian operator, $\left(\nabla\nabla^T f\right)_{i,j} = \frac{\partial^2 f}{\partial_i \partial_j}$. This can be extended to functions $g : \mathbb{R}^d \to \mathbb{R}^n$ by applying the formula to each output coordinate.

The quantity $\frac{1}{2}\Sigma(t, \boldsymbol{X}_t)\Sigma(t, \boldsymbol{X}_t)^\top$ appears very commonly, and for compactness, this will be written as $\mathbf{D}(\boldsymbol{X}_t, t)$ where appropriate from now on.

Similar to eq. (B.41) we can write the most general form of a multi-dimension stochastic differential as

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \boldsymbol{c}_i(t, \boldsymbol{X}_t) \circ_\alpha \mathrm{d}Z_t^i \tag{B.60}$$

where now $\boldsymbol{c}_i$ are time-dependent *vector fields*, not just scalar coefficients.

B.5.4. Solutions to stochastic differential equations.

An important question to ask is when a stochastic differential equation that we write down has a solution, and when these solutions are unique. This can be summarised neatly for Euclidean stochastic differential equations via the following theorem

THEOREM B.5. *Let $T > 0$ and* Existance and uniqueness of solutions to a stochastic differential equation

- $b : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$
- $\Sigma : [0, T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$

be measurable functions satisfying

$$\|b(t,x)\|_2 + \|\sigma(t,x)\|_2 \leq C(1 + \|x\|_2), \quad x \in \mathbb{R}^d, t \in [0,T] \tag{B.61}$$

for some constant C, and

$$\|b(t,x) - b(t,y)\|_2 + \|\Sigma(t,x) - \Sigma(t,y)\|_2 \leq D\|x - y\|_2 \tag{B.62}$$

for some constant D.

Let $\mathcal{F}_\infty$ be the σ-algebra generated by an m-dimension Brownian motion $(B_t)_{t\geq 0}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let Z be a random variable independent of this σ-algebra such that $\mathbb{E}\left[\|Z\|_2^2\right] < \infty$.

The stochastic differential equation

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \Sigma(t, X_t)\,\mathrm{d}B_t, \quad 0 \leq t \leq T,\ X_0 \sim Z \tag{B.63}$$

has a unique t-path-continuous solution $X_t(\omega)$ such that at t, $X_t(\omega)$ is adapted to the filtration $\mathcal{F}_t^Z$ generated by Z and $(B_s)_{s\leq t}$. In addition,

$$\mathbb{E}\left[\int_0^T \|X_t\|_2^2\,\mathrm{d}t\right] < \infty. \tag{B.64}$$

Proof. Øksendal (2003, theorem 5.2.1) ∎

The conditions of this theorem are reasonably natural. Condition B.61 tells us that the coefficients of the process do not blow up too fast, and so the process does not explode, i.e. that $\|X_t\|_2 < \infty$ for finite time. Condition B.62, effectively a joint Lipschitz condition, says that the coefficients don't change too rapidly, and guarantees the solutions to the stochastic differential equation is unique, in the sense that if $X_1(t,\omega)$ and $X_2(t,\omega)$ are two solutions to the stochastic differential equation, then

$$X_1(t,\omega) = X_2(t,\omega) \quad \text{for all } 0 \leq t \leq T, \text{ a.s.} \tag{B.65}$$

B.5.5. Weak solutions. By contrast, we can also have *weak* solutions to a given stochastic differential equation. In a strong solution to the differential equation

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \Sigma(t, X_t)\,\mathrm{d}B_t, \quad 0 \leq t \leq T,\ X_0 \sim Z \tag{B.66}$$

We specified a specific Brownian motion B_t and probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for which the solution X_t needs to be adapted to.

By contrast for a weak solution all we specify is the form of the differential equation. We simply require that for a weak solution there exists *some* Brownian motion and probability space for with the stochastic differential equation has a solution that is adapted to the filtration of the Brownian motion. It is often possible to prove the existence of weak solutions without explicitly finding this Brownian motion.

B.6. The Kolmogorov equations

We now turn to studying how functions of Itô processes, the transition density of Itô processes, and the marginal density of Itô processes evolve over time. We will see that we can connect these quantities to standard partial differential equations. Particularly the last of these studies is very useful.

A *diffusion process* is a Markov stochastic process that have almost-surely continuous sample paths. Diffusion process

We will study these in the particular setting of *Itô diffusions*. These are stochastic differential equations of the form Itô diffusions

$$\mathrm{d}X_t = b(t, X_t)\,\mathrm{d}t + \sigma(t, X_t)\,\mathrm{d}B_t, \quad 0 \leq t \leq T,\ X_0 \sim Z \tag{B.67}$$

that satisfy the conditions of theorem B.5.

Consider the Itô change of variable formula for a Borel function $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$, $f \in C^2(\mathbb{R} \times \mathbb{R}^d)$, on the stochastic differential equation $(\boldsymbol{X}_t)_{t \geq 0}$.

$$\begin{aligned} df(t, \boldsymbol{X}_t) =& \left(\frac{\partial f}{\partial t}(t, \boldsymbol{X}_t) + \boldsymbol{b}(t, \boldsymbol{X}_t) \cdot \nabla f(t, \boldsymbol{X}_t) + \mathbf{D}(t, \boldsymbol{X}_t) : \nabla^2 f(t, \boldsymbol{X}_t) \right) \mathrm{d}t && \text{(B.68)} \\ &+ \nabla f(t, \boldsymbol{X}_t)^\top \Sigma(t, \boldsymbol{X}_t)\,\mathrm{d}\mathbf{B}_t && \text{(B.69)} \end{aligned}$$

This gives us the integral formula

$$\begin{aligned} f(t, \boldsymbol{X}_t) - f(s, \boldsymbol{X}_s) =& \int_s^t \left(\frac{\partial f}{\partial t}(t, \boldsymbol{X}_t) + \boldsymbol{b}(u, \boldsymbol{X}_u) \cdot \nabla f(\boldsymbol{X}_u) + \mathbf{D}(u, \boldsymbol{X}_u) : \nabla^2 f(\boldsymbol{X}_u) \right) \mathrm{d}u \\ &+ \int_s^t \nabla f(\boldsymbol{X}_u)^\top \Sigma(u, \boldsymbol{X}_u)\,\mathrm{d}\mathbf{B}_u \end{aligned} \tag{B.70}$$

Taking expectations, and denoting $\boldsymbol{X}_t^{s,y}$ as the process $\boldsymbol{X}_t$ conditioned on $\boldsymbol{X}_s = y$, we get that

$$\begin{aligned} \mathbb{E}\left[f\left(t, \boldsymbol{X}_t^{s,y}\right)\right] - \mathbb{E}\left[f\left(s, \boldsymbol{X}_s^{s,y}\right)\right] &= \mathbb{E}\left[f\left(t, \boldsymbol{X}_t^{s,y}\right)\right] - f(s, y) \\ &= \mathbb{E}\left[\int_s^t \left[\left(\frac{\partial}{\partial u} + L \right) f \right] (u, \boldsymbol{X}_u) \right] \mathrm{d}u \end{aligned} \tag{B.71}$$

as the last term disappears due to the properties of the Itô integral, and the operator L is defined by

$$(Lf)(t, x) = \boldsymbol{b}(t, x) \cdot \nabla f(t, x) + \mathbf{D}(t, x) : \nabla^2 f(t, x) \tag{B.72}$$

B.6.1. The Kolmogorov backward equation. In short, the Kolmogorov backward equation tells us how the expectation of Borel functions of an Itô diffusion will evolve over time, and how the transition density will evolve in the conditioned argument. We study the backward equation first as it is better posed than the forward equation, and the forward equation will follow from the backward.

THEOREM B.6. *Let $f \in C_0^2(\mathbb{R}^d)$ and let $\boldsymbol{X}_t$ be an Itô diffusion. Define* The Kolmogorov backward equation I

$$u(s, y) = \mathbb{E}\left[f\left(\boldsymbol{X}_t^{s,y}\right)\right]. \tag{B.73}$$

Then $u(s,\cdot) \in C_0^2(\mathbb{R}^d)$ for all $s \leq t$ and

$$\left(\frac{\partial}{\partial s} + L\right)u = \frac{\partial u}{\partial s} + Lu = 0, \quad s \leq t \tag{B.74}$$

$$u(t,y) = f(y) \tag{B.75}$$

where Lu is interpreted as L applied to the function $x \mapsto u(s,y)$. Conversely, if eq. (B.74) and eq. (B.75) are satisfied then $u(s,y)$ is the unique solution to these equations.

Proof. Weinan et al. (2021, page 3) ∎

Note how we are differentiating with respect to the time of the *initial condition* of the diffusion we specify, not the time at which we evaluate the expectation. This is where the backward name of this equation is derived from.

One interesting way to view this result is as follows. Define a new operator $Q_{s,y}^t : f \mapsto \mathbb{E}\left[f(\boldsymbol{X}_t^{s,y})\right]$. Then both of the following are true

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(Q_{s,y}^t f\right) = Q_{s,y}^t\left(\left(\frac{\partial}{\partial s} + L\right)f\right) \qquad \text{by eqs. (B.71) and (B.73)} \tag{B.76}$$

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(Q_{s,y}^t f\right) = \left(\frac{\partial}{\partial s} + L\right)\left(Q_{s,y}^t f\right) \qquad \text{by eq. (B.74)} \tag{B.77}$$

and so in this sense the operator evaluating the expectation of f on $\boldsymbol{X}_t$ given an initial condition $\boldsymbol{X}_s = y$ and the operator $\frac{\partial}{\partial s} + L$ commute.

Assuming for the diffusion there exists a transition density, $p_{t|s}(x|y)$, and that this density has continuous bounded derivatives in s and y, then we can prove a result about how this transition density changes as we vary s.

The Kolmogorov backward equation II

THEOREM B.7. *Let $\boldsymbol{X}_t$ be an Itô diffusion and assume that it has a transition density $p_{t|s}(x|y)$ with respect to the Lebesgue measure. Then*

$$\frac{\partial p_{t|s}}{\partial s} + L_y p_{t|s} = 0, \quad s < t,\ p_{s|s}(x|y) = \delta_x(y) \tag{B.78}$$

where the operator L_y is the operator L acting on the y argument of $p_{t|s}(x|y)$

Proof. Let $u(s,y) = \int f(x) p_{t|s}(x|y)\,\mathrm{d}x$. Then

$$\left(\frac{\partial}{\partial s} + L\right)u(s,y) = \left(\frac{\partial}{\partial s} + L\right)\int f(x) p_{t|s}(x|y)\,\mathrm{d}x \tag{B.79}$$

$$= \int f(x)\left(\frac{\partial p_{t|s}}{\partial s}(x|y) + L p_{t|s}(x|y)\right)\mathrm{d}x \tag{B.80}$$

Where exchanging the integral and differential operators holds as we made assumptions about the boundedness and continuity of the derivatives. Since this holds for all test functions f, this holds as the operator and the result follows from theorem B.6. From the same the terminal condition also follows. ∎

B.6.2. The generator of a diffusion. We can arrive at the differential operator in the Kolmogorov backwards equation in a slightly different way. We can define the *generator* of a *time-homogenous* diffusion, one for which the diffusion coefficients are independent of time, as

DEFINITION B.8. *Let $(\boldsymbol{X}_t)_{t\geq 0}$ be a time-homogenous diffusion in $\mathbb{R}^d$. The (infinitesimal) generator A of $\boldsymbol{X}_t$ is defined by its action on a function $f \in C_0^2(\mathbb{R}^d)$,* The generator of a diffusion

$$(Af)(x) = \lim_{s\downarrow 0^+} \frac{\mathbb{E}\left[f\left(\boldsymbol{X}_s^{0,x}\right)\right] - f(x)}{s}. \tag{B.81}$$

The set of functions for which this converges at $x \in \mathbb{R}^d$ is denoted $\mathcal{D}_A(x)$ and the set of functions for which it converges everywhere is denoted $\mathcal{D}_A$.

For Itô diffusions, we can compute the generator in terms of the diffusion's coefficients.

THEOREM B.9. *Let $\boldsymbol{X}_t$ be a time-homogenous Itô diffusion with equation* The generator of an Itô diffusion

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(\boldsymbol{X}_t)\,\mathrm{d}t + \Sigma(\boldsymbol{X}_t)\,\mathrm{d}\mathbf{B}_t, \tag{B.82}$$

and let $f \in C_0^2(\mathbb{R}^d)$, where $C_0^2(\mathbb{R}^d)$ is the set of twice differentiable functions that vanish as $\|x\| \to infty$. Then the generator of this stochastic process is given by

$$(Af)(x) = \boldsymbol{b}(x)\cdot\nabla f(x) + \mathbf{D}(x) : \nabla^2 f(x) \tag{B.83}$$

Proof. To see this,

$$(Af)(x) = \lim_{s\downarrow 0^+} \frac{\mathbb{E}\left[f\left(\boldsymbol{X}_s^{0,x}\right)\right] - f(x)}{s} = \lim_{s\downarrow 0^+} \mathbb{E}\,\frac{1}{s}\int_0^{+s} Lf(\boldsymbol{X}_u^x)\,\mathrm{d}u = (Lf)(x) \tag{B.84}$$

Giving the desired form. The first equality follows from the argument in eq. (B.71) and removing the partial derivative with respect to t, and the second by the dominated convergence theorem (Kallenberg, 1997, Theorem 1.21). ∎

The *adjoint* operator of the generator of a diffusion, A^* or L^*, is also very interesting to us. This is the operator that for any $f \in C_0^2$ and $g \in C^2$, $\langle g, Af\rangle = \langle A^* g, f\rangle$, for $\langle\cdot,\cdot\rangle$ being the L^2 norm on the space of L^2 integrable functions.

THEOREM B.10. *The adjoint of the generator of an Itô diffusion* The adjoit of the generator of an Itô diffusion

$$(Af)(t,x) = \boldsymbol{b}(t,x)\cdot\nabla f(x) + \mathbf{D}(t,x) : \nabla^2 f(x) \tag{B.85}$$

is given by

$$(A^* f)(t,x) = -\nabla\cdot(\boldsymbol{b}f)(t,x) + \nabla^2 : (\mathbf{D}f)(t,x) \tag{B.86}$$

Proof. Using the standard chain rule

$$\boldsymbol{b}\cdot\nabla a = \nabla\cdot(a\boldsymbol{b}) - a\nabla\cdot\boldsymbol{b}, \quad a : \mathbb{R}^d \to \mathbb{R},\ \boldsymbol{b} : \mathbb{R}^d \to \mathbb{R}^d \tag{B.87}$$

and setting $a = f$ and $\boldsymbol{b} = g\boldsymbol{b}$, and using the identity

$$\mathbf{C} : \nabla^2 a = \nabla\cdot(\mathbf{C}\cdot\nabla a) - \nabla\cdot(a\nabla\cdot\mathbf{C}) + a\nabla^2 : \mathbf{B}, \quad a : \mathbb{R}^d \to \mathbb{R},\ \mathbf{C} : \mathbb{R}^d \to \mathbb{R}^{d\times d} \tag{B.88}$$

and by setting $a = f$ and $\mathbf{C} = \frac{1}{2}g\Sigma\Sigma^\top = g\mathbf{D}$, we see that

$$\langle g, Af\rangle = \int_{\mathbb{R}^d}\left[g\left(\boldsymbol{b}\cdot\operatorname{grad} f + \frac{1}{2}\Sigma\Sigma^\top : \left(\nabla^2 f\right)\right)\right](x)\,\mathrm{d}x \tag{B.89}$$

$$= \int_{\mathbb{R}^d}\left[\boldsymbol{b}\cdot\operatorname{grad} a + \mathbf{C} : \left(\nabla^2 a\right)\right](x)\,\mathrm{d}x \tag{B.90}$$

$$= \int_{\mathbb{R}^d}\left[\nabla\cdot(a\boldsymbol{b}) - a\nabla\cdot\boldsymbol{b} + \nabla\cdot(\mathbf{C}\cdot\nabla a) - \nabla\cdot(a\nabla\cdot\mathbf{C}) + a\nabla^2 : \mathbf{B}\right](x)\,\mathrm{d}x \tag{B.91}$$

$$= \int_{\mathbb{R}^d}\left(a\left(-\nabla\cdot\boldsymbol{b} + \nabla^2 : \mathbf{C}\right) + \nabla\cdot(a\boldsymbol{b}) + \nabla\cdot(\mathbf{C}\cdot\nabla a) - \nabla\cdot(a\nabla\cdot\mathbf{C})\right)(x)\,\mathrm{d}x \tag{B.92}$$

$$= \int_{\mathbb{R}^d}\left(a\left(-\nabla\cdot\boldsymbol{b} + \nabla^2 : \mathbf{C}\right)\right)(x)\,\mathrm{d}x \tag{B.93}$$

$$= \int_{\mathbb{R}^d}\left[f\left(-\nabla\cdot(\boldsymbol{b}g) + \nabla^2 : (g\mathbf{D})\right)\right](x)\,\mathrm{d}x \tag{B.94}$$

The second group of terms in eq. (B.92) are all divergences. Applying the *divergence theorem* (see theorem A.50) and assuming that all these terms tend to 0 as $\|x\| \to \infty$, the integral of these terms are zero.

This gives the form of the adjoint operator as required, and is also the form of the adjoint of the L operator. ■

Characteristic operator

It should be noted that a generalisation of the generator exists, called the *characteristic operator* (Øksendal, 2003, section 7.5). This converges on a strictly larger set of functions than the generator, and they coincide on functions where they are both defined. Importantly the form of theorem B.9 holds not just for functions $f \in C_0^2(\mathbb{R}^d)$, but for all functions $f \in C^2(\mathbb{R}^d)$. This greatly expands the space of functions we can study.

The notions of the generator of a diffusion gives us a way of defining Brownian motion via its generator. Note that in the form of the *generator of an Itô diffusion* (see theorem B.9), if we have no drift and a diffusion coefficient given by the identity, then the generator of the diffusion is simply the Laplace operator, $\Delta = \nabla^2 = \sum_{i=1}^d \frac{\partial}{\partial x_i}$. We can therefore define Brownian motion as the diffusion with the Laplace operator as its generator.

B.6.3. The Kolmogorov forward equation. The backward Kolmogorov equation told us how the expectations of Borel functions of a diffusion evolve over time, or how the transition density changes with respect to the time we are transitioning *from*. The forward equation tells us how the transition density changes with respect to the time we are transitioning *to*, and as a knock on how the *time-marginal density* of the diffusion will evolve with time given an initial density.

The Kolmogorov forward equation I

THEOREM B.11. *Let $\boldsymbol{X}_t$ be an Itô diffusion and assume that it has a transition density $p_{t|s}(x|y)$ with respect to the Lebesgue measure. Assume that the derivative of $p_{t|s}(x|y)$ with respect to t is bounded and continuous. Then*

$$\frac{\partial}{\partial t}p_{t|s} = L_x^* p_{t|s}, \quad s < t,\ p_{s|s}(x|y) = \delta_x(y) \tag{B.95}$$

where the operator L_x^* is the adjoint operator L^* acting on the x argument of $p_{t|s}(x|y)$

Proof. Under the assumption that the transition density exists,

$$\mathbb{E}\left[f\left(t, \boldsymbol{X}_t^{s,y}\right)\right] - f(s,y) = \int_{\mathbb{R}^d} f(x) p_{t|s}(x,y)\,\mathrm{d}x - f(y) \tag{B.96}$$

$$= \int_s^t \int_{\mathbb{R}^d} (Lf)(u,x) p_{u|s}(x,y)\,\mathrm{d}x\,\mathrm{d}u \tag{B.97}$$

Taking the partial derivate with respect to time of both sides,

$$\frac{\partial}{\partial t}\left(\int_{\mathbb{R}^d} f(x) p_{t|s}(x,y)\,\mathrm{d}x - f(y)\right) = \frac{\partial}{\partial t}\left(\int_s^t \int_{\mathbb{R}^d} (Lf)(u,x) p_{u|s}(x,y)\,\mathrm{d}x\,\mathrm{d}u\right) \tag{B.98}$$

$$\implies \int_{\mathbb{R}^d} f(x) \frac{\partial}{\partial t}\left(p_{t|s}(x,y)\right)\mathrm{d}x = \int_{\mathbb{R}^d} (Lf)(t,x) p_{t|s}(x,y)\,\mathrm{d}x \tag{B.99}$$

Applying the adjoint to the integral on the right hand side,

$$\implies \int_{\mathbb{R}^d} f(x) \frac{\partial}{\partial t}\left(p_{t|s}(x,y)\right)\mathrm{d}x = \int_{\mathbb{R}^d} f(x)\left(L_x^* p_{t|s}(x,y)\right)\mathrm{d}x \tag{B.100}$$

Since this holds for all test functions f, it holds as an operator. This then defines a weak solution to the problem ∎

We now apply this to the evolution of a density p_t given some initial distribution p_0

THEOREM B.12. *Let $\boldsymbol{X}_t$ be an Itô diffusion, with initial distribution $\boldsymbol{X}_0 \sim p_0$, where p_0 is a density with respect to the Lebesgue measure on $\mathbb{R}^d$. Let $p(t,x)$ denote the density of $\boldsymbol{X}_t$ at time t. Then*

The Kolmogorov forward equation II, The Fokker-Plank eqaution

$$\begin{aligned} \frac{\partial p}{\partial t} &= L^* p \\ p(0,x) &= p_0(x) \end{aligned} \tag{B.101}$$

Proof. Remember

$$p(t,x) = \int_{\mathbb{R}^d} p_{t|0}(x|y) p_0(y)\,\mathrm{d}y. \tag{B.102}$$

If we integrate both sides of eq. (B.95) by $p_0(y)$, we recover exactly eq. (B.101). ∎

Theorem B.12 was in fact discovered in physics before the discovery of the more general Kolmogorov equations. There it is known as the Fokker-Plank equation, and as such is referred to as this sometimes.

The Fokker-Plank equation is a very powerful tool in the study of stochastic processes as we will see in the next two sections.

B.7. Langevin Dynamics

Langevin dynamics is a type of stochastic differential equation for which we can show that the density to which the equation will converge over time is a distribution we can control.

Langevin dynamics
Energy function

Definition B.13. *Let $U : \mathbb{R}^d \to \mathbb{R}$ be a function often called the* energy function. *The* Langevin dynamics *defined by U is given by the time-homogenous stochastic differential equation*

$$\mathrm{d}\boldsymbol{X}_t = -\nabla_{\boldsymbol{X}_t} U(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t \tag{B.103}$$

We can prove that under these dynamics, the distribution $p(x) \propto \exp(-U(x))$ is invariant. I.e. that if the distribution of the solution of stochastic differential equation at a specific time is given by this, then it will not change over time.

Fixed point of Langevin dynamics

Theorem B.14. *For the stochastic differential equation*

$$\mathrm{d}\boldsymbol{X}_t = -\nabla_{\boldsymbol{X}_t} U(\boldsymbol{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t \tag{B.104}$$

, under the condition that

$$-\nabla U(\boldsymbol{X}) \cdot \boldsymbol{X} \leq a|\boldsymbol{X}|^2 + b \tag{B.105}$$

for some constants a, b, the density $p_U(x) = \frac{1}{Z}\exp(-U(x))$ is stable under the evolution of the stochastic differential equation.

Proof. We require that $\frac{\partial}{\partial t}p(t,x) = 0$ when $p(t,x) = p_U(x)$. Applying the Fokker-Plank equation, theorem B.12, we see that this implies that $L^* p_U = 0$

$$\begin{aligned}
L^* p_U &= -\nabla \cdot (\boldsymbol{b} p_U)(t,x) + \nabla^2 : (\mathbf{D} p_U)(t,x) && \text{(B.106)}\\
&= -\nabla \cdot \left(-\nabla U(x)\frac{1}{Z}\exp(-U(x))\right)(t,x) + \nabla^2 : \left(\frac{1}{Z}\exp(-U(x))\mathbf{I}\right)(t,x) && \text{(B.107)}\\
&\propto -\nabla \cdot (\nabla \exp(-U(x)))(t,x) + \nabla^2 : (\exp(-U(x))\mathbf{I})(t,x) && \text{(B.108)}\\
&=\propto -\nabla^2 : (\exp(-U(x)))(t,x) + \nabla^2 : (\exp(-U(x))\mathbf{I})(t,x) && \text{(B.109)}\\
&= 0 && \text{(B.110)}
\end{aligned}$$

where in eq. (B.109) we substituted $\boldsymbol{b}(t,x) = -\nabla U(x)$ and $\mathbf{D}(t,x) = \frac{1}{2}\sqrt{2}^2\mathbf{I}$, eq. (B.109) follows from the gradient of the exponential, and eq. (B.110) from the fact that for scalar functions $a : \mathbb{R}^d \to \mathbb{R}$, $\nabla \cdot \nabla a = \nabla^2 : a\mathbf{I}$. ∎

Moreover, the stochastic differential equation initialised from any point is guaranteed to converge in infinite time to p_U (Roberts and Tweedie, 1996, theorem 2.1). With slightly stronger conditions on U, this convergence will be exponentially quick (Roberts and Tweedie, 1996, theorem 2.2).

B.8. Time-reversal results

A series of very remarkable results exist show that for a given stochastic differential equation, under certain conditions, there exists a process that follows the same dynamics as the process, but in reverse.

Let us have a forward stochastic differential equation

$$\mathrm{d}\boldsymbol{X}_t^{\rightarrow} = \boldsymbol{b}^{\rightarrow}(\boldsymbol{X}_t^{\rightarrow}, t)\,\mathrm{d}t + \Sigma^{\rightarrow}(\boldsymbol{X}^{\rightarrow}, t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0^{\rightarrow} \sim \pi, \tag{B.111}$$

where $\boldsymbol{B}^{\rightarrow}$ is a Brownian motion. Let the densities of the time-marginals of this SDE be denoted p_t, and we let it run from $t = 0$ up to $t = T$.

We are then going to look for a "reverse" process

$$\mathrm{d}\boldsymbol{X}_t^{\leftarrow} = \boldsymbol{b}^{\leftarrow}(\boldsymbol{X}_t^{\leftarrow}, t)\,\mathrm{d}t + \Sigma^{\leftarrow}(\boldsymbol{X}^{\leftarrow}, t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0^{\leftarrow} \sim p_T \tag{B.112}$$

such that for all times $t \in [0, T]$, $\boldsymbol{X}_t^{\rightarrow} \overset{d}{=} \boldsymbol{X}_{T-t}^{\leftarrow}$, i.e. their time-marginal densities are equal, but in opposite time order.

Fokker-Plank time reversal

THEOREM B.15. *Let the stochastic differential equation $\boldsymbol{X}^{\rightarrow}$, called the forward stochastic differential equation, be given by*

$$\mathrm{d}\boldsymbol{X}_t^{\rightarrow} = \boldsymbol{b}(\boldsymbol{X}_t^{\rightarrow}, t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0^{\rightarrow} \sim \pi, \quad t \in [0, T], \tag{B.113}$$

with time-marginal distributions $p_t^{\rightarrow}(\boldsymbol{X})$. Let the stochastic differential equation, called the reverse stochastic differential equation, $\boldsymbol{X}^{\leftarrow}$ be given by

$$\mathrm{d}\boldsymbol{X}_t^{\leftarrow} = \left[-\boldsymbol{b}(\boldsymbol{X}_t^{\leftarrow}, t) + \sigma(t)^2 \nabla \log p_t(\boldsymbol{X}_t^{\leftarrow})\right]\mathrm{d}t + \sigma(t)\,\mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0^{\leftarrow} \sim p_T, \quad t \in [0, T], \tag{B.114}$$

with time-marginal distributions $p^{\leftarrow}{}_t(\boldsymbol{X})$.

The time-marginals of these two stochastic differential equations match, but reversed in time, in the sense that $p_t^{\rightarrow} = p_{T-t}^{\leftarrow}\ \forall\ t \in [0, T]$.

Proof. The Fokker-Plank equation for the forward stochastic differential equation is given by

$$\frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t} = -\operatorname{div}\left(\boldsymbol{b}^{\rightarrow}(\boldsymbol{X}, t) p_t^{\rightarrow}(\boldsymbol{X})\right) + \tfrac{1}{2}\sigma^2(t)\Delta p_t^{\rightarrow}(\boldsymbol{X}) \tag{B.115}$$

Looking at $\Delta p_t^{\rightarrow}(\boldsymbol{X})$ we see that

$$\begin{aligned}
\Delta p_t^{\rightarrow}(\boldsymbol{X}) &= \nabla \cdot \nabla p_t^{\rightarrow}(\boldsymbol{X}) && \text{(B.116)}\\
&= \nabla \cdot \left[\frac{p_t^{\rightarrow}(\boldsymbol{X})}{p_t^{\rightarrow}(\boldsymbol{X})} \nabla p_t^{\rightarrow}(\boldsymbol{X})\right] && \text{(B.117)}\\
&= \nabla \cdot \left[p_t^{\rightarrow}(\boldsymbol{X}) \nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right] && \text{(B.118)}\\
&= \operatorname{div}\left(p_t^{\rightarrow}(\boldsymbol{X}) \nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right) && \text{(B.119)}
\end{aligned}$$

Adding and subtracting this identity we get

$$\begin{aligned}
&\frac{\partial p_t^{\rightarrow}(\boldsymbol{X})}{\partial t} && \text{(B.120)}\\
&= \operatorname{div}\left(\boldsymbol{b}(\boldsymbol{X}, t) p_t^{\rightarrow}(\boldsymbol{X})\right) + \tfrac{1}{2}\sigma^2(t)\Delta p_t^{\rightarrow}(\boldsymbol{X}) - \sigma(t)^2 \operatorname{div}\left(p_t^{\rightarrow}(\boldsymbol{X}) \nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right) && \text{(B.121)}\\
&= \operatorname{div}\left(\left[\boldsymbol{b}(\boldsymbol{X}, t) - \sigma(t)^2 \nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right] p_t^{\rightarrow}(\boldsymbol{X})\right) - \tfrac{1}{2}\sigma^2(t)\Delta p_t^{\rightarrow}(\boldsymbol{X}) && \text{(B.122)}
\end{aligned}$$

Taking the negative of this term we see that the evolution of the time-marginal density of a stochastic differential equation, but with the evolution happing in reverse, is given by

$$\frac{\partial p_t^{\leftarrow}(\boldsymbol{X})}{\partial t} = \operatorname{div}\left(-\left[\boldsymbol{b}(\boldsymbol{X}, t) + \sigma(t)^2 \nabla \log p_t^{\rightarrow}(\boldsymbol{X})\right] p_t^{\rightarrow}(\boldsymbol{X})\right) + \tfrac{1}{2}\sigma^2(t) \Delta p_t^{\rightarrow}(\boldsymbol{X}) \tag{B.123}$$

which is exactly the Fokker-Plank equation of a stochastic differential equation of the form

$$\mathrm{d}\boldsymbol{X}_t^{\leftarrow} = \left[-\boldsymbol{b}(\boldsymbol{X}_t^{\leftarrow}, t) + \sigma(t)^2 \nabla \log p_t^{\rightarrow}(\boldsymbol{X}_t^{\leftarrow})\right] \mathrm{d}t + \sigma(t)\, \mathrm{d}\boldsymbol{B}_t. \tag{B.124}$$

The instantaneous change of the forward and reverse time-marginal densities are therefore equal.

Since by assumption the time-marginal of the reverse stochastic differential equation at time 0 is exactly $p_T(\boldsymbol{X})$, and the evolution of the time-marginals match, but in reverse by the form of the Fokker-Plank equation, it follows that $p_t^{\rightarrow} = p_{T-t}^{\leftarrow},\ t \in [0, T]$. ∎

This proof is not the same as, but is very related to, one found in Nelson (1967, chapter 13). This can easily be extended to the case where the noise is an arbitrary full rank matrix, not just $\sigma(t)\boldsymbol{I}$, and to the case where the noise also depends on the position also, although an additional term appears (Song et al., 2020b).

For the purposes of this thesis, and diffusion modelling generally, this is in fact the only time reversal result we need. There do exist however a series of stronger results that are commonly cited in the literature. The main difference between the result presented and these stronger results is that the stronger result show that not only do the time-marginals of the reverse-time stochastic differential equation match those of the forward-time stochastic differential equation, the *paths* of the forward- and backward-time also match in distribution. Variations on this result can be found in Anderson (1982), Haussmann and Pardoux (1986), Millet et al. (1989), Föllmer (1985), and Cattiaux et al. (2023)

B.9. Connections to ordinary differential equations

We will make one other note-worthy observation using a similar technique to the proof presented for the reverse-time stochastic process in the last section.

Time-marginal ordinary differential equation

Theorem B.16. *For a stochastic differential equation*

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\, \mathrm{d}t + \sigma(t)\, \mathrm{d}\boldsymbol{B}_t, \quad \boldsymbol{X}_0 \sim p_0 \tag{B.125}$$

with time-marginal distributions $p_t(\boldsymbol{X})$, the ODE

$$\mathrm{d}\boldsymbol{X}_t = \left[\boldsymbol{b}(t, \boldsymbol{X}_t) - \frac{1}{2}\sigma(t)^2 \nabla \log p_t(\boldsymbol{X}_t)\right] \mathrm{d}t, \quad \boldsymbol{X}_0 \sim p_0 \tag{B.126}$$

has the same time-marginal distributions $p_t(\boldsymbol{X})$ as the forward-time stochastic differential equation.

Proof. The Fokker-Plank equation for the forward stochastic differential equation is given by

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -\operatorname{div}(\boldsymbol{b}(\boldsymbol{X}, t)p_t(\boldsymbol{X})) + \tfrac{1}{2}\sigma^2(t)\Delta p_t(\boldsymbol{X}) \tag{B.127}$$

As before, $\Delta p_t(\boldsymbol{X}) = \operatorname{div}(p_t(\boldsymbol{X})\nabla \log p_t(\boldsymbol{X}))$. Adding and subtracting half this identity we get

$$\frac{\partial p_t(\boldsymbol{X})}{\partial t} = -\operatorname{div}\left(\left[\boldsymbol{b}(\boldsymbol{X}, t) - \frac{1}{2}\sigma(t)^2\nabla \log p_t(\boldsymbol{X})\right]p_t(\boldsymbol{X})\right) + 0 * \tfrac{1}{2}\sigma^2(t)\Delta p_t(\boldsymbol{X}) \tag{B.128}$$

This is exactly the form of the Fokker-Plank equation for a stochastic differential equation of the form

$$\begin{aligned} \mathrm{d}\boldsymbol{X}_t &= \left[\boldsymbol{b}(\boldsymbol{X}_t, t) - \frac{1}{2}\sigma(t)^2\nabla \log p_t(\boldsymbol{X}_t)\right]\mathrm{d}t + 0 * \sigma(t)\,\mathrm{d}\boldsymbol{B}_t && \text{(B.129)} \\ &= \left[\boldsymbol{b}(\boldsymbol{X}_t, t) - \frac{1}{2}\sigma(t)^2\nabla \log p_t(\boldsymbol{X}_t)\right]\mathrm{d}t. && \text{(B.130)} \end{aligned}$$

Since by assumption the time-marginal of this stochastic differential equation at time 0 is exactly $p_0(\boldsymbol{X})$, and the evolution of the time-marginals of this stochastic differential equation match those of the forward time stochastic differential equation, it follows that at all times the time-marginals of the two match. ∎

This gives us an interesting way of connecting a stochastic differential equation to a deterministic ordinary differential equation. Note however the *paths* of this ordinary differential equation are wildly different to that of the stochastic differential equation, unlike in the reverse-time case. If all one cares about is sampling independently of each time marginal however, this is not an issue. If one has access to the quantity $\nabla \log p_t(\boldsymbol{X}_t)$ and samples from p_0 it would be possible to use out of the box equation solvers with guaranteed error tolerance to sample from each marginal.

B.10. Approximate sampling from stochastic differential equations

We cannot in a computer represent the continuous paths of a stochastic differential equation exactly. Nor can we typically find closed form solutions to stochastic differential equations. Therefore, we need to approximate stochastic differential equations in discrete time.

For approximating an ordinary differential equation of the form

$$\mathrm{d}\boldsymbol{X}_t = f(t, \boldsymbol{X}_t)\,\mathrm{d}t \tag{B.131}$$

with an initial condition $\boldsymbol{X}_0$, the most simple approximate solution is *Euler integration*. Picking a temporal step size Δt, we iteratively take steps of the form

Euler integration

$$\boldsymbol{X}_{n+1} = \boldsymbol{X}_n + f(n\Delta t, \boldsymbol{X}_n)\Delta t \tag{B.132}$$

to produce an approximate solution $(\boldsymbol{X}_n)_{n=0}^N$, where $\boldsymbol{X}_n$ corresponds to time $n\Delta t$. This scheme has discretisation error of $O(\Delta t)$, and so we can increase precision arbitrarily with decreasing step size.

Analogously for a stochastic differential equation of the form

$$\mathrm{d}\boldsymbol{X}_t = \boldsymbol{b}(t, \boldsymbol{X}_t)\,\mathrm{d}t + \Sigma(t, \boldsymbol{X}_t)\,\mathrm{d}\boldsymbol{B}_t \tag{B.133}$$

Euler-Maruyama integration

with initial condition $\boldsymbol{X}_0$, we can apply the *Euler-Maruyama integration* to approximate the solution. For a chosen step size Δt we iteratively take steps of the form

$$\boldsymbol{X}_{n+1} = \boldsymbol{X}_n + \boldsymbol{b}(t, \boldsymbol{X}_n)\Delta t + \Sigma^{1/2}(t, \boldsymbol{X}_n)\boldsymbol{z}\sqrt{\Delta t} \tag{B.134}$$

where $\boldsymbol{z} \sim \mathcal{N}(0, I)$, a random Gaussian vector of the dimension of the Brownian motion.

Let $\boldsymbol{X}_T$ be the solution to the stochastic differential equation at time T, and let $\boldsymbol{X}_T^n$ be the solution to the discretised scheme of n steps from 0 to T, giving a step size of $\Delta t = T/n$. Under the condition that the elements of $\boldsymbol{b}$ and Σ are C^∞ functions with bounded derivatives, and for any measurable and bounded function f, we have that

$$\mathbb{E}[f(\boldsymbol{X}_T)] - \mathbb{E}\left[f(\boldsymbol{X}_T^n)\right] = -A\Delta t + B(\Delta t)^2 \tag{B.135}$$

where the constants A, B depend on f, T (Bally and Talay, 1996, Theorem 3.1). As such, the Euler-Maruyama scheme *converges in distribution* (see appendix B.1) to the solution to the original stochastic differential equation as we decrease step size of the solution.

C | Score-based modelling on Riemannian manifolds

Organization of the appendix

In this supplementary we first introduce notation in Appendix C.1. We gather the proof of Theorem 3.6 as well as additional derivations on score-based generative models and Riemannian manifolds. In appendix C.2, we recall basics on stochastic Riemannian geometry following Hsu (2002). In section 3.2.5, we introduce an extension to the Riemannian setting of the likelihood computation techniques in diffusion models. We present an extension of Algorithm 3.11 using predictor-corrector schemes in Section 3.2.7. In appendix B.8, we prove the extension of the time-reversal formula to manifold in Theorem 3.6. We prove the convergence of RSGM, i.e. Theorem 3.13, in Appendix C.4. The proof of the manifold implicit score matching loss, drawing links between the denoising score matching loss and the implicit score matching loss, is presented appendix C.5. Experimental details are given in appendix C.6

C.1. Notation

We refer to Appendix C.2 for more details about the basic concepts of Riemannian geometry and stochastic processes. In this section, we merely introduce the notation used in our work. We postpone an introduction to stochastic processes on manifolds to Appendix C.2.2.

In this work we always consider a smooth, connected and complete manifold $\mathcal{M}$. We focus on the case of Riemannian manifolds, namely manifolds equipped with a metric g. Metrics g are smooth scalar product on the manifold allowing us to define the notion of *distance* on a manifold. We refer to Appendix C.2 for a precise definition and a discussion on metrics. Given a smooth map $f \in \mathrm{C}^\infty(\mathcal{M}, \mathbb{R})$, the gradient ∇f is defined for any $f : \mathcal{M} \to \mathbb{R}$, $x \in \mathcal{M}, v \in \mathrm{T}_x\mathcal{M}$, $\langle \nabla f, v \rangle_g = \mathrm{d}f(v)$. The distance $d_{\mathcal{M}}(x, y)$ is defined as the infimum of the length of all the curves on $\mathcal{M}$ joining x and y. Geodesics are path defined on $\mathcal{M}$ by a second order equation (and a starting point and speed). This second order equation corresponds to the first order minimization of an *energy* functional whose minimizers also minimize the length. In Appendix C.2, we introduce the notion of geodesics using parallel transport. The exponential mapping $\exp_x : \mathsf{U}\mathcal{M} \to \mathcal{M}$ with $\mathsf{U} \subset \mathrm{T}_x\mathcal{M}$ is such that

$\exp_x(v) = \gamma(1)$ with $\gamma(1)$ the geodesics with initial condition (x, v) at time $t = 1$. Finally the volume form is a differentiable form of same degree as the dimension of $\mathcal{M}$. Since $\mathcal{M}$ is an orientable Riemannian manifold there is a natural volume form defined using the metric g, namely $\omega(x) = |g(x)|^{1/2} \mathrm{d}x_1 \wedge \ldots \mathrm{d}x_d$. In this paper, we abuse notation and denote by the volume form this natural volume form.

C.2. Preliminaries on stochastic Riemannian geometry

In this section, we recall some basic facts on Riemannian geometry and stochastic Riemannian geometry. We follow Hsu (2002), Lee (2018), and Lee (2006) and refer to Lee (2010) and Lee (2013) for a general introduction to topological and smooth manifolds. Throughout this section $\mathcal{M}$ is a d-dimensional smooth manifold, $\mathrm{T}\mathcal{M}$ its tangent bundle and $\mathrm{T}\mathcal{M}^*$ its cotangent bundle. We denote $\mathrm{C}^\infty(\mathcal{M})$ the set of real-valued smooth functions on $\mathcal{M}$ and $\mathcal{X}(\mathcal{M})$ the set of vector fields on $\mathcal{M}$.

C.2.1. Tensor field, metric, connection and transport.

Tensor field and Riemannian metric For a vector space V let $\mathrm{T}^{k,\ell}(V) = V^{\otimes k} \otimes (V^\star)^{\otimes \ell}$ with $k, \ell \in \mathbb{N}$. For any $k, \ell \in \mathbb{N}$ we define the space of (k, ℓ)-tensors as $\mathrm{T}^{k,\ell}\mathcal{M} = \sqcup_{p \in \mathcal{M}} \mathrm{T}^{k,\ell}(\mathrm{T}_p\mathcal{M})$. Note that $\Gamma(\mathcal{M}, \mathrm{T}^{0,0}\mathcal{M}) = \mathrm{C}^\infty(\mathcal{M})$, $\mathcal{X}(\mathcal{M}) = \Gamma(\mathcal{M}, \mathrm{T}^{1,0}\mathcal{M})$ and that the space of 1-form on $\mathcal{M}$ is given by $\Gamma(\mathcal{M}, \mathrm{T}^{0,1}\mathcal{M})$, where $\Gamma(\mathcal{M}, V(\mathcal{M}))$ is a section of a vector bundle $V(\mathcal{M})$ (see Lee, 2013, Chapter 10). For any $k \in \mathbb{N}$, we denote $\mathrm{T}^{|k|}\mathcal{M} = \sqcup_{j=0}^{k} \mathrm{T}^{j,k-j}\mathcal{M}$. $\mathcal{M}$ is said to be a Riemannian manifold if there exists $g \in \Gamma(\mathcal{M}, \mathrm{T}^{0,2}\mathcal{M})$ such that for any $x \in \mathcal{M}$, $g(x)$ is positive definite. g is called the Riemannian metric of $\mathcal{M}$. Every smooth manifold can be equipped with a Riemannian metric (see Lee, 2018, Proposition 2.4). In local coordinates we define $G = \{g_{i,j}\}_{1 \leq i,j \leq d} = \{g(X_i, X_j)\}_{1 \leq i,j \leq d}$, where $\{X_i\}_{i=1}^d$ is a basis of the tangent space. In what follows we consider that $\mathcal{M}$ is equipped with a metric g and for any $X, Y \in \mathcal{X}(\mathcal{M})$ we denote $\langle X, Y \rangle_{\mathcal{M}} = g(X, Y)$.

Connection A connection ∇ is a mapping which allows one to differentiate vector fields with respect to other vector fields. ∇ is a linear map $\nabla : \mathcal{X}(\mathcal{M}) \times \mathcal{X}(\mathcal{M}) \to \mathcal{X}(\mathcal{M})$. In addition, we assume that i) for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $X, Y \in \mathcal{X}(\mathcal{M})$, $\nabla_{fX}(Y) = f\nabla_X Y$, ii) for any $f \in \mathrm{C}^\infty(\mathcal{M}), X, Y \in \mathcal{X}(\mathcal{M}), \nabla_X(fY) = f\nabla_X Y + X(f)Y$. Given a system of local coordinates, the Christoffel symbols $\{\Gamma_{i,j}^k\}_{1 \leq i,j,k \leq d}$ are given for any $i, j \in \{1, \ldots, d\}$ by $\nabla_{X_i} X_j = \sum_{k=1}^d \Gamma_{i,j}^k X_k$. We also define the Levi–Civita connection ∇ by considering the additional two conditions: i) ∇ is torsion-free, i.e. for any $X, Y \in \mathcal{X}(\mathcal{M})$ we have $\nabla_X Y - \nabla_Y X = [X, Y]$, where $[X, Y]$ is the Lie bracket between X and Y, ii) ∇ is compatible with the metric g, i.e. for any $X, Y, Z \in \mathcal{X}(\mathcal{M})$, $X(\langle Y, Z \rangle_{\mathcal{M}}) = \langle \nabla_X Y, Z \rangle_{\mathcal{M}} + \langle Y, \nabla_X Z \rangle_{\mathcal{M}}$. We recall that the Levi–Civita connection is uniquely defined since for any $X, Y, Z \in \mathcal{X}(\mathcal{M})$ we have

$$2\langle \nabla_X Y, Z \rangle_{\mathcal{M}} = X(\langle Y, Z \rangle_{\mathcal{M}}) + Y(\langle Z, X \rangle_{\mathcal{M}}) - Z(\langle X, Y \rangle_{\mathcal{M}}) \tag{C.1}$$

$$+ \langle [X, Y], Z \rangle_{\mathcal{M}} - \langle [Z, X], Y \rangle_{\mathcal{M}} - \langle [Y, Z], X \rangle_{\mathcal{M}}. \tag{C.2}$$

In this case, the Christoffel symbols are given for any $i, j, k \in \{1, \dots, d\}$ by

$$\Gamma_{i,j}^k = \frac{1}{2} \sum_{m=1}^{d} g^{km}(\partial_j g_{m,i} + \partial_i g_{m,j} - \partial_m g_{i,j}), \tag{C.3}$$

where $\{g^{i,j}\}_{1 \le i,j \le d} = G^{-1}$. Note that if $\mathcal{M}$ is Euclidean then for any $i, j, k \in \{1, \dots, d\}$, $\Gamma_{i,j}^k = 0$. We also extend the connection so that for any $X \in \mathcal{X}(\mathcal{M})$ and $f \in \mathrm{C}^\infty(M)$ we have $\nabla_X f = X(f)$. In particular, we have that $\nabla_X f \in \mathrm{C}^\infty(\mathcal{M})$. In addition, we extend the connection such that for any $\alpha \in \Gamma(\mathcal{M}, \mathrm{T}^{0,1}\mathcal{M})$, $X, Y \in \mathcal{X}(\mathcal{M})$ we have $\nabla_X \alpha(Y) = \alpha(\nabla_X Y) - X(\alpha(Y))$. In particular, we have that $\nabla_X \alpha \in \Gamma(\mathcal{M}, \mathrm{T}^{1,0}\mathcal{M})$. Note that for any $X \in \mathcal{X}(\mathcal{M})$ and $\alpha, \beta \in \mathrm{T}^{|1|}\mathcal{M}$ we have $\nabla_X(\alpha \otimes \beta) = \nabla_X \alpha \otimes \beta + \alpha \otimes \nabla_X \beta$. Similarly, we can define recursively $\nabla_X \alpha$ for any $\alpha \in \Gamma(\mathcal{M}, \mathrm{T}^{k,\ell}\mathcal{M})$ with $k, \ell \in \mathbb{N}$. Such an extension is called a covariant derivative.

Parallel transport, geodesics and exponential mapping Given a connection, we can define the notion of parallel transport, which transports vector fields along a curve. Let $\gamma : [0,1] \to \mathcal{M}$ be a smooth curve. We define the covariant derivative along the curve γ by $D_{\dot\gamma} : \mathcal{X}(\gamma) \to \mathcal{X}(\gamma)$ similarly to the connection, where $\mathcal{X}(\gamma) = \Gamma(\gamma([0,1]), \mathrm{T}\mathcal{M})$. In particular if $\dot\gamma$ and $X \in \mathcal{X}(\gamma)$ can be extended to $\mathcal{X}(\mathcal{M})$ then we define $D_{\dot\gamma}(X) = \nabla_{\dot\gamma} X \in \mathcal{X}(\mathcal{M})$. In what follows, we denote $D = \nabla$ for simplicity. We say that $X \in \mathcal{X}(\gamma)$ is parallel to γ if for any $t \in [0,1]$, $\nabla_{\dot\gamma} X(t) = 0$. In local coordinates, let $X \in \mathcal{X}(\gamma)$ be given for any $t \in [0,1]$ by $X = \sum_{i=1}^d a_i(t) E_i(t)$ (assuming that $\gamma([0,1])$ is entirely contained in a local chart), then we have that for any $t \in [0,1]$ and $k \in \{1, \dots, d\}$

$$\dot a_k(t) + \sum_{i,j=1}^{d} \Gamma_{i,j}^k(x(t)) \dot x_i(t) a_j(t) = 0. \tag{C.4}$$

A curve γ on $\mathcal{M}$ is said to be a geodesic if $\dot\gamma$ is parallel to γ. Using eq. (C.4) we get that

$$\ddot x_k(t) + \sum_{i,j=1}^{d} \Gamma_{i,j}^k(x(t)) \dot x_i(t) \dot x_j(t) = 0. \tag{C.5}$$

For more details on geodesics and parallel transport, we refer to Lee (2018, Chapter 4). In addition, we have that parallel transport provides a linear isomorphism between tangent spaces. Indeed, let $v \in \mathrm{T}_x\mathcal{M}$ and $\gamma : [0,1] \to \mathcal{M}$ with $\gamma(0) = x$ a smooth curve. Then, there exists a unique vector field $X^v \in \mathcal{X}(\gamma)$ such that $X^v(x) = v$ and X^v is parallel to γ. For any $t \in [0,1]$, we denote $\Gamma_0^t : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{\gamma(t)}\mathcal{M}$ the linear isomorphism such that $\Gamma_0^t(v) = X^v(\gamma(t))$.

For any $x \in \mathcal{M}$ and $v \in \mathrm{T}_x\mathcal{M}$ we denote $\gamma^{x,v} : [0, \varepsilon^{x,v}]$ the geodesics (defined on the maximal interval $[0, \varepsilon^{x,v}]$) on $\mathcal{M}$ such that $\gamma(0) = x$ and $\dot\gamma(0) = v$. We denote $\mathrm{U}^x = \{v \in \mathrm{T}_x\mathcal{M} : \varepsilon^{x,v} \ge 1\}$. Note that $0 \in \mathrm{U}^x$. For any $x \in \mathcal{M}$, we define the exponential mapping $\exp_x : \mathrm{U}^x \to \mathcal{M}$ such that for any $v \in \mathrm{U}^x$, $\exp_x(v) = \gamma^{x,v}(1)$. If for any $x \in \mathcal{M}$, $\mathrm{U}^x = \mathrm{T}_x\mathcal{M}$, the manifold is called *geodesically complete*. As any connected compact manifold is geodesically complete, there exists a geodesic between any two points $x, y \in \mathcal{M}$ (see Lee, 2018, Lemma 6.18). For any $x, y \in \mathcal{M}$, we denote $\mathrm{Geo}_{x,y}$ the sets of geodesics γ such that $\gamma(0) = x$ and $\gamma(y) = 1$. For any $x, y \in \mathcal{M}$ we denote $\Gamma_x^y(\gamma) : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_y\mathcal{M}$ the linear isomorphism such

that for any $v \in \mathrm{T}_x\mathcal{M}$, $\Gamma_x^y(v) = X^v(\gamma(1))$, where $\gamma \in \mathrm{Geo}_{x,y}$. Note that for any $x \in \mathcal{M}$ there exists $\mathsf{V}^x \subset \mathcal{M}$ such that $x \in \mathsf{V}^x$ and for any $y \in \mathsf{V}^x$ we have that $\left|\mathrm{Geo}_{x,y}\right| = 1$. In this case, we denote $\Gamma_x^y = \Gamma_x^y(\gamma)$ with $\gamma \in \mathrm{Geo}_{x,y}$.

Orthogonal projection We will make repeated use of orthonormal projections on manifolds. Recall that since $\mathcal{M}$ is a closed Riemannian manifold we can use the Nash embedding theorem (Gunther, 1991). In the rest of this paragraph, we assume that $\mathcal{M}$ is a Riemannian submanifold of $\mathbb{R}^p$ for some $p \in \mathbb{N}$ such that its metric is induced by the Euclidean metric. In order to define the projection we introduce

$$\mathrm{unpp}(\mathcal{M}) = \left\{x \in \mathbb{R}^d \ : \ \text{there exists a unique } \xi_x \text{ such that } \|x - \xi_x\| = d(x, \mathcal{M})\right\}. \tag{C.6}$$

Let $\mathcal{E}(\mathcal{M}) = \mathrm{int}(\mathrm{unpp}(\mathcal{M}))$. By Leobacher and Steinicke (2021, Theorem 1), we have $\mathcal{M} \subset \mathcal{E}(\mathcal{M})$. We define $\tilde{p} : \mathcal{E}(\mathcal{M}) \to \mathcal{M}$ such that for any $x \in \mathcal{E}(\mathcal{M})$, $\tilde{p}(x) = \xi_x$. Using Leobacher and Steinicke (2021, Theorem 2), we have $\tilde{p} \in \mathrm{C}^\infty(\mathbb{R}^p, \mathcal{M})$ and for any $x \in \mathcal{M}$, $\tilde{P}(x) = \mathrm{d}\tilde{p}(x)$ is the orthogonal projection on $\mathrm{T}_x\mathcal{M}$. Since $\mathbb{R}^p$ is normal and $\mathcal{M}$ and $\mathcal{E}(\mathcal{M})^\mathrm{c}$ are closed, there exists f open such that $\mathcal{M} \subset \mathsf{f} \subset \mathcal{E}(\mathcal{M})$. Let $p \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^p)$ such that for any $x \in \mathsf{f}$, $p(x) = \tilde{p}(x)$ (given by Whitney extension theorem for instance). Finally, we define $P : \mathbb{R}^p \to \mathbb{R}^p$ such that for any $x \in \mathbb{R}^p$, $P(x) = \mathrm{d}p(x)$. Note that for any $x \in \mathcal{M}$, $P(x)$ is the orthogonal projection $\mathrm{T}_x\mathcal{M}$ and that $P \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^p)$.

C.2.2. Stochastic Differential Equations on manifolds.

Stratonovich integral For reasons that will become clear in the next paragraph, it is easier to define Stochastic Differential Equations (SDEs) on manifolds with respect to the Stratonovich integral (Kloeden and Platen, 2011, Part II, Chapter 3). We consider a filtered probability space $(\Omega, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. Let $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ be two real continuous semimartingales. We define the quadratic covariation $([X, Y]_t)_{t\geq 0}$ such that for any $t \geq 0$

$$[X, Y]_t = X_t Y_t - X_0 Y_0 - \int_0^t X_s \mathrm{d}Y_s - \int_0^t Y_s \mathrm{d}X_s. \tag{C.7}$$

We refer to Revuz and Yor (1999, Chapter IV) for more details on semimartingales and quadratic variations. We denote $[X] = [X, X]$. In particular, we have that $([X, Y]_t)_{t\geq 0}$ is an adapted continuous process with finite-variation and therefore $[[X, Y]] = 0$. Let $(X_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ be two real continuous semimartingales, then we define the Stratonovich integral as follows for any $t \geq 0$

$$\int_0^t X_s \circ \mathrm{d}Y_s = \int_0^t X_s \mathrm{d}Y_s + \tfrac{1}{2}[X, Y]_t. \tag{C.8}$$

In particular, denoting $(Z_t^1)_{t\geq 0}$ and $(Z_t^2)_{t\geq 0}$ the processes such that for any $t \geq 0$, $Z_t^1 = \int_0^t X_s \circ \mathrm{d}Y_s$ and $Z_t^2 = \int_0^t X_s \mathrm{d}Y_s$, we have that $[Z^1] = [Z^2]$. We refer to Kurtz et al. (1995) for more details on Stratonovich integrals. Note that if for any $t \geq 0$, $X_t = \int_0^t f(X_s) \circ \mathrm{d}Y_s$ with $\mathrm{C}^1(\mathbb{R}, \mathbb{R})$, then $[X, Y]_t = \int_0^t f(X_s) f'(X_s) \mathrm{d}Y_s$. Assuming that $f \in \mathrm{C}^3(\mathbb{R}, \mathbb{R})$ we have that (Revuz and Yor, 1999, Chapter IV, Exercise 3.15)

$$f(X_t) = f(X_0) + \int_0^t f'(X_s) \circ \mathrm{d}X_s. \tag{C.9}$$

The proof relies on the fact that for any $t \geq 0$, $\mathrm{d}[\boldsymbol{X}, f'(\boldsymbol{X})]_t = f''(\boldsymbol{X}_t)\mathrm{d}[\boldsymbol{X}]_t$. This result should be compared with Itô's lemma. In particular, Stratonovich calculus satisfies the ordinary chain rule making it a useful tool in differential geometry which makes a heavy use of diffeomorphism. Finally, we have the following correspondence between Stratonovich and Itô SDEs. Assume that $(\boldsymbol{X}_t)_{t\in[0,T]}$ is a strong solution to $\mathrm{d}\boldsymbol{X}_t = b(t, \boldsymbol{X}_t)\mathrm{d}t + \sigma(t, \boldsymbol{X}_t) \circ \mathrm{d}\boldsymbol{b}_t$, with $b \in \mathrm{C}^\infty(\mathbb{R}^d, \mathbb{R}^d)$ and $\sigma \in \mathrm{C}^\infty(\mathbb{R}^d, \mathbb{R}^{d\times d})$. Then, we have that

$$\mathrm{d}\boldsymbol{X}_t = \{b(t, \boldsymbol{X}_t) + \bar{b}(\boldsymbol{X}_t)\}\mathrm{d}t + \sigma(t, \boldsymbol{X}_t)\mathrm{d}\boldsymbol{b}_t, \qquad \bar{b} = (-1/2)[\mathrm{div}(\sigma\sigma^\top) - \sigma\mathrm{div}(\sigma^\top)]. \tag{C.10}$$

where for any $\mathrm{A} \in \mathrm{C}^\infty(\mathbb{R}^d, \mathbb{R}^{d\times d})$ we have that $\mathrm{div}(\mathrm{A}) \in \mathrm{C}^\infty(\mathbb{R}^d, \mathbb{R}^d)$ and for any $i \in \{1, \ldots, d\}$ and $x \in \mathbb{R}^d$, $\mathrm{div}(\mathrm{A})_i(x) = \sum_{j=1}^d \partial_j \mathrm{A}_{i,j}(x)$. In particular, note that if for $x_0 \in \mathbb{R}^d$, $\sigma(x_0)$ is an orthogonal projection, then $\sigma(x_0)\bar{b}(x_0) = 0$.

SDEs on manifolds We define semimartingales and SDEs on manifold through the lens of their actions on functions. A continuous $\mathcal{M}$-valued stochastic process $(\boldsymbol{X}_t)_{t\geq 0}$ is called an $\mathcal{M}$-valued semimartingale if for any $f \in \mathrm{C}^\infty(\mathcal{M})$ we have that $(f(\boldsymbol{X}_t))_{t\geq 0}$ is a real valued semimartingale. Let $\ell \in \mathbb{N}$, $V^{1:\ell} = \{V_i\}_{i=1}^\ell \in \mathcal{X}(\mathcal{M})^\ell$ and $Z^{1:\ell} = \{Z^i\}_{i=1}^\ell$ a collection of ℓ real-valued semimartingales. A $\mathcal{M}$-valued semimartingale $(\boldsymbol{X}_t)_{t\geq 0}$ is said to be the solution of $\mathrm{SDE}(V^{1:\ell}, Z^{1:\ell}, \boldsymbol{X}_0)$ up to a stopping τ with $\boldsymbol{X}_0$ a $\mathcal{M}$-valued random variable if for all $f \in \mathrm{C}^\infty(\mathcal{M})$ and $t \in [0, \tau]$ we have

$$f(\boldsymbol{X}_t) = f(\boldsymbol{X}_0) + \sum_{i=1}^\ell \int_0^t V_i(f)(\boldsymbol{X}_s) \circ \mathrm{d}\boldsymbol{Z}_s^i. \tag{C.11}$$

Since the previous SDE is defined with respect to the Stratonovich integral we have that if $(\boldsymbol{X}_t)_{t\geq 0}$ is a solution of $\mathrm{SDE}(V^{1:\ell}, Z^{1:\ell}, \boldsymbol{X}_0)$ and $\Phi : \mathcal{M} \to \mathcal{N}$ is a diffeomorphism then $(\Phi(\boldsymbol{X}_t))_{t\geq 0}$ is a solution of $\mathrm{SDE}(\Phi_\star V^{1:\ell}, Z^{1:\ell}, \Phi(\boldsymbol{X}_0))$, where $\Phi_\star$ is the pushforward operation (see Hsu, 2002, Proposition 1.2.4). Because the vector fields $\{V_i\}_{i=1}^\ell$ are smooth we have that for any $\ell \in \mathbb{N}$, $V^{1:\ell} = \{V_i\}_{i=1}^\ell \in \mathcal{X}(\mathcal{M})^\ell$ and $Z^{1:\ell} = \{Z^i\}_{i=1}^\ell$ a collection of ℓ real-valued semimartingales, there exists a unique solution to $\mathrm{SDE}(V^{1:\ell}, Z^{1:\ell}, \boldsymbol{X}_0)$ (see Hsu, 2002, Theorem 1.2.9).

C.2.3. Brownian motion on manifolds. In this section, we introduce the notion of Brownian motion on manifolds. We derive some of its basic convergence properties and provide alternative definitions (stochastic development, isometric embedding, random walk limit). These alternative definitions are the basis for our alternative methodologies to sample from the time-reversal. To simplify our discussion, we assume that $\mathcal{M}$ is a connected compact orientable Riemannian manifold equipped with the Levi–Civita connection ∇. We denote p_{ref}^m the Hausdorff measure of the manifold (which coincides with the measure associated with the Riemannian volume form (see Federer, 2014, Theorem 2.10.10) and $p_{\mathrm{ref}} = p_{\mathrm{ref}}^m / p_{\mathrm{ref}}(\mathcal{M})$ the associated probability measure.

Gradient, divergence and Laplace operators Let $f \in \mathrm{C}^\infty(\mathcal{M})$. We define $\nabla f \in \mathcal{X}(\mathcal{M})$ such that for any $X \in \mathcal{X}(\mathcal{M})$ we have $\langle X, \nabla f\rangle_{\mathcal{M}} = X(f)$. Let $\{X_i\}_{i=1}^d \in \mathcal{X}(\mathcal{M})^d$ such that for any $x \in \mathcal{M}$, $\{X_i(x)\}_{i=1}^d$ is an orthonormal basis of $\mathrm{T}_x\mathcal{M}$. Then, we define $\mathrm{div} : \mathcal{X}(\mathcal{M}) \to \mathrm{C}^\infty(\mathcal{M})$ (linear) such that for any $X \in \mathcal{X}(\mathcal{M})$, $\mathrm{div}(X) = \sum_{i=1}^d \langle \nabla_{X_i} X, X_i \rangle_{\mathcal{M}}$. The following Stokes formula (also

called divergence theorem, see Lee (2018, p.51)) holds for any $f \in \mathrm{C}^\infty(\mathcal{M})$ and $X \in \mathcal{X}(\mathcal{M})$, $\int_M \mathrm{div}(X)(x) f(x) \mathrm{d}p_{\mathrm{ref}}(x) = -\int_M X(f)(x) \mathrm{d}p_{\mathrm{ref}}(x)$. Let $X = \sum_{i=1}^d a_i X_i$ in local coordinates. Using the Stokes formula and the definition of the gradient we get that in local coordinates

$$\nabla f = \sum_{i,j=1}^d g^{i,j} \partial_i f X_j, \qquad \mathrm{div}(X) = \det(G)^{-1/2} \sum_{i=1}^d \partial_i (\det(G)^{1/2} a_i). \tag{C.12}$$

The Laplace–Beltrami operator is given by $\Delta_{\mathcal{M}} : \mathrm{C}^\infty(M) \to \mathrm{C}^\infty(M)$ and for any $f \in \mathrm{C}^\infty(M)$ by $\Delta_{\mathcal{M}}(f) = \mathrm{div}(\mathrm{grad}(f))$. In local coordinates we obtain $\Delta_{\mathcal{M}}(f) = \det(G)^{-1/2} \sum_{i=1}^d \partial_i (\det(G)^{1/2} \sum_{j=1}^d g^{i,j} \partial_j f)$. Using the Nash isometric embedding theorem (Gunther, 1991) we will see that $\Delta_{\mathcal{M}}$ can always be written as a sum of squared operators. However, this result requires an *extrinsic* point of view as it relies on the existence of projection operators. In contrast, if we consider the orthonormal bundle $\mathrm{O}\mathcal{M}$, see (Hsu, 2002, Chapter 2), we can define the Laplace–Bochner operator $\Delta_{\mathrm{O}\mathcal{M}} : \mathrm{C}^\infty(\mathrm{O}\mathcal{M}) \to \mathrm{C}^\infty(\mathrm{O}\mathcal{M})$ as $\Delta_{\mathrm{O}\mathcal{M}} = \sum_{i=1}^d H_i^2$, where we recall that for any $i \in \{1, \dots, d\}$, H_i is the horizontal lift of e_i. In this case, $\Delta_{\mathrm{O}\mathcal{M}}$ is a sum of squared operators and we have that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $\Delta_{\mathrm{O}\mathcal{M}}(f \circ \pi) = \Delta_{\mathcal{M}}(f)$ (see Hsu, 2002, Proposition 3.1.2). Being able to express the various Laplace operators as a sum of squared operators is key to express the associated diffusion process as the solution of an SDE.

Alternatives definitions of Brownian motion We are now ready to define a Brownian motion on the manifold $\mathcal{M}$. Using the Laplace–Beltrami operator, we can introduce the Brownian motion through the lens of diffusion processes.

Brownian motion

DEFINITION C.1. *Let $(\boldsymbol{b}_t^{\mathcal{M}})_{t\geq 0}$ be an $\mathcal{M}$-valued semimartingale. $(\boldsymbol{b}_t^{\mathcal{M}})_{t\geq 0}$ is a Brownian motion on $\mathcal{M}$ if for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $(\boldsymbol{M}_t^f)_{t\geq 0}$ is a local martingale where for any $t \geq 0$*

$$\boldsymbol{M}_t^f = f(\boldsymbol{b}_t^{\mathcal{M}}) - f(\boldsymbol{b}_0^{\mathcal{M}}) - \tfrac{1}{2} \int_0^t \Delta_{\mathcal{M}} f(\boldsymbol{b}_s^{\mathcal{M}}) \mathrm{d}s. \tag{C.13}$$

Note that this definition is in accordance with the definition of the Brownian motion as a diffusion process in the Euclidean space $\mathbb{R}^d$, since in this case $\Delta_{\mathcal{M}} = \Delta$. A key property of frame bundles and orthonormal bundles is that any semimartingale on $\mathcal{M}$ can be associated to a process on $\mathrm{F}\mathcal{M}$ (or $\mathrm{O}\mathcal{M}$) and a process on $\mathbb{R}^d$. The proof of the following result can be found in Hsu (2002, Propositions 3.2.1 and 3.2.2).

Intrinsic view of Brownian motion

PROPOSITION C.2. *Let $(\boldsymbol{b}_t^{\mathcal{M}})_{t\geq 0}$ be an $\mathcal{M}$-valued semimartingales. Then $(\boldsymbol{b}_t^{\mathcal{M}})_{t\geq 0}$ is a Brownian motion on $\mathcal{M}$ if and only on the following conditions hold:*

a) The horizontal lift $(\boldsymbol{U}_t)_{t\geq 0}$ is a $\Delta_{\mathrm{O}\mathcal{M}}/2$ diffusion process, i.e. for any $f \in \mathrm{C}^\infty(\mathrm{O}\mathcal{M})$, we have that $(\boldsymbol{M}_t^f)_{t\geq 0}$ is a local martingale where for any $t \geq 0$

$$\boldsymbol{M}_t^f = f(\boldsymbol{U}_t) - f(\boldsymbol{U}_0) - \tfrac{1}{2} \int_0^t \Delta_{\mathrm{O}\mathcal{M}} f(\boldsymbol{U}_s) \mathrm{d}s. \tag{C.14}$$

b) The stochastic antidevelopment of $(\boldsymbol{b}_t^{\mathcal{M}})_{t\geq 0}$ is a $\mathbb{R}^d$-valued Brownian motion $(\boldsymbol{b}_t)_{t\geq 0}$.

In particular the previous proposition provides us with an *intrinsic* way to sample the Brownian motion on $\mathcal{M}$ with initial condition $\boldsymbol{b}_0^{\mathcal{M}}$. First sample $(\boldsymbol{U}_t)_{t\geq 0}$ solution of $\mathrm{SDE}(H^{1:d}, \boldsymbol{b}^{1:d}, \boldsymbol{U}_0)$ with $H^{1:d} = \{H_i\}_{i=1}^d$ and $\pi(\boldsymbol{U}_0) = \boldsymbol{b}_0^{\mathcal{M}}$ and $\boldsymbol{b}^{1:d}$ the Euclidean d-dimensional Brownian motion. Then, we recover the $\mathcal{M}$-valued Brownian motion $(\boldsymbol{b}_t^{\mathcal{M}})_{t\geq 0}$ upon letting $(\boldsymbol{b}_t^{\mathcal{M}})_{t\geq 0} = (\pi(\boldsymbol{U}_t))_{t\geq 0}$.

We now consider an *extrinsic* approach to the sampling of Brownian motions on $\mathcal{M}$. Using the Nash embedding theorem (Gunther, 1991), there exists $p \in \mathbb{N}$ such that without loss of generality we can assume that $\mathcal{M} \subset \mathbb{R}^p$. For any $x \in \mathcal{M}$, we denote $\mathrm{P}(x) : \mathbb{R}^p \to \mathrm{T}_x\mathcal{M}$ the projection operator. In addition for any $x \in \mathcal{M}$, we denote $\{\mathrm{P}_i(x)\}_{i=1}^p = \{\mathrm{P}(x)e_i\}_{i=1}^p$, where $\{e_i\}_{i=1}^p$ is the canonical basis of $\mathbb{R}^p$. For any $i \in \{1, \ldots, p\}$, we smoothly extend P_i to $\mathbb{R}^p$. In this case, we have the following proposition (Hsu, 2002, Theorem 3.1.4):

Extrinsic view of Brownian motion

PROPOSITION C.3. *For any $f \in \mathrm{C}^\infty(\mathcal{M})$ we have that $\Delta_{\mathcal{M}}(f) = \sum_{i=1}^p \mathrm{P}_i(\mathrm{P}_i(f))$. Hence, we have that $(\boldsymbol{b}_t^{\mathcal{M}})_{t\geq 0}$ solution of $\mathrm{SDE}(\{\mathrm{P}_i\}_{i=1}^p, \boldsymbol{b}^{1:p}, \boldsymbol{b}_0^{\mathcal{M}})$ with $\boldsymbol{b}_0^{\mathcal{M}}$ a $\mathcal{M}$-valued random variable and $\boldsymbol{b}^{1:p}$ a $\mathbb{R}^p$-valued Brownian motion.*

The second part of this proposition, stems from the fact that any solution of $\mathrm{SDE}(\{V_i\}_{i=1}^\ell, \boldsymbol{b}^{1:\ell}, \boldsymbol{X}_0)$, where $\boldsymbol{X}_0$ is a $\mathcal{M}$-valued random variable and $\boldsymbol{b}^{1:\ell}$ a $\mathbb{R}^\ell$-valued Brownian motion is a diffusion process with generator $\mathcal{A}$ such that for any $f \in \mathrm{C}^\infty(\mathcal{M})$, $\mathcal{A}(f) = \sum_{i=1}^\ell V_i(V_i(f))$. The *extrinsic* approach is particularly convenient since the SDE appearing in proposition C.3 can be seen as an SDE on the Euclidean space $\mathbb{R}^p$.

We finish this paragraph, by investigating the behaviour of the Brownian motion in local coordinates. For simplicity, we assume here that we have access to a system of global coordinates. In the case where the coordinates are strictly local then we refer to Ikeda and Watanabe (1989, Chapter 5, Theorem 1) for a construction of a global solution by patching local solutions. We denote $\{X_k, X_{i,j}\}_{1\leq i,j,k\leq d}$ such that for any $u \in \mathrm{F}\mathcal{M}$, $\{X_k(u), X_{i,j}(u)\}_{1\leq i,j,k\leq d}$ is a basis of $\mathrm{T}_u\mathrm{F}\mathcal{M}$, Using properties of the horizontal lift, see (Hsu, 2002, Chapter 2), we get that $(\boldsymbol{U}_t)_{t\geq 0} = (\{\boldsymbol{X}_t^k, \boldsymbol{E}_t^{i,j}\}_{1\leq i,j,k\leq d})$ obtained in proposition C.2 is given in the global coordinates for any $i, j, k \in \{1, \ldots, d\}$ by

$$\mathrm{d}\boldsymbol{X}_t^k = \sum_{j=1}^d \boldsymbol{E}_t^{k,j} \circ \mathrm{d}\boldsymbol{b}_t^k, \qquad \mathrm{d}\boldsymbol{E}_t^{i,j} = -\sum_{n=1}^d \{\sum_{\ell,m=1}^d \boldsymbol{E}_t^{\ell,n}\boldsymbol{E}_t^{m,j}\Gamma_{\ell,m}^i(\boldsymbol{X}_t)\} \circ \mathrm{d}\boldsymbol{b}_t^n. \quad \text{(C.15)}$$

By definition of the Stratonovich integral we have that for any $k \in \{1, \ldots, d\}$

$$\mathrm{d}\boldsymbol{X}_t^k = \sum_{j=1}^d \{\boldsymbol{E}_t^{k,j}\mathrm{d}\boldsymbol{b}_t^k + \tfrac{1}{2}\mathrm{d}[\boldsymbol{E}_t^{k,j}, \boldsymbol{b}_t^j]_t\}. \quad \text{(C.16)}$$

Let $(\boldsymbol{M}_t)_{t\geq 0} = (\{\boldsymbol{M}_t^k\}_{k=1}^d)_{t\geq 0}$ such that for any $t \geq 0$ and $k \in \{1, \ldots, d\}$ $\boldsymbol{M}_t^k = \sum_{j=1}^d \int_0^t \boldsymbol{E}_t^{k,j}\mathrm{d}\boldsymbol{b}_t^k$. We obtain that $\mathrm{d}\boldsymbol{M}_t = G(\boldsymbol{X}_t)^{-1/2}\mathrm{d}\boldsymbol{b}_t$ for some d-dimensional Brownian motion $(\boldsymbol{b}_t)_{t\geq 0}$, using Lévy's characterization of Brownian motion. In addition, we have that for any $k, j \in \{1, \ldots, d\}$

$$[\boldsymbol{E}^{k,j}, \boldsymbol{b}^j]_t = -\sum_{\ell,m=1}^d \int_0^t \boldsymbol{E}_t^{\ell,j}\boldsymbol{E}_t^{m,j}\Gamma_{\ell,m}^k(\boldsymbol{X}_t)\mathrm{d}t \quad \text{(C.17)}$$

Hence, using this result and the fact that $\sum_{j=1}^{d} \boldsymbol{E}_t^{\ell,j} \boldsymbol{E}_t^{m,j} = g^{\ell,m}(\boldsymbol{X}_t)$, we get that for any $k \in \{1, \dots, d\}$

$$\mathrm{d}\boldsymbol{X}_t^k = -\tfrac{1}{2} \sum_{\ell,m=1}^{d} g^{\ell,m}(\boldsymbol{X}_t) \Gamma_{\ell,m}^k(\boldsymbol{X}_t) \mathrm{d}t + (G(\boldsymbol{X}_t)^{-1/2} \mathrm{d}\boldsymbol{b}_t)^k. \tag{C.18}$$

Note that this result could also have been obtained using the expression of the Laplace–Beltrami in local coordinates.

Brownian motion and random walks In the previous paragraph we consider three SDEs to obtain a Brownian motion on $\mathcal{M}$ (stochastic development, isometric embedding and local coordinates). In this section, we summarize results from Jørgensen (1975) establishing the limiting behaviour of Geodesic Random Walks (GRWs) when the step size of the random walk goes to 0. This will be of particular interest when considering the time-reversal process. We start by defining the geodesic random walk on $\mathcal{M}$, following Jørgensen (1975, Section 2).

Let $\{\nu_x\}_{x \in \mathcal{M}}$ such that for any $x \in \mathcal{M}$, $\nu_x : \lfloor(\mathrm{T}_x\mathcal{M}) \to [0,1]$ with $\nu_x(\mathrm{T}_x\mathcal{M}) = 1$, i.e. for any $x \in \mathcal{M}$, ν_x is a probability measure on $\mathrm{T}_x\mathcal{M}$. Assume that for any $x \in \mathcal{M}$, $\int_{\mathcal{M}} \|v\|^3 \mathrm{d}\nu_x(v) < +\infty$. In addition assume that there exists $\mu^{(1)} \in \mathcal{X}(\mathcal{M})$ and $\mu^{(2)} \in \mathcal{X}(\mathcal{M})deux$, where $\mathcal{X}(\mathcal{M})deux$ is the section $\Gamma(\mathcal{M}, \sqcup_{x \in \mathcal{M}} \mathcal{L}(\mathrm{T}_x\mathcal{M}))$, such that for any $x \in \mathcal{M}$, $\int_{\mathcal{M}} v \mathrm{d}\nu_x(v) = \mu^{(1)}(x)$ and $\int_{\mathcal{M}} v \otimes v \mathrm{d}\nu_x(v) = \mu^{(2)}(x)$. In addition, we assume that for any $x \in \mathcal{M}$, $\Sigma(x) = \mu^{(2)}(x) - \mu^{(1)}(x) \otimes \mu^{(1)}(x)$ is strictly positive definite and that there exists $\mathtt{m}L \geq$ such that for any $x, y \in \mathcal{M}$, $\left\|\nu_x - \nu_y\right\|_{\mathrm{TV}} \leq \mathtt{m}L d_{\mathcal{M}}(x,y)$. Where we have that for any $\nu_1 \in \mathcal{P}(\mathrm{T}_x\mathcal{M})$ and $\nu_2 \in \mathcal{P}(\mathrm{T}_y\mathcal{M})$,

$$\left\|\nu_x - \nu_y\right\|_{\mathrm{TV}} = \sup\left\{\nu_1[f] - \Gamma_x^y(\gamma)_{\#}\nu_2[f] \ : \ \gamma \in \mathrm{Geo}_{x,y},\ f \in \mathrm{C}(\mathrm{T}_x\mathcal{M})\right\}. \tag{C.19}$$

Note that if $d_{\mathcal{M}}(x,y) \leq \varepsilon$ then for some $\varepsilon > 0$ we have that $\left|\mathrm{Geo}_{x,y}\right| = 1$.

Geodesic random walk

Definition C.4. *Let X_0 be a $\mathcal{M}$-valued random variable. For any $\gamma > 0$, we define $(\boldsymbol{X}_t^\gamma)_{t \geq 0}$ such that $\boldsymbol{X}_0^\gamma = X_0$ and for any $n \in \mathbb{N}$ and $t \in [0, \gamma]$, $\boldsymbol{X}_{n\gamma+t} = \exp_{\boldsymbol{X}_{n\gamma}}[t\gamma\{\mu_n + (1/\sqrt{\gamma})(V_n - \mu_n)\}]$, where $(V_n)_{n \in \mathbb{N}}$ is a sequence of random variables in such that for any $n \in \mathbb{N}$, V_n has distribution $\nu_{\boldsymbol{X}_{n\gamma}}$ conditionally to $\boldsymbol{X}_{n\gamma}$.*

For any $\gamma > 0$, the process $(X_n^\gamma)_{n \in \mathbb{N}} = (\boldsymbol{X}_{n\gamma}^\gamma)_{n \in \mathbb{N}}$ is called a geodesic random walk. In particular, for any $\gamma > 0$ we denote $(\mathrm{R}_n^\gamma)_{n \in \mathbb{N}}$ the sequence of Markov kernels such that for any $n \in \mathbb{N}$, $x \in \mathcal{M}$ and $\mathsf{A} \in \lfloor(\mathcal{M})$ we have that $\delta\mathrm{R}(\mathsf{A}) = \mathbb{P}(X_n^\gamma \in \mathsf{A})$, with $X_0^\gamma = x$. The following theorem establishes that the limiting dynamics of a geodesic random walk is associated with a diffusion process on $\mathcal{M}$ whose coefficients only depends on the properties of ν (see Jørgensen, 1975, Theorem 2.1).

Convergence of geodesic random walks

Theorem C.5. *For any $t \geq 0$, $f \in \mathrm{C}(\mathcal{M})$ and $x \in \mathcal{M}$ we have that*

$$\lim_{\gamma \to 0} \left\| \mathrm{R}_\gamma^{\lceil t/\gamma \rceil}[f] - \mathrm{Pker}_t[f] \right\|_\infty = 0 \tag{C.20}$$

, where $(\mathrm{Pker}_t)_{t \geq 0}$ is the semi-group associated with the infinitesimal generator $\mathcal{A} : \mathrm{C}^\infty(\mathcal{M}) \to \mathrm{C}^\infty(\mathcal{M})$ given for any $f \in \mathrm{C}^\infty(\mathcal{M})$ by $\mathcal{A}(f) = \langle \mu^{(1)}, \nabla f \rangle_{\mathcal{M}} + \frac{1}{2}\langle \Sigma, \nabla^2 f \rangle_{\mathcal{M}}$.

In particular if $\mu^{(1)} = 0$ and $\mu^{(2)} = \mathbf{I}$ then the random walk converges towards a Brownian motion on $\mathcal{M}$ in the sense of the convergence of semi-groups. For any $x \in \mathcal{M}$ in local coordinates we have that $\Phi_{\#}\nu_x$ has zero mean and covariance matrix $G(x)$, where Φ is a local chart around x and $G(x) = (g_{i,j}(x))_{1 \leq i,j \leq d}$ the coordinates of the metric in that chart.

Convergence of Brownian motion We finish this section with a few considerations regarding the convergence of the Brownian motion on $\mathcal{M}$. Since we have assumed that $\mathcal{M}$ is compact we have that there exist $(\Phi_k)_{k \in \mathbb{N}}$ an orthonormal basis of $-\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(p_{\mathrm{ref}})$, $(\lambda_k)_{k \in \mathbb{N}}$ such that for any $i, j \in \mathbb{N}$, $i \leq j$, $\lambda_i \leq \lambda_j$ and $\lambda_0 = 0$, $\Phi_0 = 1$ and for any $k \in \mathbb{N}$, $\Delta_{\mathcal{M}}\Phi_k = -\lambda_k \Phi_k$. For any $t \geq 0$ and $x, y \in \mathcal{M}$, $p_{t|0}(y|x) = \sum_{k \in \mathbb{N}} \mathrm{e}^{-\lambda_k t}\Phi_k(x)\Phi_k(y)$ where for any $f \in \mathrm{C}^\infty$ we have

$$\mathbb{E}\left[f(\boldsymbol{b}_t^{\mathcal{M},x})\right] = \int_{\mathcal{M}} p_{t|0}(x,y) f(y) \mathrm{d}p_{\mathrm{ref}}(y), \tag{C.21}$$

where $(\boldsymbol{b}_t^{\mathcal{M},x})_{t \geq 0}$ is the Brownian motion on $\mathcal{M}$ with $\boldsymbol{b}_0^{\mathcal{M},x} = x$ and p_{ref} is the probability measure associated with the Hausdorff measure on $\mathcal{M}$. We also have the following result (see Urakawa, 2006, Proposition 2.6).

Convergence of Brownian motion

PROPOSITION C.6. *For any $t > 0$, Pker_t admits a density $p_{t|0}$ with respect to p_{ref} and $p_{\mathrm{ref}}\mathrm{Pker}_t = p_{\mathrm{ref}}$, i.e. p_{ref} is an invariant measure for $(\mathrm{Pker}_t)_{t \geq 0}$. In addition, if there exists $C, \alpha \geq 0$ such that for any $t \in [0,1]$, $p_{t|0}(x|x) \leq Ct^{-\alpha/2}$ then for any $p_0 \in \mathcal{P}(\mathcal{M})$ and for any $t \geq 1/2$ we have*

$$\|p_0\mathrm{Pker}_t - p_{\mathrm{ref}}\|_{\mathrm{TV}} \leq C^{1/2}\mathrm{e}^{\lambda_1/2}\mathrm{e}^{-\lambda_1 t}, \tag{C.22}$$

where λ_1 is the first non-negative eigenvalue of $-\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(p_{\mathrm{ref}})$ and we recall that $(\mathrm{Pker}_t)_{t \geq 0}$ is the semi-group of the Brownian motion.

A review on lower bounds on the first positive eigenvalue of the Laplace–Beltrami operator can be found in (He, 2013). These lower bounds usually depend on the Ricci curvature of the manifold or its diameter. We conclude this section by noting that in the non-compact case (Li, 1986) establishes similar estimates in the case of a manifold with non-negative Ricci curvature and maximal volume growth.

C.3. Time-reversal formula: extension to Riemannian manifolds

In this section, we provide the proof of Theorem 3.6. The proof follows the arguments of Cattiaux et al. (2023, Theorem 4.9). We could have also applied the abstract results of Cattiaux et al. (2023, Theorem 5.7) to obtain our results. Note that the time-reversal on manifold could also be obtained by readily extending arguments from Haussmann and Pardoux (1986), however the entropic conditions found by Cattiaux et al. (2023) are more natural when it comes to the study of the Schrödinger Bridge problem. For the interested reader we provide an informal derivation of the time-reversal formula obtained by Haussmann and Pardoux (1986) in appendix C.3.1. The proof of Theorem 3.6 is given in appendix C.3.2. Finally, we emphasize that García-Zelada and Huguet (2021) have developed a Girsanov theory for stochastic processes defined on compact manifolds with boundary in order to study the Brenier-Schrödinger problem.

C.3.1. Informal derivation. In this section, we provide a non-rigorous derivation of Theorem 3.6 following the approach of Haussmann and Pardoux (1986). Let $(\boldsymbol{X}_t)_{t\in[0,T]}$ be a continuous process such that for any $f \in \mathrm{C}^2(\mathcal{M})$ we have that $(\boldsymbol{M}_t^{X,f})_{t\in[0,T]}$ is a $\boldsymbol{X}$-martingale where for any $t \in [0,T]$

$$\boldsymbol{M}_t^{X,f} = f(\boldsymbol{X}_t) - \int_0^t \{\langle b(\boldsymbol{X}_s), \nabla f(\boldsymbol{X}_s)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} f(\boldsymbol{X}_s)\}\mathrm{d}s. \tag{C.23}$$

Let $(\boldsymbol{Y}_t)_{t\in[0,T]} = (\boldsymbol{X}_{T-t})_{t\in[0,T]}$. Our goal is to show that for any $f \in \mathrm{C}^2(\mathcal{M})$, $(\boldsymbol{M}_t^{Y,f})_{t\in[0,T]}$ is a $\boldsymbol{Y}$-martingale where for any $t \in [0,T]$

$$\boldsymbol{M}_t^{Y,f} = f(\boldsymbol{Y}_t) - \int_0^t \{\langle -b(\boldsymbol{Y}_s) + \nabla \log p_{T-s}(\boldsymbol{Y}_s), \nabla f(\boldsymbol{Y}_s)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} f(\boldsymbol{Y}_s)\}\mathrm{d}s. \tag{C.24}$$

Note that here we implicitly assume that for any $t \in [0,T]$, $\boldsymbol{X}_t$ admits a smooth positive density with respect to. p_{ref} denoted p_t. In other words, we want to show that for any $g \in \mathrm{C}^2(\mathcal{M})$ and $s,t \in [0,T]$ with $t \geq s$ we have

$$\mathbb{E}[g(\boldsymbol{Y}_s)(f(\boldsymbol{Y}_t) - f(\boldsymbol{Y}_s))] \tag{C.25}$$

$$= \mathbb{E}\left[g(\boldsymbol{Y}_s)\int_s^t \{\langle -b(\boldsymbol{Y}_u) + \nabla \log p_{T-u}(\boldsymbol{Y}_u), \nabla f(\boldsymbol{Y}_u)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} f(\boldsymbol{Y}_u)\}\mathrm{d}u\right]. \tag{C.26}$$

We introduce the infinitesimal generator $\mathcal{A} : \mathrm{C}^2(\mathcal{M}) \to \mathrm{C}(\mathcal{M})$ given for any $f \in \mathrm{C}^2(\mathcal{M})$ and $x \in \mathcal{M}$ by

$$\mathcal{A}(f)(x) = \langle b(x), \nabla f(x)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} f(x). \tag{C.27}$$

Similarly, we introduce the infinitesimal generator $\mathcal{A}t : [0,T] \times \mathrm{C}^2(\mathcal{M}) \to \mathrm{C}(\mathcal{M})$ given for any $f \in \mathrm{C}^2(\mathcal{M})$, $t \in [0,T]$ and $x \in \mathcal{M}$ by

$$\mathcal{A}t(t,f)(x) = \langle -b(x) + \nabla \log p_{T-t}(x), \nabla f(x)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} f(x). \tag{C.28}$$

With these notations, (C.29) can be written as follows: we want to show that for any $g \in \mathrm{C}^2(\mathcal{M})$ and $s,t \in [0,T]$ with $t \geq s$ we have

$$\mathbb{E}[g(\boldsymbol{Y}_s)(f(\boldsymbol{Y}_t) - f(\boldsymbol{Y}_s))] = \mathbb{E}\left[g(\boldsymbol{Y}_s)\int_s^t \mathcal{A}t(u, \boldsymbol{Y}_u)\mathrm{d}u\right]. \tag{C.29}$$

The rest of this section follows the first part of the proof of Haussmann and Pardoux (1986, Theorem 2.1). Let $t,s \in [0,T]$ with $t \geq s$. We have

$$\mathbb{E}[g(\boldsymbol{Y}_s)(f(\boldsymbol{Y}_t) - f(\boldsymbol{Y}_s))] \tag{C.30}$$

$$= \mathbb{E}[g(\boldsymbol{X}_{T-s})(f(\boldsymbol{X}_{T-t}) - f(\boldsymbol{X}_{T-t}))] \tag{C.31}$$

$$= \mathbb{E}[\mathbb{E}[g(\boldsymbol{X}_{T-s}) \mid \boldsymbol{X}_{T-t}]f(\boldsymbol{X}_{T-t})] - \mathbb{E}[g(\boldsymbol{X}_{T-s})f(\boldsymbol{X}_{T-s})] \tag{C.32}$$

$$= \mathbb{E}[v(T-t, \boldsymbol{X}_{T-t})f(\boldsymbol{X}_{T-t})] - \mathbb{E}[v(T-s, \boldsymbol{X}_{T-s})f(\boldsymbol{X}_{T-s})], \tag{C.33}$$

with $v : [0, T-s] \times \mathcal{M} \to \mathbb{R}$ given for any $u \in [0, T-s]$ and $x \in \mathcal{M}$ by $v(u,x) = \mathbb{E}[g(\boldsymbol{X}_{T-s}) \mid \boldsymbol{X}_u = x]$. We have that v satisfies the backward Kolmogorov equation, i.e. we have for any $u \in [0, T-s]$ and $x \in \mathcal{M}$

$$\partial_u v(u,x) = -\mathcal{A}v(u,x). \tag{C.34}$$

Note that it is not trivial to show that v is regular enough to satisfy the backward Kolmogorov equation. In this informal derivation, we assume that v is regular enough and will provide a different rigorous proof of the time-reversal formula in appendix C.3.2. However, note that it is possible to show that v indeed satisfies the backward Kolmogorov equation by adapting arguments from Haussmann and Pardoux (1986) to the manifold framework.

Let $h : [0, T-s] \times \mathcal{M} \to \mathbb{R}$ given for any $u \in [0, T-s]$ and $x \in \mathcal{M}$ by $h(u,x) = v(u,x)f(x)$. Using (C.34), we have for any $u \in [0, T-s]$ and $x \in \mathcal{M}$

$$\partial_u h(u,x) + \mathcal{A}h(u,x) \tag{C.35}$$
$$= f(x)\partial_u v(u,x) + f(x)\mathcal{A}v(u,x) + v(u,x)\mathcal{A}f(x) + \langle \nabla f(x), \nabla v(u,x)\rangle \tag{C.36}$$
$$= v(u,x)\mathcal{A}f(x) + \langle \nabla f(x), \nabla v(u,x)\rangle. \tag{C.37}$$

In addition, using the divergence theorem (see Lee, 2018, p.51), we have for any $u \in [0, T-s]$

$$\mathbb{E}[\langle \nabla f(\boldsymbol{X}_u), \nabla v(u, \boldsymbol{X}_u)\rangle] = \int_{\mathcal{M}} \langle \nabla f(x_u), \nabla v(u,x_u) p_u(x_u)\rangle \mathrm{d}p_{\mathrm{ref}}(x_u) \tag{C.38}$$
$$= -\int_{\mathcal{M}} v(u,x_u)\mathrm{div}(p_u \nabla f)(x_u)\mathrm{d}p_{\mathrm{ref}}(x_u) \tag{C.39}$$
$$= -\int_{\mathcal{M}} v(u,x_u)\Delta_{\mathcal{M}} f(x_u) p_u(x_u)\mathrm{d}p_{\mathrm{ref}}(x_u) \tag{C.40}$$
$$- \int_{\mathcal{M}} v(u,x_u)\langle \nabla f(x_u), \nabla \log p_u(x_u)\rangle p_u(x_u)\mathrm{d}p_{\mathrm{ref}}(x_u) \tag{C.41}$$
$$= -\mathbb{E}[v(u,\boldsymbol{X}_u)\Delta_{\mathcal{M}} f(\boldsymbol{X}_u)] - \mathbb{E}[v(u,\boldsymbol{X}_u)\langle \nabla f(\boldsymbol{X}_u), \nabla \log p_u(\boldsymbol{X}_u)\rangle]. \tag{C.42}$$

Therefore, using this result and (C.37) we get that for any $u \in [0, T-s]$

$$\mathbb{E}[\partial_u h(u, \boldsymbol{X}_u) + \mathcal{A}h(u, \boldsymbol{X}_u)] \tag{C.43}$$
$$= \mathbb{E}\left[v(u,\boldsymbol{X}_u)\{\langle b(\boldsymbol{X}_u) - \nabla \log p_u(\boldsymbol{X}_u), \nabla f(\boldsymbol{X}_u)\rangle - \tfrac{1}{2}\Delta_{\mathcal{M}} f(\boldsymbol{X}_u)\}\right] \tag{C.44}$$
$$= -\mathbb{E}[v(u,\boldsymbol{X}_u)\mathcal{A}t(T-u,f)(\boldsymbol{X}_u)]. \tag{C.45}$$

Combining this result and (C.23) and that for any $u \in [0, T-s]$ and $x \in \mathcal{M}$, $v(u,x) = \mathbb{E}[g(\boldsymbol{X}_{T-s}) \mid \boldsymbol{X}_u = x]$ we get

$$\mathbb{E}[v(T-t, \boldsymbol{X}_{T-t})f(\boldsymbol{X}_{T-t})] - \mathbb{E}[v(T-s, \boldsymbol{X}_{T-s})f(\boldsymbol{X}_{T-s})] \tag{C.46}$$
$$= \mathbb{E}[h(T-t, \boldsymbol{X}_{T-t}) - h(T-s, \boldsymbol{X}_{T-s})] \tag{C.47}$$
$$= \int_{T-t}^{T-s} \mathbb{E}[v(u,\boldsymbol{X}_u)\mathcal{A}t(T-u,\boldsymbol{X}_u)]\mathrm{d}u \tag{C.48}$$
$$= \mathbb{E}\left[g(\boldsymbol{X}_{T-s})\int_{T-t}^{T-s} \mathcal{A}t(T-u,\boldsymbol{X}_u)\mathrm{d}u\right]. \tag{C.49}$$

Using this result, (C.33) and the change of variable $u \mapsto T-u$ we obtain

$$\mathbb{E}[g(\boldsymbol{Y}_s)(f(\boldsymbol{Y}_t) - f(\boldsymbol{Y}_s))] = \mathbb{E}\left[g(\boldsymbol{X}_{T-s})\int_{T-t}^{T-s} \mathcal{A}t(T-u,\boldsymbol{X}_u)\mathrm{d}u\right] \tag{C.50}$$
$$= \mathbb{E}\left[g(\boldsymbol{Y}_s)\int_s^t \mathcal{A}t(u,\boldsymbol{Y}_u)\mathrm{d}u\right]. \tag{C.51}$$

Hence, (C.25) holds and we have proved Theorem 3.6. Again, we emphasize that in order to make the proof completely rigorous one needs to derive regularity properties of v.

C.3.2. Proof of Theorem 3.6. In this section, we follow another approach to prove the time-reversal formula. We are going to use the integration by part formula of Cattiaux et al. (2023, Theorem 3.17) in a similar spirit as Cattiaux et al. (2023, Theorem 4.9) in the Euclidean setting. In order to adapt arguments from Cattiaux et al. (2023) to our Riemannian setting, we use the Nash embedding theorem in order to embed our processes in a Euclidean space and leverage tools from Girsanov theory. The rest of the section is organized as follows. First in appendix C.3.2, we recall basic properties of infinitesimal generators and recall the integration by part formula of Cattiaux et al. (2023, Theorem 3.17). Then in appendix C.3.2, we extend some Girsanov theory to compact Riemannian manifolds using the Nash embedding theorem. We conclude the proof in appendix C.3.2.

Diffusion processes and integration by part formula

In this section, we state a simplified version of Cattiaux et al. (2023, Theorem 3.17) for Markov continuous path (probability) measure on Polish spaces. Let $(\mathsf{X}, \mathcal{X})$ be a Polish space. We say that $\mathbb{P}$ is a path measure if $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T],\mathsf{X}))$. Let $(\boldsymbol{X}_t)_{t\in[0,T]}$ with distribution $\mathbb{P}$. We denote $(\mathcal{F}_t)_{t\in[0,T]}$ the filtration such that for any $t \in [0,T]$, $\mathcal{F}_t = \sigma(\boldsymbol{X}_s,\ s \in [0,t])$. Let $(\boldsymbol{M}_t)_{t\in[0,T]}$ be a Polish-valued stochastic process. We say that $(\boldsymbol{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-local martingale if it is a local martingale with respect to. the filtration $(\mathcal{F}_t)_{t\in[0,T]}$. A function $u : [0,T]\times\mathsf{X} \to \mathbb{R}$ is said to be in the domain of the extended generator of $\mathbb{P}$ if there exists a process $(\mathcal{A}b_{\mathbb{P}}u(t,\boldsymbol{X}_{[0,t]}))_{t\in[0,T]}$ such that:

(a) $(\mathcal{A}b_{\mathbb{P}}u(t,\boldsymbol{X}_{[0,t]}))_{t\in[0,T]}$ is adapted with respect to. $(\mathcal{F}_t)_{t\in[0,T]}$.

(b) $\int_0^T \left|\mathcal{A}b_{\mathbb{P}}u(t,\boldsymbol{X}_{[0,t]})\right|\mathrm{d}t < +\infty$, $\mathbb{P}$-a.s.

(c) The process $(\boldsymbol{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-local martingale, where for any $t \in [0,T]$

$$\boldsymbol{M}_t = u(t,\boldsymbol{X}_t) - u(0,\boldsymbol{X}_0) - \int_0^t \mathcal{A}b_{\mathbb{P}}u(s,\boldsymbol{X}_{[0,s]})\mathrm{d}s. \tag{C.52}$$

The domain of the extended generator is denoted $\mathrm{dom}(\mathcal{A}\mathrm{b}_{\mathbb{P}})$. We say that (u,v) with $u,v : [0,T]\times\mathsf{X} \to \mathbb{R}$ is in the domain of the carré du champ if $u, v, uv \in \mathrm{dom}(\mathcal{A}\mathrm{b}_{\mathbb{P}})$. In this case, we define the carré du champ $\bar{\Upsilon}_{\mathbb{P}}$ as

$$\bar{\Upsilon}_{\mathbb{P}}(u,v) = \mathcal{A}b_{\mathbb{P}}(uv) - \mathcal{A}b_{\mathbb{P}}(u)v - \mathcal{A}b_{\mathbb{P}}(v)u. \tag{C.53}$$

Note that if $\mathsf{X} = \mathcal{M}$ is a Riemannian manifold, $\mathrm{C}^2(\mathcal{M}) \subset \mathrm{dom}(\mathcal{A}\mathrm{b}_{\mathbb{P}})$ and for any $u \in \mathrm{C}^2(\mathcal{M})$ $\mathcal{A}b_{\mathbb{P}}(u) = \langle \nabla u, X\rangle + \frac{1}{2}\Delta_{\mathcal{M}}u$ with $X \in \Gamma(\mathrm{T}\mathcal{M})$ then we have that $\mathrm{C}^2(\mathcal{M})\times\mathrm{C}^2(\mathcal{M}) \subset \mathrm{dom}(\bar{\Upsilon}_{\mathbb{P}})$ and for any $u,v \in \mathrm{C}^2(\mathcal{M})$, $\bar{\Upsilon}_{\mathbb{P}}(u,v) = \langle \nabla u, \nabla v\rangle$. Assume that there exists $\mathcal{U}_{\mathbb{P}} \subset \mathrm{dom}(\mathcal{A}\mathrm{b}_{\mathbb{P}}) \cap \mathrm{C}_{\mathrm{b}}(\mathsf{X})$ such that $\mathcal{U}_{\mathbb{P}}$ is an algebra. We denote $\mathcal{U}_{\mathbb{P},2}$ such that

$$\mathcal{U}_{\mathbb{P},2} = \left\{u \in \mathcal{U}_{\mathbb{P}} \ :\ \mathcal{A}b_{\mathbb{P}}u \in \mathrm{L}^2(\mathbb{P}),\ \bar{\Upsilon}_{\mathbb{P}}(u,u) \in \mathrm{L}^1(\mathbb{P})\right\}. \tag{C.54}$$

Finally we denote $R(\mathbb{P})$ the time-reverse path measure, i.e. for any $\mathsf{A} \in \lfloor(\mathrm{C}([0,T],\mathsf{X}))$ we have $R(\mathbb{P})(\mathsf{A}) = \mathbb{P}(R(\mathsf{A}))$, where $R(\mathsf{A}) = \{t \mapsto \omega_{T-t} \ : \ \omega \in \mathsf{A}\}$. In what follows, we assume $\mathbb{P}$ is Markov. It is well-known, see (Léonard et al., 2014, Theorem 1.2) for instance, that in this case $R(\mathbb{P})$ is also Markov. In addition, since $\mathbb{P}$ is Markov, for any $u \in \mathrm{dom}(\mathcal{A}b_{\mathbb{P}})$ and $t \in [0,T]$ there exists $\mathcal{A}_{\mathbb{P}}$ such that $\mathcal{A}b_{\mathbb{P}}u(t, \boldsymbol{X}_{[0,t]}) = \mathcal{A}_{\mathbb{P}}u(t, \boldsymbol{X}_t)$ with $\mathcal{A}_{\mathbb{P}}u : \ [0,T] \times \mathsf{X} \to \mathbb{R}$. Similarly, we define $\Upsilon_{\mathbb{P}}(u,v) : \ [0,T] \times \mathsf{X} \to \mathbb{R}$ from $\bar{\Upsilon}_{\mathbb{P}}(u,v)$.

We are now ready to state the integration by part formula, (Cattiaux et al., 2023, Theorem 3.17).

THEOREM C.7. *Let $u, v \in \mathcal{U}_{\mathbb{P},2}$. The following hold:*

(a) If $u \in \mathrm{dom}(\mathcal{A}_{\mathrm{R}(\mathbb{P})})$ and $\mathcal{A}_{R(\mathbb{P})}u \in \mathrm{L}^1(\mathbb{P})$ then for almost any $t \in [0,T]$

$$\mathbb{E}\Big[\{\mathcal{A}_{\mathbb{P}}u(t, \boldsymbol{X}_t) + \mathcal{A}_{R(\mathbb{P})}u(T-t, \boldsymbol{X}_t)\}v(\boldsymbol{X}_t) + \Upsilon_{\mathbb{P}}(u,u)(t, \boldsymbol{X}_t)\Big] = 0. \tag{C.55}$$

(b) If the following hold:

i) $\Upsilon_{\mathbb{P}}(u,v) \in \mathrm{C}([0,T] \times \mathsf{X}, \mathbb{R})$.

ii) $\mathcal{U}_{2,\mathbb{P}}$ determines the weak convergence of Borel measures.

iii) μ defines a finite measure on $[0,T] \times \mathsf{X}$ where for any $\omega \in \bar{\mathcal{U}}_{2,\mathbb{P}}$ we have

$$\mu[\omega] = \mathbb{E}\bigg[\int_0^T \Upsilon_{\mathbb{P}}(u, \omega_t)(t, \boldsymbol{X}_t)\mathrm{d}t\bigg], \tag{C.56}$$

where $\bar{\mathcal{U}}_{2,\mathbb{P}} = \big\{\omega \in \mathrm{C}([0,T] \times \mathsf{X}, \mathbb{R}) \ : \ \omega(\mathrm{t}, \cdot) \in \mathcal{U}_{2,\mathbb{P}} \ \text{ for any } t \in [0,T]\big\}$.

Then $u \in \mathrm{dom}(\mathcal{A}_{\mathrm{R}(\mathbb{P})})$ and $\mathcal{A}_{R(\mathbb{P})}u \in \mathrm{L}^1(\mathbb{P})$.

Note that this theorem is a simplified version of Cattiaux et al. (2023, Theorem 3.17) where we restrict ourselves to the case of Markov path measures. In what follows, we wish to apply Theorem C.7 to diffusion processes on manifolds. To do so, we will verify that under a finite entropy assumption, the conditions $u \in \mathrm{dom}(\mathcal{A}_{\mathrm{R}(\mathbb{P})})$ and $\mathcal{A}_{R(\mathbb{P})}u \in \mathrm{L}^1(\mathbb{P})$ are fulfilled for a class of regular functions u. These integrability results are obtained using Girsanov theory.

Girsanov theory on compact Riemannian manifolds

In this section, we will consider two types of martingale problems: one on Euclidean spaces and one on the compact Riemannian manifold $\mathcal{M}$. Let $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathbb{R}^p))$. We say that $\mathbb{P}$ satisfies the (Euclidean) martingale problem with infinitesimal generator $\mathcal{A} : \ [0,T] \times \mathrm{C}^2(\mathbb{R}^p) \times \mathbb{R}^p \to \mathbb{R}$ if for any $u \in \mathrm{C}^2_c(\mathbb{R}^p)$, $(\boldsymbol{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-martingale where for any $t \in [0,T]$ we have

$$\boldsymbol{M}_t = \boldsymbol{M}_0 + \int_0^t \mathcal{A}(t,u)(\boldsymbol{X}_s)\mathrm{d}s, \tag{C.57}$$

where $(\boldsymbol{X}_t)_{t\in[0,T]}$ has distribution $\mathbb{P}$ and $\int_0^T |\mathcal{A}(t,u)(\boldsymbol{X}_s)\mathrm{d}t| < +\infty$, $\mathbb{P}$-a.s. Let $\mathbb{P} \in \mathcal{P}(\mathrm{C}([0,T], \mathcal{M}))$. We say that $\mathbb{P}$ satisfies the (Riemannian) martingale problem

with infinitesimal generator $\mathcal{A}t : [0,T] \times \mathrm{C}^2(\mathcal{M}) \times \mathcal{M} \to \mathbb{R}$ if for any $u \in \mathrm{C}^2(\mathcal{M})$, $(\boldsymbol{M}_t)_{t\in[0,T]}$ is a $\mathbb{P}$-martingale where for any $t \in [0,T]$ we have

$$\boldsymbol{M}_t = \boldsymbol{M}_0 + \int_0^t \mathcal{A}t(t,u)(\boldsymbol{X}_s)\mathrm{d}s, \tag{C.58}$$

where $(\boldsymbol{X}_t)_{t\in[0,T]}$ has distribution $\mathbb{P}$ and $\int_0^T |\mathcal{A}t(t,u)(\boldsymbol{X}_s)\mathrm{d}t| < +\infty$, $\mathbb{P}$-a.s. We now prove the following theorem.

Girsanov Theroy on Manifolds

Proposition C.8. *Let $\mathbb{Q}$ be the path measure of a Brownian motion on $\mathcal{M}$. Let $\mathbb{P}$ be a Markov path measure on $\mathrm{C}([0,T],\mathcal{M})$ such that $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) < +\infty$. Then there exists β such that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\beta(t,x) \in \mathrm{T}_x\mathcal{M}$ and we have that $\mathbb{P}$ satisfies the martingale problem with infinitesimal generator $\mathcal{A}$ where for any $t \in [0,T]$, $u \in \mathrm{C}^2(\mathcal{M})$ and $x \in \mathcal{M}$ we have*

$$\mathcal{A}(t,u)(x) = \langle \beta(t,x), \nabla u(x)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} u(x). \tag{C.59}$$

In addition, we have that

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = \mathrm{KL}(\mathbb{P}_0 \mid \mathbb{Q}_0) + \tfrac{1}{2}\int_0^T \mathbb{E}\left[\|\beta(t,\boldsymbol{X}_t)\|^2\right]\mathrm{d}t, \tag{C.60}$$

where $(\boldsymbol{X}_t)_{t\in[0,T]}$ has distribution $\mathbb{P}$.

Proof. First, we extend $(\boldsymbol{b}_t^{\mathcal{M}})_{t\in[0,T]}$ to $\mathbb{R}^p$ using the Nash embedding theorem (see Gunther, 1991). $(\boldsymbol{b}_t^{\mathcal{M}})_{t\in[0,T]}$ can be seen as a process on $\mathbb{R}^p$ (for some $p \in \mathbb{N}$) which satisfies in a weak sense

$$\mathrm{d}\boldsymbol{b}_t^{\mathcal{M}} = \sum_{i=1}^p \mathrm{P}_i(\boldsymbol{b}_t^{\mathcal{M}}) \circ \mathrm{d}\boldsymbol{b}_t^i = P(\boldsymbol{b}_t^{\mathcal{M}}) \circ \mathrm{d}\boldsymbol{b}_t, \tag{C.61}$$

where $(\boldsymbol{b}_t)_{t\in[0,T]}$ is a p-dimensional Brownian motion and $\mathrm{P} \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^{p\times p})$ is such that for any $x \in \mathcal{M}$, $\mathrm{P}(x)$ is the projection onto $\mathrm{T}_x\mathcal{M}$ and for any $i \in \{1,\dots,p\}$, $\mathrm{P}_i \in \mathrm{C}^\infty(\mathbb{R}^p,\mathbb{R}^p)$ with $\mathrm{P}_i = \mathrm{P}e_i$ where $\{e_j\}_{j=1}^p$ is the canonical basis of $\mathbb{R}^p$. We refer to Appendix C.2.1 for more details on the projection operator and its extension to $\mathbb{R}^p$. Using the link between Stratonovich and Itô integral, there exists $\bar{b} \in \mathrm{C}^\infty(\mathbb{R}^p,\mathbb{R}^p)$ such that $(\boldsymbol{b}_t^{\mathcal{M}})_{t\in[0,T]}$ can be seen as a process on $\mathbb{R}^p$ which satisfies in a weak sense

$$\mathrm{d}\boldsymbol{b}_t^{\mathcal{M}} = \bar{b}(\boldsymbol{b}_t^{\mathcal{M}})\mathrm{d}t + \mathrm{P}(\boldsymbol{b}_t^{\mathcal{M}})\mathrm{d}\boldsymbol{b}_t, \tag{C.62}$$

where $\bar{b}$ is given in (C.10) and satisfies $\mathrm{P}\bar{b}(x) = 0$ for any $x \in \mathcal{M}$, see the remark following (C.10). For any $u, v \in \mathrm{C}_c^2(\mathcal{M})$, we consider $\bar{u}, \bar{v}$ extensions to $\mathrm{C}_c^2(\mathbb{R}^p)$ and we have for any $s,t \in [0,T]$

$$\mathbb{E}\left[\bar{v}(\boldsymbol{b}_s^{\mathcal{M}}) \int_s^t \tfrac{1}{2}\Delta_{\mathcal{M}} u(\boldsymbol{b}_u^{\mathcal{M}})\mathrm{d}u\right] \tag{C.63}$$

$$= \mathbb{E}\left[\bar{v}(\boldsymbol{b}_s^{\mathcal{M}}) \int_s^t \{\langle \nabla\bar{u}(\boldsymbol{b}_w^{\mathcal{M}}), \bar{b}(\boldsymbol{b}_w^{\mathcal{M}})\rangle + \tfrac{1}{2}\langle \mathrm{P}(\boldsymbol{b}_w^{\mathcal{M}}), \nabla^2\bar{u}(\boldsymbol{b}_w^{\mathcal{M}})\rangle\}\mathrm{d}w\right]. \tag{C.64}$$

In particular, we get that for any $x \in \mathcal{M}$, $\Delta_{\mathcal{M}} u(x) = 2\langle \nabla\bar{u}(x), \bar{b}(x)\rangle + \langle \mathrm{P}(x), \nabla^2\bar{u}(x)\rangle$. Note that $(\boldsymbol{b}_t^{\mathcal{M}})_{t\in[0,T]}$ (seen as a process on $\mathbb{R}^p$) satisfies the condition (U) in

Léonard (2012), i.e. uniqueness of the trajectories given an initial condition. Therefore applying (Léonard, 2012, Theorem 2.1), (Cattiaux et al., 2023, Claim 4.5), there exists $\bar{\beta} : [0,T] \times \mathbb{R}^p \to \mathbb{R}^p$ such that

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = \mathrm{KL}(\mathbb{P}_0 \mid \mathbb{Q}_0) + \tfrac{1}{2} \int_0^T \mathbb{E}\Big[\big\|\mathrm{P}(\boldsymbol{X}_t)\bar{\beta}(t, \boldsymbol{X}_t)\big\|^2\Big] \mathrm{d}t. \tag{C.65}$$

In addition, $\mathbb{P}$ (seen as a process on $\mathbb{R}^p$) satisfies a martingale problem with infinitesimal generator $\mathcal{A}b : [0,T] \times \mathrm{C}_c^2(\mathbb{R}^p) \times \mathbb{R}^p \to \mathbb{R}$ such that for any $t \in [0,T]$, $\bar{u} \in \mathrm{C}_c^2(\mathbb{R}^p)$ and $x \in \mathbb{R}^p$

$$\mathcal{A}b(t, \bar{u})(x) = \langle \bar{b}(x) + \mathrm{P}(x)\bar{\beta}(t,x), \nabla \bar{u}(x) \rangle + \tfrac{1}{2} \langle \mathrm{P}(x), \nabla^2 \bar{u}(x) \rangle. \tag{C.66}$$

Let $\beta : [0,T] \times \mathcal{M}$ such that for any $t \in [0,T]$ and $x \in \mathcal{M}$ we have $\beta(t,x) = \mathrm{P}(x)\bar{\beta}(t,x)$. In particular, we have that for any $x \in \mathcal{M}$, $\beta(t,x) \in \mathrm{T}_x\mathcal{M}$. Let $u \in \mathrm{C}_c^2(\mathcal{M})$ and consider an extension $\bar{u}$ to $\mathrm{C}_c^2(\mathbb{R}^p)$. For any $t \in [0,T]$ and $x \in \mathcal{M}$ we have

$$\begin{aligned}
\mathcal{A}b(t, \bar{u})(x) &= \langle \bar{b}(x) + \mathrm{P}(x)\bar{\beta}(t,x), \nabla \bar{u}(x) \rangle + \tfrac{1}{2} \langle \mathrm{P}(x), \nabla^2 \bar{u}(x) \rangle && \text{(C.67)} \\
&= \langle \beta(t,x), \nabla \bar{u}(x) \rangle + \tfrac{1}{2} \Delta_{\mathcal{M}} u(x) && \text{(C.68)} \\
&= \langle \beta(t,x), \nabla u(x) \rangle + \tfrac{1}{2} \Delta_{\mathcal{M}} u(x). && \text{(C.69)}
\end{aligned}$$

In particular, we have that $\mathbb{P}$ (seen as a process on $\mathcal{M}$) satisfies a martingale problem with infinitesimal generator $\mathcal{A} : [0,T] \times \mathrm{C}_c^2(\mathcal{M}) \times \mathcal{M} \to \mathbb{R}$ such that for any $t \in [0,T]$, $u \in \mathrm{C}^2(\mathbb{R}^p)$ and $x \in \mathcal{M}$

$$\mathcal{A}(t, \bar{u})(x) = \langle \beta(t,x), \nabla u(x) \rangle + \tfrac{1}{2} \Delta_{\mathcal{M}} u(x). \tag{C.70}$$

In addition, rewriting (C.65) we have

$$\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = \mathrm{KL}(\mathbb{P}_0 \mid \mathbb{Q}_0) + \tfrac{1}{2} \int_0^T \mathbb{E}\big[\|\beta(t, \boldsymbol{X}_t)\|^2\big] \mathrm{d}t, \tag{C.71}$$

which concludes the proof. ∎

We also derive the following useful lemma, which will be used in the proof of convergence of RSGM.

Corollary C.9. *Assume Assumption 3.12. Let $\mathbb{P}^1, \mathbb{P}^2$ be a Markov path measure on $\mathrm{C}([0,T], \mathcal{M})$ with $\mathbb{P}_0^1 = \mathbb{P}_0^2$. In addition, assume that there exist $b_1, b_2 \in \mathrm{C}^\infty([0,T], \mathcal{X}(\mathcal{M}))$ such that $(\boldsymbol{X}_t^1)_{t\in[0,T]}$ and $(\boldsymbol{X}_t^2)_{t\in[0,T]}$ are associated to $\mathbb{P}^1$ and $\mathbb{P}^2$ respectively and satisfy weakly $\mathrm{d}\boldsymbol{X}_t^i = b_1(t, \boldsymbol{X}_t^i)\mathrm{d}t + \mathrm{d}\boldsymbol{b}_t$ for $i \in \{1,2\}$. Then, we have that*

$$\mathrm{KL}\big(\mathbb{P}^1 \,\big|\, \mathbb{P}^2\big) = \tfrac{1}{2} \int_0^T \mathbb{E}\Big[\big\|b_1(t, \boldsymbol{X}_t^1) - b_2(t, \boldsymbol{X}_t^1)\big\|^2\Big] \mathrm{d}t. \tag{C.72}$$

Proof. Upon, using the Nash embedding theorem (see Gunther, 1991), we can assume that $\mathcal{M}$ is a sub-manifold of $\mathbb{R}^p$ with $p \in \mathbb{N}$ such that the Riemannian metric on $\mathcal{M}$ is induced by the Euclidean metric on $\mathbb{R}^p$. Since $\mathcal{M}$ is compact, there exists $R > 0$ such that $\mathcal{M} \subset \bar{\mathrm{B}}(0,R)$. Let $\varphi \in \mathrm{C}^\infty(\mathbb{R}^p, [0,1])$ such that for any $x \in \bar{\mathrm{B}}(0,R)$, $\varphi(x) = 1$ and for any $x \in \mathbb{R}^p$ with $\|x\| \geq R+1$, $\varphi(x) = 0$. Consider

$\bar{b}_1, \bar{b}_2 \in \mathrm{C}_c^2([0,T] \times \mathbb{R}^p, \mathbb{R}^p)$ such that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\bar{b}_i(x) = b_i(x)$ with $i \in \{1,2\}$. Consider $(\bar{X}_t^i)_{t\in[0,T]}$ such that for any $i \in \{1,2\}$

$$\mathrm{d}\bar{X}_t^i = \varphi(\bar{X}_t^i)\{\mathrm{P}(\bar{X}_t^i)\bar{b}^i(t,\bar{X}_t^i) + \bar{b}(\bar{X}_t)\}\mathrm{d}t + \varphi(\bar{X}_t^i)\mathrm{P}(\bar{X}_t^i)\mathrm{d}\boldsymbol{b}_t, \tag{C.73}$$

where $\bar{b} \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R}^p)$ is defined in the proof of Proposition C.8. Let $\bar{X}_0^i \sim \mathbb{P}_0^1$ for any $i \in \{1,2\}$ then for any $i \in \{1,2\}$, $(\bar{X}_t^i)_{t\in[0,T]}$ (seen as a process on $\mathcal{M}$) is such that $\mathcal{L}((\bar{X}_t^i)_{t\in[0,T]}) = \mathbb{P}^i$. Indeed, denote $\{\mathcal{A}b_t^i\}_{t\in[0,T]}$ the generator of $(\bar{X}_t^i)_{t\in[0,T]}$ for any $i \in \{1,2\}$. Let $f \in \mathrm{C}^\infty(\mathcal{M}, \mathbb{R})$ and $\bar{f} \in \mathrm{C}^\infty(\mathbb{R}^p, \mathbb{R})$ an extension to $\mathbb{R}^p$. We have that for any $i \in \{1,2\}$, $x \in \mathcal{M}$ and $t \in [0,T]$

$$\mathcal{A}b_t^i(\bar{f})(x) = \langle \bar{b}^i(t,x) + \bar{b}(x), \nabla\bar{f}(x)\rangle + \tfrac{1}{2}\langle \mathrm{P}(x), \nabla^2\bar{f}(x)\rangle \tag{C.74}$$

$$= \langle b^i(t,x), \nabla f(x)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} f(x). \tag{C.75}$$

Hence, for any $i \in \{1,2\}$, $(\bar{X}_t^i)_{t\in[0,T]}$ (seen as a process on $\mathcal{M}$) and $(X_t^i)_{t\in[0,T]}$ have the same infinitesimal generators. Hence, $\mathcal{L}((\bar{X}_t^i)_{t\in[0,T]}) = \mathbb{P}^i$ for any $i \in \{1,2\}$. For any $i \in \{1,2\}$, denote $\bar{\mathbb{P}}^i = \mathcal{L}((\bar{X}_t^i)_{t\in[0,T]})$ (seen as a process on $\mathbb{R}^p$). Note that since for any $x \in \mathbb{R}^p$ with $\|x\| \geq R+1$, $\varphi(x) = 0$ we have that (Liptser and Shiryaev, 2001, Equation (7.137)) is satisfied. In addition, since for any $x \in \mathbb{R}^p$ with $\|x\| \geq R+1$, $\varphi(x) + \|\nabla\varphi(x)\| = 0$, we have that (Liptser and Shiryaev, 2001, Equation (4.110), Equation (4.111)) are satisfied. In addition, letting for any $t \in [0,T]$ and $x \in \mathbb{R}^p$, $\alpha(t,x) = \bar{b}^1(t,x) - \bar{b}^2(t,x) = \mathrm{P}(x)(\bar{b}^1(t,x) - \bar{b}^2(t,x))$, we have that for any $t \in [0,T]$, $\mathrm{P}(x)\alpha(t,x) = \mathrm{P}(x)(\bar{b}^1(t,x) - \bar{b}^2(t,x))$. Therefore, we can apply (Liptser and Shiryaev, 2001, Section 7.6.4) and using that $\mathrm{P}(x)\bar{b}(x) = 0$ for any $x \in \mathcal{M}$ (see the proof of Proposition C.8), we have that

$$(\mathrm{d}\bar{\mathbb{P}}^1/\mathrm{d}\bar{\mathbb{P}}^2)((\bar{X}_t^1)_{t\in[0,T]}) = \exp\Big[\int_0^T \langle \bar{b}^1(t,\bar{X}_t^1) - \bar{b}^2(t,\bar{X}_t^1), \mathrm{P}(\bar{X}_t^1)\mathrm{d}\bar{X}_t^1\rangle \tag{C.76}$$

$$-\tfrac{1}{2}\int_0^T \langle \bar{b}^1(t,\bar{X}_t^1) - \bar{b}^2(t,\bar{X}_t^1), \mathrm{P}(\bar{X}_t^1)(\bar{b}^1(t,\bar{X}_t^1) + \bar{b}^2(t,\bar{X}_t^1))\rangle \mathrm{d}t\Big] \tag{C.77}$$

$$= \exp\Big[\int_0^T \langle \bar{b}^1(t,\bar{X}_t^1) - \bar{b}^2(t,\bar{X}_t^1), \mathrm{P}(\bar{X}_t^1)\{\bar{b}^1(t,\bar{X}_t^1) + \bar{b}(\bar{X}_t^1)\}\rangle \mathrm{d}t \tag{C.78}$$

$$+ \int_0^T \langle \bar{b}^1(t,\bar{X}_t^1) - \bar{b}^2(t,\bar{X}_t^1), \mathrm{P}(\bar{X}_t^1)\mathrm{d}\boldsymbol{b}_t\rangle \tag{C.79}$$

$$-\tfrac{1}{2}\int_0^T \langle \bar{b}^1(t,\bar{X}_t^1) - \bar{b}^2(t,\bar{X}_t^1), \mathrm{P}(\bar{X}_t^1)(\bar{b}^1(t,\bar{X}_t^1) + \bar{b}^2(t,\bar{X}_t^1))\rangle \mathrm{d}t\Big] \tag{C.80}$$

$$= \exp\Big[\tfrac{1}{2}\int_0^T \big\|\bar{b}^1(t,\bar{X}_t^1) - \bar{b}^2(t,\bar{X}_t^1)\big\|^2 \mathrm{d}t + \int_0^T \langle \bar{b}^1(t,\bar{X}_t^1) - \bar{b}^2(t,\bar{X}_t^1), \mathrm{P}(\bar{X}_t^1)\mathrm{d}\boldsymbol{b}_t\rangle\Big]. \tag{C.81}$$

Therefore, we have that

$$\mathrm{KL}\big(\bar{\mathbb{P}}^1 \,\big|\, \bar{\mathbb{P}}^2\big) = \tfrac{1}{2}\int_0^T \mathbb{E}\Big[\big\|\bar{b}^1(t,\bar{X}_t^1) - \bar{b}^2(t,\bar{X}_t^1)\big\|^2\Big]\mathrm{d}t. \tag{C.82}$$

Hence, we get

$$\mathrm{KL}\big(\bar{\mathbb{P}}^1 \,\big|\, \bar{\mathbb{P}}^2\big) = \tfrac{1}{2}\int_0^T \mathbb{E}\Big[\big\|b^1(t,X_t^1) - b^2(t,X_t^1)\big\|^2\Big]\mathrm{d}t. \tag{C.83}$$

which concludes the proof. ∎

Once Proposition C.8 is established, we can obtain the following straightforward extension of Cattiaux et al. (2023, Proposition 4.6).

PROPOSITION C.10. *Assume Assumption 3.12. Let $\mathbb{Q}$ be a Brownian motion with $\mathbb{Q}_0 = p_{ref}$ and $\mathbb{P}$ a path measure on $\mathrm{C}([0,T],\mathcal{M})$ such that $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) < +\infty$. Then, there exist $\beta_{\mathbb{P}}, \beta_{R(\mathbb{P})} : [0,T] \times \mathcal{M} \to$ such that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\beta_{\mathbb{P}}(t,x), \beta_{R(\mathbb{P})}(t,x) \in \mathrm{T}_x\mathcal{M}$. In addition, we have that $\mathbb{P}$ and $R(\mathbb{P})$ satisfy martingale problems with infinitesimal generator $\mathcal{A}_{\mathbb{P}}$, respectively $\mathcal{A}_{R(\mathbb{P})}$ where for any $t \in [0,T]$, $u \in \mathrm{C}^2(\mathcal{M})$ and $x \in \mathcal{M}$ we have*

$$\mathcal{A}_{\mathbb{P}}(t,u)(x) = \langle \beta_{\mathbb{P}}(t,x), \nabla u(x)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} u(x), \tag{C.84}$$

$$\mathcal{A}_{R(\mathbb{P})}(t,u)(x) = \langle \beta_{R(\mathbb{P})}(t,x), \nabla u(x)\rangle + \tfrac{1}{2}\Delta_{\mathcal{M}} u(x). \tag{C.85}$$

Finally, we have that

$$\int_0^T \mathbb{E}\left[\|\beta_{\mathbb{P}}(t, \boldsymbol{X}_t)\|^2\right] \mathrm{d}t + \int_0^T \mathbb{E}\left[\left\|\beta_{R(\mathbb{P})}(t, \boldsymbol{X}_{T-t})\right\|^2\right] \mathrm{d}t < +\infty, \tag{C.86}$$

where $(\boldsymbol{X}_t)_{t\in[0,T]}$ has distribution $\mathbb{P}$.

Proof. The proof is straightforward upon combining proposition C.8 and the fact that $\mathrm{KL}(\mathbb{P} \mid \mathbb{Q}) = \mathrm{KL}(R(\mathbb{P}) \mid R(\mathbb{Q})) = \mathrm{KL}(R(\mathbb{P}) \mid \mathbb{Q}) < +\infty$, using that $\mathbb{Q}$ is stationary. ∎

We conclude this section, with the following application of Theorem C.7.

PROPOSITION C.11. *For any $u, v \in \mathrm{C}_c^\infty(\mathcal{M})$, we have that for almost any $t \in [0,T]$*

$$\mathbb{E}\left[v(\boldsymbol{X}_t)(\langle \beta_{\mathbb{P}}(t,\boldsymbol{X}_t) + \beta_{R(\mathbb{P})}(T-t,\boldsymbol{X}_t), \nabla u(\boldsymbol{X}_t)\rangle + \Delta_{\mathcal{M}} u(\boldsymbol{X}_t)) + \langle \nabla u(\boldsymbol{X}_t), \nabla v(\boldsymbol{X}_t)\rangle\right] = 0. \tag{C.87}$$

Proof. Remark that $\mathrm{C}_c^2(\mathcal{M}) \subset \mathrm{dom}(\Upsilon_{\mathbb{P}})$ and $\mathrm{C}_c^2(\mathcal{M}) \subset \mathrm{dom}(\Upsilon_{\mathrm{R}(\mathbb{P})})$. In addition, we have that for any $u, v \in \mathrm{C}_c^2(\mathcal{M})$, $\Upsilon_{\mathbb{P}}(u,v) = \Upsilon_{R(\mathbb{P})}(u,v) = \langle u, v\rangle$. Note that by Proposition C.10 and Theorem C.7 we have that for any $u, v \in \mathrm{C}_c^\infty(\mathcal{M})$, (C.87) holds. ∎

Concluding the proof

Using proposition C.11 we can now conclude the proof of Theorem 3.6. First, remark that we can identify $\beta_{\mathbb{P}} = b$. Let $u, v \in \mathrm{C}^\infty(\mathcal{M})$, we have that

$$\mathbb{E}\left[v(\boldsymbol{X}_t)\langle b(\boldsymbol{X}_t) + \beta_{R(\mathbb{P})}(T-t,\boldsymbol{X}_t), \nabla u(\boldsymbol{X}_t)\rangle + \Delta_{\mathcal{M}} u(\boldsymbol{X}_t) v(\boldsymbol{X}_t) + \langle \nabla u(\boldsymbol{X}_t), \nabla v(\boldsymbol{X}_t)\rangle\right] = 0. \tag{C.88}$$

Using that for any $t \in [0,T]$, $\mathbb{P}_t$ admits a smooth positive density with respect to. p_{ref} denoted p_t and the divergence theorem, see (Lee, 2018, p.51), we have that for

any $t \in [0,T]$,

$$\int_{\mathcal{M}} \{\langle \beta_{R(\mathbb{P})}(T-t,x), \nabla u(x)\rangle + \langle b(x), \nabla u(x)\rangle\} v(x) p_t(x) \mathrm{d}p_{\mathrm{ref}}(x) \tag{C.89}$$

$$= -\int_{\mathcal{M}} \langle \nabla u(x) p_t(x), \nabla v(x)\rangle \mathrm{d}p_{\mathrm{ref}}(x) - \int_{\mathcal{M}} \Delta_{\mathcal{M}} u(x) v(x) p_t(x) \mathrm{d}p_{\mathrm{ref}}(x) \tag{C.90}$$

$$= \int_{\mathcal{M}} \langle \nabla \log p_t(x), \nabla u(x) v(x) p_t(x) \mathrm{d}p_{\mathrm{ref}}(x). \tag{C.91}$$

Therefore, we get that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\langle \beta_{R(\mathbb{P})}(T-t,x), \nabla u(x)\rangle = \langle -b(x) + \nabla \log p_t(x), \nabla u(x)\rangle$, which concludes the proof.

C.4. Convergence of RSGM

In this section, we study the convergence of RSGM and prove Theorem 3.13. We state our main results in Appendix C.4.1 and give discretization bounds following the recent work of Cheng et al. (2022) in sec:discr-bounds-grw.

C.4.1. Main results. In this section, we prove Theorem 3.13. We start by recalling the sequence considered in RSGM. Let $(Y_k)_{k\in\{0,\dots,N\}}$ be given by $Y_0 \sim p_{\mathrm{ref}}$ and for any $k \in \{0,\dots,N-1\}$

$$Y_{k+1} = \exp_{Y_k}[\gamma \mathbf{s}_{\theta^\star}(T-n\gamma, Y_k) + \sqrt{2} Z_{k+1}], \tag{C.92}$$

where $\{Z_k\})_{n\in\mathbb{N}}$ is a sequence of independent square integrable random variables with zero mean and identity covariance matrix. For ease of reading, we restate Theorem 3.13.

Theorem C.12. *Assume Assumption 3.12, that p_0 is smooth and positive and that there exists $\mathcal{M}tt \geq 0$ such that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\|\mathbf{s}_{\theta^\star}(t,x) - \nabla \log p_t(x)\| \leq \mathcal{M}tt$, with $\mathbf{s}_{\theta^\star} \in \mathrm{C}([0,T], \mathcal{X}(\mathcal{M}))$. Then if $T > 1/2$, there exists $C \geq 0$ independent on T such that*

$$\boldsymbol{W}_1(\mathcal{L}(Y_N), p_0) = C(\mathrm{e}^{-\lambda_1 T} + \sqrt{T/2}\mathtt{M} + \mathrm{e}^T \gamma^{1/2}), \tag{C.93}$$

where $\boldsymbol{W}_1$ is the Wasserstein distance of order one on the probability measures on $\mathcal{M}$.

Proof. For any $k \in \{1,\dots,N\}$, denote R_k such that for any $x \in \mathbb{R}^d$, $\mathsf{A} \in \lfloor(\mathbb{R}^d)$ and $k \in \{0,\dots,N-1\}$ we have

$$\mathbb{E}[\mathrm{R}_{k+1}(Y_k, \mathsf{A})] = \mathbb{E}[\mathbb{1}_{\mathsf{A}}(Y_{k+1})]. \tag{C.94}$$

Define for any $k_0, k_1 \in \{1,\dots,N\}$ with $k_1 \geq k_0$ $Q_{k_0,k_1} = \prod_{\ell=k_0}^{k_1} \mathrm{R}_{k_1+k_0-\ell}$. Finally, for ease of notation, we also define for any $k \in \{1,\dots,N\}$, $Q_k = Q_{k+1,N}$. Note that for any $k \in \{1,\dots,N\}$, Y_k has distribution $\pi_\infty Q_k$, where $\pi_\infty \in \mathcal{P}(\mathcal{M})$ with density with respect to. the Hausdorff measure p_{ref}. Let $\mathbb{P} \in \mathcal{P}(C)$ be the probability measure associated with $(\boldsymbol{b}_t)_{t\in[0,T]}$ with $\boldsymbol{b}_0 \sim \pi_0$, where $\pi_0 \in \mathcal{P}(\mathcal{M})$ admits a density with respect to. the Hausdorff measure given by p_0. We denote $(\hat{Y}_t)_{t\in[0,T]}$ the process defined by the diffusion $\mathrm{d}\hat{Y}_t = \mathbf{s}_{\theta^\star}(T-t, \hat{Y}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{b}_t$ and $\hat{Y}_0 \sim \pi_\infty$. We

also denote $\hat{\mathbb{P}}^R \in \mathcal{P}(C)$ the probability measure associated with $(\hat{Y}_t)_{t\in[0,T]}$. First note that using that $\mathbb{P}_0 = \pi_0$ we have for any $\mathrm{A} \in \lfloor(\mathcal{M})$

$$\pi_0 \mathbb{P}_{T|0}(\mathbb{P}^R)_{T|0}(\mathrm{A}) = \mathbb{P}_T(\mathbb{P}^R)_{T|0}(\mathrm{A}) = (\mathbb{P}^R)_0(\mathbb{P}^R)_{T|0}(\mathrm{A}) = (\mathbb{P}^R)_T(\mathrm{A}) = \pi_0(\mathrm{A}). \tag{C.95}$$

Hence we have that

$$\pi_0 = \pi_0 \mathbb{P}_{T|0}(\mathbb{P}^R)_{T|0}. \tag{C.96}$$

Let $\varphi \in \mathrm{C}(\mathcal{M})$ with is 1-Lipschitz, i.e. for any $x, y \in \mathcal{M}$, $|\varphi(x) - \varphi(y)| \leq d(x, y)$. Since $\mathcal{M}$ is compact, we have that φ is bounded. Using this result, (C.96), the data processing theorem (Kullback, 1997, Theorem 4.1) and Pinsker's inequality (Bakry et al., 2014, Equation 5.2.2) we have

$$\left|\mathbb{E}[\varphi(Y_N)] - \int_{\mathcal{M}} \varphi(x) p_0(x) \mathrm{d}\mu(x)\right| \tag{C.97}$$

$$\leq |\mathbb{E}[\varphi(\boldsymbol{b}_0)] - \mathbb{E}[\varphi(Y_T)]| + \left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_T)]\right|\left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_N)]\right| \tag{C.98}$$

$$\leq \|\varphi\|_\infty \left\|\pi_0 - \pi_\infty(\mathbb{P}^R)_{T|0}\right\|_{\mathrm{TV}} + \left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_T)]\right| + \left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_N)]\right| \tag{C.99}$$

$$\leq \|\varphi\|_\infty \left\|\pi_0 \mathbb{P}_{T|0}(\mathbb{P}^R)_{T|0} - \pi_\infty(\mathbb{P}^R)_{T|0}\right\|_{\mathrm{TV}} \tag{C.100}$$

$$+ \left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_T)]\right| + \left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_N)]\right| \tag{C.101}$$

$$\leq \|\varphi\|_\infty \left\|\pi_0 \mathbb{P}_{T|0} - \pi_\infty\right\|_{\mathrm{TV}} + \left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_T)]\right| + \left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_N)]\right| \tag{C.102}$$

$$\leq \|\varphi\|_\infty \left\|\pi_0 \mathbb{P}_{T|0} - \pi_\infty\right\|_{\mathrm{TV}} + \sqrt{2}\|\varphi\|_\infty \mathrm{KL}^{1/2}\, \pi_\infty \mathbb{P}^R_{|0} \pi_\infty \hat{\mathbb{P}}^R_{|0} + \left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_N)]\right|. \tag{C.103}$$

We now control each one of these terms. The first term can be easily controlled using the geometric ergodicity of the Brownian motion on compact manifolds. The second term can be controlled using the Girsanov theory on isometrically embedded manifolds. For the last term, we rely on the convergence of the GRW to its associated diffusion as presented in Appendix C.4.2. We now control each one of these terms.

(a) Using Proposition C.6, we have that $\left\|\pi_0 \mathbb{P}_{T|0} - \pi_\infty\right\|_{\mathrm{TV}} \leq C^{1/2} \mathrm{e}^{\lambda_1/2} \mathrm{e}^{-\lambda_1 T}$ where λ_1 is the first positive eigenvalue of $-\Delta_{\mathcal{M}}$ in $\mathrm{L}^2(\pi_\infty)$. Therefore, we get that

$$\|\varphi\|_\infty \left\|\pi_0 \mathbb{P}_{T|0} - \pi_\infty\right\|_{\mathrm{TV}} \leq C^{1/2} \mathrm{e}^{\lambda_1/2} \|\varphi\|_\infty \mathrm{e}^{-\lambda_1 T}. \tag{C.104}$$

(b) Recall that we have that $\mathbb{P}^R_{|0}$ is associated with the process $\mathrm{d}\boldsymbol{Y}_t = \nabla \log p_{T-t}(\boldsymbol{Y}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{b}^{\mathcal{M}}_t$ and that $\hat{\mathbb{P}}^R_{|0}$ is associated with the process $\mathrm{d}\hat{\boldsymbol{Y}}_t = \mathbf{s}_{\theta^\star}(T-t, \hat{\boldsymbol{Y}}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{b}^{\mathcal{M}}_t$. Using Corollary C.9 we have that

$$\mathrm{KL}\left(\pi_\infty \mathbb{P}^R_{|0} \,\middle|\, \pi_\infty \hat{\mathbb{P}}^R_{|0}\right) = \tfrac{1}{2} \int_0^T \mathbb{E}\left[\|\mathbf{s}_{\theta^\star}(T-t, Y_t) - \nabla \log p_{T-t}(Y_t)\|^2\right] \leq M^2 T. \tag{C.105}$$

(c) Let us define $\{\bar{Y}^k\}_{k=0}^N$ such that for any $k \in \{0, \ldots, N\}$, $\bar{Y}^k_0 = \hat{Y}_0 = Y_0$ and for any $t \in [0, k\gamma]$ we have that $\bar{Y}^0_t = \hat{Y}_t$. For any $t \in [k\gamma, T]$, we have that $\bar{Y}^k_t = Y_{t,k}$,

where $Y_{k\gamma,k} = \hat{Y}_{k\gamma}$ and for any $j \in \{k, \dots, N-1\}$ and $t \in [0, \gamma]$

$$Y_{j\gamma+t,k} = \exp_{Y_{j\gamma,k}}[t\mathbf{s}_{\theta^\star}(T - j\gamma, Y_{j\gamma,k}) + \sqrt{t}E_j^k Z_j], \tag{C.106}$$

where $\{Z_j\}_{j=0}^{N-1}$ are independent Gaussian random variables with identity covariance matrix and zero mean and E_j^k is a frame of $\mathrm{T}_{Y_{j\gamma,k}}\mathcal{M}$ such that for any $j \in \{k+1, \dots, N-1\}$, $E_j^{k+1} = \Gamma_{Y_{j\gamma,k}}^{Y_{j\gamma,k+1}} E_j^k$ and $\{E_j^0\}_{j=0}^{N-1}$ is such that for any $j \in \{0, \dots, N-1\}$, E_j^0 is a frame of $\mathrm{T}_{Y_{j\gamma}}\mathcal{M}$. One $[0, k\gamma]$, we define $(\hat{Y}_t^k)_{t\in[0,k\gamma]}$ as follows. For any $k \in \{0, \dots, N-1\}$, we set $(Y_t^{k+1})_{t\in[0,k\gamma]} = (Y_t^k)_{t\in[0,k\gamma]}$. For any $k \in \{0, \dots, N-1\}$, we set $(Y_t)_{k\gamma,(k+1)\gamma}$ as in Proposition C.17 (taking the notations of Proposition C.17, $X_1^0 = \hat{Y}_{(k+1)\gamma}^k$ and $X_\gamma = \hat{Y}_{k\gamma}^k$). Note that we have $\{\bar{Y}_{j\gamma,0}^N\}_{j=0}^N = \{Y_j^N\}_{j=0}^N$ and $\{\bar{Y}_{t,N}\}_{t\in[0,T]} = \{\hat{Y}_t\}_{t\in[0,T]}$. Therefore, we have that

$$\left|\varphi(\hat{Y}_T) - \varphi(Y_N)\right| = \left|\varphi(\bar{Y}_T^0) - \varphi(\bar{Y}_T^N)\right| \tag{C.107}$$

$$\leq \sum_{k=0}^{N-1} \left|\varphi(\bar{Y}_T^k) - \varphi(\bar{Y}_T^{k+1})\right| \leq \|\nabla\varphi\|_\infty \sum_{k=0}^{N-1} d(\bar{Y}_T^k, \bar{Y}_T^{k+1}). \tag{C.108}$$

In addition, using Proposition C.17 and Proposition C.18, we have that there exists $C \geq 0$ such that for any $k \in \{0, \dots, N-1\}$

$$\mathbb{E}\left[d(\bar{Y}_{k,T}, \bar{Y}_{k+1,T})\right] \leq C \exp[(N-k)\gamma]\gamma^{3/2}. \tag{C.109}$$

Therefore, we get that there exists $C \geq 0$ such that

$$\left|\mathbb{E}\left[\varphi(\hat{Y}_T)\right] - \mathbb{E}[\varphi(Y_N)]\right| \leq C \exp[T]\gamma^{1/2}, \tag{C.110}$$

Therefore, we get that there exists $C \geq 0$ such that for any $\varphi \in \mathrm{C}(\mathcal{M})$ which is 1-Lipschitz, we have

$$\mathbb{E}[\varphi(Y_N)] - \int_{\mathcal{M}} \varphi(x) p_0(x) \mathrm{d}p_{\mathrm{ref}}(x) \leq C(\mathrm{e}^{\lambda_1/2}\|\varphi\|_\infty \mathrm{e}^{-\lambda_1 T} + \sqrt{T/2}\|\varphi\|_\infty \mathtt{M} + \mathrm{e}^T\gamma^{1/2}). \tag{C.111}$$

Let $x_0 \in \mathcal{M}$. Let $\mathrm{Lip}(\mathcal{M})$ the set of Lipschitz functions on $\mathcal{M}$ with Lipschitz constant equal to 1. Let $\mathrm{Lip}(\mathcal{M})_0$ the set of Lipschitz functions on $\mathcal{M}$ with Lipschitz constant equal to 1 and such that for any $\varphi \in \mathrm{Lip}(\mathcal{M})_0$, $\varphi(x_0) = 0$. Note that in this case, we have that $\|\varphi\|_\infty \leq \mathrm{diam}(\mathcal{M})$. Using (C.111), we have

$$\boldsymbol{W}_1(\mathcal{L}(Y_N), p_0) = \sup\left\{\mathbb{E}[\varphi(Y_N)] - \int_{\mathcal{M}} \varphi(x) p_0(x) \mathrm{d}p_{\mathrm{ref}}(x) \ : \ \varphi \in \mathrm{Lip}(\mathcal{M})\right\} \tag{C.112}$$

$$= \sup\left\{\mathbb{E}[\varphi(Y_N)] - \int_{\mathcal{M}} \varphi(x) p_0(x) \mathrm{d}p_{\mathrm{ref}}(x) \ : \ \varphi \in \mathrm{Lip}(\mathcal{M})_0\right\} \tag{C.113}$$

$$\leq C(\mathrm{e}^{\lambda_1/2}\mathrm{diam}(\mathcal{M})\mathrm{e}^{-\lambda_1 T} + \sqrt{T/2}\mathrm{diam}(\mathcal{M})\mathtt{M} + \mathrm{e}^T\gamma^{1/2}), \tag{C.114}$$

which concludes the proof. ∎

We now state a result regarding the continuous-time process (i.e. we now longer consider discretization errors). We recall that we denote $(\hat{Y}_t)_{t\in[0,T]}$ the process defined by the diffusion $\mathrm{d}\hat{Y}_t = \mathbf{s}_{\theta^\star}(T-t,\hat{Y}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{b}_t$ and $\hat{Y}_0 \sim \pi_\infty$.

THEOREM C.13. *Assume Assumption 3.12, that p_0 is smooth and positive and that there exists $\mathcal{M}tt \geq 0$ such that for any $t \in [0,T]$ and $x \in \mathcal{M}$, $\|\mathbf{s}_{\theta^\star}(t,x) - \nabla \log p_t(x)\| \leq \mathcal{M}tt$, with $\mathbf{s}_{\theta^\star} \in \mathrm{C}([0,T],\mathcal{X}(\mathcal{M}))$. Then if $T > 1/2$, there exists $C \geq 0$ independent on T such that*

$$\left\|\mathcal{L}(\hat{Y}_T) - p_0\right\|_{\mathrm{TV}} = C(\mathrm{e}^{-\lambda_1 T} + \sqrt{T/2\mathtt{M}}). \tag{C.115}$$

Proof. The proof is identical to the one of Theorem C.12, except that we do not have to deal with the discretization error. We use that for any $\mu, \nu \in \mathcal{P}(\mathcal{M})$

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup\left\{\mu[f] - \nu[f] \ : \ f \in \mathrm{C}(\mathcal{M}), \|f\|_\infty \leq 1\right\}. \tag{C.116}$$

■

The result of Theorem C.13 should be compared with the one of (Rozen et al., 2021, Theorem 3). With our result we control a L^1 bound between the density of $\hat{Y}_T$ and the one of p_0. In (Rozen et al., 2021, Theorem 3) a L^∞ bound between the densities is recovered. It can be shown that $\hat{p}_T = \mathcal{L}(\hat{Y}_T)$. Let κ be the modulus of continuity of $\hat{p}_T - p_0$, i.e. for any $\varepsilon \geq 0$

$$\kappa(\varepsilon) = \sup\{|\hat{p}_T(x) - p_0(x) - \hat{p}_T(y) + p_0(y)| \ : \ x,y \in \mathcal{M},\ d(x,y) \leq \varepsilon\}. \tag{C.117}$$

Let $x_0 \in \mathcal{M}$ such that

$$|\hat{p}_T(x_0) - p_0(x_0)| = M = \sup\{|\hat{p}_T(x) - p_0(x)| \ : \ x \in \mathcal{M}\}. \tag{C.118}$$

For any $x \in \bar{\mathrm{B}}(x_0, \kappa(M/2))$, we have $|\hat{p}_T(x) - p_0(x)| \geq M/2$. Hence, denoting $\mathrm{Vol}_\kappa = \int_{\bar{\mathrm{B}}(x_0,\kappa(M/2))} \mathrm{d}p_{\mathrm{ref}}(x) > 0$, we have

$$(2/\mathrm{Vol}_\kappa)\int_{\mathcal{M}} |\hat{p}_T(x) - p_0(x)|\mathrm{d}p_{\mathrm{ref}}(x) \geq \|\hat{p}_T - p_0\|_\infty. \tag{C.119}$$

Hence, there exists $C \geq 0$ such that for any $T > 1/2$

$$\|\hat{p}_T - p_0\|_\infty \leq C(\mathrm{e}^{-\lambda_1 T} + \sqrt{T/2\mathtt{M}}). \tag{C.120}$$

Therefore, we recover the same guarantees as Theorem C.13 (note that $\mathtt{M}$ is not explicitly controlled using network properties in our work, but we could use universal approximation properties as in Rozen et al. (2021) in order to obtain a similar result).

C.4.2. Discretization bounds for GRW. In this section, we establish discretization bounds for GRW. Our results are a straightforward extension of Cheng et al. (2022) to the case where the drift term in the GRW is time-inhomogeneous.

Since $\mathcal{M}$ is compact, we have that for any $x_1, x_2 \in M$, there exists a minimizing geodesic such that $\gamma \in \mathrm{C}^\infty([0,1],\mathcal{M})$ and $\gamma(0) = x_1$ and $\gamma(1) = x_2$. When this choice is not unique we fix a minimizing geodesic. We denote $\Gamma_{x_1}^{x_2} : \mathrm{T}_{x_1}\mathcal{M} \to \mathrm{T}_{x_2}\mathcal{M}$ the associated parallel transport. Let $b \in \mathrm{C}^\infty([0,T],\mathcal{X}(\mathcal{M}))$.

We start by introducing a family of GRWs defined on progressively finer grids. Let $\gamma > 0, X_0 \in \mathcal{M}, E_0 \in \mathrm{F}_{X_0}\mathcal{M}$ (the vector space of frames at X_0) and consider the families $\left\{E_k^\ell \ : \ k \in \{0, \dots, 2^\ell\}, \ \ell \in \mathbb{N}\right\}$, $\left\{X_k^\ell \ : \ k \in \{0, \dots, 2^\ell\}, \ \ell \in \mathbb{N}\right\}$ such that $X_0^0 = X_0$, $X_1^0 = \exp_{X_0^0}[\gamma b(0, X_0^0) + \sqrt{\gamma}(\boldsymbol{b}_1 - \boldsymbol{b}_0)E_0^0]$ and $E_1^0 = \Gamma_{X_0^0}^{X_1^0} E_0^0$ (note that $E_{2^\ell}^\ell$ is not used in the proof but defined for completeness). In addition, we have that for any $\ell \in \mathbb{N}$ with $\ell \geq 1$, $X_0^\ell = X_0$, $E_0^\ell = E_0$ and for any $k \in \{0, \dots, 2^{\ell-1} - 1\}$

$$X_{2k+1}^\ell = \exp_{X_{2k}^\ell}[\gamma_\ell b(2k\gamma_\ell, X_{2k}^\ell) + E_{2k}^\ell(\boldsymbol{b}_{(2k+1)\gamma_\ell} - \boldsymbol{b}_{2k\gamma_\ell})], \tag{C.121}$$

$$E_{2k+1}^\ell = \Gamma_{X_{2k}^\ell}^{X_{2k+1}^\ell} E_{2k}^\ell, \tag{C.122}$$

$$X_{2k+2}^\ell = \exp_{X_{2k+1}^\ell}[\gamma_\ell b((2k+1)\gamma_\ell, X_{2k+1}^\ell) + E_{2k+1}^\ell(\boldsymbol{b}_{(2k+2)\gamma_\ell} - \boldsymbol{b}_{(2k+1)\gamma_\ell})], \tag{C.123}$$

$$E_{2k+2}^\ell = \Gamma_{X_{k+1}^{\ell-1}}^{X_{2k+2}^\ell} E_{k+1}^{\ell-1}, \tag{C.124}$$

where $\gamma_\ell = \gamma/2^\ell$. For any $\ell \in \mathbb{N}$, we also define $(\boldsymbol{X}_t^\ell)_{t\in[0,\gamma]}$ such that for any $\ell \in \mathbb{N}, k \in \{0, \dots, 2^\ell - 1\}$, we have for any $t \in [k\gamma_\ell, (k+1)\gamma_\ell)$, $\boldsymbol{X}_t^\ell = \exp_{X_k^\ell}[(t - k\gamma_\ell)b(k\gamma_\ell, X_k^\ell) + E_k^\ell(\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell})]$. Note that for any $\ell \in \mathbb{N}$ and $k \in \{0, \dots, 2^\ell - 1\}$, $\boldsymbol{X}_{k\gamma_\ell}^\ell = X_k$.

We are going to use the following useful lemma, see (Cheng et al., 2022, Lemma 62).

Lemma C.14. *Assume Assumption 3.12. Then, there exists $C \geq 0$ such that for any $x, y \in \mathcal{M}$, $\gamma : [0,1] \to \mathcal{M}$ minimizing geodesic with $\gamma(0) = x$, $\gamma(1) = y$ and $u \in \mathrm{T}_x\mathcal{M}, v \in \mathrm{T}_y\mathcal{M}$ we have*

$$d(\exp_y[v], \exp_x[u])^2 \leq (1+C\kappa^2 \exp[4\kappa])d(x,y)^2 + C\exp[4\kappa]\left\|\Gamma_y^x v - u\right\|^2 + 2\langle \gamma'(0), \Gamma_y^x v - u\rangle, \tag{C.125}$$

with $\kappa = \|u\| + \|v\|$.

We are now ready to state the main result of this section.

Proposition C.15. *Assume Assumption 3.12. Then, there exists $C \geq 0$ such that for any $\ell \in \mathbb{N}$*

$$\mathbb{E}\left[\sup_{t\in[0,\gamma]} d(\boldsymbol{X}_t^\ell, \boldsymbol{X}_t^{\ell+1})^2\right] \leq C\gamma^3 2^{-2\ell}. \tag{C.126}$$

Proof. Let $\ell \in \mathbb{N}$, $k \in \{0, \dots, 2^\ell - 1\}$ and $t \in [k\gamma_\ell, (k+1)\gamma_\ell]$. We define $U_k^t = d(\boldsymbol{X}_t^\ell, \boldsymbol{X}_t^{\ell+1})^2$, $U_k = \sup\left\{U_k^t \ : \ t \in [k\gamma_\ell, (k+1)\gamma_\ell]\right\}$ and $U_{-1} = 0$. We also introduce for any $j \in \{0, \dots, 2^\ell - 1\}$ and for $t \in [k\gamma_\ell, (2k+1)\gamma_{\ell+1})$, $\bar{\boldsymbol{X}}_t^{\ell+1} = \boldsymbol{X}_t^{\ell+1}$ and for $t \in [(2k+1)\gamma_{\ell+1}, (k+1)\gamma_\ell)$

$$\bar{\boldsymbol{X}}_t^{\ell+1} = \exp_{X_{2j}^{\ell+1}}[\gamma_{\ell+1} b(2j\gamma_{\ell+1}, X_{2j}^{\ell+1}) \tag{C.127}$$

$$+ (t - (2k+1)\gamma_{\ell+1})b((2j+1)\gamma_{\ell+1}, X_{2j}^{\ell+1}) + (\boldsymbol{b}_t - \boldsymbol{b}_{j\gamma_\ell})E_{2j}^{\ell+1}]. \tag{C.128}$$

Using this result and that for any $a, b \geq 0$, $(a+b)^2 \leq (1+2^{-\ell})a^2 + (1+2^\ell)b^2$, we have that for any $t \in [k\gamma_\ell, (k+1)\gamma_\ell]$

$$U_{k+1}^t \leq (1+2^{-\ell})d(\boldsymbol{X}_t^\ell, \bar{\boldsymbol{X}}_t^{\ell+1})^2 + (1+2^\ell)d(\bar{\boldsymbol{X}}_t^{\ell+1}, \boldsymbol{X}_t^{\ell+1})^2. \tag{C.129}$$

Note that for $t \in [k\gamma_\ell, (2k+1)\gamma_{\ell+1}]$, the second term in (C.129) is zero. We now bound each one of these terms:

(a) First, we assume that $t \in [(k+1)\gamma_\ell, (2k+1)\gamma_{\ell+1}]$. Recall that

$$\bar{X}_t^{\ell+1} = \exp_{X_{2k}^{\ell+1}}[\gamma_{\ell+1} b(k\gamma_\ell, X_{2k}^{\ell+1}) \tag{C.130}$$

$$(t-(2k+1)\gamma_{\ell+1}) b((2k+1)\gamma_{\ell+1}, X_{2k}^{\ell+1}) + (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell}) E_{2k}^{\ell+1}], \tag{C.131}$$

$$\boldsymbol{X}_t^\ell = \exp_{X_k^\ell}[(t-k\gamma_\ell) b(k\gamma_\ell, X_k^\ell) + (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell}) E_k^\ell]. \tag{C.132}$$

Hence, using Lemma C.14, we have that

$$d(\bar{X}_t^{\ell+1}, \boldsymbol{X}_t^\ell)^2 \le (1 + C\kappa_k^2 \exp[4\kappa_k]) d(X_k^\ell, X_{2k}^{\ell+1})^2 \tag{C.133}$$

$$+ C\exp[4\kappa_k] \left\| \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} v_k - u_k \right\|^2 + 2\langle w'(0), \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} v_k - u_k \rangle, \tag{C.134}$$

with $w : [0,1] \to \mathcal{M}$ a minimizing geodesic between X_k^ℓ and $X_{2k}^{\ell+1}$

$$\kappa_k = \|u_k\| + \|v_k\|, \tag{C.135}$$

$$u_k^1 = (t-k\gamma_\ell) b(k\gamma_\ell, X_k^\ell), \tag{C.136}$$

$$v_k^1 = \gamma_{\ell+1} b(2k\gamma_{\ell+1}, X_{2k}^{\ell+1}) + (t-(2k+1)\gamma_{\ell+1}) b((2k+1)\gamma_{\ell+1}, X_{2k}^{\ell+1}), \tag{C.137}$$

$$u_k^2 = (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell}) E_k^\ell, \qquad v_k^2 = (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell}) E_{2k}^{\ell+1}, \tag{C.138}$$

$$u_k = u_k^1 + u_k^2, \qquad v_k = v_k^1 + v_k^2. \tag{C.139}$$

In particular, since $E_k^\ell = \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} E_{2k}^{\ell+1}$ using (C.124), we have that $u_k^2 = \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} v_k^2$. Therefore, combining this result and that $t-(2k+1)\gamma_{\ell+1} + \gamma_{\ell+1} = t - k\gamma_\ell$, we get that

$$\left\| \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} v_k^1 - u_k^1 \right\| \le \gamma_{\ell+1} \left\| b(k\gamma_\ell, X_k^\ell) - \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} b(k\gamma_\ell, X_{2k}^{\ell+1}) \right\| \tag{C.140}$$

$$+ \gamma_{\ell+1} \left\| b(k\gamma_\ell, X_k^\ell) - \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} b((2k+1)\gamma_{\ell+1}, X_{2k}^{\ell+1}) \right\| \tag{C.141}$$

$$\le \gamma_\ell \left\| b(k\gamma_\ell, X_k^\ell) - \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} b(k\gamma_\ell, X_{2k}^{\ell+1}) \right\| + \mathfrak{m} L_2 \gamma_\ell^2 \tag{C.142}$$

$$\le \mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1}) + \mathfrak{m} L_2 \gamma_\ell^2. \tag{C.143}$$

Therefore, we get that $\|u_k - v_k\| \le \mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1}) + \mathfrak{m} L_2 \gamma_\ell^2$. In addition, we have that $\|w'(0)\| \le d(X_k^\ell, X_{2k}^{\ell+1})$ since w is a minimizing geodesic. Combining these results and (C.133) we get that

$$d(\bar{X}_t^{\ell+1}, \boldsymbol{X}_t^\ell)^2 \le (1 + C\kappa_k^2 \exp[4\kappa_k]) d(X_k^\ell, X_{2k}^{\ell+1})^2 \tag{C.144}$$

$$+ C\exp[4\kappa_k](\mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1}) + \mathfrak{m} L_2 \gamma_\ell^2)^2 \tag{C.145}$$

$$+ 2(\mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1}) + \mathfrak{m} L_2 \gamma_\ell^2) d(X_k^\ell, X_{2k}^{\ell+1}) \tag{C.146}$$

$$\le (1 + C\kappa_k^2 \exp[4\kappa_k] + 2C\exp[4\kappa_k] \mathfrak{m} L_1^2 \gamma_\ell^2) d(X_k^\ell, X_{2k}^{\ell+1})^2 \tag{C.147}$$

$$+ 2(\mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1}) + \mathfrak{m} L_2 \gamma_\ell^2) d(X_k^\ell, X_{2k}^{\ell+1}) + 2\mathfrak{m} L_2^2 \gamma_\ell^4 \tag{C.148}$$

$$\le (1 + C\kappa_k^2 \exp[4\kappa_k] + 2C\exp[4\kappa_k] \mathfrak{m} L_1^2 \gamma_\ell^2 + 2\mathfrak{m} L_1 \gamma_\ell + 4\mathfrak{m} L_2 \gamma_\ell) d(X_k^\ell, X_{2k}^{\ell+1})^2 + 8\mathfrak{m} L_2 \gamma_\ell^3, \tag{C.149}$$

Hence, there exists $C_1 \ge 0$ (not dependent on k or ℓ) such that

$$(1+2^{-\ell}) d(\bar{X}_t^{\ell+1}, \boldsymbol{X}_t^\ell)^2 \le (1 + C_1\{\kappa_k^2 \exp[4\kappa_k] + \gamma_\ell^2 \exp[4\kappa_k] + 2^{-\ell}\}) d(X_k^\ell, X_{2k}^{\ell+1})^2 + C_1 \gamma_\ell^3. \tag{C.150}$$

Next, we assume that $t \in [k\gamma_\ell, (2k+1)\gamma_{\ell+1}]$. Recall that

$$\bar{\boldsymbol{X}}_t^{\ell+1} = \exp_{X_{2k}^{\ell+1}}[(t-k\gamma_\ell)b(k\gamma_\ell, X_{2k}^{\ell+1}) + (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell})E_{2k}^{\ell+1}], \tag{C.151}$$

$$\boldsymbol{X}_t^{\ell} = \exp_{X_k^{\ell}}[(t-k\gamma_\ell)b(k\gamma_\ell, X_k^{\ell}) + (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell})E_k^{\ell}]. \tag{C.152}$$

Hence, using Lemma C.14, we have that

$$d(\bar{\boldsymbol{X}}_t^{\ell+1}, \boldsymbol{X}_t^{\ell})^2 \leq (1 + C\kappa_k^2 \exp[4\kappa_k]) d(X_k^\ell, X_{2k}^{\ell+1})^2 \tag{C.153}$$

$$+ C\exp[4\kappa_k] \left\| \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} v_k - u_k \right\|^2 + 2\langle w'(0), \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} v_k - u_k \rangle, \tag{C.154}$$

with $w : [0,1] \to \mathcal{M}$ a minimizing geodesic between X_k^ℓ and $X_{2k}^{\ell+1}$

$$\kappa_k = \|u_k\| + \|v_k\|, \tag{C.155}$$

$$u_k^1 = (t-k\gamma_\ell)b(k\gamma_\ell, X_k^\ell), \tag{C.156}$$

$$v_k^1 = (t-k\gamma_\ell)b(k\gamma_\ell, X_{2k}^{\ell+1}), \tag{C.157}$$

$$u_k^2 = (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell})E_k^\ell, \qquad v_k^2 = (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell})E_{2k}^{\ell+1}, \tag{C.158}$$

$$u_k = u_k^1 + u_k^2, \qquad v_k = v_k^1 + v_k^2. \tag{C.159}$$

In particular, since $E_k^\ell = \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} E_{2k}^{\ell+1}$ using (C.124) and $t - (2k+1)\gamma_{\ell+1} + \gamma_{\ell+1} = t - k\gamma_\ell$, we have that $u_k^2 = \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} v_k^2$. Therefore, we get that

$$\left\| \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} v_k^1 - u_k^1 \right\| \leq \gamma_{\ell+1} \left\| b(k\gamma_\ell, X_k^\ell) - \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} b(k\gamma_\ell, X_{2k}^{\ell+1}) \right\| \tag{C.160}$$

$$\leq \gamma_\ell \left\| b(k\gamma_\ell, X_k^\ell) - \Gamma_{X_{2k}^{\ell+1}}^{X_k^\ell} b(k\gamma_\ell, X_{2k}^{\ell+1}) \right\| + \mathfrak{m} L_2 \gamma_\ell^2 \tag{C.161}$$

$$\leq \mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1}). \tag{C.162}$$

Therefore, we get that $\|u_k - v_k\| \leq \mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1})$. In addition, we have that $\|w'(0)\| \leq d(X_k^\ell, X_{2k}^{\ell+1})$ since w is a minimizing geodesic. Combining these results and (C.153) we get that

$$d(\bar{\boldsymbol{X}}_t^{\ell+1}, \boldsymbol{X}_t^{\ell})^2 \leq (1 + C\kappa_k^2 \exp[4\kappa_k]) d(X_k^\ell, X_{2k}^{\ell+1})^2 \tag{C.163}$$

$$+ C\exp[4\kappa_k] \mathfrak{m} L_1^2 \gamma_\ell^2 d(X_k^\ell, X_{2k}^{\ell+1})^2 \tag{C.164}$$

$$+ 2\mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1}) d(X_k^\ell, X_{2k}^{\ell+1}) \tag{C.165}$$

$$\leq (1 + C\kappa_k^2 \exp[4\kappa_k] + 2C\exp[4\kappa_k] \mathfrak{m} L_1^2 \gamma_\ell^2) d(X_k^\ell, X_{2k}^{\ell+1})^2 \tag{C.166}$$

$$+ 2\mathfrak{m} L_1 \gamma_\ell d(X_k^\ell, X_{2k}^{\ell+1})^2 + 2\mathfrak{m} L_2^2 \gamma_\ell^4 \tag{C.167}$$

$$\leq (1 + C\kappa_k^2 \exp[4\kappa_k] + 2C\exp[4\kappa_k] \mathfrak{m} L_1^2 \gamma_\ell^2 + 2\mathfrak{m} L_1 \gamma_\ell) d(X_k^\ell, X_{2k}^{\ell+1})^2. \tag{C.168}$$

Hence, there exists $C_1 \geq 0$ (not dependent on k or ℓ) such that for any $t \in [k\gamma_\ell, (k+1)\gamma_\ell]$

$$(1+2^{-\ell}) d(\bar{\boldsymbol{X}}_t^{\ell+1}, \boldsymbol{X}_t^{\ell})^2 \leq (1 + C_1\{\kappa_k^2 \exp[4\kappa_k] + \gamma_\ell^2 \exp[4\kappa_k] + 2^{-\ell}\}) d(X_k^\ell, X_{2k}^{\ell+1})^2 + C_1 \gamma_\ell^3. \tag{C.169}$$

(b) We recall that if $t \in [k\gamma_\ell, (2k+1)\gamma_{\ell+1}]$ the second term in (C.129) is zero. Therefore in what follows, we assume $t \in [(2k+1)\gamma_{\ell+1}, (k+1)\gamma_\ell]$. We introduce

$$\hat{\boldsymbol{X}}_t^{\ell+1} = \exp_{X_{2k+1}^{\ell+1}}[(t-(2k+1)\gamma_{\ell+1})\Gamma_{X_{2k}^{\ell+1}}^{X_{2k+1}^{\ell+1}} b((2k+1)\gamma_{\ell+1}, X_{2k}^{\ell+1}) \tag{C.170}$$

$$(\boldsymbol{b}_t - \boldsymbol{b}_{(2k+1)\gamma_{\ell+1}})E_{2k+1}^{\ell+1}]. \tag{C.171}$$

In what follows, we provide an upper-bound for $d(\bar{\boldsymbol{X}}_t^{\ell+1}, \boldsymbol{X}_t^{\ell+1})$. First, we have that

$$d(\bar{\boldsymbol{X}}_t^{\ell+1}, \boldsymbol{X}_t^{\ell+1}) \le d(\bar{\boldsymbol{X}}_t^{\ell+1}, \hat{\boldsymbol{X}}_t^{\ell+1}) + d(\hat{\boldsymbol{X}}_t^{\ell+1}, \boldsymbol{X}_t^{\ell+1}). \tag{C.172}$$

We recall that

$$\bar{\boldsymbol{X}}_t^{\ell+1} = \exp_{X_{2k}^{\ell+1}}[\gamma_{\ell+1} b(2k\gamma_{\ell+1}, X_{2k}^{\ell+1}) \tag{C.173}$$

$$+ (t-(2k+1)\gamma_{\ell+1}) b((2k+1)\gamma_{\ell+1}, X_{2k}^{\ell+1}) + (\boldsymbol{b}_t - \boldsymbol{b}_{k\gamma_\ell})E_{2k}^{\ell+1}]. \tag{C.174}$$

Denote a_k, b_k such that

$$a_k = b(2k\gamma_{\ell+1}, X_{2k}^{\ell+1}) + (\boldsymbol{b}_{(2k+1)\gamma_{\ell+1}} - \boldsymbol{b}_{k\gamma_\ell})E_{2k}^{\ell+1}, \tag{C.175}$$

$$b_k = (t-(2k+1)\gamma_{\ell+1}) b((2k+1)\gamma_{\ell+1}, X_{2k}^{\ell+1}) + (\boldsymbol{b}_t - \boldsymbol{b}_{(2k+1)\gamma_{\ell+1}})E_{2k}^{\ell+1}. \tag{C.176}$$

Using (C.124), (C.171) and (C.174) we have that

$$X_{2k+1}^{\ell+1} = \exp_{X_{2k}^{\ell+1}}[a_k], \qquad \hat{\boldsymbol{X}}_t^{\ell+1} = \exp_{X_{2k+1}^{\ell+1}}[\Gamma_{X_{2k}^{\ell+1}}^{X_{2k+1}^{\ell+1}} b_k], \qquad \bar{\boldsymbol{X}}_t^{\ell+1} = \exp_{X_{2k}^{\ell+1}}[a_k + b_k]. \tag{C.177}$$

Using this result and (Sun et al., 2019, Lemma 3), there exists $C_2 \ge 0$ (not dependent on k or ℓ) such that

$$d(\hat{\boldsymbol{X}}_t^{\ell+1}, \bar{\boldsymbol{X}}_t^{\ell+1}) \le C_2(\|a_k\| + \|b_k\|)^3. \tag{C.178}$$

Using this result and that for any $t \in [0, \gamma]$ and $x \in \mathcal{M}$, $\|b(t,x)\| \le \mathtt{m}K$ we get that there exists $C_3 \ge 0$ (not dependent on k or ℓ) such that

$$d(\hat{\boldsymbol{X}}_t^{\ell+1}, \bar{\boldsymbol{X}}_t^{\ell+1})^2 \le C_3(\gamma_{\ell+1}^6 + \left\|\boldsymbol{b}_t - \boldsymbol{b}_{(2k+1)\gamma_{\ell+1}}\right\|^6 + \left\|\boldsymbol{b}_{(2k+1)\gamma_\ell} - \boldsymbol{b}_{(k+1)\gamma_\ell}\right\|^6). \tag{C.179}$$

Finally, we recall that

$$\hat{\boldsymbol{X}}_t^{\ell+1} = \exp_{X_{2k+1}^{\ell+1}}[(t-(2k+1)\gamma_{\ell+1})\Gamma_{X_{2k}^{\ell+1}}^{X_{2k+1}^{\ell+1}} b((2k+1)\gamma_{\ell+1}, X_{2k}^{\ell+1}) \tag{C.180}$$

$$+ (\boldsymbol{b}_t - \boldsymbol{b}_{(2k+1)\gamma_{\ell+1}})E_{2k+1}^{\ell+1}], \tag{C.181}$$

$$\boldsymbol{X}_t^{\ell+1} = \exp_{X_{2k+1}^{\ell+1}}[(t-(2k+1)\gamma_{\ell+1}) b((2k+1)\gamma_{\ell+1}, X_{2k+1}^{\ell+1}) + (\boldsymbol{b}_t - \boldsymbol{b}_{(2k+1)\gamma_{\ell+1}})E_{2k+1}^{\ell+1}]. \tag{C.182}$$

Let us define

$$\tau_k = \|c_k\| + \|d_k\|, \tag{C.183}$$

$$c_k = c_k^1 + c_k^2, \qquad d_k = d_k^1 + d_k^2, \tag{C.184}$$

$$c_k^1 = (t-(2k+1)\gamma_{\ell+1}) b((2k+1)\gamma_{\ell+1}, X_{2k+1}^{\ell+1}), \tag{C.185}$$

$$d_k^1 = (t-(2k+1)\gamma_{\ell+1})\Gamma_{X_{2k}^{\ell+1}}^{X_{2k+1}^{\ell+1}} b((2k+1)\gamma_{\ell+1}, X_{2k}^{\ell+1}), \tag{C.186}$$

$$c_k^2 = d_k^2 = (\boldsymbol{b}_t - \boldsymbol{b}_{(2k+1)\gamma_{\ell+1}})E_{2k+1}^{\ell+1}. \tag{C.187}$$

Using Lemma C.14, we get that

$$d(\boldsymbol{X}_t^{\ell+1}, \hat{\boldsymbol{X}}_t^{\ell+1})^2 \leq C \exp[4\tau_k] \|c_k - d_k\|^2 \leq C \mathfrak{m} L_2^2 \gamma_{\ell+1}^2 \exp[4\tau_k] d(X_{2k+1}^{\ell+1}, X_{2k}^{\ell+1})^2. \tag{C.188}$$

In addition, using Lemma C.14, we get that

$$d(X_{2k+1}^{\ell+1}, X_{2k}^{\ell+1})^2 \leq \exp[4\|e_k\|] \|e_k\|, \tag{C.189}$$

with $e_k = \gamma_{\ell+1} b(k\gamma_\ell, X_{2k}^{\ell+1}) + (\boldsymbol{b}_{(2k+1)\gamma_{\ell+1}} - \boldsymbol{b}_{k\gamma_\ell}) E_{2k}^{\ell+1}$. Combining this result and (C.188), we get that

$$d(\boldsymbol{X}_t^{\ell+1}, \hat{\boldsymbol{X}}_t^{\ell+1})^2 \leq C_3 \gamma_{\ell+1}^2 (\gamma_{\ell+1}^2 + \left\|\boldsymbol{b}_{(2k+1)\gamma_{\ell+1}} - \boldsymbol{b}_{k\gamma_\ell}\right\|^2) \exp[4\tau_k + \|e_k\|]. \tag{C.190}$$

Combining (C.179) and (C.190), there exists C_5 such that

$$d(\bar{\boldsymbol{X}}_t^{\ell+1}, \boldsymbol{X}_t^{\ell+1})^2 \leq C_5 \gamma_{\ell+1}^2 (\gamma_{\ell+1}^2 + \left\|\boldsymbol{b}_{(2k+1)\gamma_{\ell+1}} - \boldsymbol{b}_{k\gamma_\ell}\right\|^2) \exp[4\tau_k + \|e_k\|] \tag{C.191}$$

$$+ C_5 (\gamma_{\ell+1}^6 + \left\|\boldsymbol{b}_t - \boldsymbol{b}_{(2k+1)\gamma_{\ell+1}}\right\|^6 + \left\|\boldsymbol{b}_{(2k+1)\gamma_\ell} - \boldsymbol{b}_{(k+1)\gamma_\ell}\right\|^6). \tag{C.192}$$

In what follows, we denote

$$\alpha_k = C_1 \{(\kappa_k^+)^2 \exp[4\kappa_k] + \gamma_\ell^2 \exp[4\kappa_k^+] + 2^{-\ell}\}. \tag{C.193}$$

$$\beta_k = C_1 \gamma_\ell^3 + C_5 (1 + 2^\ell) \gamma_{\ell+1}^2 (\gamma_{\ell+1}^2 + \left\|\boldsymbol{b}_{(2k+1)\gamma_{\ell+1}} - \boldsymbol{b}_{k\gamma_\ell}\right\|^2) \exp[4\tau_k^+ + \|e_k\|] \tag{C.194}$$

$$+ C_5 (1 + 2^\ell) \, (\gamma_{\ell+1}^6 + \sup_{t \in [k\gamma_\ell, (k+1)\gamma_\ell]} \{\left\|\boldsymbol{b}_t - \boldsymbol{b}_{(2k+1)\gamma_{\ell+1}}\right\|^6\} \tag{C.195}$$

$$+ \left\|\boldsymbol{b}_{(2k+1)\gamma_\ell} - \boldsymbol{b}_{(k+1)\gamma_\ell}\right\|^6), \tag{C.196}$$

with $\tau_k^+ = \sup\{\|c_k\| + \|d_k\| \ : \ t \in [k\gamma_\ell, (k+1)\gamma_\ell]\}$, see (C.187). Therefore, using (C.129), (C.169) and (C.192), we get that for any $k \in \{0, \ldots, 2^\ell - 1\}$

$$U_{k+1} \leq (1 + \alpha_k) U_k + \beta_k. \tag{C.197}$$

Let $\{R_k\}_{k=-1}^{2^\ell}$ such that $R_{-1} = 0$ and for any $k \in \{0, \ldots, 2^\ell - 1\}$

$$R_{k+1} = (1 + \alpha_k) R_k + \beta_k. \tag{C.198}$$

Then, for any $k \in \{0, \ldots, 2^\ell - 1\}$, we have that $R_{2^\ell - 1} \geq R_k \geq U_k$. Therefore

$$\mathbb{E}[R_{2^\ell}] \geq \mathbb{E}\big[\sup\big\{U_k \ : \ k \in \{0, \ldots, 2^\ell\}\big\}\big] \geq \mathbb{E}\big[\sup\big\{d(\boldsymbol{X}_t^\ell, \boldsymbol{X}_t^{\ell+1})^2 \ : \ t \in [0, \gamma]\big\}\big]. \tag{C.199}$$

In addition, using that for any $k \in \{0, \ldots, 2^\ell - 1\}$, $\mathbb{E}[\alpha_k \mid \mathcal{F}_k] = \bar{\alpha}_k$ and $\mathbb{E}[\beta_k \mid \mathcal{F}_k] = \bar{\beta}_k$ are constant, where $\mathcal{F}_k = \sigma(\{\boldsymbol{b}_t \ : \ t \in [0, k\gamma_\ell]\})$. Therefore, we get that for any $k \in \{0, \ldots, 2^\ell - 1\}$

$$\mathbb{E}[R_{k+1}] = (1 + \bar{\alpha}_k) \, \mathbb{E}[R_k] + \bar{\beta}_k. \tag{C.200}$$

Therefore, using the discrete Grönwall lemma we get that for any $k \in \{0, \ldots, 2^\ell - 1\}$

$$\mathbb{E}[R_{2^\ell}] \leq \bar{\beta}_{2^\ell - 1} + \exp[\sum_{n=0}^{2^\ell - 1} \bar{\alpha}_n] \sum_{j=0}^{2^\ell - 1} \bar{\beta}_j \bar{\alpha}_j. \tag{C.201}$$

In addition, there exists $C_8 \geq 0$ such that for any $k \in \{0, \dots, 2^\ell\}$, $\bar{\alpha}_k \leq C_8 2^{-\ell}$ and $\bar{\beta}_k \leq C_8 \gamma^3 2^{-2\ell}$. Hence, there exists $C_9 \geq 0$ such that

$$\mathbb{E}[R_{2^\ell}] \leq C_9 \gamma^3 2^{-2\ell}, \tag{C.202}$$

which concludes the proof upon using (C.199). ■

PROPOSITION C.16. *Assume Assumption 3.12. Then, there exists $(\boldsymbol{X}_t)_{t\in[0,\gamma]}$ such that $\lim_{\ell\to+\infty} \sup\{d(\boldsymbol{X}_t^\ell, \boldsymbol{X}_t) \ : \ t \in [0,\gamma]\} = 0$ and $(\boldsymbol{X}_t)_{t\in[0,\gamma]}$ is a weak solution to $\mathrm{d}\boldsymbol{X}_t = b(t, \boldsymbol{X}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{b}_t^{\mathcal{M}}$.*

Proof. The proof is a straightforward application of Proposition C.15 and (Cheng et al., 2022, A.1 (Step 2 and Step 3), A.2). ■

PROPOSITION C.17. *Assume Assumption 3.12. Then, there exists $C \geq 0$ such that $\mathbb{E}\left[d(X_1^0, \boldsymbol{X}_\gamma)^2\right] \leq C\gamma^{3/2}$.*

Proof. Using Proposition C.15, there exists $C \geq 0$ such that for any $\ell \in \mathbb{N}$

$$\mathbb{E}\left[\sup_{t\in[0,\gamma]} d(\boldsymbol{X}_t^\ell, \boldsymbol{X}_t^{\ell+1})\right] \leq C\gamma^{3/2} 2^{-\ell}. \tag{C.203}$$

Therefore, combining this result and Proposition C.16 we get that for any $\ell \in \mathbb{N}$

$$\mathbb{E}\left[\sup_{t\in[0,\gamma]} d(\boldsymbol{X}_t^\ell, \boldsymbol{X}_t)\right] \leq 2C\gamma^{3/2}, \tag{C.204}$$

which concludes the proof. ■

Finally, we consider the two following processes $(X_k^1, X_k^2)_{k\in\mathbb{N}}$ such that for any $k \in \mathbb{N}$ and $i \in \{1, 2\}$

$$X_{k+1}^i = \exp_{X_k^i}[\gamma b(k\gamma, X_k^i) + \sqrt{\gamma} E_k^i Z_k], \tag{C.205}$$

where $\{Z_k\}_{k\in\mathbb{N}}$ is a family of independent Gaussian random variables with zero mean and identity covariance matrix, and for any $k \in \mathbb{N}$, E_k^1 is a frame for $\mathrm{T}_{X_k^1}\mathcal{M}$ and $E_k^2 = \Gamma_{X_k^1}^{X_k^2} E_k^1$.

PROPOSITION C.18. *Assume Assumption 3.12. Then, there exists $C \geq 0$ such that for any $k \in \mathbb{N}$*

$$\mathbb{E}\left[d(X_k^1, X_k^2)\right] \leq \exp[Ck\gamma]\, \mathbb{E}\left[d(X_0^1, X_0^2)\right]. \tag{C.206}$$

Proof. Let $k \in \mathbb{N}$. Using Lemma C.14, there exists $D \geq 0$ such that

$$d(X_{k+1}^1, X_{k+1}^2)^2 \leq (1 + D\kappa_k^2 \exp[4\kappa_k]) d(X_k^1, X_k^2)^2 \tag{C.207}$$

$$+ D\exp[4\kappa_k] \left\| \Gamma_{X_k^2}^{X_k^1} v_k - u_k \right\|^2 + 2\langle w'(0), \Gamma_{X_k^2}^{X_k^1} v_k - u_k \rangle, \tag{C.208}$$

with $w: [0,1] \to \mathcal{M}$ a minimizing geodesic between X_k^1 and X_k^2

$$\kappa_k = \|u_k\| + \|v_k\|, \tag{C.209}$$
$$u_k^1 = \gamma b(k\gamma, X_k^1), \tag{C.210}$$
$$v_k^1 = \gamma b(k\gamma, X_k^2), \tag{C.211}$$
$$u_k^2 = \sqrt{\gamma} Z_k E_k^1, \qquad v_k^2 = \sqrt{\gamma} Z_k E_k^2, \tag{C.212}$$
$$u_k = u_k^1 + u_k^2, \qquad v_k = v_k^1 + v_k^2. \tag{C.213}$$

We have that $\Gamma_{X_k^2}^{X_k^1} v_k^2 = v_k$ and

$$\left\| \Gamma_{X_k^2}^{X_k^1} v_k^1 - u_k^1 \right\| \leq \mathfrak{m} L_1 \gamma d(X_k^1, X_k^2). \tag{C.214}$$

In addition, $\|w'(0)\| \leq d(X_k^1, X_k^2)$. Therefore, we get that

$$d(X_{k+1}^1, X_{k+1}^2)^2 \leq (1 + D\kappa_k^2 \exp[4\kappa_k] + D\gamma^2 \exp[4\kappa_k] + 2\gamma) d(X_k^1, X_k^2)^2. \tag{C.215}$$

Hence, using that for any $t \geq 0$, $\sqrt{1+t} \leq 1 + t/2$, we have

$$d(X_{k+1}^1, X_{k+1}^2) \leq (1 + D\kappa_k^2 \exp[4\kappa_k] + D\gamma^2 \exp[4\kappa_k] + 2\gamma) d(X_k^1, X_k^2). \tag{C.216}$$

Therefore, we get that there exists $C \geq 0$ such that

$$\mathbb{E}\left[d(X_{k+1}^1, X_{k+1}^2)\right] \leq (1 + C\gamma)\, \mathbb{E}\left[d(X_k^1, X_k^2)\right], \tag{C.217}$$

which concludes the proof. ∎

C.5. Proof of the implicit score-matching loss on manifolds, proposition 3.7

The regularity conditions on $p_{t|s}(x_t|x_s)s_t(x_t)$ are

- $p_{t|s}(x_t|x_s)s_t(x_t)$ is a vector field in C_1 $\forall x_s$.
- $|p_{t|s}(x_t|x_s)s_t(x_t)| \in L_1$ $\forall x_s$.
- $\operatorname{div}(p_{t|s}(x_t|x_s)s_t(x_t)) \in L_1$ $\forall x_s$.

These conditions are not difficult to show. We can manually control s_t by our choice of score network, and $p_{t|s}(x_t|x_s)$ is controlled by choice of noising process. Under these conditions we can prove the statement.

Proof. Let $t \in [0,T]$ and $s_t \in \mathrm{C}^\infty(\mathcal{M})$. Using a divergence theorem for non-compact manifolds (see Gaffney, 1954, p.2), we have

$$\ell_{t|s}(s_t) = \int_{\mathcal{M}\times\mathcal{M}} \left\| \nabla \log p_{t|s}(x_t|x_s) \right\|^2 \mathrm{d}\mathbb{P}_{s,t}(x_s, x_t) + \int_{\mathcal{M}} \|s_t(x_t)\|^2 \mathrm{d}\mathbb{P}_t(x_t) \tag{C.218}$$
$$-2 \int_{\mathcal{M}\times\mathcal{M}} \langle \nabla \log p_{t|s}(x_t|x_s), s_t(x_t) \rangle_{\mathcal{M}} \mathrm{d}\mathbb{P}_{s,t}(x_s, x_t) \tag{C.219}$$

Looking at the last term

$$\int_{\mathcal{M}\times\mathcal{M}} \langle \nabla \log p_{t|s}(x_t|x_s), s_t(x_t)\rangle_{\mathcal{M}} \mathrm{d}\mathbb{P}_{s,t}(x_s, x_t) \tag{C.220}$$

$$= \int_{\mathcal{M}\times\mathcal{M}} \langle \nabla \log p_{t|s}(x_t|x_s), s_t(x_t)\rangle_{\mathcal{M}} p_{t|s}(x_t|x_s) p_s(x_s) \mathrm{d}(p_{\mathrm{ref}} \otimes p_{\mathrm{ref}})(x_s, x_t) \tag{C.221}$$

$$= \int_{\mathcal{M}} \left\{ \int_{\mathcal{M}} \langle \nabla p_{t|s}(x_t|x_s), s_t(x_t)\rangle_{\mathcal{M}} \mathrm{d}p_{\mathrm{ref}}(x_t) \right\} p_s(x_s) \mathrm{d}p_{\mathrm{ref}}(x_s) \tag{C.222}$$

$$= -\int_{\mathcal{M}} \left\{ \int_{\mathcal{M}} \mathrm{div}(s_t)(x_t) p_{t|s}(x_t|x_s) \mathrm{d}p_{\mathrm{ref}}(x_t) \right\} p_s(x_s) \mathrm{d}p_{\mathrm{ref}}(x_s) \quad \text{by the divergence theorem} \tag{C.223}$$

$$= -\int_{\mathcal{M}\times\mathcal{M}} \mathrm{div}(s_t)(x_t) \mathrm{d}\mathbb{P}_{s,t}(x_s, x_t) = -\int_{\mathcal{M}\times\mathcal{M}} \mathrm{div}(s_t)(x_t) \mathrm{d}\mathbb{P}_t(x_t) \tag{C.224}$$

Therefore

$$\ell_{t|s}(s_t) = \int_{\mathcal{M}\times\mathcal{M}} \left\| \nabla \log p_{t|s}(x_t|x_s) \right\|^2 \mathrm{d}\mathbb{P}_{s,t}(x_s, x_t) + \int_{\mathcal{M}} \|s_t(x_t)\|^2 \mathrm{d}\mathbb{P}_t(x_t) \tag{C.225}$$

$$+2 \int_{\mathcal{M}} \mathrm{div}(s_t)(x_t) \mathrm{d}\mathbb{P}_t(x_t), \tag{C.226}$$

which concludes the proof. ∎

C.6. Experimental details

In what follows we describe the experimental settings used to generate results introduced in section 3.4. The models and experiments have been implemented in Jax (Bradbury et al., 2018), using a modified version of the Riemannian geometry library Geomstats (Miolane et al., 2020).

Models Following Song et al. (2020b), the score-based generative models (SGMs) diffusion coefficient is parametrized as $g(t) = \sqrt{\beta(t)}$ with $\beta : \ t \mapsto \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot t$.

Architecture The architecture of the score network $\boldsymbol{s}_\theta$ is given by a multilayer perceptron with 5 hidden layers for the Earth and $SO(3)$ experiments, and 3 for the high-dimension experiments with 512 units each. We use sinusoidal activation functions. We decompose the output of the score network on the set of divergence free vector fields as per section 3.2.4.

Loss Where not specified, SGMs are trained with the sliced score matching (SSM) loss ℓ_t^{im}, relying on the Hutchinson estimator for computing the divergence with Rademacher noise described in section 3.2.4. We found that training with the denoising score matching (DSM) loss $\ell_{t|0}$ gave similar results. Regarding the weighting function, for DSM loss $\ell_{t|0}$ we use $\lambda_t = \mathrm{Var}[X_t|X_0]$ (where we rely on the closed-form standard deviation available in the Euclidean setting as a proxy for the compact manifold setting), while for the ISM/SSM losses ℓ_t^{im} we use $\lambda_t = g(t)^2 = \beta(t)$.

Optimization All models are trained by the stochastic optimizer Adam (Kingma, 2014) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch-size of 512 data-points. The learning rate is annealed with a linear ramp from 0 to 1000 and from then with a cosine schedule.

Likelihood evaluation and sample drawing We rely on the Dormand-Prince solver (Dormand and Prince, 1980), an adaptive Runge-Kutta 4(5) solver, with absolute and relative tolerance of $1e{-}5$ to compute approximate numerical solutions of any ODEs. For the rollouts of the SGM SDEs we use a Euler Maruyama predictor and no corrector. Unless stated we use 100 step rollouts.

Hardware Models are trained on a cluster with a mixture of GeForce RTX 1080, 1080 Ti and 2080 Ti GPU cards.

C.6.1. Sphere.

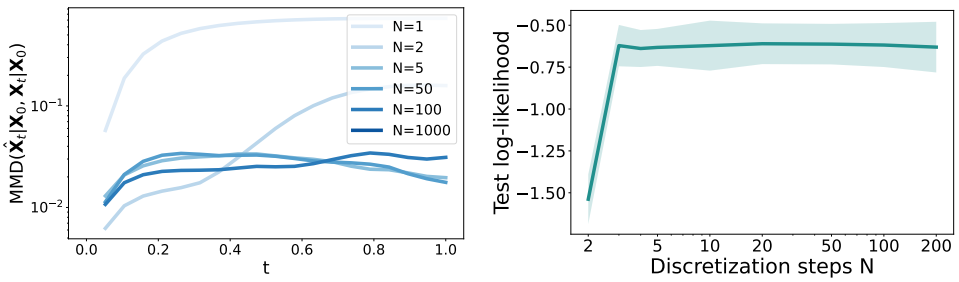
Data We randomly split the datasets intro training, validation and test datasets with $(0.8, 0.1, 0.1)$ proportions. In each case the earth is approximated as a perfect sphere.

Models The mixture of Kent distributions (Peel et al., 2001) were optimised using the EM algorithm and the number of components were selected from a grid search over the range 5, 10, 15, 20, 25, 30, 40, 50, 75, 100, based on validation set likelihood and 250 EM iterations. The number of components selected were: Volcano 25, Earthquake 50, Flood 100 and Fire 100.

For the stereographic SGM–which is a standard SGM with an Ornstein–Uhlenbeck process followed with the inverse stereographic projection–we found $\beta_{\text{min}} = 0.001$ and $\beta_{\text{max}} = 2$ to work best.

Optimization The score-based models are trained for $600k$ iterations for all datasets but 'Flood' where $300k$ performed best.

Approximate forward sampling Standard Euclidean SGMs rely on a Ornstein–Uhlenbeck (OU) forward process which can easily be simulated since $\boldsymbol{X}_t|\boldsymbol{X}_0$ is Gaussian. In contrast, for most manifolds one has to rely on an approximate sampling scheme—see section 3.2.6. First, we directly assess the quality of the approximate samples $\hat{\boldsymbol{X}}_t|\boldsymbol{X}_0$ obtained via geodesic random walk (GRW), against 'exact' samples $\boldsymbol{X}_t|\boldsymbol{X}_0$ which are obtained by using a high number of discretization steps ($N = 1000$). We report on figure C.1a the discrepancy between these distributions for different values of discretization steps N, as measured by maximum mean discrepancy (MMD) (Gretton et al., 2012). We see that from $N = 5$ the approximate samples are very closely distributed to the true samples. Then, in order to assess the impact of this approximation on the RSGMs' performance, we report on figure C.1b the log-likelihood when varying the number of discretization steps N. We similarly observe that apart from very small values of N, the models' performance is very robust to the approximation quality of the forward sampling samples.



(a) Maximum mean discrepancy (MMD) distance between *'exact'* (i.e. approximated with $N = 1000$ steps) $\boldsymbol{X}_t|\boldsymbol{X}_0$ and *approximate* $\hat{\boldsymbol{X}}_t|\boldsymbol{X}_0$ at for every $t \in [0, 1]$.

(b) Test log-likelihood of trained RSGMs on the Flood dataset while varying the number of discretization steps N when simulating forward sampling $\boldsymbol{X}_t|\boldsymbol{X}_0$.

Figure C.1. Ablation study on the impact of the forward sampling approximation quality on $\mathbb{S}^2$.

DSM loss $\ell_{t|0}$ On figure C.2, we show how the test log-likelihood varies with respect to the two hyparameters of the DSM loss, by training RSGMs over a grid of values for τ and J on the Flood dataset. We can see that the Varadhan approximation by itself ($\tau = 1$) yields descent performance, although a wise combination of Varadhan approximation with a truncation of the heat kernel can give even better results. The performance is relatively robust to the choice of such hyperparameters as long as τ and J are high enough.

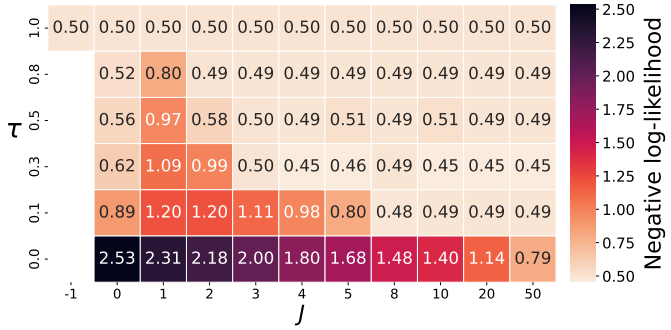


Figure C.2. Ablation study on the denoising score matching (DSM) loss $\ell_{t|0}$ when combining the heat kernel truncation and the Varadhan approximation: $\nabla_{x_t} \log p_{t|0}(x_t|x_0) \approx \mathbb{1}(t \leq \tau) \exp_{x_t}^{-1}(x_0) + \mathbb{1}(t > \tau) S_{J,t}(x_0, x_t)$.

C.6.2. Torus.

Data The synthetic data trained on consists of a wrapped Gaussian distribution on T^n with uniformly chosen random mean and standard deviation of 0.2. Such a distribution is defined by taking the density of a Normal distribution in the tangent space of the manifold at the mean and passing it through the exponential map at the mean.

Architecture To parametrize the vector field on $\mathbb{T}^n$ we use a single filed per dimension pointing in a consistent direction around the i^{th} component in the product, with unit norm.

Models All models were trained with the same 3 layer, 512 units per layer MLP across different dimension sizes.

Optimization The models are optimized for $50k$ iterations. The RSGM models are trained with both the implicit score-matching loss and the sliced score-matching loss.

C.6.3. Special Orthogonal group. Applications of orthogonal constraints span various fields, such as protein docking with ligands binding pose prediction (Ganea et al., 2022), robotics and Computer vision with rigid body transformation estimation (Barfoot et al., 2011; Prokudin et al., 2018), and medical imaging for data alignment (Hou et al., 2018).

Data We consider the synthetic dataset consisting of samples in $\mathrm{SO}_3(\mathbb{R}^d)$[1] from the mixture distribution with density $p(\mathrm{Q}) = \frac{1}{K}\sum_{k=1}^{K} \mathrm{N}^W(\mathrm{Q}|\mathrm{Q}_k, \sigma_k^2)$ with $K \in \mathbb{N}$, where for any $k \in \{1, \dots, K\}$, we have that $\mathrm{Q} = \mathrm{Q}_k \exp_{\mathrm{I}}[\sigma_k \hat{z}]$ with $z \sim \mathrm{N}(0, \mathbf{I}_{\mathbb{R}^3})$ satisfies $Q \sim \mathrm{N}^W(\mathrm{Q}_k, \sigma_k)$ and $(\cdot)^\wedge : \mathbb{R}^3 \to \mathfrak{so}(3)$. For any $k \in \{1, \dots, K\}$, we set $\mathrm{Q}_k \sim \mu$ where μ is the uniform distribution on $\mathrm{SO}_3(\mathbb{R})$ and $\sigma_k^2 \sim \mathrm{IG}(\alpha = 100, \beta = 1)$, where IG is the inverse Gaussian distribution. We choose $K = 32$ mixture components. We showcase a conditional sampling extension of our model—see section 3.2.10 for more details— by targeting individual mixture components $p(Q|k)$. Our model is trained using the $\ell_{t|0}$ (DSM) loss along with the Varadhan asymptotic approximation, see (3.78).

Architecture To parametrize the vector field, we rely on the basis of the Lie group, $\mathfrak{so}(n) = \{\mathrm{A} \in \mathrm{M}_d(\mathbb{R}) \,:\, \mathrm{A}^\top = -\mathrm{A}\}$ given by $\mathrm{E}_{ij} = \mathrm{U}_{ij} - \mathrm{U}_{ji}$ for $i, j \in \{1, \dots, d\}$ with $i < j$ and $\mathrm{U}_{ij} = (\delta_{ij}(k, \ell))_{1 \le k, \ell \le d}$, which induces a basis on the tangent spaces $\mathrm{T}_{\mathrm{Q}}\mathrm{SO}_d$ for any $\mathrm{Q} \in \mathrm{SO}_d(\mathbb{R})$ given by $\{\mathrm{QE}_{ij}\}_{1 \le i < j \le d}$. This is the divergence-free vector field approach described in section 3.2.4.

Models We compare our proposed approach against Moser flows (Rozen et al., 2021) and a wrapped-exponential baseline (Falorsi et al., 2019) defined as the pushforward along the transformation $\mathbb{R}^3 \xrightarrow{F_\theta^{-1}} \mathbb{R}^3 \xrightarrow{g} \mathbb{R}^3 \xrightarrow{\wedge} \mathfrak{so}(3) \xrightarrow{\exp} \mathrm{SO}_3(\mathbb{R})$ with F_θ^{-1} denoting the approximate time-reversed diffusion, g denoting the radial operator defined by $g : x \mapsto 2\pi \tanh(\|x\|)x/\|x\|$, $(\cdot)^\wedge : \mathbb{R}^3 \to \mathfrak{so}(n)$ the isomorphism given by the basis on $\mathfrak{so}(3)$ and exp the matrix exponential. The radial g operator's constant 2π is chosen as the injectivity radius of the group so that the transformation $\tanh \circ \wedge \circ \exp$ is injective (the set of elements with no preimage is then only the cut locus which is known to have measure zero). Henceforth, this wrapped-exponential transformation cannot be bijective, it is either injective *or* surjective depending on the choice of radius in the radial operator g.

[1] This manifold is 3-dimensional.

Optimization Models are trained for $100k$ iterations. The Riemannian SGM is trained with the Varhadan approximation of the denoising score-matching loss (DSM) section 3.2.4, and the wrapped-exponential model relies on the exact DSM loss. After a first hyperparameter exploration, a grid search is performed over `learning_rate` $\in [2e-5, 4e-5]$, for SGMs over $\beta_f \in [0.5, 1, 2, 4, 6, 8, 10]$ and for Moser flows over $K \in [1000, 10000]$ and $\lambda_{\min} \in [1, 10, 100]$.

D | Score-based modelling on constrained domains

Organization of the appendix

In appendix D.1 we remind the expression of the Brownian motion in local coordinates. Details about the geodesic Brownian motion are given in Appendix D.2. Background on the Skorokhod problem is given in Appendix D.3. In appendix D.4 we derive the implicit score matching loss. We give details about the likelihood evaluation in Appendix D.5. Then in appendix D.6 we prove the time-reversal formula of reflected Brownian motion.

Additional results, training and miscellaneous experimental details are reported in Appendix D.7.

D.1. Brownian motion in local coordinates

We consider a smooth function $f \in C^\infty(\mathcal{M})$. The Laplace-Beltrami operator $\Delta_\mathcal{M}$ is given by $\Delta_\mathcal{M}(f) = \mathrm{div}(\mathrm{grad}(f)))$. In local coordinates we have

$$\mathrm{div}(X) = (\det(\mathfrak{g})^{-1/2}) \sum_{i=1}^{d} \partial_i(\det(\mathfrak{g})^{1/2} X_i), \tag{D.1}$$

as well as

$$\mathrm{grad}(f) = \mathfrak{g}^{-1}\nabla f. \tag{D.2}$$

Therefore, the Laplace-Beltrami operator is given by

$$\Delta_\mathcal{M}(f) = \sum_{i,j=1}^{d} \mathfrak{g}^{-1}_{i,j}\partial_{i,j} f + (\det(\mathfrak{g})^{-1/2}) \sum_{i,j=1}^{d} \partial_i(\det(\mathfrak{g})^{1/2}\mathfrak{g}^{-1}_{i,j})\partial_j f. \tag{D.3}$$

Therefore, in local coordinates the infinitesimal generator associated with the Laplace-Beltrami operator is given by

$$\mathcal{A}(f) = \sum_{i,j=1}^{d} \mathfrak{g}^{-1}_{i,j}\partial_{i,j} f + \langle b^i, \nabla f\rangle, \tag{D.4}$$

with

$$b^i = (\det(\mathfrak{g})^{-1/2}) \sum_{j=1}^{d} \partial_j(\det(\mathfrak{g})^{1/2}\mathfrak{g}^{-1}_{i,j}). \tag{D.5}$$

Therefore, the dual operator associated with $\mathcal{A}$ is given by

$$\mathcal{A}^\star(f) = \sum_{i,j=1}^{d} \partial_{i,j}(\mathfrak{g}^{-1}_{i,j} f) - \sum_{i=1}^{d} \partial_i(b^i f). \tag{D.6}$$

Note that by letting $f = \det(\mathfrak{g})^{1/2}$ we get that $\mathcal{A}^\star(f) = 0$ and therefore we recover that $p \propto \det(\mathfrak{g})$ is the invariant distribution of the Brownian motion.

Langevin dynamics on $\mathcal{M}$. We know that the Brownian motion targets $\det(\mathfrak{g})^{1/2}$. Therefore in order to correct and sample from the uniform distribution we consider the Langevin dynamics

$$\mathrm{d}\boldsymbol{X}_t = -\mathrm{grad}\log(\det(\mathfrak{g})^{1/2})(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}^{\mathcal{M}}_t. \tag{D.7}$$

Note that in the previous equation grad and $\boldsymbol{B}^{\mathcal{M}}_t$ are defined with respect to. the metric of the manifold. In local coordinates we have

$$\mathrm{d}\boldsymbol{X}_t = \{b - \mathrm{grad}\log(\det(\mathfrak{g})^{1/2})\}(\boldsymbol{X}_t)\mathrm{d}t + \sqrt{2}\mathfrak{g}(\boldsymbol{X}_t)^{-1/2}\mathrm{d}\boldsymbol{B}_t. \tag{D.8}$$

where $b = \{b^i\}_{i=1}^d$ is given in (D.5). In addition, we have

$$\mathrm{grad}\log(\det(\mathfrak{g})^{1/2}) = \det(\mathfrak{g})^{-1/2}\mathfrak{g}^{-1}\nabla\det(\mathfrak{g}). \tag{D.9}$$

Using (D.5) we have

$$b^i = (\det(\mathfrak{g})^{-1/2})\sum_{j=1}^{d}\partial_j(\det(\mathfrak{g})^{1/2}\mathfrak{g}^{-1}_{i,j}) = \sum_{j=1}^{d}\partial_j\mathfrak{g}^{-1}_{i,j} + \mathrm{grad}\log(\det(\mathfrak{g})^{1/2})_i \tag{D.10}$$

This can also be rewritten as

$$\mathrm{div}_{\mathcal{M}}(\mathfrak{g}^{-1}) = \mu + \mathrm{grad}\log(\det(\mathfrak{g})^{1/2}), \tag{D.11}$$

with

$$\mu_i = \sum_{j=1}^{d}\partial_j\mathfrak{g}^{-1}_{i,j}. \tag{D.12}$$

Combining this result and (D.9) we get that (D.8) can be rewritten as

$$\mathrm{d}\boldsymbol{X}_t = \mu(\boldsymbol{X}_t) + \sqrt{2}\mathrm{d}\boldsymbol{B}_t. \tag{D.13}$$

Note that (up to a factor 2) this is the same SDE as the one considered in Lee and Vempala (2017).

D.2. Geodesic Brownian Motion

In this section, we provide some details on the geodesics Brownian motion introduced in Section 4.2. In the rest of the section, we make the following assumption.

Assumption D.1. *$\overline{\mathcal{M}} \subset \mathbb{R}^d$ is compact and $\mathfrak{g}^{-1} : \mathcal{M} \to \mathcal{S}^{++}_d$ can be $\mathrm{C}^\infty(\mathbb{R}^d, \mathbb{R}^{d\times d})$.*

First, we start by showing that the process $(\boldsymbol{X}_t)_{t\geq 0}$ defined in (4.9) exists and that we have for any $t \geq 0$, $\boldsymbol{X}_t \in \mathcal{M}$. We recall that $\mathfrak{g} = \nabla^2\phi$ and $\lim_{x\to\partial\mathcal{M}}\Phi(x) = +\infty$.

PROPOSITION D.2. *Assume* Assumption D.1. *For any $x_0 \in \mathcal{M}$, there exists a unique strong solution to* (4.9) *denoted $(\boldsymbol{X}_t)_{t\geq 0}$. In addition, we have that for any $t \geq 0$, $\boldsymbol{X}_t \in \mathcal{M}$ almost surely. More precisely, we have $\mathbb{E}[\phi(\boldsymbol{X}_t)] \leq \phi(x_0) + t$.*

Proof. A unique strong solution $(\boldsymbol{X}_t)_{t\geq 0}$ of (4.9) with starting point $x_0 \in \mathcal{M}$ exists since the coefficients are smooth, see (Ikeda and Watanabe, 2014, Theorem 3.1, p.165). For any $A \geq 0$, we define $\tau_A = \inf\{t \geq 0 \ : \ \Phi(\boldsymbol{X}_t) \geq A\}$. Note that for any $t \in \{0, \tau_A\}$, $\Phi(\boldsymbol{X}_t) \in \mathcal{M}$. Using Itô formula, we have for any $t \geq 0$

$$\mathbb{E}\left[\Phi(\boldsymbol{X}_{t\wedge\tau_A})\right] = \Phi(x_0) + \mathbb{E}\left[\int_0^{t\wedge\tau_A} \mathrm{Tr}(\mathfrak{g}^{-1}(\boldsymbol{X}_s)\nabla^2\Phi(\boldsymbol{X}_s))\mathrm{d}s\right] = \Phi(x_0) + \mathbb{E}[t \wedge \tau_A]. \tag{D.14}$$

Using Fatou's lemma, and letting $A \to +\infty$, we conclude the proof. ∎

In the next result, we show that the uniform distribution is the *unique* invariant probability distribution for $(\boldsymbol{X}_t)_{t\geq 0}$ and that $(\boldsymbol{X}_t)_{t\geq 0}$ converges to this invariant distribution. We refer to (Meyn and Tweedie, 1993, Section 2, p.490) for a definition of irreducibility. We recall that the total variation of a finite (not necessarily non-negative) measure μ over $\mathbb{R}^d$ is given by $\|\mu\|_{\mathrm{TV}} = \sup\{\mu(\mathsf{A}) \ : \ \mathsf{A} \in \mathcal{B}(\mathbb{R}^d)\}$.

PROPOSITION D.3. *Assume* Assumption D.1. *$(\boldsymbol{X}_t)_{t\geq 0}$ is π-irreducible, the uniform distribution over $\mathcal{M}$ is the only invariant probability distribution and $\lim_{t\to+\infty} \|\mathrm{P}_t - \pi\|_{\mathrm{TV}} = 0$, where P_t is the distribution of $\boldsymbol{X}_t$ for any $t \geq 0$ and π is the uniform distribution over $\mathcal{M}$.*

Proof. Since $x \mapsto \mathrm{div}(\mathfrak{g}^{-1})(x)$ and $x \mapsto \mathfrak{g}^{-1}$ are smooth and $\mathfrak{g}^{-1}(x)$ is positive definite for any $x \in \mathcal{M}$, we have that $(\boldsymbol{X}_t)_{t\geq 0}$ is π-irreducible, extending (Bhattacharya, 1978, Lemma 1.4) to $\mathcal{M}$ and using (Meyn and Tweedie, 1993, Proposition 2.1). In addition, $(\boldsymbol{X}_t)_{t\geq 0}$ is T-Feller using (Meyn and Tweedie, 1993, Proposition 3.3). Combining these results and the fact that $\mathcal{M}$ is bound, we get that $(\boldsymbol{X}_t)_{t\geq 0}$ is positive Harris recurrent (Meyn and Tweedie, 1993, Theorem 3.2). The uniform distribution π is an invariant distribution for (4.9). Since $(\boldsymbol{X}_t)_{t\geq 0}$ is π-irreducible, we get that this invariant measure is unique. Hence, we conclude using (Meyn and Tweedie, 1993, Theorem 6.1). ∎

Note that the convergence result in total variation could be improved. In particular, quantitative geometric results could be derived. We finish this section, by applying results from the Malliavin calculus to show that for any $t > 0$, $\boldsymbol{X}_t$ admits a density with respect to. the Lebesgue measure.

PROPOSITION D.4. *Assume* Assumption D.1. *Then, for any $t \geq 0$, $\boldsymbol{X}_t$ admits a smooth density p_t with respect to. the Lebesgue measure.*

Proof. This is a direct consequence of (Nualart, 2006, Theorem 2.3.3). ∎

D.3. Reflected Brownian Motion and Skorokhod problems

In this section, we provide the basic definitions and results to derive the time-reversal of the reflected Brownian motion in Appendix D.6. We follow closely the presentation of (Lions and Sznitman, 1984) and (Burdzy et al., 2004). We first define the *Skorokhod problem* for deterministic problems. We consider $\mathcal{M}$ to be a smooth open bounded domain. We recall that the normal vector n is defined on $\partial\mathcal{M}$ and we set $n(x) = 0$ for any $x \notin \partial\mathcal{M}$.

Before giving the definition of the *Skorokhod problem*, we recall what is the space of functions of *bounded variations*.

Definition D.5. *Let $a, b \in (-\infty, +\infty)$ and $f :\in \mathrm{C}(\{a, b\}, \mathbb{R})$. We define the* total variation *of f as*

$$\mathrm{V}_{a,b}(f) = \sup\left\{\sum_{i=0}^{n-1} \|f(x_{i+1}) - f(x_i)\| \ : \ (x_i)_{i=0}^{n-1},\ a = x_0 \leq \cdots \leq x_n = b,\ n \in \mathbb{N}\right\}. \tag{D.15}$$

f has bounded variations over $\{a, b\}$ if $\mathrm{V}_{a,b}(f) < +\infty$. Let $f \in \mathrm{C}([0, +\infty), \mathbb{R})$. f has bounded variations over $[0, +\infty)$ if for any $b > 0$, f has bounded variations over $\{0, b\}$.

The notion of bounded variation is a relaxation of the differentiability requirement. In particular, if $f \in \mathrm{C}^1(\{a, b\}, \mathbb{R})$, we have $\mathrm{V}_{a,b}(f) = \int_a^b \|f'(t)\| \mathrm{d}t$. In the definition of the *Skorokhod problem*, we will see that this relaxation is necessary, even in the deterministic setting.

For any function of bounded variation $f \in \mathrm{C}(\{a, b\}, \mathbb{R})$ on $\{a, b\}$, we define $|f| : \{a, b\} \to [0, +\infty)$ given for any $t \in \{a, b\}$ by $|f|_t = \mathrm{V}_{a,t}(f)$. Note that $|f|$ is non-decreasing and right-continuous. Therefore, we can define the measure $\mu_{|f|}$ on $\{a, b\}$, given for any $s, t \in \{a, b\}$ with $t \geq s$ by $\mu_{|f|}([s, t]) = |f|(t) - |f|(s)$. In particular, for any $\varphi : \{a, b\} \to \mathbb{R}_+$, we define

$$\int_a^b \varphi(t) \mathrm{d}|f|_t = \int_a^b \varphi(t) \mathrm{d}\mu_{|f|}(t). \tag{D.16}$$

In addition, f can be decomposed in a difference of two non-decreasing processes right continuous processes g_1, g_2, where for any $t \in \{a, b\}$, $f(t) = g_1(t) - g_2(t)$, $g_1(t) = |f|_t$ and $g_2(t) = |f|_t - f(t)$. Hence, for every φ bounded on $\{a, b\}$, we can define

$$\int_a^b \varphi(t) \mathrm{d}f(t) = \int_a^b \varphi(t) \mathrm{d}g_1(t) - \int_a^b \varphi(t) \mathrm{d}g_2(t). \tag{D.17}$$

Note that these definitions can be extended to the setting where $f : \{a, b\} \to \mathbb{R}^d$.

We begin with the following result, see (Lions and Sznitman, 1984).

Theorem D.6. *Let $(x_t)_{t \geq 0} \in \mathrm{C}([0, +\infty), \mathbb{R})$. Then, there exists a unique couple $(\bar{x}_t, k_t)_{t \geq 0}$ such that*

(a) $(k_t)_{t \geq 0}$ has bounded variation over $[0, +\infty)$.

(b) $(\bar{x}_t)_{t\geq 0} \in \mathrm{C}([0,+\infty), \overline{\mathcal{M}})$.

(c) For any $t \geq 0$, $x_t + k_t = \bar{x}_t$.

(d) For any $t \geq 0$, $|k|_t = \int_0^t \mathbf{1}_{\bar{x}_s \in \partial\mathcal{M}}(\bar{x}_s)\mathrm{d}|k|_s$ *and* $k_t = \int_0^t n(\bar{x}_s)\mathrm{d}|k|_s$.

Let us discuss Theorem D.6. First, Theorem D.6-(c) states the original (unconstrained) process $(x_t)_{t\geq 0}$ can be decomposed into a constrained version $(\bar{x}_t)_{t\geq 0}$ and a bounded variation process $(k_t)_{t\geq 0}$. The process $(|k|_t)_{t\geq 0}$ counts the number of times the constrained process $(\bar{x}_t)_{t\geq 0}$ hits the boundary. More formally, we have $|k|_t = \int_0^t \mathbf{1}_{x\in\partial\mathcal{M}}(\bar{x}_s)\mathrm{d}|k|_s$. When, we hit the boundary, we reflect the process. This condition is expressed in $k_t = \int_0^t n(\bar{x}_s)\mathrm{d}|k|_s$.

We now consider the extension to stochastic processes. We are given $(\boldsymbol{X}_t)_{t\geq 0}$ such that

$$\mathrm{d}\boldsymbol{X}_t = b(\boldsymbol{X}_t)\mathrm{d}t + \sigma(t)\mathrm{d}\boldsymbol{B}_t, \tag{D.18}$$

where $(\boldsymbol{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion. We also assume that b and σ are Lipschitz which implies the existence and strong uniqueness of $(\boldsymbol{X}_t)_{t\geq 0}$. We have the following result (Lions and Sznitman, 1984).

THEOREM D.7. *There exists a unique process* $(\bar{\boldsymbol{X}}_t, \boldsymbol{k}_t)_{t\geq 0}$ *such that*

(a) $(\boldsymbol{k}_t)_{t\geq 0}$ *has bounded variation over* $[0,+\infty)$ *almost surely.*

(b) $(\bar{\boldsymbol{X}}_t)_{t\geq 0} \in \mathrm{C}([0,+\infty), \overline{\mathcal{M}})$.

(c) For any $t \geq 0$, $\bar{\boldsymbol{X}}_t = \bar{\boldsymbol{X}}_0 + \int_0^t b(\bar{\boldsymbol{X}}_s)\mathrm{d}s + \int_0^t \sigma(\bar{\boldsymbol{X}}_s)\mathrm{d}\boldsymbol{B}_s - \boldsymbol{k}_t$.

(d) For any $t \geq 0$, $|\boldsymbol{k}|_t = \int_0^t \mathbf{1}_{\bar{x}_s\in\partial\mathcal{M}}(\bar{x}_s)\mathrm{d}|\boldsymbol{k}|_s$ *and* $\boldsymbol{k}_t = \int_0^t n(\bar{x}_s)\mathrm{d}|\boldsymbol{k}|_s$.

The process $(\boldsymbol{X}_t)_{t\geq 0}$ is almost surely continuous, so we could apply the previous theorem almost surely for all the realizations of the process,. However, this does not tell us if the obtained solutions $(\bar{\boldsymbol{X}}_t, \boldsymbol{k}_t)_{t\geq 0}$ form themselves a process. The main difference with Theorem D.6 is in Theorem D.7-(c) which differs from Theorem D.7-(c). Note that in the case where $b = 0$ and $\sigma = \mathbf{I}$ we recover Theorem D.7-(c). This is not true in the general case. However, it can be seen that for any realization of the process $(\bar{\boldsymbol{X}}_t)_{t\geq 0}$, we have that $(\bar{\boldsymbol{X}}_t, \boldsymbol{k}_t)_{t\geq 0}$ is solution of the *deterministic* Skorokhod problem by letting $x_t = \bar{\boldsymbol{X}}_0 + \int_0^t b(\bar{\boldsymbol{X}}_s)\mathrm{d}s + \int_0^t \sigma(\bar{\boldsymbol{X}}_s)\mathrm{d}\boldsymbol{B}_s$. The backward and forward Kolmogorov equations can be found in (Burdzy et al., 2004). Note that the presence of the process $(\boldsymbol{k}_t)_{t\geq 0}$ incurs notable complications compared to unconstrained processes. In particular, there is no martingale problem associated with weak solutions of reflected SDEs but only sub-martingale problems, see (Kang and Ramanan, 2017) for instance.

D.4. Implicit Score Matching Loss

D.4.1. Proof of ISM. Using the divergence theorem, we have

$$(1/2)\int_{\mathcal{M}} \|\mathbf{s}_{t,\theta}(x) - \nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x) \tag{D.19}$$

$$= (1/2)\int_{\mathcal{M}} \|\mathbf{s}_{t,\theta}(x)\|^2 p_t(x)\mathrm{d}\mu(x) - \int_{\mathcal{M}} \langle \mathbf{s}_{t,\theta}(x), \nabla \log p_t(x)\rangle p_t(x)\mathrm{d}\mu(x) \tag{D.20}$$

$$+ (1/2)\int_{\mathcal{M}} \|\nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x) \tag{D.21}$$

$$= (1/2)\int_{\mathcal{M}} \|\mathbf{s}_{t,\theta}(x)\|^2 p_t(x)\mathrm{d}\mu(x) - \int_{\partial\mathcal{M}} \langle \mathbf{s}_{t,\theta}(x), \mathbf{n}\rangle p_t(x)\mathrm{d}\nu(x) \tag{D.22}$$

$$+ \int_{\mathcal{M}} \mathrm{div}(\mathbf{s}_{t,\theta})(x) p_t(x)\mathrm{d}\mu(x) + (1/2)\int_{\mathcal{M}} \|\nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x). \tag{D.23}$$

Using that $\left\langle \mathbf{s}_{t,\theta}(x), \boldsymbol{n}(x)\right\rangle = 0$ for all $x \in \partial\mathcal{M}$, we get that

$$(1/2)\int_{\mathcal{M}} \|\mathbf{s}_{t,\theta}(x) - \nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x) \tag{D.24}$$

$$= (1/2)\int_{\mathcal{M}} \|\mathbf{s}_{t,\theta}(x)\|^2 p_t(x)\mathrm{d}\mu(x) \tag{D.25}$$

$$+ \int_{\mathcal{M}} \mathrm{div}(\mathbf{s}_{t,\theta})(x) p_t(x)\mathrm{d}\mu(x) + (1/2)\int_{\mathcal{M}} \|\nabla \log p_t(x)\|^2 p_t(x)\mathrm{d}\mu(x), \tag{D.26}$$

which concludes the proof.

D.4.2. Importance of scaling function. As discussed in Section 4.4, we include a monotone scaling function h which is zero close to the boundary to ensure the relevant conditions are met for the score matching loss and the boundary conditions. This may seem like a technical detail, but is of significant practical importance. Upon removal of the scaling function, we observe that the learned score functions behave strangely around the boundary in the reverse process, leading to samples that do not match the forward process. The problems are apparent when comparing the top three plots of Fig. D.1 and Fig. D.2. Interestingly, we found that these issues early on in the sampling are smoothed out by the end of the reverse process, but still lead to a failure to recover the target density.

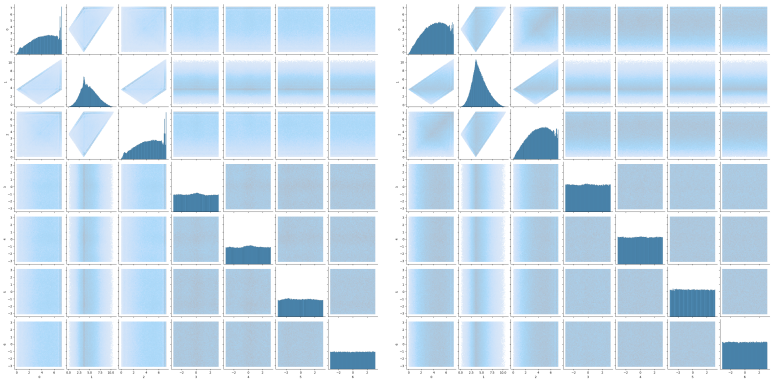


Figure D.1. Reverse process samples for the cyclic peptide dataset from Section 4.6.3 at $t = 1.0, 0.9$ (left and right respectively) trained without the scaling function.

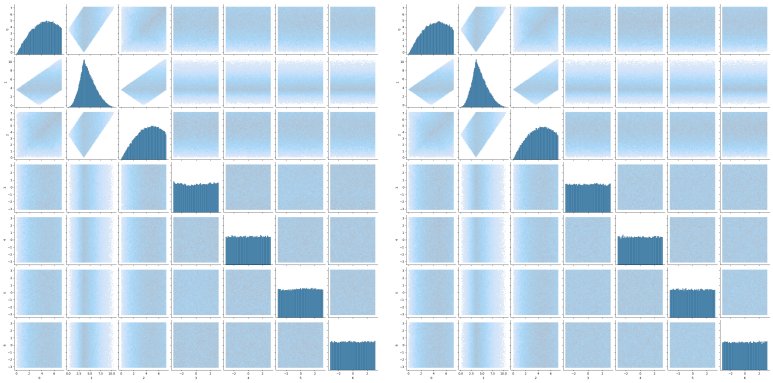


Figure D.2. Reverse process samples for the cyclic peptide dataset from Section 4.6.3 at $t = 1.0, 0.9$ (left and right respectively) trained with the scaling function.

D.5. Proof of proposition 4.5

Proof. Since the distributions of $(\bar{\mathbf{B}}_t)_{t\in\{0,T\}}$ and $(\mathbf{X}_t)_{t\in\{0,T\}}$ satisfy the same Fokker-Planck equation whenever these processes are well-defined. Therefore, we first show that the process $(\mathbf{X}_t)_{t\in\{0,T\}}$ is well-defined and stay in $\mathcal{M}$ at all times. Using (Burdzy et al., 2004, Theorem 2.2), we have that $\partial p_t(x) = \frac{1}{2}\mathrm{div}(\nabla \log p_t)(x)$, for any $t > 0$ and $x \in \mathcal{M}$. Next, we define $\mathrm{d}\boldsymbol{X}_t = \frac{1}{2}\nabla \log p_t(\boldsymbol{X}_t)\mathrm{d}t$. Note that $(\boldsymbol{X}_t)_{t\geq 0}$ is defined up to an explosion time T_∞ after which we fix $\boldsymbol{X}_t = \infty$. Denote T_0 the first time such that $\boldsymbol{X}_t \in \partial\mathcal{M}$. Note that since p_0 is supported on $\mathcal{M}$ we have $T_0 > 0$. We denote $(\boldsymbol{Y}_t)_{t\in\{0,T_0\}} = (\boldsymbol{X}_{T_0-t})_{t\in\{0,T_0\}}$. We have that for any $t \in \{0, T_0\}$, $\mathrm{d}\boldsymbol{Y}_t = -\frac{1}{2}\nabla \log p_{T_0-t}(\boldsymbol{Y}_t)\mathrm{d}t$. Since $(t, x) \mapsto \nabla \log p_t(x)$ is smooth on $[0, +\infty) \times \partial\mathcal{M}$, we get that for any $t \in \{0, T_0\}$, $\boldsymbol{Y}_t \in \partial\mathcal{M}$. In particular, we have that $\boldsymbol{Y}_{T_0/2} = \boldsymbol{X}_{T_0/2} \in \partial\mathcal{M}$ which is absurd. Therefore $T_0 = +\infty$ (which also implies that $T_\infty = +\infty$). Hence, $(\boldsymbol{X}_t)_{t\geq 0}$ is a flow on $\mathcal{M}$ and therefore for any $t \geq 0$, the density q_t of $\boldsymbol{X}_t$ is smooth and satisfies $\partial_t q_t(x) = -\frac{1}{2}\mathrm{div}(q_t \nabla \log p_t)(x)$. We conclude using the uniqueness of the solutions to the transport equation for smooth initialisation and coefficients on $\mathbb{R}^d$. ∎

D.6. Time-reversal for reflected Brownian motion

We start with the following definition.

Definition D.8. *Let $\mathcal{M} \subset \mathbb{R}^d$ be an open set. $\mathcal{M}$ has a smooth boundary if for any $x \in \partial\mathcal{M}$, there exists $\mathrm{U} \subset \mathbb{R}^d$ open and $f \in \mathrm{C}^\infty(\mathrm{U}, \mathbb{R})$ such that $x \in \mathrm{U}$ and (a) $\mathrm{cl}(\mathcal{M}) \cap \mathrm{U} = \{x \in \mathrm{U} : f(x) \leq 0\}$, (b) $\nabla f(x) \neq 0$ for any $x \in \mathrm{U}$ where $\mathrm{cl}(\mathcal{M})$ is the closure of $\mathcal{M}$.*

We will make the use of the following lemma which is a straightforward extension of Burdzy et al. (2004, Theorem 2.6). The surface measure is defined in (Lee, 2006, Proposition 2.43). Under mild regularity assumptions, it corresponds to the Hausdorff measure of $\partial\mathcal{M}$, see (Evans and Gariepy, 2015).

Lemma D.9. *Let u such that $s \mapsto u(s,x) \in \mathrm{C}^1((0,T), \mathbb{R})$, for any $s \in (0,T)$, $x \mapsto u(s,x) \in \mathrm{C}^2(\mathcal{M}, \mathbb{R})$ and $u \in \mathrm{C}^1(\mathrm{cl}(\mathcal{M}), \mathbb{R})$. Then for any $T \geq 0$, $s, t \in \{0, T\}$, we have*

$$\mathbb{E}\left[\int_s^t u(w, \bar{\boldsymbol{B}}_w) \mathrm{d}\boldsymbol{k}_w\right] = \tfrac{1}{2} \int_s^t \int_{\partial\mathcal{M}} u(x) p_w(x) \mathrm{d}\sigma(x) \mathrm{d}w. \tag{D.27}$$

Note that we recover Burdzy et al. (2004, Theorem 2.6) if we set $u = 1$. We also emphasize that the result of Burdzy et al. (2004, Theorem 2.6) is stronger than Lemma D.9 as it holds not only in expectation but also in L^2 and almost surely.

We are now ready to prove Theorem 4.2. We follow the approach of (Petit, 1997) which itself is based on an extension of (Haussmann and Pardoux, 1986). We refer to Cattiaux et al. (2023) for recent entropic approaches of time-reversal. Recall that $(\bar{\boldsymbol{B}}_t, \boldsymbol{k}_t)_{t\geq 0}$ is a solution to the *Skorokhod problem* (Skorokhod, 1961) if $(\boldsymbol{k}_t)_{t\geq 0}$ a bounded variation process and $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ a continuous adapted process such that for any $t \geq 0$, $\boldsymbol{B}_t = \bar{\boldsymbol{B}}_t + \boldsymbol{k}_t \in \mathcal{M}$, $(\bar{\boldsymbol{B}}_t)_{t\geq 0}$ and

$$|\boldsymbol{k}|_t = \int_0^t \mathbf{1}_{\bar{\boldsymbol{B}}_s \in \partial\mathcal{M}} \mathrm{d}|\boldsymbol{k}|_s, \quad \boldsymbol{k}_t = \int_0^t \boldsymbol{n}(\bar{\boldsymbol{B}}_s) \mathrm{d}|\boldsymbol{k}|_s, \tag{D.28}$$

In what follows, we define $(\boldsymbol{Y}_t)_{t\in\{0,T\}}$ such that for any $t \in \{0, T\}$, $\boldsymbol{Y}_t = \bar{\boldsymbol{B}}_{T-t}$. Let us consider the process $(\tilde{\boldsymbol{B}}_t)_{t\in\{0,T\}}$ defined for any $t \in \{0, T\}$ by

$$\tilde{\boldsymbol{B}}_t = -\bar{\boldsymbol{B}}_T + \bar{\boldsymbol{B}}_{T-t} + \boldsymbol{k}_T - \boldsymbol{k}_{T-t} - \int_{T-t}^T \nabla \log p_s(\bar{\boldsymbol{B}}_s) \mathrm{d}s. \tag{D.29}$$

First, note that $t \mapsto \tilde{\boldsymbol{B}}_t$ is continuous. Denote by $\mathcal{F}$, the filtration associated with $(\bar{\boldsymbol{B}}_{T-t})_{t\in\{0,T\}}$. We have that $(\tilde{\boldsymbol{B}}_t)_{t\in\{0,T\}}$ is adapted to $(\bar{\boldsymbol{B}}_{T-t})_{t\in\{0,T\}}$. Even more so, we have that $(\tilde{\boldsymbol{B}}_t)_{t\in\{0,T\}}$ satisfies the strong Markov property since $(\bar{\boldsymbol{B}}_t)_{t\in\{0,T\}}$ also satisfies the strong Markov property. Let $g \in \mathrm{C}_c^\infty(\mathrm{cl}(\mathcal{M}))$ and consider for any $0 \leq s \leq t \leq T$, $\mathbb{E}\left[(\tilde{\boldsymbol{B}}_t - \tilde{\boldsymbol{B}}_s) g(\bar{\boldsymbol{B}}_{T-t})\right]$. For any $0 \leq s \leq t \leq T$ we have

$$\mathbb{E}\left[(\tilde{\boldsymbol{B}}_t - \tilde{\boldsymbol{B}}_s) g(\bar{\boldsymbol{B}}_{T-t})\right] = \mathbb{E}\left[(-\bar{\boldsymbol{B}}_{T-s} + \bar{\boldsymbol{B}}_{T-t} + \boldsymbol{k}_{T-s} - \boldsymbol{k}_{T-t} - \int_{T-t}^{T-s} \nabla \log p_u(\bar{\boldsymbol{B}}_u) \mathrm{d}u) g(\bar{\boldsymbol{B}}_{T-t})\right]. \tag{D.30}$$

In what follows, we prove that for any $0 \leq s \leq t \leq T$ we have $\mathbb{E}\left[(\tilde{\boldsymbol{B}}_t - \tilde{\boldsymbol{B}}_s) g(\bar{\boldsymbol{B}}_{T-t})\right] = 0$. Therefore, we only need to prove that for any $0 \leq s \leq t \leq T$ we have

$$\mathbb{E}\left[(-\bar{\boldsymbol{B}}_t + \bar{\boldsymbol{B}}_s + \boldsymbol{k}_t - \boldsymbol{k}_s - \int_s^t \nabla \log p_u(\bar{\boldsymbol{B}}_u) \mathrm{d}u) g(\bar{\boldsymbol{B}}_t)\right] = 0. \tag{D.31}$$

Let $t \in (0, T]$. We introduce $u : \{0, t\} \times \mathcal{M}$ such that for any $s \in \{0, t\}$ and $x \in \mathcal{M}$, $u(s, x) = \mathbb{E}\left[g(\bar{\boldsymbol{B}}_t) | \bar{\boldsymbol{B}}_s = x\right]$. Using Burdzy et al. (2004, Theorem 2.8) we get that for any $x \in \mathcal{M}$, $s \mapsto u(s, x) \in \mathrm{C}^1((0, t), \mathbb{R})$ and for any $s \in (0, t)$, $x \mapsto u(s, x) \in \mathrm{C}^2(\mathcal{M}, \mathbb{R})$ and $x \mapsto u(s, x) \in \mathrm{C}^1(\mathrm{cl}(\mathcal{M}), \mathbb{R})$. In addition, we have that for any $s \in (0, t)$ and for any $x \in \mathcal{M}$ and $x_0 \in \partial\mathcal{M}$

$$\partial_s u(s, x) + \tfrac{1}{2}\Delta u(s, x) = 0, \qquad \langle \nabla u(s, x_0), \boldsymbol{n}(x_0) \rangle = 0. \tag{D.32}$$

This equation is called the backward Kolmogorov equation. Using (D.32), $\bar{\boldsymbol{B}}_t = \boldsymbol{B}_t - \boldsymbol{k}_t$ for any $t \geq 0$ and the Itô formula for semimartingale (Revuz and Yor, 2013, Chapter IV, Theorem 3.3) we have that for any $s \in (0, t)$

$$\mathbb{E}\left[u(t, \bar{\boldsymbol{B}}_t)\bar{\boldsymbol{B}}_t\right] = \mathbb{E}\left[u(s, \bar{\boldsymbol{B}}_s)\bar{\boldsymbol{B}}_s\right] + \mathbb{E}\left[\tfrac{1}{2}\int_s^t \bar{\boldsymbol{B}}_w \Delta u(w, \bar{\boldsymbol{B}}_w)\mathrm{d}w\right] + \mathbb{E}\left[\int_s^t \nabla u(w, \bar{\boldsymbol{B}}_w)\mathrm{d}w\right] \tag{D.33}$$

$$- \mathbb{E}\left[\int_s^t \bar{\boldsymbol{B}}_w \langle \nabla u(w, \bar{\boldsymbol{B}}_w), \boldsymbol{n}(\bar{\boldsymbol{B}}_w)\rangle \mathrm{d}|\boldsymbol{k}|_w\right] \tag{D.34}$$

$$- \mathbb{E}\left[\int_s^t u(w, \bar{\boldsymbol{B}}_w)\boldsymbol{n}(\bar{\boldsymbol{B}}_w)\mathrm{d}|\boldsymbol{k}|_w\right] \tag{D.35}$$

$$+ \mathbb{E}\left[\int_s^t \bar{\boldsymbol{B}}_w \partial_w u(w, \bar{\boldsymbol{B}}_w)\mathrm{d}w\right] \tag{D.36}$$

$$= \mathbb{E}\left[u(s, \bar{\boldsymbol{B}}_s)\bar{\boldsymbol{B}}_s\right] + \mathbb{E}\left[\int_s^t \nabla u(w, \bar{\boldsymbol{B}}_w)\mathrm{d}w\right] - \mathbb{E}\left[\int_s^t u(w, \bar{\boldsymbol{B}}_w)\boldsymbol{n}(\bar{\boldsymbol{B}}_w)\mathrm{d}|\boldsymbol{k}|_w\right] \tag{D.37}$$

In addition, using the Fubini theorem and the definition of $\boldsymbol{k}_t$ we have that for any $s \in (0, t)$

$$\mathbb{E}\left[\int_s^t u(w, \bar{\boldsymbol{B}}_w)\boldsymbol{n}(\bar{\boldsymbol{B}}_w)\mathrm{d}|\boldsymbol{k}|_w\right] = \mathbb{E}\left[\int_s^t \mathbb{E}\left[g(\bar{\boldsymbol{B}}_t)|\bar{\boldsymbol{B}}_w\right]\boldsymbol{n}(\bar{\boldsymbol{B}}_w)\mathrm{d}|\boldsymbol{k}|_w\right] = \mathbb{E}\left[g(\bar{\boldsymbol{B}}_t)(\boldsymbol{k}_t - \boldsymbol{k}_s)\right]. \tag{D.38}$$

Finally, using the divergence theorem and Burdzy et al. (2004, Theorem 2.6) we have that for any $s \in (0, t)$

$$\mathbb{E}\left[\int_s^t \nabla u(w, \bar{\boldsymbol{B}}_w)\mathrm{d}w\right] = \int_s^t \int_{\mathcal{M}} \nabla u(w, x) p_w(x)\mathrm{d}x\mathrm{d}w \tag{D.39}$$

$$= -\int_s^t \int_{\mathcal{M}} u(w, x)\nabla \log(p_w(x)) p_w(x)\mathrm{d}x\mathrm{d}w + \int_s^t \int_{\partial\mathcal{M}} u(w, x) p_w(x)\mathrm{d}x\mathrm{d}\sigma(w), \tag{D.40}$$

where σ is the surface area measure on $\partial\mathcal{M}$, see Burdzy et al. (2004). Using

Lemma D.9 and the Fubini theorem we get that

$$\mathbb{E}\left[\int_s^t \nabla u(w, \bar{\boldsymbol{B}}_w)\mathrm{d}w\right] = -\int_s^t \int_{\mathcal{M}} u(w,x)\nabla \log(p_w(x))p_w(x)\mathrm{d}x\mathrm{d}w + \mathbb{E}\left[\int_s^t u(w, \bar{\boldsymbol{B}}_w)\mathrm{d}\boldsymbol{k}_w\right] \tag{D.41}$$

$$= -\mathbb{E}\left[\int_s^t g(\bar{\boldsymbol{B}}_t)\nabla \log(p_w(\bar{\boldsymbol{B}}_w))\mathrm{d}w\right] + 2\,\mathbb{E}\left[g(\bar{\boldsymbol{B}}_t)(\boldsymbol{k}_t - \boldsymbol{k}_s)\right] \tag{D.42}$$

Combining (D.37), (D.38) and (D.42) we get that

$$\mathbb{E}\left[u(t, \bar{\boldsymbol{B}}_t)\right] = \mathbb{E}\left[u(s, \bar{\boldsymbol{B}}_s)\right] - \mathbb{E}\left[g(\bar{\boldsymbol{B}}_t)\int_s^t \nabla \log p_w(\bar{\boldsymbol{B}}_w)\mathrm{d}w\right] + \mathbb{E}\left[g(\bar{\boldsymbol{B}}_t)(\boldsymbol{k}_t - \boldsymbol{k}_s)\right]. \tag{D.43}$$

Therefore, we get (D.31) and (D.30). Hence, $(\tilde{\boldsymbol{B}}_t)_{t\in\{0,T\}}$ is a continuous martingale. In addition, we have that for any $t \in \{0, T\}$, $\mathbb{E}\left[\tilde{\boldsymbol{B}}_t\tilde{\boldsymbol{B}}_t^\top\right] = t\mathbf{I}$ and therefore, $(\tilde{\boldsymbol{B}}_t)_{t\in\{0,T\}}$ is a Brownian motion using the Lévy characterisation of Brownian motion (Revuz and Yor, 2013, Chapter IV, Theorem 3.6). Denote $(\boldsymbol{j}_t)_{t\in\{0,T\}} = (\boldsymbol{k}_T - \boldsymbol{k}_{T-t})_{t\in\{0,T\}}$. Using (D.30), we have that for any $t \in \{0, T\}$

$$\bar{\boldsymbol{B}}_{T-t} = \bar{Y}_0 + \tilde{\boldsymbol{B}}_t + \int_0^t \nabla \log p_{T-s}(Y_s)\mathrm{d}s - \boldsymbol{j}_t. \tag{D.44}$$

Using (D.28), we have for any $t \in \{0, T\}$

$$|\boldsymbol{j}|_t = \int_0^t \mathbf{1}_{Y_s\in\partial\mathcal{M}}\mathrm{d}|\boldsymbol{j}|_s, \quad \boldsymbol{j}_t = \int_0^t \boldsymbol{n}(\bar{Y}_s)\mathrm{d}|\boldsymbol{j}|_s, \tag{D.45}$$

which concludes the proof.

D.7. Experimental details

In what follows we describe the experimental settings used to generate results introduced in section 4.6. The models and experiments have been implemented in Jax (Bradbury et al., 2018), using a modified version of the Riemannian geometry library Geomstats (Miolane et al., 2020).

Architecture. The architecture of the score network $\boldsymbol{s}_\theta$ is given by a multilayer perceptron with 6 hidden layers with 512 units each. We use sinusoidal activation functions.

Training. All models are trained by the stochastic optimizer Adam (Kingma, 2014) with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch-size of 256 data-points. The learning rate is annealed with a linear ramp from 0 to 1000 steps, reaching the maximum value of $2e-4$, and from then with a cosine schedule down to 0 after $100k$ iterations in total.

Diffusion. Following Song et al. (2020b), the diffusion models diffusion coefficient is parametrized as $g(t) = \sqrt{\beta(t)}$ with $\beta : t \mapsto \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot t$, where we found $\beta_{\min} = 0.001$ and $\beta_{\max} = 6$ to work best.

Metrics. We measure the performance of trained models via the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), which is a kernel based metric between two distributions P and Q. The MMD can be empirically approximated with the following U-statistics $\mathrm{MMD}^2(P,Q) = \frac{1}{m(m-1)} \sum_i \sum_{j\neq i} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_i \sum_{j\neq i} k(y_i, y_j) - 2\frac{1}{m^2} \sum_i \sum_j k(x_i, y_j)$ with $x_i \sim P$ and $y_i \sim Q$, where k is a kernel. For synthetic experiments we use a sum of weighted RBF kernels matching the generating distributions for the Gaussian mixtures. For the other experiments we use an RBF kernel. For all experiments we use 100,000 samples to compute the MMD.

D.7.1. Synthetic data on polytopes.

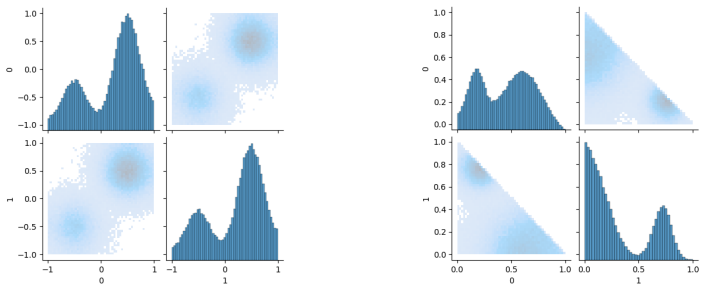
Hypercube The hypercube is a specific instance of a convex polytope where the affine constraints are given by the following coefficients:

$$A = \begin{pmatrix} 1 & \dots & 0 \\ -1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ 0 & \dots & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}. \tag{D.46}$$

Where A is a $2n \times n$ matrix and b an n dimensional vector.

We construct the training and test datasets by sampling for both 100000 points from a mixture of 'wrapped normal' distributions illustrated in figure D.3a and which density is given by

$$p_0(x) = 0.7\ \mathrm{ReflectedStep}[(0.5, 0.5), \cdot, \{f_i\}_{i\in\mathcal{I}}]\#\mathcal{N}(0, 0.25) \tag{D.47}$$
$$+ 0.3\ \mathrm{ReflectedStep}[(-0.5, -0.5), \cdot, \{f_i\}_{i\in\mathcal{I}}]\#\mathcal{N}(0, 0.25). \tag{D.48}$$



(a) On the hypercube. (b) On the simplex.

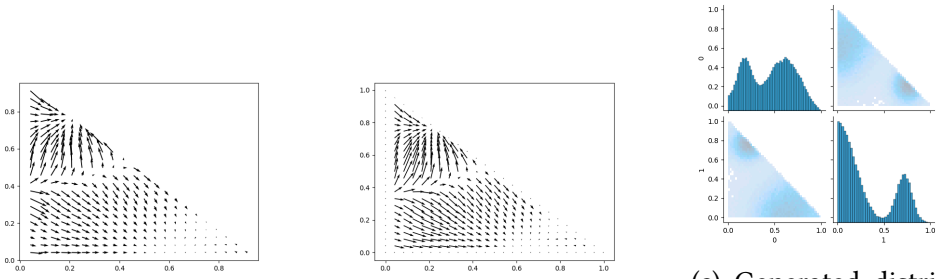
Figure D.3. Pairwise and marginals samples from the synthetic data distribution.

Simplex Δ^n. Similarly, to parametrise the simplex as a convex polytope we set the matrix and constraints to be given by

$$A = \begin{pmatrix} -1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}. \tag{D.49}$$

Where A will be a $n-1 \times n$ matrix. Essentially we perform diffusion over the first $n-1$ components of the simplex, allowing the last component to be determined by the one minus the sum of the first $n-1$.

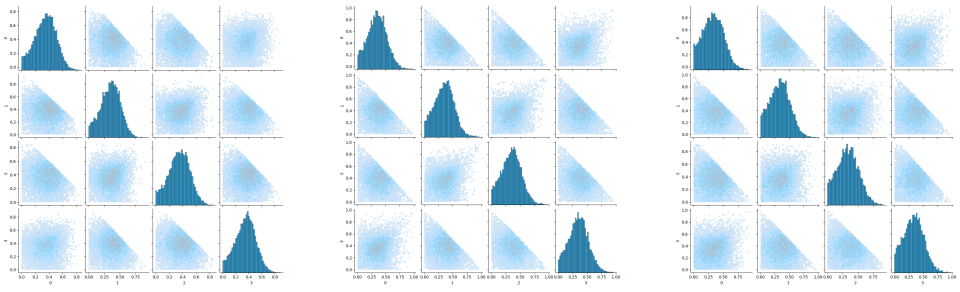
Similarly than for the hypercube, we construct the training and test datasets from generated data points which are illustrated in figure D.3b. The score network at different times is illustrated in Figure D.4.



(a) $t = 0.01$. (b) $t = 0.5$. (c) Generated distribution.

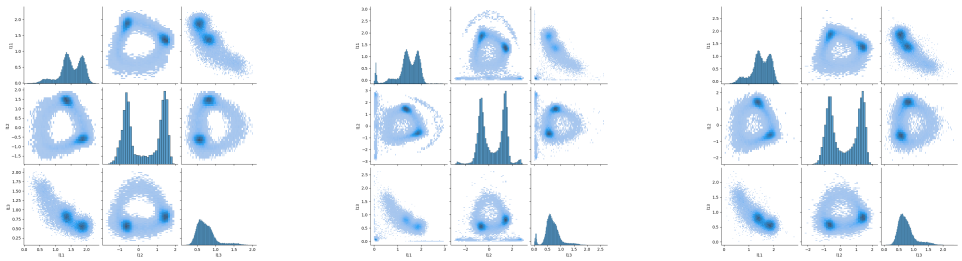
Figure D.4. Evolution of the score on the simplex and generated distribution.

The Birkhoff polytope. The Birkhoff polytope is the space of doubly stochastic matrices, i.e. $B_n = \{\mathrm{P} \in [0,1]^{n \times n} : \sum_i^n P_{i,j} = 1, \sum_j^n P_{i,j} = 1\}$. It is a convex polytope in $\mathbb{R}^{n^2}$ and has dimension $d = (n-1)^2$.



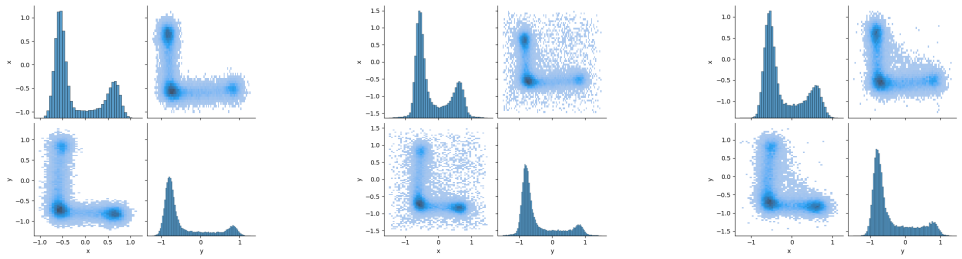
(a) Data. (b) Log-barrier. (c) Reflected.

Figure D.5. Pairwise and marginals samples on the Birkhoff polytope from synthetic data distribution and from trained constrained diffusion models.



(a) Data. (b) Log-barrier. (c) Reflected.

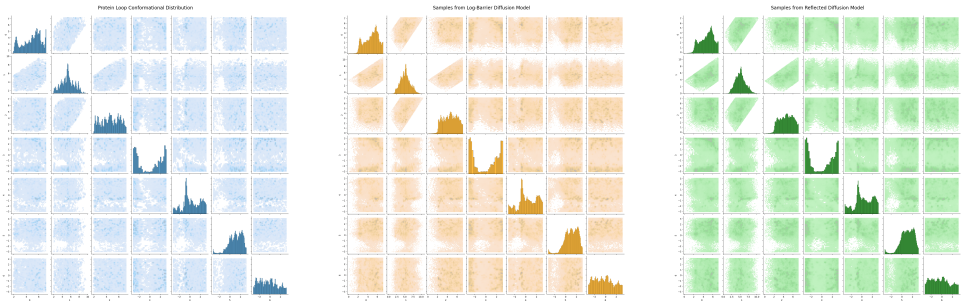
Figure D.6. Pairwise and marginals distributions over the coefficients L_{11}, L_{21}, L_{22} of the lower triangle matrix parametrising SPD matrices $M = LL^\top$ (which represent the manipulability ellipsoids of the robotic arms).



(a) Data. (b) Log-barrier. (c) Reflected.

Figure D.7. Pairwise and marginals distributions over the (x, y) locations of the robotic arms.

D.7.2. Constrained SPD matrices for robotic arms modelling.



(a) Data. (b) Log-barrier. (c) Reflected.

Figure D.8. Pairwise and marginals distributions over the dimensions of the polytope and torus used to model the conformational ensembles of cyclic peptides generated by the reflected diffusion model.

D.7.3. Conformational modelling of polypeptide backbones under anchor point constraints.

E | Efficient score-based modelling on constrained domains

Organization of the appendix

In Appendix E.1, we prove the convergence of the rejection and Metropolis discretisations to the true reflected Brownian Motion. The geospatial dataset with non-convex constraints based on wildfire incidence rates in the continental United States is presented Appendix E.2. All supplementary experimental details and empirical results are gathered in Appendix E.3.

E.1. Convergence to the reflected process

In this note, we assume that $\mathcal{M} = \left\{x \in \mathbb{R}^d \,:\, \Phi(x) > 0\right\}$ is compact, with $\Phi \in C^2(\mathbb{R}^d, \mathbb{R})$. We have that $\partial\mathcal{M} = \left\{x \in \mathbb{R}^d \,:\, \Phi(x) = 0\right\}$. In addition, we assume that for any $x \in \partial\mathcal{M}$, $\|\nabla\Phi(x)\| = 1$ and that Φ is concave. The closure of $\mathcal{M}$ is denoted $\overline{\mathcal{M}}$. The assumption that Φ is concave is only used in Theorem E.2-(d) and can be dropped. We consider it for simplicity.

Let $(\hat{X}_k^\gamma)_{k\in\mathbb{N}}$ given for any $\gamma > 0$ and $k \in \mathbb{N}$ by $\hat{X}_0^\gamma = x \in \overline{\mathcal{M}}$ and for $\hat{X}_{k+1}^\gamma = \hat{X}_k^\gamma + \sqrt{\gamma}Z_k^\gamma$ with Z_k^γ a Gaussian random variable conditioned on $\hat{X}_k^\gamma + \sqrt{\gamma}Z_k^\gamma \in \overline{\mathcal{M}}$. In practice, Z_k^γ is sampled using rejection sampling. We define $\hat{\boldsymbol{X}}^\gamma : \mathbb{R}_+ \to \overline{\mathcal{M}}$ given for any $k \in \mathbb{N}$ by $\hat{\boldsymbol{X}}_{k\gamma}^\gamma = \hat{X}_k^\gamma$ and for any $t \in [k\gamma, (k+1)\gamma)$, $\hat{\boldsymbol{X}}_t^\gamma = \hat{X}_k^\gamma$. Note that $(\boldsymbol{X}_t)_{t\in\{0,T\}}$ is a $\mathrm{D}(\{0,T\}, \overline{\mathcal{M}})$ valued random variable, where $\mathrm{D}(\{0,T\}, \overline{\mathcal{M}})$ is the space of right-continuous with left-limit processes which take values in $\overline{\mathcal{M}}$. We denote $\hat{\mathbb{P}}^\gamma$ the distribution of $(\hat{\boldsymbol{X}}_t^\gamma)_{t\in\{0,T\}}$ on $\mathrm{D}(\{0,T\}, \overline{\mathcal{M}})$.

Our goal is to show the following theorem.

Theorem E.1. *For any $T \geq 0$, $(\hat{\boldsymbol{X}}_t^\gamma)_{t\in\{0,T\}}$ weakly converges to $(\boldsymbol{X}_t)_{t\in\{0,T\}}$ such that for any $t \in \{0,T\}$*

$$\boldsymbol{X}_t = x + \boldsymbol{B}_t - \boldsymbol{k}_t, \qquad |\boldsymbol{k}|_t = \int_0^t \mathbf{1}_{\boldsymbol{X}_s \in \partial\mathcal{M}} \mathrm{d}|\boldsymbol{k}|_s, \qquad \boldsymbol{k}_t = \int_0^t \boldsymbol{n}(\boldsymbol{X}_s)\mathrm{d}|\boldsymbol{k}|_s. \quad \text{(E.1)}$$

Proof. In order to prove the result, we prove that the distribution of the Markov chain seen as an element of $\mathrm{D}(\{0,T\}, \overline{\mathcal{M}})$ converges to a solution of the Skorokhod

problem (E.1). In particular, we first show that the limiting distribution satisfies a submartingale problem following (Stroock and Varadhan, 1971, Theorem 6.3). The transition from a solution of a submartingale problem to a weak solution of the Skorokhod problem is given by (Kang and Ramanan, 2017, Theorem 1, Proposition 2.12) and (Ramanan, 2006, Corollary 2.10). In order to apply (Stroock and Varadhan, 1971, Theorem 6.3), we define an intermediate drift and diffusion matrix, see (E.53) and (E.49). To prove the theorem one needs to control the drift and diffusion matrix inside $\mathcal{M}$ and show that it converges to 0 and $\mathbf{I}$ respectively. The technical part of the proof comes from the control of the drift coefficient near the boundary. In particular, we show that if the intermediate drift is large then we are close to the boundary and the intermediate drift is pointing inward. To investigate the local properties of the drift near the boundary we rely on the notion of tubular neighbourhood, see (Lee, 2013, Theorem 6.24). ■

Some key properties of the tubular neighbourhood are stated in Appendix E.1.1. We then establish a few technical lemmas about the tail probability of some distributions in Appendix E.1.2. Controls on the diffusion matrix and lower bounds on the probability of belonging in $\mathcal{M}$ are given in Appendix E.1.3. Properties of large drift terms are given in Appendix E.1.4. The convergence of the drift and diffusion matrix on compact sets is given in Appendix E.1.5. The convergence of the boundary terms is investigated in Appendix E.1.6. Finally, we conclude the proof in Appendix E.1.7.

E.1.1. Properties of the tubular neighbourhood. Using the results of (Lee, 2013), we establish the existence of an open set of $\overline{\mathcal{M}}$ (for the induced topology of $\mathbb{R}^d$) satisfying several important properties.

THEOREM E.2. *There exist* $\mathsf{U} \subset \overline{\mathcal{M}}$ *open and* $C \geq 1, \bar{r} > 0$ *such that for any* $\gamma \in (0, \bar{\gamma})$ *with* $\bar{\gamma} = 1$ *the following properties hold:*

(a) For any $x \in \mathsf{U}$*, there exist a unique* $\bar{x} \in \partial\mathcal{M}$ *and* $\bar{\alpha} > 0$ *such that* $x = \bar{x} + \bar{\alpha}\nabla\Phi(\bar{x})$.

(b) For any $\bar{\alpha} \in \{0, \bar{r}\}$ *and* $\bar{x} \in \partial\mathcal{M}$ *such that* $\bar{x}+\bar{\alpha}\nabla\Phi(\bar{x}) \in \overline{\mathcal{M}}$*, let* $x = \bar{x}+\bar{\alpha}\nabla\Phi(\bar{x})$ *and* $\mathsf{C}(\mathsf{x}, \gamma)$ *such that* $x + \sqrt{\gamma}z \in \mathsf{C}(\mathsf{x}, \gamma)$ *if*

$$-\bar{\alpha}\gamma^{-1/2} \leq \alpha < \bar{r}\gamma^{-1/2}, \qquad \|v\|^2 \leq (\alpha\gamma^{1/2} + \bar{\alpha})/(C\gamma), \tag{E.2}$$

with $z = \alpha\nabla\Phi(\bar{x}) + v$*, with* $v \perp \nabla\Phi(\bar{x})$*. Then* $\mathsf{C}(\mathsf{x}, \gamma) \subset \overline{\mathcal{M}}$.

(c) $\mathsf{V} = \{\cancel{x} + \alpha\nabla^{\vee}(\cancel{x}) \ : \ \cancel{x} \in \partial\mathcal{M}, \ \alpha \in [0, \cancel{r})\}$ *is open in* $\overline{\mathcal{M}}$.

(d) For any $x \in \mathsf{U}$*,* $x + \sqrt{\gamma}z \in \overline{\mathcal{M}} \cap \mathsf{C}(\mathsf{x}, \gamma)^{\mathrm{c}}$ *then* $\alpha \geq \bar{r}\gamma^{-1/2}$ *or* $\|v\|^2 \geq (\alpha\gamma^{1/2} + \bar{\alpha})/(C\gamma)$ *and* $\alpha\gamma^{1/2} + \bar{\alpha} \geq 0$*, with* $z = \alpha\nabla\Phi(\bar{x}) + v$*, with* $\bar{x}$ *given in (a) and* $v \perp \nabla\Phi(\bar{x})$. .

(e) There exists $R > 0$ *such that* $\left\{x \in \overline{\mathcal{M}} \ : \ d(x, \partial\mathcal{M}) \leq 2R\right\} \subset \mathsf{V}$.

Proof. Let $\gamma \in (0, \bar{\gamma})$ with $\bar{\gamma} = 1$. First, note that for any $\bar{x} \in \partial\mathcal{M}$, the normal space is given by $\{\alpha\nabla\Phi(\bar{x}) \ : \ \alpha \in \mathbb{R}\}$. Using this result and (Lee, 2013, Theorem 6.24)

there exists $\tilde{r}_0 > 0$ such that $\mathsf{U}_0 = \{\bar{x} + \alpha\nabla^{\circ}(\bar{x}) \,:\, \bar{x} \in \partial\mathcal{M},\ \alpha \in (-\tilde{r}_0, \tilde{r}_0)\} \subset \mathbb{R}^{\mathsf{d}}$ is open[1]. We have that for any $\alpha \in [-r_0, 0)$ and $\bar{x} \in \partial\mathcal{M}$

$$\Phi(\bar{x} + \alpha\nabla\Phi(\bar{x})) = \Phi(\bar{x}) + \alpha\|\nabla\Phi(\bar{x})\|^2 + \int_0^1 \nabla^2\Phi(\bar{x} + t\alpha\nabla\Phi(\bar{x}))(\alpha\nabla\Phi(\bar{x}))^{\otimes 2}\mathrm{d}t \tag{E.3}$$

$$\leq \alpha + \tilde{C}_0\alpha^2 < 0, \tag{E.4}$$

with $r_0 = \min(\tilde{r}_0, 1/(2\tilde{C}_0))$, where we have used that $\Phi(\bar{x}) = 0$, $\|\nabla\Phi(\bar{x})\| = 1$ and defined $\tilde{C}_0 = \sup\{\|\nabla^2\Phi(\bar{x} + \alpha\nabla\Phi(\bar{x}))\| \,:\, \bar{x} \in \partial\mathcal{M},\ \alpha \in \{-\tilde{r}_0, \tilde{r}_0\}\}$. Reciprocally, we have for any $\alpha \in [0, r_0)$ and $\bar{x} \in \partial\mathcal{M}$

$$\Phi(\bar{x} + \alpha\nabla\Phi(\bar{x})) = \Phi(\bar{x}) + \alpha\|\nabla\Phi(\bar{x})\|^2 + \int_0^1 \nabla^2\Phi(\bar{x} + t\alpha\nabla\Phi(\bar{x}))(\alpha\nabla\Phi(\bar{x}))^{\otimes 2}\mathrm{d}t \geq \alpha - C_0\alpha^2, \tag{E.5}$$

where we have used that $\Phi(\bar{x}) = 0$, $\|\nabla\Phi(\bar{x})\| = 1$ and defined

$$C_0 = \sup\{\|\nabla^2\Phi(\bar{x} + \alpha\nabla\Phi(\bar{x}))\| \,:\, \bar{x} \in \partial\mathcal{M},\ \alpha \in \{-r_0, r_0\}\}. \tag{E.6}$$

Let $r_1 = \min(r_0, 1/(2C_0))$. Then, $\mathsf{U}_1 = \{\bar{x} + \alpha\nabla^{\circ}(\bar{x}) \,:\, \bar{x} \in \partial\mathcal{M},\ \alpha \in (-\mathsf{r}_1, \mathsf{r}_1)\} \subset \mathbb{R}^{\mathsf{d}}$ is open and

$$\mathsf{U}_1 \cap \overline{\mathcal{M}} = \{\bar{x} + \alpha\nabla^{\circ}(\bar{x}) \,:\, \bar{x} \in \partial\mathcal{M},\ \alpha \in [0, \mathsf{r}_1)\}. \tag{E.7}$$

In what follows, we define $\mathsf{U} = \mathsf{U}_1 \cap \overline{\mathcal{M}}$. Note that U is open for the induced topology and that $\partial\mathcal{M} \subset \mathsf{U}$. In particular, $\partial\mathcal{M}$ is compact, U^{c} is closed and $\partial\mathcal{M} \cap \mathsf{U}^{\mathrm{c}} = \emptyset$. Hence, there exists $r > 0$ such that $d(\partial\mathcal{M}, \mathsf{U}^{\mathrm{c}}) \geq 4\mathsf{r}$. Without loss of generality we can assume that $r \leq 1/2$. We also have $\left\{x \in \overline{\mathcal{M}} \,:\, d(x, \partial\mathcal{M}) \leq 2r\right\} \subset \mathsf{U}$. The proof of (a) follows from the definition of U_0. In the rest of the proof, we define

$$C^{1/2} = 2\max(1, \sup\{\|\nabla^2\Phi(\bar{x} + u)\| \,:\, \bar{x} \in \partial\mathcal{M},\ \|u\|^2 \leq r(r+1)\}), \quad \bar{r} = \min(1/(2C^{1/2}), r/2). \tag{E.8}$$

Let us prove (b). Consider $x + \sqrt{\gamma}z \in \mathsf{C}(\mathsf{x}, \gamma)$ with $\mathsf{C}(\mathsf{x}, \gamma)$ given by (E.2) and $x = \bar{x} + \bar{\alpha}\nabla\Phi(\bar{x})$ and $z = \alpha\nabla\Phi(\bar{x}) + v$ with $v \perp \nabla\Phi(\bar{x})$. In particular, we recall that we have

$$-\bar{\alpha}\gamma^{-1/2} \leq \alpha < \bar{r}\gamma^{-1/2}, \qquad \|v\|^2 \leq (\alpha\gamma^{1/2} + \bar{\alpha})/(C\gamma). \tag{E.9}$$

This implies that

$$\bar{\alpha} + \sqrt{\gamma}\alpha \leq 2\bar{r}, \qquad \gamma\|v\|^2 \leq 2\bar{r}/C. \tag{E.10}$$

First, using that $C \geq 1$, $\|\nabla\Phi(\bar{x})\| = 1$, (E.10) and (E.8), we have

$$\left\|x + \sqrt{\gamma}z - \bar{x}\right\|^2 = (\bar{\alpha} + \sqrt{\gamma}\alpha)^2 + \gamma\|v\|^2 \leq r^2 + r/C \leq r(r+1). \tag{E.11}$$

Then, we have that

$$\Phi(x + \sqrt{\gamma}z) = \Phi(\bar{x}) + \bar{\alpha} + \sqrt{\gamma}\alpha + \int_0^1 \nabla^2\Phi(\bar{x} + t(x + \sqrt{\gamma}z - \bar{x}))(x + \sqrt{\gamma}z - \bar{x})^{\otimes 2}\mathrm{d}t \tag{E.12}$$

$$\geq \bar{\alpha} + \sqrt{\gamma}\alpha - (C^{1/2}/2)((\bar{\alpha} + \sqrt{\gamma}\alpha)^2 + \gamma\|v\|^2), \tag{E.13}$$

[1]This is the tubular neighbourhood theorem which is key to the rest of the proof.

where we recall that

$$C^{1/2} = 2\max(1, \sup\{\|\nabla^2\Phi(\bar{x}+u)\| \ : \ \bar{x}\in\partial\mathcal{M},\ \|u\|^2 \le r(r+1)\}), \tag{E.14}$$

$$\bar{r} = \min(1/(2C^{1/2}), r/2). \tag{E.15}$$

First, using that $r \le 1/2$ and (E.10), we have $\bar{\alpha} + \sqrt{\gamma}\alpha \le 2r \le 1$. Since, $\|v\|^2 \le (\bar{\alpha}+\sqrt{\gamma}\alpha)/(C\gamma)$ and we have that $\|v\|^2 < 1/(C\gamma)$. Let $P(X) = X - (C^{1/2}/2)X^2 - (C^{1/2}/2)\gamma\|v\|^2$. We have that $P(x) \ge 0$ if and only if $x \in \{x_{\min}, x_{\max}\}$ with

$$x_{\min} = (1-(1-C\gamma\|v\|^2)^{1/2})/C^{1/2}, \qquad x_{\max} = (1+(1-C\gamma\|v\|^2)^{1/2})/C^{1/2}. \tag{E.16}$$

Using that for any $t \in (0,1)$, $(1-t)^{1/2} \ge 1-t$ we have that

$$x_{\min} \le \gamma C\|v\|^2/2, \qquad x_{\max} \ge 1/C^{1/2}. \tag{E.17}$$

Since $\|v\|^2 \le (\sqrt{\gamma}\alpha+\bar{\alpha})/(\gamma C)$, we have that $\bar{\alpha}+\sqrt{\gamma}\alpha \ge x_{\min}$. In addition, using that $\bar{\alpha}+\sqrt{\gamma}\alpha \le 2\bar{r} \le 1/C^{1/2} \le x_{\max}$, we get that $P(\bar{\alpha}+\sqrt{\gamma}\alpha) \ge 0$ and therefore $x+\sqrt{\gamma}z \in \overline{\mathcal{M}}$ since $\Phi(x+\sqrt{\gamma}z) \ge 0$. This concludes the proof of (b). Note that the condition $\alpha \ge -\gamma^{-1/2}\bar{\alpha}$ is implied by the condition $\|v\|^2 \le (\sqrt{\gamma}\alpha+\bar{\alpha})/(\gamma C)$. Using that $\mathsf{V} \subset \left\{\mathsf{x}\in\overline{\mathcal{M}} \ : \ \mathsf{d}(\mathsf{x},\partial\mathcal{M}) \le 2\mathsf{r}\right\} \subset \mathsf{U}$, (c) is a direct consequence of (Lee, 2013, Theorem 6.24)]. Next, we prove (d). Let $x+\sqrt{\gamma}z \in \overline{\mathcal{M}} \cap \mathsf{C}(\mathsf{x},\gamma)^{\mathrm{c}}$. If $\alpha < -\bar{\alpha}\gamma^{-1/2}$ then since Φ is concave, we have

$$\Phi(x+\sqrt{\gamma}z) = \Phi(\bar{x}) + \bar{\alpha} + \sqrt{\gamma}\alpha + \int_0^1 \nabla^2\Phi(\bar{x}+t(x+\sqrt{\gamma}z-\bar{x}))(x+\sqrt{\gamma}z-\bar{x})^{\otimes 2}\mathrm{d}t < 0, \tag{E.18}$$

where we have used that $\Phi(\bar{x}) = 0$. This is absurd, hence either $\alpha \ge \bar{r}\gamma^{-1/2}$ or $\|v\|^2 \ge (\alpha\gamma^{1/2}+\bar{\alpha})/(C\gamma)$ and $\alpha\gamma^{1/2}+\bar{\alpha} \ge 0$, which concludes the proof. The proof of (e) is similar to the proof that $\left\{x\in\overline{\mathcal{M}} \ : \ d(x,\partial\mathcal{M}) \le 2r\right\} \subset \mathsf{U}$.

■

The main message of Theorem E.2 is that using Theorem E.2-(d), if we move in the direction of $\nabla\Phi(\bar{x})$ (the inward normal) with magnitude α then we are allowed to move in the orthonormal direction with magnitude $\alpha^{1/2}$. In the next paragraph, we discuss this fact in details and shows it is necessary for the rest of our study.

The necessity of Theorem E.2-(b). At first sight one can wonder if the statement of Theorem E.2-(b) could be simplify. In particular, it would be simpler to replace this statement with: for any $\bar{\alpha} \in \{0, \bar{r}\}$ and $\bar{x} \in \partial\mathcal{M}$ such that $\bar{x}+\bar{\alpha}\nabla\Phi(\bar{x}) \in \overline{\mathcal{M}}$, let $x = \bar{x}+\bar{\alpha}\nabla\Phi(\bar{x})$ and $\mathsf{C}(\mathsf{x},\gamma)$ such that $x+\sqrt{\gamma}z \in \mathsf{C}(\mathsf{x},\gamma)$ if

$$-\bar{\alpha}\gamma^{-1/2} \le \alpha < \bar{r}\gamma^{-1/2}, \qquad \|v\|^2 \le (\alpha\gamma^{1/2}+\bar{\alpha})^2/(C\gamma), \tag{E.19}$$

with $z = \alpha\nabla\Phi(\bar{x}) + v$, with $v \perp \nabla\Phi(\bar{x})$. Then $\mathsf{C}(\mathsf{x},\gamma) \subset \overline{\mathcal{M}}$. Note that $\|v\|^2 \le (\alpha\gamma^{1/2}+\bar{\alpha})/(C\gamma)$ is replaced by $\|v\|^2 \le (\alpha\gamma^{1/2}+\bar{\alpha})^2/(C\gamma)$, see Figure E.1 for an illustration. However, in that case Theorem E.2-(d) becomes: in addition, if $x+\sqrt{\gamma}z \in \overline{\mathcal{M}} \cap \mathsf{C}(\mathsf{x},\gamma)^{\mathrm{c}}$ then $\alpha \ge r\gamma^{-1/2}$ or $\|v\|^2 \ge (\alpha\gamma^{1/2}+\bar{\alpha})^2/(C\gamma)$ and $\alpha\gamma^{1/2}+\bar{\alpha} \ge 0$.

In what follows, when controlling the properties of large drift, see the proof of Proposition E.13 and the proof of Proposition E.16, we need to control quantities

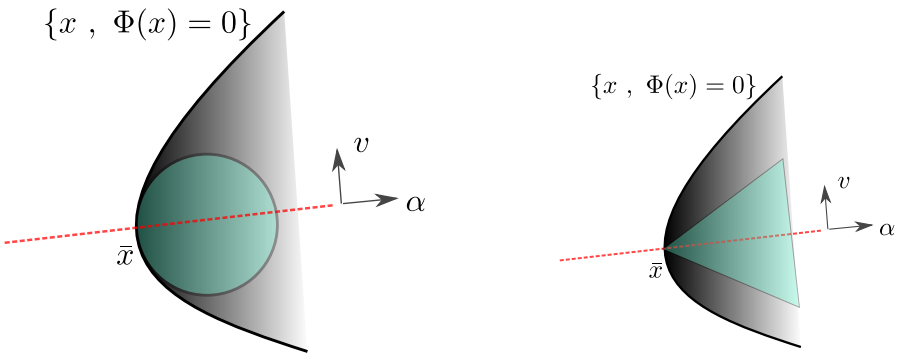


Figure E.1. The grey shaded area represents $\overline{\mathcal{M}}$ while the blue shaded area represents $\mathrm{C}(\mathrm{x},\gamma)$ for an arbitrary value of γ and $x=\bar{x}\in\partial\mathcal{M}$.

of the form $\mathbb{P}x+\sqrt{\gamma}Z\in\mathrm{C}(\mathrm{x},\gamma)^{\mathrm{c}}\cap\overline{\mathcal{M}}/\sqrt{\gamma}$[2] Using the original Theorem E.2-(d) it is possible to show that this quantity is bounded. However, if one uses the updated version of Theorem E.2-(d) then one needs to show that there exists $M\geq 0$ and $\bar{\gamma}>0$ such that for any $\gamma\in(0,\bar{\gamma})$ (here we have assumed that $\bar{\alpha}=0$, i.e. $x\in\partial\mathcal{M}$ for simplicity)

$$\int_0^{r/\gamma^{-1/2}}\int_{\nabla\Phi(\bar{x})^\perp}\mathbf{1}_{\|v\|^2\geq\alpha^2}\varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha\leq M\sqrt{\gamma},\tag{E.20}$$

which is absurd.

E.1.2. Technical lemmas. We start with a few technical lemmas which will allow us to control some Gaussian probabilities outside of a compact set. We denote $\Psi:\mathbb{R}_+\times\mathbb{N}\to\{0,1\}$ such that for any $k\in\mathbb{N}$, $\Psi(\cdot,k)$ is the tail probability of a χ-squared random variable with parameter k, i.e. for any $k\in\mathbb{N}$ and $t\geq 0$ we have

$$\Psi(t,k)=\mathbb{P}\|Z\|^2\geq t,\tag{E.21}$$

with Z a Gaussian random variable in $\mathbb{R}^k$ with zero mean and identity covariance matrix. We will make extensive use of the following lemma which is a direct consequence of (Laurent and Massart, 2000, Section 4, Lemma 1).

Lemma E.3. *For any $k\in\mathbb{N}$ and $t\in\mathbb{R}_+$ with $t\geq 5k$, $\Psi(t,k)\leq\exp[-t/5]$.*

Proof. Let $k\in\mathbb{N}$. First, note that for any $x\geq k$, we have that $k+2(kx)^{1/2}+2x\leq 5x$. Combining this result and (Laurent and Massart, 2000, Section 4, Lemma 1, Equation (4.3)), we have that for any $x\geq k$

$$\mathbb{P}\|X\|^2\geq 5x\leq\exp[-x],\tag{E.22}$$

with X a $\mathbb{R}^k$-valued Gaussian random variable with zero mean and identity covariance matrix. This concludes the proof upon letting $t=5x$. ∎

Let $\varphi:\mathbb{R}^p\to\mathbb{R}_+$ given for any $u\in\mathbb{R}$ by $\varphi(u)=(2\pi)^{-p/2}\exp[-\|u\|^2/2]$[3], i.e. the density of a real Gaussian random variable with zero mean and unit variance.

[2] The division by $\sqrt{\gamma}$ comes from the definition of the intermediate drift (E.53).

[3] In the rest of the supplementary, we never precise the dimension $p\in\mathbb{N}$ which can be deduced from the variable.

While Lemma E.4 appears technical, it will be central to provide quantitative upper bounds on the *rejection* probability, see Lemma E.7 for instance.

LEMMA E.4. *For any $k \in \mathbb{N}$, $\alpha_0 > 0$, $\beta_0 \in (0,1]$ and $\delta > 0$ we have*

$$\psi(\delta) = \sup\left\{\int_0^{+\infty} \Psi(\alpha_0 t/\delta, k)^{\beta_0} \varphi(t - t_0/\delta)\mathrm{d}t \ : \ t_0 \geq 0\right\} \leq C_0\delta, \tag{E.23}$$

with $C_0 = 5(2\pi)^{-1/2}(k+1)/(\alpha_0\beta_0)$.

Proof. Let $k \in \mathbb{N}$, $\alpha_0 > 0$, $\beta_0 \in (0,1]$ and $\delta > 0$. Let $t_\delta = 5k\delta/\alpha_0$. Note that if $t \geq t_\delta$ then, $\alpha_0 t/\delta \geq 5k$. In addition, we have

$$\begin{aligned}\int_0^{+\infty} \Psi(\alpha_0 t/\delta, k)^{\beta_0} \varphi(t - t_0/\delta)\mathrm{d}t &\leq (2\pi)^{-1/2} \int_0^{+\infty} \Psi(\alpha_0 t/\delta, k)^{\beta_0}\mathrm{d}t &\text{(E.24)}\\ &\leq (2\pi)^{-1/2} \int_0^{t_\delta} \Psi(\alpha_0 t/\delta, k)^{\beta_0}\mathrm{d}t &\text{(E.25)}\\ &+ (2\pi)^{-1/2} \int_{t_\delta}^{+\infty} \Psi(\alpha_0 t/\delta, k)^{\beta_0}\mathrm{d}t. &\text{(E.26)}\end{aligned}$$

Using that for any $w > 0$, $\int_0^{+\infty} \exp[-wt]\mathrm{d}t \leq (1/w)$, that for any $u \geq 0$, $\Psi(u,k) \leq 1$ and that if $u \geq 5k$, $\Psi(u,k) \leq \exp[-u/5]$, we get for any $t_0 \geq 0$

$$\int_0^{+\infty} \Psi(\alpha_0 t/\delta, k)\varphi(t - t_0/\delta) \leq (2\pi)^{-1/2}[5k\delta/\alpha_0 + 5\delta/(\alpha_0\beta_0)] \leq (5(2\pi)^{-1/2}(k+1)/(\alpha_0\beta_0))\delta, \tag{E.27}$$

which concludes the proof. ∎

Finally, we have the following lemma, which is similar to Lemma E.3 but will be used to control quantities related to the norm.

LEMMA E.5. *For any $k \in \mathbb{N}$, $\alpha_0 > 0$, $\beta_0 \in (0,1]$ and $\delta > 0$ we have*

$$\psi(\delta) = \int_0^{+\infty} \Psi(\alpha_0 t/\delta, k)^{\beta_0} t\varphi(t)\mathrm{d}t \leq C_0\delta^2, \tag{E.28}$$

with $C_0 = 25(2\pi)^{-1}(k^2+1)/(\alpha_0\beta_0)^2$.

Proof. Let $k \in \mathbb{N}$, $\alpha_0 > 0$, $\beta_0 \in (0,1]$ and $\delta > 0$. Let $t_\delta = 5k\delta/\alpha_0$. Note that if $t \geq t_\delta$ then, $\alpha_0 t/\delta \geq 5k$. In addition, we have

$$\int_0^{+\infty} \Psi(\alpha_0 t/\delta, k)^{\beta_0} t\varphi(t)\mathrm{d}t \leq (2\pi)^{-1}\int_0^{t_\delta} \Psi(\alpha_0 t/\delta, k)^{\beta_0} t\mathrm{d}t + (2\pi)^{-1}\int_{t_\delta}^{+\infty} \Psi(\alpha_0 t/\delta, k)^{\beta_0} t\mathrm{d}t. \tag{E.29}$$

In addition, using that if $u \geq 5k$ then $\Psi(u,k) \leq \exp[-u/5]$, we get

$$(2\pi)^{-1}\int_{t_\delta}^{+\infty} \Psi(\alpha_0 t/\delta, k)^{\beta_0} t\mathrm{d}t \leq (2\pi)^{-1}\int_0^{+\infty} \exp[-\alpha_0\beta_0 t/(5\delta)]t\mathrm{d}t \leq (2\pi)^{-1}25\delta^2/(\alpha_0\beta_0)^2. \tag{E.30}$$

Finally, using that for any $u \geq 0$, $\Psi(u,k) \leq 1$, we have

$$(2\pi)^{-1}\int_0^{t_\delta} \Psi(\alpha_0 t/\delta, k)^{\beta_0} t\mathrm{d}t \leq (2\pi)^{-1}25k^2\delta^2/\alpha_0^2, \tag{E.31}$$

which concludes the proof. ∎

E.1.3. Lower bound on the inside probability and control of moments of order two and higher.

Lower bound on the inside probability. We begin with the following lemma which controls the expectation of $1 + \|Z\|$ *outside* of $\mathrm{C}(\mathrm{x}, \gamma)$. We recall that V is defined in Theorem E.2-(c).

LEMMA E.6. *Let $\bar{\gamma} = 1$. Let $x \in \mathsf{V}$, $Z \in\sim \mathrm{N}(0, \mathbf{I})$ and $\gamma \in (0, \bar{\gamma})$ then we have*

$$\max(\mathbb{E}\Big[\mathbf{1}_{x+\sqrt{\gamma}Z\in\overline{\mathcal{M}}\cap\mathrm{C}(\mathrm{x},\gamma)^{\mathrm{c}}}\Big], \mathbb{E}\Big[\langle Z, \nabla\Phi(\bar{x})\rangle\mathbf{1}_{x+\sqrt{\gamma}Z\in\overline{\mathcal{M}}\cap\mathrm{C}(\mathrm{x},\gamma)^{\mathrm{c}}}\Big]) \leq \psi(\gamma), \tag{E.32}$$

with $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\limsup_{t\to 0} \psi(t)/t^{1/2} < +\infty$.

Proof. Let $\bar{r} > 0$ given by Theorem E.2. First, we have that

$$\int_{\mathbb{R}}\int_{\mathbb{R}^{d-1}}(1 + |\alpha| + \|v\|)\mathbf{1}_{\alpha\geq\bar{r}/\gamma^{1/2}}\varphi(\alpha)\varphi(v)\mathrm{d}\alpha\mathrm{d}v \tag{E.33}$$

$$\leq d\int_{\mathbb{R}}(1 + |\alpha|)\mathbf{1}_{\alpha\geq\bar{r}/\gamma^{1/2}}\varphi(\alpha)\mathrm{d}\alpha \leq d(\Psi(\bar{r}^2/\gamma, 1) + \exp[-\bar{r}^2/(2\gamma)]). \tag{E.34}$$

Second, using Lemma E.4, we have that

$$\int_{\mathbb{R}}\int_{\mathbb{R}^{d-1}}\mathbf{1}_{\|v\|^2\geq(\bar{\alpha}+\sqrt{\gamma}\alpha)/(C\gamma)}\mathbf{1}_{\bar{\alpha}+\sqrt{\gamma}\alpha\geq 0}\varphi(\alpha)\varphi(v)\mathrm{d}\alpha\mathrm{d}v \tag{E.35}$$

$$\leq \int_{\mathbb{R}}\mathbf{1}_{\bar{\alpha}+\sqrt{\gamma}\alpha\geq 0}\Psi((\bar{\alpha} + \sqrt{\gamma}\alpha)/(C\gamma), d-1)\varphi(\alpha)\mathrm{d}\alpha \tag{E.36}$$

$$\leq \int_0^{+\infty}\Psi(\alpha/C\gamma^{1/2}, d-1)\varphi(\alpha - \bar{\alpha}/\gamma^{1/2})\mathrm{d}\alpha \leq \Psi_1(\gamma^{1/2}). \tag{E.37}$$

Second, using Lemma E.5, we have that

$$\int_{\mathbb{R}}\int_{\mathbb{R}^{d-1}}\alpha\mathbf{1}_{\|v\|^2\geq(\bar{\alpha}+\sqrt{\gamma}\alpha)/(C\gamma)}\mathbf{1}_{\bar{\alpha}+\sqrt{\gamma}\alpha\geq 0}\varphi(\alpha)\varphi(v)\mathrm{d}\alpha\mathrm{d}v \tag{E.38}$$

$$= \int_{\mathbb{R}}\alpha\Psi((\bar{\alpha} + \sqrt{\gamma}\alpha)/(C\gamma), d-1)\mathbf{1}_{\bar{\alpha}+\sqrt{\gamma}\alpha\geq 0}\varphi(\alpha)\mathrm{d}\alpha \tag{E.39}$$

$$\leq \int_0^{+\infty}\Psi(\alpha/C\gamma^{1/2}, d-1)\alpha\varphi(\alpha)\mathrm{d}\alpha \leq \Psi_2(\gamma^{1/2}). \tag{E.40}$$

Note that we have $\limsup_{\gamma\to 0}\Psi_2(\gamma^{1/2})+\Psi_1(\gamma^{1/2}) < +\infty$. We conclude upon combining (E.34), (E.37) and (E.40) with Theorem E.2-(d) and the fact that $\|\Phi(\bar{x})\| = 1$. ∎

The following lemma allow us to give a lower bound to the quantity $\mathbb{E}\Big[\mathbf{1}_{x+\sqrt{\gamma}Z\in\overline{\mathcal{M}}}\Big]$ uniformly with respect to $x \in \overline{\mathcal{M}}$.

LEMMA E.7. *There exists $\bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma})$ and for any $x \in \overline{\mathcal{M}}$, $\gamma \in (0, \bar{\gamma})$ and $Z \sim \mathrm{N}(0, \mathbf{I})$ we have*

$$\mathbb{E}\Big[\mathbf{1}_{x+\sqrt{\gamma}Z\in\overline{\mathcal{M}}}\Big] \geq 1/4. \tag{E.41}$$

Proof. Let $\gamma \in (0, \bar{\gamma})$. If $x \notin \mathsf{V}$ then $\mathrm{B}(x, 2R) \subset \mathcal{M}$ using Theorem E.2-(e) and therefore $\mathbb{E}\Big[\mathbf{1}_{x+\sqrt{\gamma}Z\in\overline{\mathcal{M}}}\Big] \geq 1/4$ for $\bar{\gamma} > 0$ small enough. Now, assume that $x \in \mathsf{V}$. Using Lemma E.6, we have that $\mathbb{E}\Big[\mathbf{1}_{x+\sqrt{\gamma}Z\in\overline{\mathcal{M}}\cap\mathrm{C}(\mathrm{x},\gamma)^{\mathrm{c}}}\Big] \leq \psi(\gamma)$. In addition, using

Theorem E.2-(b), we have that for any $\gamma > 0$

$$\mathbb{E}\left[\mathbf{1}_{x+\sqrt{\gamma}Z\in\overline{\mathcal{M}}}\right] \geq \mathbb{E}\left[\mathbf{1}_{x+\sqrt{\gamma}Z\in \mathrm{C}(\mathrm{x},\gamma)}\right] \tag{E.42}$$

$$\geq \int_{-\bar{\alpha}\gamma^{-1/2}}^{r\gamma^{-1/2}} \int_{\nabla\Phi(\bar{x})^{\perp}} \mathbf{1}_{\|v\|^2\leq(\bar{\alpha}+\gamma^{1/2}\alpha)/(C\gamma)}\varphi(\alpha)\varphi(v)\mathrm{d}\alpha\mathrm{d}v \tag{E.43}$$

$$\geq \int_{-\bar{\alpha}\gamma^{-1/2}}^{r\gamma^{-1/2}} (1-\Psi((\bar{\alpha}+\gamma^{1/2}\alpha)/(C\gamma), d-1))\varphi(\alpha)\mathrm{d}\alpha \tag{E.44}$$

$$\geq (1/2) - \Psi(r^2/\gamma, 1) - \int_{-\bar{\alpha}\gamma^{-1/2}}^{+\infty} \Psi((\bar{\alpha}+\gamma^{1/2}\alpha)/(C\gamma), d-1)\varphi(\alpha)\mathrm{d}\alpha. \tag{E.45}$$

Hence, using Lemma E.3 and Lemma E.4, there exists $\bar{\gamma} > 0$ such that for any $\gamma \in (0,\bar{\gamma})$, $\Psi(r^2/\gamma,1)+\int_0^{+\infty}\Psi(\alpha/(C\gamma^{1/2}),d)\varphi(\alpha-\gamma^{1/2}\bar{\alpha})\mathrm{d}\alpha \leq 1/4$, which concludes the proof. ■

Note that the result of Lemma E.7 can be improved to $1/2-\varepsilon$ for any $\varepsilon > 0$. In particular this result tells us that for $\gamma > 0$ small enough, $\overline{\mathcal{M}}$ looks like the *hyperplane* from the point of view of the Gaussian with variance γ centred on $\partial\mathcal{M}$.

Bound on moments of order two and higher. In what follows, we define for any $\gamma > 0$, $\Delta^{\gamma} : \overline{\mathcal{M}} \to \mathbb{R}_+$ given for any $x \in \overline{\mathcal{M}}$ by

$$\Delta^{\gamma}(x) = (1/\gamma)\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\left\|\sqrt{\gamma}z\right\|^4\varphi(z)\mathrm{d}z / \int_{\mathbb{R}^d}\mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\varphi(z)\mathrm{d}z. \tag{E.46}$$

Proposition E.8. *We have* $\lim_{\gamma\to 0}\sup\left\{\Delta^{\gamma}(x) \ : \ x\in\overline{\mathcal{M}}\right\} = 0$.

Proof. Let $\bar{\gamma} > 0$ given by Lemma E.7. Let $x \in \overline{\mathcal{M}}$ and $\gamma \in (0,\bar{\gamma})$. We have using Lemma E.7

$$\int_{\mathbb{R}^d}\mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\varphi(z)\mathrm{d}z \geq 1/4. \tag{E.47}$$

We also have that

$$(1/\gamma)\int_{\mathbb{R}^d}\mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\left\|\sqrt{\gamma}z\right\|^4\varphi(z)\mathrm{d}z \leq 3\gamma d^2. \tag{E.48}$$

Therefore, we get that for any $\gamma \in (0,\bar{\gamma})$, $\Delta^{\gamma}(x) \leq 12\gamma d^2$, which concludes the proof. ■

In what follows, we define for any $\gamma > 0$, $\hat{\Sigma}^{\gamma} : \overline{\mathcal{M}} \to \mathrm{S}_d^+(\mathbb{R})$ given for any $x \in \overline{\mathcal{M}}$ by

$$\hat{\Sigma}^{\gamma}(x) = \int_{\mathbb{R}^d}\mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}} z\otimes z\varphi(z)\mathrm{d}z / \int_{\mathbb{R}^d}\mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\varphi(z)\mathrm{d}z. \tag{E.49}$$

Proposition E.9. *There exists* $\bar{\gamma} > 0$ *such that for any* $x \in \overline{\mathcal{M}}$ *and* $\gamma \in (0,\bar{\gamma})$ *we have*

$$\left\|\hat{\Sigma}^{\gamma}(x)\right\| \leq 4d. \tag{E.50}$$

Proof. Let $x \in \overline{\mathcal{M}}$ and $\bar{\gamma} > 0$ given by Lemma E.7. For any $\gamma \in (0,\bar{\gamma})$, we have using Lemma E.7

$$\int_{\mathbb{R}^d}\mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\varphi(z)\mathrm{d}z \geq 1/4. \tag{E.51}$$

We also have that

$$\int_{\mathbb{R}^d}\mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\|z\|^2\varphi(z)\mathrm{d}z \leq d, \tag{E.52}$$

which concludes the proof. ■

E.1.4. Properties of large drift terms. Finally, we define for any $\gamma > 0$, $\hat{b}^\gamma : \overline{\mathcal{M}} \to \mathbb{R}^d$ given for any $x \in \overline{\mathcal{M}}$ by

$$\hat{b}^\gamma(x) = \gamma^{-1/2} \int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} z\varphi(z)\mathrm{d}z / \int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} \varphi(z)\mathrm{d}z. \tag{E.53}$$

First, we show away from the boundary the drift $\hat{b}^\gamma$ converges to zero.

PROPOSITION E.10. *There exists $\bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma})$, $r > 0$ and $x \in \overline{\mathcal{M}}$ such that $d(x, \partial\mathcal{M}) \geq r$ we have $\left\|\hat{b}^\gamma(x)\right\| \leq 2d\Psi(r/\gamma, d)^{1/2}/\gamma^{1/2}$.*

Proof. Let $x \in \overline{\mathcal{M}}$ and $\bar{\gamma} > 0$ given by Lemma E.7. For any $\gamma \in (0, \bar{\gamma})$ we have using Lemma E.7

$$\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} \varphi(z)\mathrm{d}z \geq 1/4. \tag{E.54}$$

We also have that

$$\left\|\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} z\varphi(z)\mathrm{d}z\right\| \leq \left\|\int_{\mathbb{R}^d} \mathbf{1}_{\|z\| \leq r/\gamma^{1/2}} z\varphi(z)\mathrm{d}z\right\| + \int_{\mathbb{R}^d} \mathbf{1}_{\|z\| \geq r/\gamma^{1/2}} \|z\|\varphi(z)\mathrm{d}z \tag{E.55}$$

$$\leq 2\int_{\mathbb{R}^d} \mathbf{1}_{\|z\| \geq r/\gamma^{1/2}} \|z\|\varphi(z)\mathrm{d}z \leq 2d\Psi(r/\gamma, d)^{1/2}/\gamma^{1/2}, \tag{E.56}$$

which concludes the proof. ∎

We have the following corollary.

COROLLARY E.11. *There exists $\bar{\gamma} > 0$ such that for any $\delta > 0$ there exists $M_\delta > 0$ such that for any $\gamma \in (0, \bar{\gamma})$ and $x \in \overline{\mathcal{M}}$, $\left\|\hat{b}^\gamma(x)\right\| \geq M_\delta$, then $\Phi(x) \leq \delta$.*

Proof. Let $\bar{\gamma} > 0$ given by Lemma E.7. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ given for any $r > 0$ by $f(r) = \sup\left\{\gamma > 0 \ : \ \Psi(r/\gamma, 1)^{1/2}/\gamma^{1/2}\right\}$. We have that f is non-increasing and $\lim_{r \to 0} f(r) = +\infty$. Let $\delta > 0$ and $M_\delta = 2df(\delta/C)$ with $C = \sup\left\{\|\nabla\Phi(x)\| \ : \ x \in \overline{\mathcal{M}}\right\}$. Let $\gamma \in (0, \bar{\gamma})$ and $x \in \overline{\mathcal{M}}$ such that $\left\|\hat{b}^\gamma(x)\right\| \geq M_\delta$ then using Proposition E.10 we have that $d(x, \partial\mathcal{M}) \leq \delta/C$. Let $\bar{x} \in \partial\mathcal{M}$ such that $\|x - \bar{x}\| = d(x, \partial\mathcal{M})$. We have

$$\Phi(x) \leq \Phi(\bar{x}) + \int_0^1 \langle \nabla\Phi(\bar{x} + t(x - \bar{x})), x - \bar{x}\rangle \mathrm{d}t \leq \delta, \tag{E.57}$$

which concludes the proof. ∎

For ease of notation, for any $\gamma > 0$, we define $\bar{b}^\gamma = \gamma^{1/2}\hat{b}^\gamma$, the *renormalized* version of the drift. First, we have the following result which will ensure that the drift projected on the normal component does not vanish.

LEMMA E.12. *There exists $\bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma})$ and $x \in \mathsf{V}$ we have*

$$\langle \bar{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle \geq \left\|\bar{b}^\gamma(x)\right\| - \psi(\gamma), \tag{E.58}$$

with $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\limsup_{\gamma \to 0} \psi(\gamma)/\sqrt{\gamma} < +\infty$.

Proof. Let $x \in \overline{\mathcal{M}}$ and $\bar{\gamma} > 0$ given by Lemma E.7. For any $\gamma \in (0, \bar{\gamma})$ we have using Lemma E.7

$$\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} \varphi(z) \mathrm{d}z \geq 1/4. \tag{E.59}$$

In addition, we have

$$\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} \langle z, \nabla\Phi(\bar{x})\rangle \varphi(z)\mathrm{d}z \geq \int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, \nabla\Phi(\bar{x})\rangle \varphi(z)\mathrm{d}z \tag{E.60}$$

$$- \int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M} \cap \mathrm{C}(\mathrm{x},\gamma)^{\mathrm{c}}} \langle z, \nabla\Phi(\bar{x})\rangle \varphi(z). \tag{E.61}$$

Using Lemma E.6, we get that

$$\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} \langle z, \nabla\Phi(\bar{x})\rangle \varphi(z)\mathrm{d}z \geq \int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, \nabla\Phi(\bar{x})\rangle \varphi(z)\mathrm{d}z - \psi(\gamma). \tag{E.62}$$

Let $\{e_i\}_{i=1}^{d-1}$ a basis of $\nabla\Phi(\bar{x})^{\perp}$. Using Theorem E.2-(b), we have that for any $i \in \{1, \ldots, d-1\}$

$$\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, e_i\rangle \varphi(z)\mathrm{d}z = \int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^{\perp}} \mathbf{1}_{\|v\|^2 \leq (\gamma^{1/2}\alpha+\bar{\alpha})/\gamma} \langle v, e_i\rangle \varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha. \tag{E.63}$$

Hence, combining this result and the Cauchy-Schwarz inequality we have for any $i \in \{1, \ldots, d-1\}$

$$\Big(\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, e_i\rangle \varphi(z)\mathrm{d}z\Big)^2 = \Big(\int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^{\perp}} \mathbf{1}_{\|v\|^2 \geq (\gamma^{1/2}\alpha+\bar{\alpha})/\gamma} \langle v, e_i\rangle \varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha\Big)^2 \tag{E.64}$$

$$\leq \int_{\nabla\Phi(\bar{x})^{\perp}} \langle v, e_i\rangle^2 \varphi(v)\mathrm{d}v \Big(\int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \Psi((\bar{\alpha}+\alpha\gamma^{1/2})/\gamma, d-1)^{1/2} \varphi(\alpha)\mathrm{d}\alpha\Big)^2 \tag{E.65}$$

$$\leq \Big(\int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \Psi((\bar{\alpha}+\alpha\gamma^{1/2})/\gamma, d-1)^{1/2} \varphi(\alpha)\mathrm{d}\alpha\Big)^2. \tag{E.66}$$

Hence, using Lemma E.4, we get that

$$\sum_{i=1}^{d-1} \Big(\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, e_i\rangle \varphi(z)\mathrm{d}z\Big)^2 \leq (d-1)\psi^2(\gamma), \tag{E.67}$$

with ψ given by Lemma E.4 with $\beta_0 = 1/2$. Therefore, we get that

$$\Big(\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, \nabla\Phi(\bar{z})\rangle \varphi(z)\mathrm{d}z\Big)^2 \tag{E.68}$$

$$= \Big(\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} \varphi(z)\mathrm{d}z\Big)^2 \Big\|\bar{b}^{\gamma}(x)\Big\|^2 - \sum_{i=1}^{d-1} \Big(\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, e_i\rangle \varphi(z)\mathrm{d}z\Big)^2 \tag{E.69}$$

$$\geq \Big(\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathcal{M}} \varphi(z)\mathrm{d}z\Big)^2 \Big\|\bar{b}^{\gamma}(x)\Big\|^2 - \psi(\gamma)^2. \tag{E.70}$$

We conclude the proof upon using that for any $a, b \geq 0$, $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$ and (E.59). ∎

We are now ready to state the following lower bound on the drift.

PROPOSITION E.13. *There exist $\bar{\gamma} > 0$, $M \geq 0$ and $c > 0$ such that for any $x \in \overline{\mathcal{M}}$ and $\gamma \in (0, \bar{\gamma})$ if $\Big\|\hat{b}^{\gamma}(x)\Big\| \geq M$ then $x \in \mathsf{V}$ and*

$$\min(\langle \hat{b}^{\gamma}(x), \nabla\Phi(x)\rangle, \langle \hat{b}^{\gamma}(x), \nabla\Phi(\bar{x})\rangle) \geq c\Big\|\hat{b}^{\gamma}(x)\Big\|. \tag{E.71}$$

Proof. Let $\bar{\gamma} > 0$ given by Lemma E.7 and $M_0 = 4\sup\{\psi(\gamma)/\gamma^{1/2} \ : \ \gamma \in (0, \bar{\gamma}]\}$. In addition, let $c = 1/4$. Using Proposition E.10 and Theorem E.2-(e), there exists $M_1 \geq 0$ such that for any any $x \in \overline{\mathcal{M}}$, if $\left\|\hat{b}^\gamma(x)\right\| \geq M_1$ then $x \in \mathsf{V}$ and $x = \bar{x} + \alpha\nabla\Phi(\bar{x})$ with $\alpha \leq 1/(4C)$ and $C = \sup\left\{\left\|\nabla^2\Phi(x)\right\| \ : \ x \in \overline{\mathcal{M}}\right\}$. We denote $M = \max(M_0, M_1)$. Let $\gamma \in (0, \bar{\gamma})$ and $x \in \overline{\mathcal{M}}$ such that $\left\|\hat{b}^\gamma(x)\right\| \geq M$. Using Lemma E.12, we have that

$$\langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle \geq \left\|\hat{b}^\gamma(x)\right\| - \psi(\gamma)/\gamma^{1/2}. \tag{E.72}$$

Using that $\psi(\gamma)/\gamma^{1/2} \leq M/2 \leq \left\|\hat{b}^\gamma(x)\right\|/2$, we have

$$\langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle \geq (1/2)\left\|\hat{b}^\gamma(x)\right\|. \tag{E.73}$$

Since $\|x - \bar{x}\| \leq \alpha \leq 1/(4C)$ we have $\langle \hat{b}^\gamma(x), \nabla\Phi(x)\rangle \geq (1/2 - C\alpha)\left\|\hat{b}^\gamma(x)\right\| \geq \left\|\hat{b}^\gamma(x)\right\|/4$, which concludes the proof. ∎

E.1.5. Convergence on compact sets. In this section, we show the convergence of the drift and diffusion matrix on compact sets. We recall that $\mathcal{M}$ does *not* include its boundary $\partial\mathcal{M}$.

Proposition E.14. *For any compact set $\mathsf{K} \subset \mathcal{M}$ and $\varepsilon > 0$, there exists $\bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma})$ we have for any $x \in \mathsf{K}$*

$$\left\|\hat{b}^\gamma(x)\right\| \leq \varepsilon, \qquad \left\|\hat{\Sigma}^\gamma(x) - \mathbf{I}\right\| \leq \varepsilon. \tag{E.74}$$

Proof. Let $\mathsf{K} \subset \mathcal{M}$ be a compact set and $\gamma > 0$. Since $\mathsf{K} \cap \partial\mathcal{M} = \varnothing$, there exists $r > 0$ such that for any $x \in \mathsf{K}$, $d(x, \partial\mathcal{M}) > r$. Therefore, we have that for any $x \in \mathsf{K}$

$$\left\|\hat{b}^\gamma(x)\right\| = \gamma^{-1/2}\left\|\textstyle\int_{x+\sqrt{\gamma}z\in\mathcal{M}} z\varphi(z)\mathrm{d}z\right\| / \int_{x+\sqrt{\gamma}z\in\mathcal{M}} \varphi(z)\mathrm{d}z. \tag{E.75}$$

In addition, using the Cauchy-Schwarz inequality we have

$$\left\|\textstyle\int_{x+\sqrt{\gamma}z\in\mathcal{M}} z\varphi(z)\mathrm{d}z\right\| \leq \left\|\textstyle\int_{\mathbb{R}^d} z\varphi(z)\mathrm{d}z\right\| + \int_{\mathcal{M}^\mathrm{c}} \|z\|\varphi(z)\mathrm{d}z \tag{E.76}$$

$$\leq \textstyle\int_{\mathbb{R}^d} \mathbf{1}_{\|z\|\geq r/\gamma^{1/2}}\|z\|\varphi(z)\mathrm{d}z \leq \sqrt{d}\Psi(r^2/\gamma, d)^{1/2}. \tag{E.77}$$

Using Lemma E.3 and Lemma E.7, there exists $\bar{\gamma}_0 > 0$ such that for any $\gamma \in (0, \bar{\gamma}_0)$ we have that for any $x \in \mathsf{K}$

$$\left\|\hat{b}^\gamma(x)\right\| \leq 4d\Psi(r^2/\gamma, 1)^{1/2}/\gamma^{1/2} \leq \varepsilon, \tag{E.78}$$

which concludes the first part of the proof. Similarly, we have that for any $x \in \mathsf{K}$

$$\left\|\textstyle\int_{x+\sqrt{\gamma}z\in\mathcal{M}} (z\otimes z - \mathbf{I})\varphi(z)\mathrm{d}z\right\| \leq \left\|\textstyle\int_{\mathbb{R}^d} (z\otimes z - \mathbf{I})\varphi(z)\mathrm{d}z\right\| + \int_{\mathcal{M}^\mathrm{c}} \|z\|\varphi(z)\mathrm{d}z \tag{E.79}$$

$$\leq \textstyle\int_{\mathbb{R}^d} \mathbf{1}_{\|z\|\geq r/\gamma^{1/2}}\|z\otimes z - \mathbf{I}\|\varphi(z)\mathrm{d}z \tag{E.80}$$

$$\leq \sqrt{2}(1 + 3d^2)^{1/2}\Psi(r^2/\gamma, d)^{1/2}. \tag{E.81}$$

Using Lemma E.3 and Lemma E.7, there exists $\bar{\gamma}_1 > 0$ such that for any $\gamma \in (0, \bar{\gamma}_1)$, we have that for any $x \in \mathsf{K}$

$$\left\|\hat{\Sigma}^\gamma(x) - \mathbf{I}\right\| \leq 4\sqrt{2}(1+3d^2)^{1/2}\Psi(r^2/\gamma, 1)^{1/2} \leq \varepsilon, \tag{E.82}$$

which concludes the proof upon letting $\bar{\gamma} = \min(\bar{\gamma}_0, \bar{\gamma}_1)$. ∎

E.1.6. Convergence on the boundary. Finally, we investigate the behaviour at the boundary of the diffusion matrix and the drift. First, we show that there is a lower bound to the diffusion matrix near the boundary. Second, we show that the renormalized drift converges to the outward normal.

PROPOSITION E.15. *There exist $c > 0$ and $\bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma})$, $u \in \mathbb{R}^d$ and $x \in \mathsf{V}$ we have*

$$\langle u, \hat{\Sigma}^\gamma(x)u\rangle \geq c\|u\|^2. \tag{E.83}$$

In particular, there exist $r, \varepsilon > 0$ such that for any $\gamma \in (0, \bar{\gamma})$ and $x \in \overline{\mathcal{M}}$ with $d(x, \partial\mathcal{M}) \leq r$

$$\langle \nabla\Phi(x), \hat{\Sigma}^\gamma(x)\nabla\Phi(x)\rangle \geq \varepsilon. \tag{E.84}$$

Proof. First, we show (E.83). Let $x \in \mathsf{V}$. We have for any $u \in \mathbb{R}^d$

$$\langle u, \hat{\Sigma}^\gamma(x)u\rangle = \int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\langle z, u\rangle^2\varphi(z)\mathrm{d}z / \int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z\in\mathcal{M}}\mathrm{d}z \tag{E.85}$$

$$\geq \int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z\in\mathsf{C}(\mathsf{x},\gamma)}\langle z, u\rangle^2\varphi(z)\mathrm{d}z. \tag{E.86}$$

For any $u \in \mathbb{R}^d$, let $\alpha_u = \langle u, \nabla\Phi(\bar{x})\rangle$. Using Theorem E.2-(b) we have for any $u \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z\in\mathsf{C}(\mathsf{x},\gamma)}\langle z, u\rangle^2\varphi(z)\mathrm{d}z \tag{E.87}$$

$$= \int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^\perp} (\langle u, v\rangle + \alpha\alpha_u)^2 \mathbf{1}_{\|v\|^2\leq(\alpha\gamma^{1/2}+\bar{\alpha})/\gamma}\varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha \tag{E.88}$$

$$\geq \int_0^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^\perp} (\langle u, v\rangle^2 + \alpha^2\alpha_u^2) \mathbf{1}_{\|v\|^2\leq(\alpha\gamma^{1/2}+\bar{\alpha})/\gamma}\varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha \tag{E.89}$$

$$\geq \alpha_u^2 \int_0^{r/\gamma^{1/2}} \alpha^2\varphi(\alpha)\mathrm{d}\alpha + \int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^\perp} \langle u, v\rangle^2 \mathbf{1}_{\|v\|^2\leq(\alpha\gamma^{1/2}+\bar{\alpha})/\gamma}\varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha. \tag{E.90}$$

Using Cauchy-Schwarz inequality, we have

$$\int_0^{r/\gamma^{1/2}} \alpha^2\varphi(\alpha)\mathrm{d}\alpha = (1/2) - \int_{r/\gamma^{1/2}}^{+\infty} \alpha^2\varphi(\alpha)\mathrm{d}\alpha \geq (1/2) - 3\Phi(r^2/\gamma, 1)^{1/2}. \tag{E.91}$$

In addition, using the Cauchy-Schwarz inequality, we have that

$$\int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^\perp} \langle u, v\rangle^2 \mathbf{1}_{\|v\|^2\leq(\alpha\gamma^{1/2}+\bar{\alpha})/\gamma}\varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha \tag{E.92}$$

$$= \int_{\nabla\Phi(\bar{x})^\perp} \langle u, v\rangle^2\varphi(v)\mathrm{d}v \int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \varphi(\alpha)\mathrm{d}\alpha \tag{E.93}$$

$$- \int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^\perp} \langle u, v\rangle^2 \mathbf{1}_{\|v\|^2\geq(\alpha\gamma^{1/2}+\bar{\alpha})/\gamma}\varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha \tag{E.94}$$

$$\geq (\|u\|^2 - \alpha_u^2)((1/2) - \Phi(r^2/\gamma, 1)) \tag{E.95}$$

$$- \sqrt{3}(d-1)\|u\|^2 \int_0^{+\infty} \Phi(\alpha/\gamma^{1/2}, d-1)^{1/2}\varphi(\alpha - \bar{\alpha}/\gamma^{1/2})\mathrm{d}\alpha. \tag{E.96}$$

Combining this result, (E.91), (E.90) and Lemma E.4 there exists $\bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma}]$ and $u \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, u\rangle^2 \varphi(z) \mathrm{d}z \geq (1/4)\|u\|^2, \tag{E.97}$$

which concludes the proof of (E.83). Finally, using Theorem E.2-(e), we have that for any $x \in \overline{\mathcal{M}}$ if $d(x, \partial\mathcal{M}) \leq R$ then $x \in \mathsf{V}$. Let $r = \min(R, 1/(2C))$ with $C = \sup\left\{\left\|\nabla^2\Phi(x)\right\| \, : \, x \in \overline{\mathcal{M}}\right\}$. We have that for any $x \in \overline{\mathcal{M}}$ such that $d(x, \partial\mathcal{M}) \leq r$

$$\|\nabla\Phi(x)\| \geq \|\nabla\Phi(\bar{x}_0)\| - Cr \geq 1/2, \tag{E.98}$$

where $\bar{x}_0$ is such that $\|x - \bar{x}_0\| \leq r$ and $\bar{x}_0 \in \partial\mathcal{M}$. Combining this result and (E.97) concludes the proof upon letting $\varepsilon = 1/16$. ∎

Finally, we investigate the behaviour of the normalized drift near the boundary.

Proposition E.16. *For any $\bar{x}_0 \in \partial\mathcal{M}$ and $\varepsilon > 0$, there exist $\bar{\gamma}, r, M > 0$ such that for any $x \in \overline{\mathcal{M}}$ and $\gamma \in (0, \bar{\gamma})$ with $\|x - \bar{x}_0\| \leq r$ and $\left\|\hat{b}^\gamma(x)\right\| \geq M$*

$$\left\|\hat{b}^\gamma(x)/\langle\hat{b}^\gamma(x), \nabla\Phi(x)\rangle - \nabla\Phi(\bar{x}_0)\right\| \leq \varepsilon. \tag{E.99}$$

Proof. Let $\bar{\gamma}$ be given by Proposition E.13. Let ψ given by Lemma E.4 and $M_0 = \sup\left\{\psi(\gamma)/\gamma^{1/2} \, : \, \gamma \in (0, \bar{\gamma})\right\} < +\infty$. Let $M = 16M_0/(c\varepsilon^{1/2})$ with c given in Proposition E.13. Let $R > 0$ given by Theorem E.2-(e) such that for any $x \in \overline{\mathcal{M}}$ with $d(x, \partial\mathcal{M})$ there exist $\bar{x} \in \partial\mathcal{M}$ and $\alpha \in \{0, c\varepsilon/(4C)\}$ such that $x = \bar{x} + \alpha\nabla\Phi(\bar{x})$ with $C = \sup\left\{\left\|\nabla^2\Phi(x)\right\| \, : \, x \in \overline{\mathcal{M}}\right\}$ and c given in Proposition E.13. Let $r = \min(\bar{r}, c\varepsilon/4, R)$ and $x \in \overline{M}$ with $\|x - \bar{x}_0\| \leq r$. First, since $d(x, \partial\mathcal{M}) \leq R$, there exist $\bar{x} \in \partial\mathcal{M}$ and $\alpha \in \{0, \varepsilon/(4C)\}$ such that $x = \bar{x} + \alpha\nabla\Phi(\bar{x})$. Therefore, we get that $\|\bar{x} - \bar{x}_0\| \leq \varepsilon/(2C)$ and therefore $\|\nabla\Phi(\bar{x}_0) - \nabla\Phi(\bar{x})\| \leq \varepsilon/2$. In addition, we have that

$$\left\|\hat{b}^\gamma(x)/\langle\hat{b}^\gamma(x), \nabla\Phi(x)\rangle - \hat{b}^\gamma(x)/\langle\hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle\right\| \tag{E.100}$$

$$\leq \left\|\hat{b}^\gamma(x)\right\|^2 \|\nabla\Phi(x) - \nabla\Phi(\bar{x})\|/(\langle\hat{b}^\gamma(x), \nabla\Phi(x)\rangle\langle\hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle). \tag{E.101}$$

Using Proposition E.13, we get that

$$\left\|\hat{b}^\gamma(x)/\langle\hat{b}^\gamma(x), \nabla\Phi(x)\rangle - \hat{b}^\gamma(x)/\langle\hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle\right\| \leq \varepsilon/4. \tag{E.102}$$

In what follows, we show that

$$\left\|\hat{b}^\gamma(x)/\langle\hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle - \nabla\Phi(\bar{x})\right\|^2 \leq \varepsilon/2. \tag{E.103}$$

In particular, we show that for any $u \in \nabla\Phi(\bar{x})^\perp$ with $\|u\| = 1$,

$$\langle\hat{b}^\gamma(x), u\rangle^2 \leq (\varepsilon/16)\langle\hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle^2. \tag{E.104}$$

Assuming (E.104), letting $u = (\hat{b}^\gamma(x) - \langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle)/(\left\|\hat{b}^\gamma(x)\right\|^2 - \langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle^2)^{1/2}$ and using that $\hat{b}^\gamma(x) = \langle \hat{b}^\gamma(x), u\rangle u + \langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle \nabla\Phi(\bar{x})$ we have

$$\left\|\hat{b}^\gamma(x)/\langle \hat{b}^\gamma(x), \nabla\Phi(x)\rangle - \nabla\Phi(\bar{x})\right\| \leq \left\|\hat{b}^\gamma(x)/\langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle - \nabla\Phi(\bar{x})\right\| \tag{E.105}$$

$$+ \left\|\hat{b}^\gamma(x)/\langle \hat{b}^\gamma(x), \nabla\Phi(x)\rangle - \hat{b}^\gamma(x)/\langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle\right\| \tag{E.106}$$

$$\leq \left|\langle \hat{b}^\gamma(x), u\rangle/\langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle\right| + \varepsilon/4 \leq \varepsilon/2, \tag{E.107}$$

which concludes the proof. Let $u \in \nabla\Phi(\bar{x})^\perp$ with $\|u\| = 1$ and $\{e_i\}_{i=1}^{d-1}$ an orthonormal basis of $\nabla\Phi(\bar{x})^\perp$. There exist $\{a_i\}_{i=1}^{d-1}$ such that $\sum_{i=1}^{d-1} a_i^2 = 1$ and $u = \sum_{i=1}^{d-1} a_i e_i$. Using Theorem E.2-(b), we have that for any $i \in \{1, \ldots, d-1\}$

$$\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, e_i\rangle \varphi(z)\mathrm{d}z = \int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^\perp} \mathbf{1}_{\|v\|^2 \leq (\gamma^{1/2}\alpha+\bar{\alpha})/\gamma} \langle v, e_i\rangle \varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha \tag{E.108}$$

$$= \int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^\perp} \mathbf{1}_{\|v\|^2 \geq (\gamma^{1/2}\alpha+\bar{\alpha})/\gamma} \langle v, e_i\rangle \varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha \tag{E.109}$$

Hence, combining this result and the Cauchy-Schwarz inequality we have for any $i \in \{1, \ldots, d-1\}$

$$\left(\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, e_i\rangle \varphi(z)\mathrm{d}z\right)^2 = \left(\int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \int_{\nabla\Phi(\bar{x})^\perp} \mathbf{1}_{\|v\|^2 \geq (\gamma^{1/2}\alpha+\bar{\alpha})/\gamma} \langle v, e_i\rangle \varphi(v)\varphi(\alpha)\mathrm{d}v\mathrm{d}\alpha\right)^2 \tag{E.110}$$

$$\leq \int_{\nabla\Phi(\bar{x})^\perp} \langle v, e_i\rangle^2 \varphi(v)\mathrm{d}v \left(\int_{-\bar{\alpha}/\gamma^{1/2}}^{r/\gamma^{1/2}} \Psi((\bar{\alpha} + \alpha\gamma^{1/2})/\gamma, d-1)^{1/2}\varphi(\alpha)\mathrm{d}\alpha\right)^2. \tag{E.111}$$

Hence, we get that

$$\sum_{i=1}^{d-1} a_i^2 \left(\int_{\mathbb{R}^d} \mathbf{1}_{x+\sqrt{\gamma}z \in \mathrm{C}(\mathrm{x},\gamma)} \langle z, e_i\rangle \varphi(z)\mathrm{d}z\right)^2 \leq \|u\|^2 \psi^2(\gamma), \tag{E.112}$$

with ψ given by Lemma E.4. Recalling that $\left\|\hat{b}^\gamma(x)\right\| \geq M$ we have

$$\langle \hat{b}^\gamma(x), u\rangle^2 \leq 16\psi(\gamma)^2/\gamma \leq c^2(\varepsilon/16)M^2 \leq (\varepsilon/16)\langle \hat{b}^\gamma(x), \nabla\Phi(\bar{x})\rangle^2, \tag{E.113}$$

which concludes the proof. ∎

E.1.7. Submartingale problem and weak solution. We are now ready to conclude the proof.

THEOREM E.17. *There exists $\mathbb{P}^\star$ a distribution on $\mathrm{D}(\{0,T\}, \overline{\mathcal{M}})$ such that $\lim_{\gamma\to 0} \hat{\mathbb{P}}^\gamma = \mathbb{P}^\star$. In addition, for any $f \in \mathcal{C}^{1,2}(\{0,T\} \times \overline{\mathcal{M}}, \mathbb{R})$ with $\langle \nabla\Phi(\bar{x}), \nabla f(x)\rangle \geq 0$ for any $t \in \{0,T\}$ and $x \in \partial\mathcal{M}$, we have that the process $(f(t,\omega(t)))_{t\in\{0,T\}}$ given for any $t \in \{0,T\}$*

$$f(t,\omega(t)) - \int_0^t (\partial_s f(s,\omega(s)) + \tfrac{1}{2}\Delta f(s,\omega(s)))\mathbf{1}_{\mathcal{M}}(\omega(s))\mathrm{d}s, \tag{E.114}$$

is a $\mathbb{P}$ submartingale.

Proof. Condition (A) (Stroock and Varadhan, 1971, p.197) is a consequence of Proposition E.8. Condition (B) (Stroock and Varadhan, 1971, p.197) is a consequence of Proposition E.13. Condition (C) (Stroock and Varadhan, 1971, p.198) is a consequence of Corollary E.11. Condition (D) (Stroock and Varadhan, 1971, p.198) is a consequence of Proposition E.9. We fix $\rho = 0$ and condition (1) (Stroock and Varadhan, 1971, p.203) is a consequence of Proposition E.14. Condition (2)-(iii) (Stroock and Varadhan, 1971, p.203) is a consequence of Proposition E.15. Condition (2)-(iv) (Stroock and Varadhan, 1971, p.203) is a consequence of Proposition E.16. We conclude upon using (Stroock and Varadhan, 1971, Theorem 6.3) and (Stroock and Varadhan, 1971, Theorem 5.8). ∎

We finally conclude the proof of Theorem E.1 upon using the results of (Kang and Ramanan, 2017) which establish the link between a weak solution to the reflected SDE and the solution to a submartingale problem.

THEOREM E.18. *For any $T \geq 0$, $(\hat{\boldsymbol{X}}^\gamma_t)_{t\in\{0,T\}}$ weakly converges to $(\boldsymbol{X}_t)_{t\in\{0,T\}}$ such that for any $t \in \{0, T\}$*

$$\boldsymbol{X}_t = x + \boldsymbol{B}_t - \boldsymbol{k}_t, \qquad |\boldsymbol{k}|_t = \int_0^t \mathbf{1}_{\boldsymbol{X}_s\in\partial\mathcal{M}} \mathrm{d}|\boldsymbol{k}|_s, \qquad \boldsymbol{k}_t = \int_0^t \boldsymbol{n}(\boldsymbol{X}_s)\mathrm{d}|\boldsymbol{k}|_s. \quad \text{(E.115)}$$

Proof. Using Theorem E.17 and (Kang and Ramanan, 2017, Theorem 1, Proposition 2.12), we have that $\mathbb{P}$ in Theorem E.17is associated with a solution to the extended Skorokhod problem. We conclude that a solution to the extended Skorokhod problem is a solution to the Skorokhod problem using (Ramanan, 2006, Corollary 2.10). ∎

E.1.8. Extension to the Metropolis process. We recall that the Metropolis process is defined as follows. Let $(X^\gamma_k)_{k\in\mathbb{N}}$ given for any $\gamma > 0$ and $k \in \mathbb{N}$ by $X^\gamma_0 = x \in \overline{\mathcal{M}}$ and for $X^\gamma_{k+1} = X^\gamma_k + \sqrt{\gamma}Z_k$ if $X^\gamma_k + \sqrt{\gamma}Z^\gamma_k \in \overline{\mathcal{M}}$ and X^γ_k otherwise, $Z_k \sim \mathrm{N}(0, \mathbf{I})$. We recall that $\hat{b}^\gamma$, $\hat{\Sigma}^\gamma$ and $\hat{\Delta}^\gamma$ are given by (E.46), (E.49) and (E.53). In particular, denoting $\hat{\mathrm{K}}^\gamma$ the Markov kernel associated with $(\hat{X}^\gamma_k)_{k\in\mathbb{N}}$, i.e. $\hat{\mathrm{K}}^\gamma : \mathcal{M} \times \mathcal{B}(\mathcal{M}) \to \{0, 1\}$ such that for any $x \in \mathcal{M}$, $\hat{\mathrm{K}}^\gamma(x, \cdot)$ is a probability measure, for any $\mathsf{A} \in \mathcal{B}(\mathcal{M})$, $\hat{\mathrm{K}}^\gamma(\cdot, \mathsf{A})$ is a measurable function and $\mathbb{E}\left[\mathbf{1}_{\mathsf{A}}(\hat{X}^\gamma_1) \mid \hat{X}^\gamma_0 = x\right] = \hat{\mathrm{K}}^\gamma(x, \mathsf{A})$. We have that for any $\gamma > 0$ and $x \in \mathcal{M}$

$$\hat{b}^\gamma(x) = (1/\gamma)\int_{\mathcal{M}}(y - x)\hat{\mathrm{K}}^\gamma(x, \mathrm{d}y), \quad \text{(E.116)}$$

$$\hat{\Sigma}^\gamma(x) = (1/\gamma)\int_{\mathcal{M}}(y - x)^{\otimes 2}\hat{\mathrm{K}}^\gamma(x, \mathrm{d}y), \quad \text{(E.117)}$$

$$\hat{\Delta}^\gamma(x) = (1/\gamma)\int_{\mathcal{M}}\|y - x\|^4\hat{\mathrm{K}}^\gamma(x, \mathrm{d}y). \quad \text{(E.118)}$$

In what follows, we denote $a^\gamma(x) = \mathbb{E}\left[\mathbf{1}_{x+\sqrt{\gamma}Z_0\in\mathcal{M}}\right]$. Denote K^γ the kernel associated with $(X^\gamma_k)_{k\in\mathbb{N}}$. We have that for any $\mathsf{A} \in \mathcal{B}(\mathcal{M})$, $\gamma > 0$ and $x \in \mathcal{M}$

$$\mathrm{K}^\gamma(x, \mathsf{A}) = \mathbb{E}\left[\mathbf{1}_{X^\gamma_{k+1}\in\mathsf{A}}\mathbf{1}_{x+\sqrt{\gamma}Z_{k+1}\in\mathcal{M}}\right] + (1 - a^\gamma(x))\mathbf{1}_{\mathsf{A}}(x) \quad \text{(E.119)}$$

$$= a^\gamma(x)\hat{\mathrm{K}}^\gamma(x, \mathsf{A}) + (1 - \mathrm{a}^\gamma(\mathrm{x}))1_{\mathsf{A}}(\mathrm{x}). \quad \text{(E.120)}$$

We define for any $\gamma > 0$ and $x \in \mathcal{M}$

$$b^\gamma(x) = (1/\gamma) \int_{\mathcal{M}} (y - x) \mathrm{K}^\gamma(x, \mathrm{d}y), \tag{E.121}$$

$$\Sigma^\gamma(x) = (1/\gamma) \int_{\mathcal{M}} (y - x)^{\otimes 2} \mathrm{K}^\gamma(x, \mathrm{d}y), \tag{E.122}$$

$$\Delta^\gamma(x) = (1/\gamma) \int_{\mathcal{M}} \|y - x\|^4 \mathrm{K}^\gamma(x, \mathrm{d}y). \tag{E.123}$$

Using (E.120), we get that for any $\gamma > 0$ and $x \in \mathcal{M}$

$$b^\gamma(x) = a^\gamma(x) \hat{b}^\gamma(x), \qquad \Sigma^\gamma(x) = a^\gamma(x) \hat{\Sigma}^\gamma(x), \qquad \Delta^\gamma(x) = a^\gamma(x) \hat{\Delta}^\gamma(x). \tag{E.124}$$

Using Lemma E.7, we have that for any $\gamma \in (0, \bar{\gamma})$ and $x \in \mathcal{M}$, $a^\gamma(x) \geq 1/4$.

In order to conclude for the convergence of the Metropolis process we adapt Theorem E.17 and Theorem E.18. We define $\boldsymbol{X}^\gamma : \mathbb{R}_+ \to \overline{\mathcal{M}}$ given for any $k \in \mathbb{N}$ by $\boldsymbol{X}^\gamma_{k\gamma} = X^\gamma_k$ and for any $t \in [k\gamma, (k+1)\gamma)$, $\boldsymbol{X}^\gamma_t = X^\gamma_k$. Note that $(\boldsymbol{X}_t)_{t \in \{0,T\}}$ is a $\mathrm{D}(\{0, T\}, \overline{\mathcal{M}})$ valued random variable, where $\mathrm{D}(\{0, T\}, \overline{\mathcal{M}})$ is the space of right-continuous with left-limit processes which take values in $\overline{\mathcal{M}}$. We denote $\mathbb{P}^\gamma$ the distribution of $(\boldsymbol{X}_t)_{t \in \{0,T\}}$ on $\mathrm{D}(\{0, T\}, \overline{\mathcal{M}})$.

THEOREM E.19. *There exists $\mathbb{P}^\star$ a distribution on $\mathrm{D}(\{0, T\}, \overline{\mathcal{M}})$ such that $\lim_{\gamma \to 0} \mathbb{P}^\gamma = \mathbb{P}^\star$. In addition, for any $f \in \mathcal{C}^{1,2}(\{0, T\} \times \overline{\mathcal{M}}, \mathbb{R})$ with $\langle \nabla \Phi(\bar{x}), \nabla f(x) \rangle \geq 0$ for any $t \in \{0, T\}$ and $x \in \partial\mathcal{M}$, we have that the process $(f(t, \omega(t)))_{t \in \{0,T\}}$ given for any $t \in \{0, T\}$*

$$f(t, \omega(t)) - \int_0^t (\partial_s f(s, \omega(s)) + \tfrac{1}{2} \Delta f(s, \omega(s))) \mathbf{1}_{\mathcal{M}}(\omega(s)) \mathrm{d}s, \tag{E.125}$$

is a $\mathbb{P}$ submartingale.

Proof. Condition (A) (Stroock and Varadhan, 1971, p.197) is a consequence of Proposition E.8 and (E.124). Condition (B) (Stroock and Varadhan, 1971, p.197) is a consequence of Proposition E.13 and (E.124). Condition (C) (Stroock and Varadhan, 1971, p.198) is a consequence of Corollary E.11 and (E.124). Condition (D) (Stroock and Varadhan, 1971, p.198) is a consequence of Proposition E.9 and (E.124). We fix $\rho = 0$ and condition (1) (Stroock and Varadhan, 1971, p.203) is a consequence of Proposition E.14 and that $\lim_{\gamma \to 0} a^\gamma = 1$ uniformly on compact subsets $\mathrm{K} \subset \mathcal{M}$. Condition (2)-(iii) (Stroock and Varadhan, 1971, p.203) is a consequence of Proposition E.15 and (E.124). Condition (2)-(iv) (Stroock and Varadhan, 1971, p.203) is a consequence of Proposition E.16 and (E.124). We conclude upon using (Stroock and Varadhan, 1971, Theorem 6.3) and (Stroock and Varadhan, 1971, Theorem 5.8). ∎

THEOREM E.20. *For any $T \geq 0$, $(\boldsymbol{X}^\gamma_t)_{t \in \{0,T\}}$ weakly converges to $(\boldsymbol{X}_t)_{t \in \{0,T\}}$ such that for any $t \in \{0, T\}$*

$$\boldsymbol{X}_t = x + \boldsymbol{B}_t - \boldsymbol{k}_t, \qquad |\boldsymbol{k}|_t = \int_0^t \mathbf{1}_{X_s \in \partial\mathcal{M}} \mathrm{d}|\boldsymbol{k}|_s, \qquad \boldsymbol{k}_t = \int_0^t \boldsymbol{n}(\boldsymbol{X}_s) \mathrm{d}|\boldsymbol{k}|_s. \tag{E.126}$$

Proof. The proof is identical to Theorem E.18. ∎

E.2. Modelling geospatial data within non-convex boundaries

To demonstrate the ability of the proposed method to model distributions whose support is restricted to manifolds with highly non-convex boundaries, we derived a geospatial dataset based on the historical wildfire incidence rate within the continental United States (described in in Appendix E.2.1) and, using the corresponding country borders, trained a constrained diffusion model by adapting the point-in-spherical-polytope conditions outlined in (Ketzner et al., 2022) (described in Appendix E.2.2).

E.2.1. Derivation of bounded geospatial dataset. Specifically, we retrieved the rasterised version of the wildfire data provided by Welty and Jeffries (2020), converted it to a spherical geodetic coordinate system using the Cartopy library (of the United Kingdom, 2015), and drew a weighted subsample of size 1×10^6. We then retrieved the country borders of the continental United States from (Natural Earth, 2023) and mapped them to the same geodetic reference frame as the wildfire data. A visualization of the resulting dataset is presented in Figure E.2.

Figure E.2. Orthographic projection of the wildfire dataset described in Appendix E.2. The projection is aligned with the centroid of the continental United States and zoomed in ten-fold for visual clarity. All visualisations of geospatial data were generated using the GeoViews (Rudiger et al., 2023) and DataShader (Bednar et al., 2023) libraries.

E.2.2. Point-in-spherical-polytope algorithms. The support of the data-generating distribution we aim to approximate is thus restricted to a highly non-convex spherical polytope $\mathbb{P} \in \mathcal{S}^2$ given by the country borders of the continental United States. To determine whether a query point $q \in \mathcal{S}^2$ is within $\mathbb{P}$, we adapt an efficient reformulation of the point-in-spherical-polygon algorithm (Bevis and Chatelain, 1989) presented in (Ketzner et al., 2022). The algorithm requires the provision of a reference point $r \in \mathcal{S}^2$ known to be located in $\mathbb{P}$ and determines whether q is inside or outside the polygon by checking whether the geodesic between r and q crosses the polygon an even or odd number of times. Letting $\hat{x} \in \mathbb{R}^3$ denote the

Cartesian coordinates of a point $x \in \mathcal{S}^2$, (Ketzner et al., 2022) rely on a Cartesian reference coordinate system $\hat{Q}$ (with its z-axis given by $\hat{r}$) and the corresponding spherical coordinate system Q to decompose the edge-crossing condition of Bevis and Chatelain (1989) into two efficiently computable parts. That is, the geodesic between q and r crosses an edge $e_i = (v_i, v_j)$ of the polygon if:

(i) the longitude of q in Q is bounded by the longitudes of v_i and v_j in Q, i.e.

$$\phi_Q(q) \in [\min(\phi_Q(v_i), \phi_Q(v_j)), \max(\phi_Q(v_i), \phi_Q(v_j))],$$

(ii) the plane specified by the normal vector $\hat{p}_i = \hat{v}_i \times \hat{v}_j$ represents an equator that separates q and r into two different hemispheres, i.e.

$$\operatorname{sign}(\langle \hat{p}_i, \hat{r} \rangle \cdot \langle \hat{p}_i, \hat{q} \rangle) = -1.$$

Especially when $\mathbb{P}$ is fixed and the corresponding coordinate transformations and normal vectors can be precomputed for each edge, this algorithm affords an efficient and parallelisable approach to determining whether any given point on $\mathcal{S}^2$ is contained by a spherical polytope.

E.3. Supplementary Experimental Results

E.3.1. Evaluating log-barrier models. Following chapter 4, we approached the empirical evaluation of our Metropolis model by computing the maximum mean discrepancy (MMD) (Gretton et al., 2012) between samples from the true distribution and the trained diffusion models. The MMD is a statistic that quantifies the similarity of two samples by computing the distance of their respective mean embeddings in a reproducing kernel Hilbert space. For this, we use an RBF kernel with the same length scales as the standard deviations of the normal distributions used to generate the synthetic distribution. We sum these RBF kernels by the weights of the corresponding components of the synthetic Gaussian mixture model.

Table E.1. Maximum mean discrepancy (MMD) (↓) of a held-out test set from a synthetic bimodal distribution over convex subsets of $\mathbb{R}^d$ bounded by the hypercube $[-1, 1]^d$ and unit simplex Δ^d. Means and standard deviations are computed over 3 different runs.

Constraint	d	Log-Barrier	Reflected	Metropolis
$[-1, 1]^d$	2	0.066 ± 0.006	0.055 ± 0.015	$\mathbf{0.029 \pm 0.002}$
	3	0.209 ± 0.077	0.080 ± 0.004	$\mathbf{0.047 \pm 0.006}$
	10	0.330 ± 0.004	0.313 ± 0.048	$\mathbf{0.226 \pm 0.006}$
Δ^d	2	0.050 ± 0.012	0.043 ± 0.002	$\mathbf{0.032 \pm 0.007}$
	3	0.238 ± 0.009	0.181 ± 0.007	$\mathbf{0.121 \pm 0.010}$
	10	0.275 ± 0.015	0.290 ± 0.009	$\mathbf{0.267 \pm 0.003}$

From the results in Table E.1, it is clear that the log-barrier approach performs significantly worse than the Reflected model across all but one and worse than

the Metropolis models across all settings. This, in conjunction with numerical instabilities we encountered when attempting to evaluate sample likelihoods with the log-barrier models as presented in chapter 4, motivated us to focus on the Reflected and Metropolis models in the main text.

E.3.2. Experimental details. We use the same architecture in all of our experiments: a 6-layer MLP with 512 hidden units and sine activation functions, except in the output layer, which uses a linear activation function. Following Liu et al. (2022a) and chapter 4, we implement a simple linear function that scales the score by the distance to the boundary, approaching zero within $\epsilon = 0.01$ of the boundary. This ensures the score obeys the Neumann boundary conditions required by the reflected Brownian Motion. For the geospatial dataset within non-convex country borders, we do not use distance rescaling. Instead, we substitute it with a series of step functions to rescale the score. This is a proof-of-concept to show that even when computing the distance is hard, simple and efficient approximations suffice. When constructing Riemannian diffusion models on the torus and sphere for the protein and geospatial datasets, we follow (De Bortoli et al., 2022) and include an additional preconditioner for the score on the manifold. We *do not* use the residual trick or the standard deviation trick, which are both common score-rescaling functions in image model architectures; in our setting, we find that they adversely affect model training.

For the forward/reverse process we always set $T = 1$, $\beta_0 = 1 \times 10^{-3}$ and then tune β_1 to ensure that the forward process just reaches the invariant distribution with a linear β-schedule. We use a learning rate of 2×10^{-4} with a cosine learning rate schedule and an ism loss with a modified loss weighting function of $(1 + t)$, a batch size of 256 and 8 repeats per batch. At sampling time we use $N = 100$ steps of the discretised process. We discretise the training process by selecting a random N between 0 and 100 for each example, rolling out to that time point. This lets us cheaply implement a simple variance reduction technique: we take multiple samples from this trajectory by selecting multiple random N to save for each example. For all experiments, we use the ism loss with a modified weighting function of $(1+t)$, which we found to be essential to model training. All experiments use a batch size of 256 with 8 repeats per batch. For training, we use a learning rate of 2×10^{-4} with a cosine learning rate schedule. We trained for 100,000 batches on the synthetic examples and 300,000 batches on the real-world examples (robotics, proteins, wildfires).

We selected these hyperparameters from a systematic search over learning rates (6×10^{-4}, 2×10^{-4}, 6×10^{-5}, 2×10^{-5}), learning rate schedules (cosine, log-linear), and batch sizes (128, 256, 512, 1024) on synthetic examples for the reflected and log-barrier models. Similar parameters worked well for both, and we used those for our Metropolis models to allow a straightforward comparison. We tried $N = 100, 1000$ for several synthetic examples but found that very large rollout times actually hurt performance for the Metropolis model, though the log-barrier performed a bit better with longer rollouts and the reflected was the same.

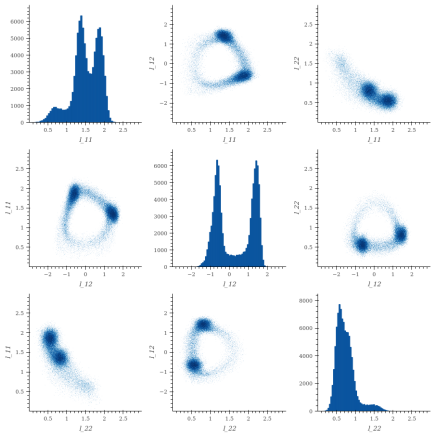
All models were trained on a single NVIDIA GeForce GTX 1080 GPU. All of the Metropolis models presented here can easily be trained on this hardware in under 4 hours. The runtime for the log-barrier and reflected models is considerably longer.

E.3.3. Constrained Configurational Modelling of Robotic Arms. The following univariate marginal and pairwise bivariate plots visualise the distribution of different samples in

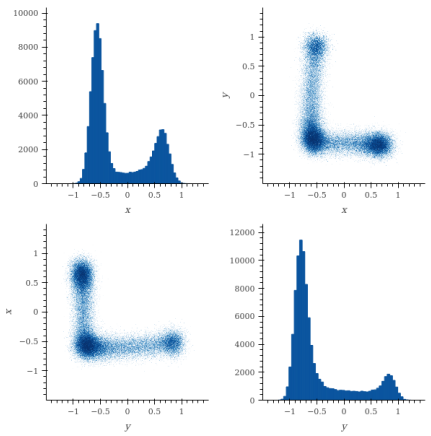
(i) the three dimensions needed to describe an ellipsoid $M = \begin{bmatrix} l_1 & l_2 \\ l_2 & l_3 \end{bmatrix} \in \mathcal{S}^2_{++}$ and

(ii) the two dimensions needed to describe a location in $\mathbb{R}^2$.

Visualisation of samples from the data distribution



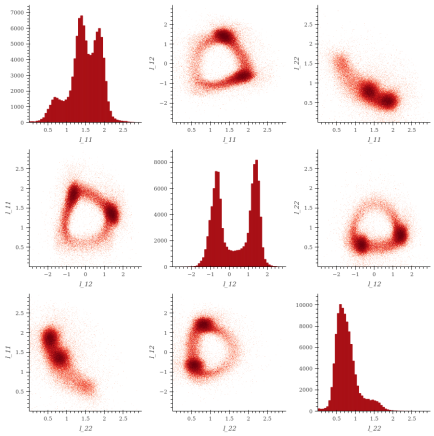
(a) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from the data distribution in $\mathcal{S}^2_{++}$.



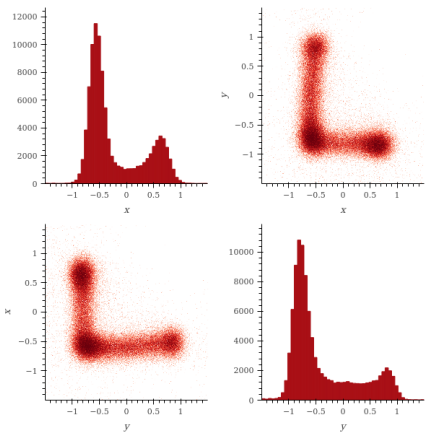
(b) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from the data distribution in $\mathbb{R}^2$.

Figure E.3. Visualisation of the data distribution in $\mathcal{S}^2_{++} \times \mathbb{R}^2$ using univariate marginal and pairwise bivariate plots.

Visualisation of samples from our Metropolis sampling-based diffusion model



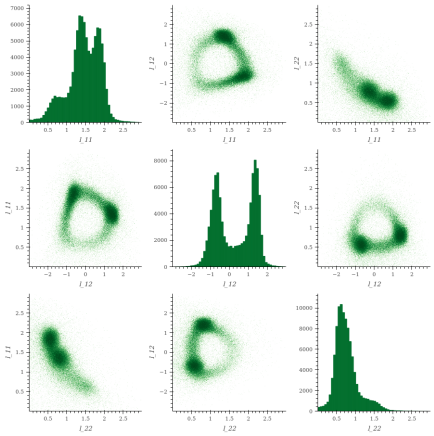
(a) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from our Metropolis sampling-based diffusion model in $\mathcal{S}^2_{++}$.



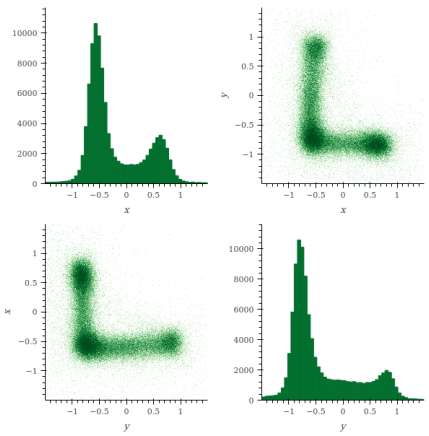
(b) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from our Metropolis sampling-based diffusion model in $\mathbb{R}^2$.

Figure E.4. Visualisation of the distribution learned by our Metropolis sampling-based diffusion model in $\mathcal{S}^2_{++} \times \mathbb{R}^2$ using univariate marginal and pairwise bivariate plots.

Visualisation of samples from a reflected Brownian motion-based diffusion model



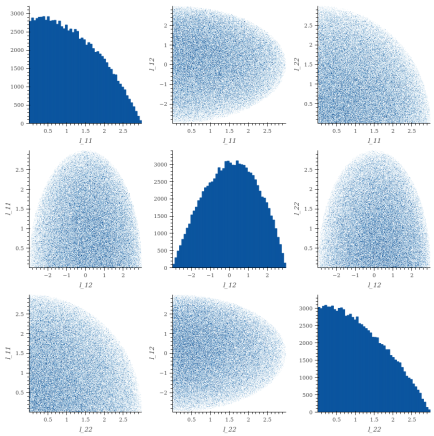
(a) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from a reflected Brownian motion-based diffusion model in $\mathcal{S}^2_{++}$.



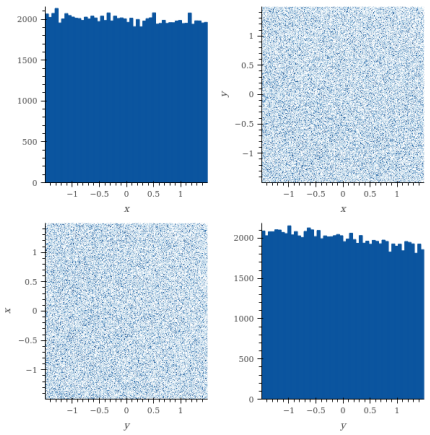
(b) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from a reflected Brownian motion-based diffusion model in $\mathbb{R}^2$.

Figure E.5. Visualisation of the distribution learned by a reflected Brownian motion-based diffusion model in $\mathcal{S}^2_{++} \times \mathbb{R}^2$ using univariate marginal and pairwise bivariate plots.

Visualisation of samples from the uniform distribution



(a) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from the uniform distribution in $\mathcal{S}^2_{++}$.

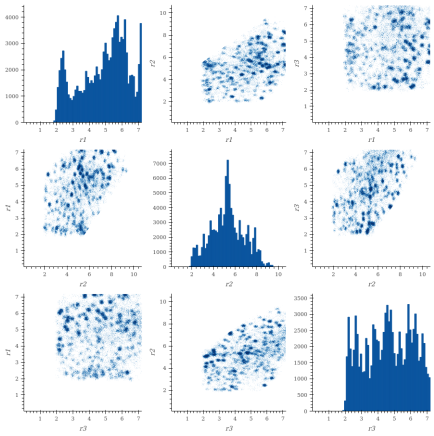


(b) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from the uniform distribution in $\mathbb{R}^2$.

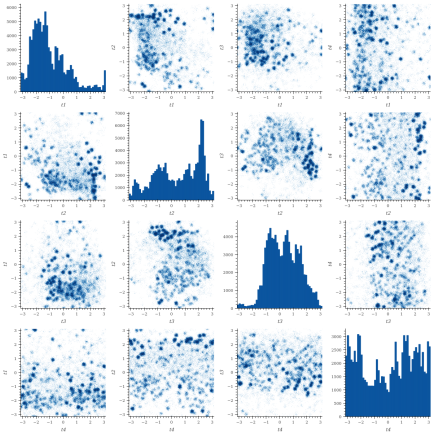
Figure E.6. Visualisation of the uniform distribution in $\mathcal{S}^2_{++} \times \mathbb{R}^2$ using univariate marginal and pairwise bivariate plots.

E.3.4. Conformational Modelling of Protein Backbones. The following univariate marginal and pairwise bivariate plots visualise the distribution of different samples in (i) the polytope $\mathbb{P} \subset \mathbb{R}^3$ and (ii) the torus $\mathbb{T}^4$ used to parametrise the conformations of a polypeptide chain of length $N = 6$ with coinciding endpoints. We refer to (Han and Rudolph, 2006) for full detail on the reparametrisation and to section 4.6.3 for a full description of the dataset.

Visualisation of samples from the data distribution



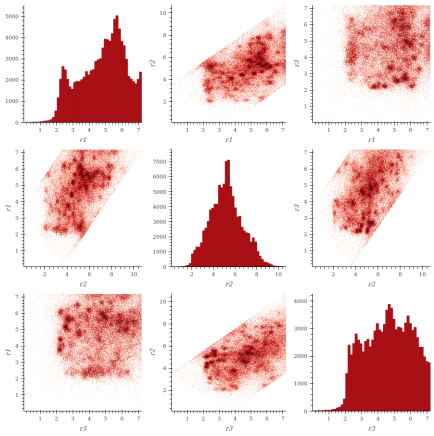
(a) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from the data distribution in $\mathbb{P} \subset \mathbb{R}^3$.



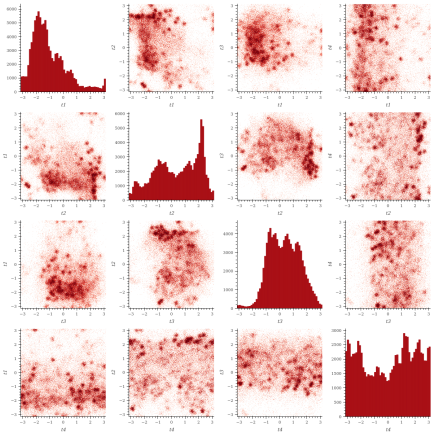
(b) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from the data distribution in $\mathbb{T}^4$.

Figure E.7. Visualisation of the data distribution in $\mathbb{P} \subset \mathbb{R}^3 \times \mathbb{T}^4$ using univariate marginal and pairwise bivariate plots.

Visualisation of samples from our Metropolis sampling-based diffusion model



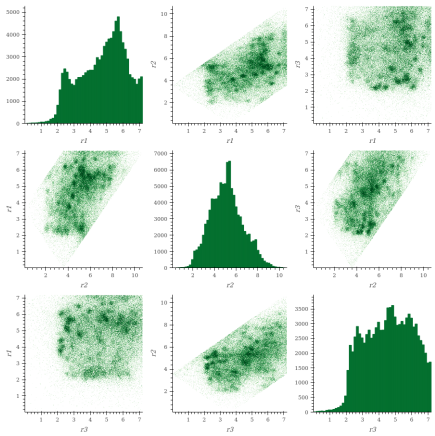
(a) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from our Metropolis model in $\mathbb{P} \subset \mathbb{R}^3$.



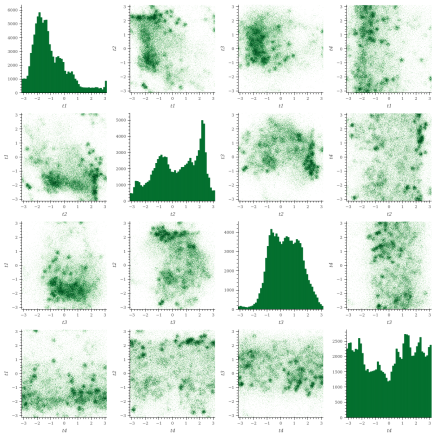
(b) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from our Metropolis model in $\mathbb{T}^4$.

Figure E.8. Visualisation of the distribution learned by our Metropolis model in $\mathbb{P} \subset \mathbb{R}^3 \times \mathbb{T}^4$ using univariate marginal and pairwise bivariate plots.

Visualisation of samples from a reflected Brownian motion-based diffusion model



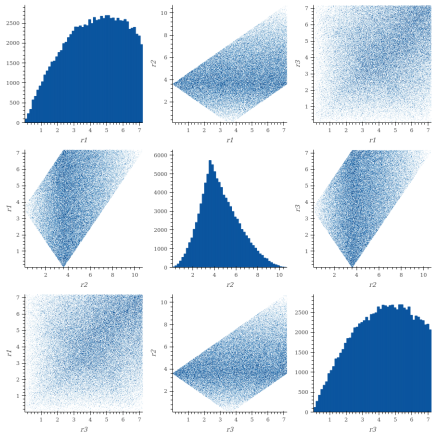
(a) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from a reflected Brownian motion-based diffusion model in $\mathbb{P} \subset \mathbb{R}^3$.



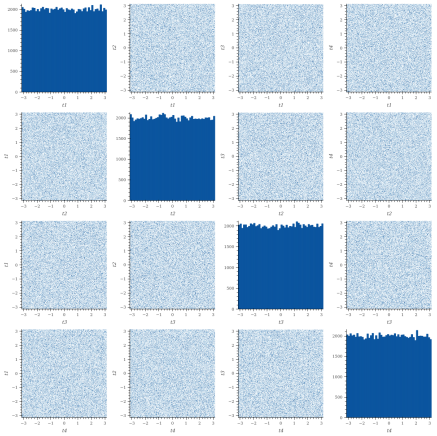
(b) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from a reflected Brownian motion-based diffusion model in $\mathbb{T}^4$.

Figure E.9. Visualisation of the distribution learned by a reflected Brownian motion-based diffusion model in $\mathbb{P} \subset \mathbb{R}^3 \times \mathbb{T}^4$ using univariate marginal and pairwise bivariate plots.

Visualisation of samples from the uniform distribution



(a) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from the uniform distribution in $\mathbb{P} \subset \mathbb{R}^3$.



(b) Plots of the univariate marginal and pairwise bivariate distributions of 1×10^5 samples from the uniform distribution in $\mathbb{T}^4$.

Figure E.10. Visualisation of the uniform distribution in $\mathbb{P} \subset \mathbb{R}^3 \times \mathbb{T}^4$ using univariate marginal and pairwise bivariate plots.

F | Score-based modelling on geometric structures on Riemannian manifolds

F.1. Organisation of appendices

In this supplementary, we first introduce in appendix F.2 an Ornstein Uhlenbeck process on function space (via finite marginals) along with several score approximations.

Later in appendix F.3, we derive sufficient conditions for this introduced model to yield a group invariant process. Finally in appendix F.4, we give a thorough description of experimental settings along with additional empirical results.

F.2. Ornstein Uhlenbeck on function space

F.2.1. Multivariate Ornstein-Uhlenbeck process. First, we aim to show that we can define a stochastic process on an infinite dimensional function space, by defining the joint finite marginals $Y(x)$ as the solution of a multidimensional Ornstein-Uhlenbeck process. In particular, for any set of input $x = (x_1, \cdots, x_k) \in \mathcal{X}^k$, we define the joint marginal as the solution of the following SDE

$$\mathrm{d}Y_t = \frac{1}{2}\beta_t(m(x) - Y_t)\,\mathrm{d}t + \sqrt{\beta_t K(x,x)} \tag{F.1}$$

where m is a mean function and K a covariance kernel.

Proposition F.1. *(Phillips et al., 2022) We assume we are given a data process $(Y_0(x))_{x\in\mathcal{X}}$ and we denote by $\mathbf{G} \sim \mathrm{GP}(0,k)$ a Gaussian process with zero mean and covariance. Then let's define*

$$\boldsymbol{Y_t} \triangleq e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\, Y_0 + \left(1 - e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\right) m + \left(1 - e^{-\int_{s=0}^{t}\beta_s ds}\right)^{1/2} \mathbf{G}.$$

Then $(\boldsymbol{Y_t}(x))_{x\in\mathcal{X}}$ is a stochastic process (by virtue of being a linear combination of stochastic processes). We thus have that $\boldsymbol{Y_t} \xrightarrow[t\to 0]{a.s.} Y_0$ and $\boldsymbol{Y_t} \xrightarrow[t\to\infty]{a.s.} Y_\infty$ with $Y_\infty \sim \mathrm{GP}(m,k)$, so effectively $(\boldsymbol{Y_t}(x))_{t\in\mathbb{R}_+, x\in\mathcal{X}}$ interpolates between the data process and this limiting Gaussian process. Additionally, $\mathcal{L}(Y_t|Y_0 = y_0) = \mathrm{GP}(m_t, K_t)$ with

$m_t = e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\, y_0 + \left(1 - e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\right) m$ *and* $\Sigma_t = \left(1 - e^{-\int_{s=0}^{t}\beta_s ds}\right) K$. *Furthermore,* $(\boldsymbol{Y_t}(x))_{t\in\mathbb{R}_+, x\in\mathcal{X}}$ *is the solution of the SDE in* (F.1).

Proof. We aim to compute the mean and covariance of the process $(\boldsymbol{Y_t})_{t\geq 0}$ described by the SDE (6.4). First let's recall the time evolution of the mean and covariance of the solution from a multivariate Ornstein-Uhlenbeck process given by

$$\mathrm{d}\boldsymbol{Y_t} = f(\boldsymbol{Y_t}, t)\mathrm{d}t + L(\boldsymbol{Y_t}, t)\mathrm{d}\boldsymbol{B_t}. \tag{F.2}$$

We know that the time evolution of the mean and the covariance are given respectively by Särkkä and Solin (2019)

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = \mathbb{E}[f(\boldsymbol{Y_t}, t)] \tag{F.3}$$

$$\frac{\mathrm{d}\Sigma_t}{\mathrm{d}t} = \mathbb{E}[f(\boldsymbol{Y_t}, t)(m_t - \boldsymbol{Y_t})^\top] + \mathbb{E}[(m_t - \boldsymbol{Y_t})f(\boldsymbol{Y_t}, t)^\top] + \mathbb{E}[L(\boldsymbol{Y_t}, t)L(\boldsymbol{Y_t}, t)^\top]. \tag{F.4}$$

Plugging in the drift $f(\boldsymbol{Y_t}, t) = 1/2 \cdot (m - \boldsymbol{Y_t})\beta_t$ and diffusion term $L(\boldsymbol{Y_t}, t) = \sqrt{\beta_t K}$ from (6.4), we get

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = 1/2 \cdot (m - \boldsymbol{Y_t})\beta_t \tag{F.5}$$

$$\frac{\mathrm{d}\Sigma_t}{\mathrm{d}t} = \beta_t \left[K - \Sigma_t\right]. \tag{F.6}$$

Solving these two ODEs we get

$$m_t = e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds} m_0 + \left(1 - e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\right) m \tag{F.7}$$

$$\tag{F.8}$$

with $m_0 \triangleq \mathbb{E}[\boldsymbol{Y}_0]$ and $\Sigma_0 \triangleq \mathrm{Cov}[\boldsymbol{Y}_0]$.

Now let's compute the first two moments of $(\boldsymbol{Y_t}(x))_{x\in\mathcal{X}}$. We have

$$\mathbb{E}[\boldsymbol{Y_t}] = \mathbb{E}\left[e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\, \boldsymbol{Y}_0 + \left(1 - e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\right) m + \left(1 - e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\right) \mathbf{G}\right] \tag{F.9}$$

$$= e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds} m_0 + \left(1 - e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\right) m \tag{F.10}$$

$$= m_t \tag{F.11}$$

$$\mathrm{Cov}[\boldsymbol{Y_t}] = \mathrm{Cov}\left[e^{-\frac{1}{2}\cdot\int_{s=0}^{t}\beta_s ds}\, \boldsymbol{Y}_0\right] + \mathrm{Cov}\left[\left(1 - e^{-\int_{s=0}^{t}\beta_s ds}\right)^{1/2} \mathbf{G}\right] \tag{F.12}$$

$$= e^{-\int_{s=0}^{t}\beta_s ds}\, \Sigma_0 + \left(1 - e^{-\int_{s=0}^{t}\beta_s ds}\right) K \tag{F.13}$$

$$= K + e^{-\int_{s=0}^{t}\beta_s ds}\, (\Sigma_0 - K) \tag{F.14}$$

$$= \Sigma_t. \tag{F.15}$$

■

F.2.2. Conditional score. Hence, condition on Y_0 the score is the gradient of the log Gaussian characterised by mean $m_{t|0} = e^{-\frac{1}{2}B(t)}Y_0$ and $\Sigma_{t|0} = (1 - e^{-B(t)})\mathrm{K}$ with $B(t) = \int_0^t \beta(s)ds$ which can be derived from the above marginal mean and covariance with $m_0 = Y_0$ and $\Sigma_0 = 0$.

$$\begin{aligned}
\nabla_{Y_t} \log p_t(Y_t|Y_0) &= \nabla_{Y_t} \log \mathcal{N}\left(Y_t|m_{t|0}, \Sigma_{t|0}\right) && \text{(F.16)}\\
&= \nabla_{Y_t} - 1/2(\boldsymbol{Y_t} - m_{t|0})^\top \Sigma_{t|0}^{-1}(\boldsymbol{Y_t} - m_{t|0}) + c && \text{(F.17)}\\
&= -\Sigma_{t|0}^{-1}(\boldsymbol{Y_t} - m_{t|0}) && \text{(F.18)}\\
&= -\mathrm{L}_{t|0}^{-\top}\mathrm{L}_{t|0}^{-1}\mathrm{L}_{t|0}\epsilon && \text{(F.19)}\\
&= -\mathrm{L}_{t|0}^{-\top}\epsilon && \text{(F.20)}
\end{aligned}$$

where $\mathrm{L}_{t|0}$ denotes the Cholesky decomposition of $\Sigma_{t|0} = \mathrm{L}_{t|0}\mathrm{L}_{t|0}^\top$, and $\boldsymbol{Y_t} = m_{t|0} + \mathrm{L}_{t|0}\epsilon$.

Then we can plugin our learnt (preconditioned) score into the backward SDE 6.9 which gives

$$\mathrm{d}\bar{Y}_t|x = \left[-(m(x) - \bar{Y}_t)/2 + \mathrm{K}(x,x)\nabla_{\bar{Y}_t} \log p_{T-t}(t, x, \bar{Y}_t)\right]\mathrm{d}t + \sqrt{\beta_t \mathrm{K}(x,x)}\beta_t \mathrm{d}\boldsymbol{B_t} \tag{F.21}$$

F.2.3. Several score parametrisations. In this section, we discuss several parametrisations of the neural network and the objective.

For the sake of versatility, we opt to employ the symbol D_θ for the network instead of s_θ as mentioned in the primary text, as it allows us to approximate not only the score but also other quantities from which the score can be derived. In full generality, we use a residual connection, weighted by $c_{\text{out}}, c_{\text{skip}} : \mathbb{R} \to \mathbb{R}$, to parametrise the network

$$D_\theta(t, Y_t) = c_{\text{skip}}(t)Y_t + c_{\text{out}}(t)F_\theta(t, Y_t). \tag{F.22}$$

We recall that the input to the network is time t, and the noised vector $Y_t = \boldsymbol{\mu}_{t|0} + \boldsymbol{n}$, where $\boldsymbol{\mu}_{t|0} = \mathrm{e}^{-B(t)/2}Y_0$ and $\boldsymbol{n} \sim \mathcal{N}(0, \Sigma_{t|0})$ with $\Sigma_{t|0} = (1 - \mathrm{e}^{-B(t)})K$. The gram matrix K corresponds to $k(X, X)$ with k the limiting kernel. We denote by $\mathrm{L}_{t|0}$ and S respectively the Cholesky decomposition of $\Sigma_{t|0} = \mathrm{L}_{t|0}\mathrm{L}_{t|0}^\top$ and $\mathrm{K} = \mathrm{SS}^\top$.

The denoising score matching loss weighted by $\Lambda(t)$ is given by

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|D_\theta(t, Y_t) - \nabla_{Y_t} \log p_t(Y_t|Y_0)\|^2_{\Lambda(t)}\right] \tag{F.23}$$

No preconditioning By reparametrisation, let $Y_t = \boldsymbol{\mu}_{t|0} + L_{t|0}\boldsymbol{z}$, where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the loss from eq. (F.23) can be written as

$$\begin{aligned}
\mathcal{L}(\theta) &= \mathbb{E}\left[\|D_\theta(t, Y_t) + \Sigma_{t|0}^{-1}(Y_t - \boldsymbol{\mu}_{t|0})\|^2_{\Lambda(t)}\right] && \text{(F.24)}\\
&= \mathbb{E}\left[\|D_\theta(t, Y_t) + \Sigma_{t|0}^{-1}L_{t|0}\boldsymbol{z}\|^2_{\Lambda(t)}\right] && \text{(F.25)}\\
&= \mathbb{E}\left[\|D_\theta(t, Y_t) + L_{t|0}^{-\top}\boldsymbol{z}\|^2_{\Lambda(t)}\right] && \text{(F.26)}\\
& && \text{(F.27)}
\end{aligned}$$

	No precond.	Precond. K	Precond. $S^\top$	Predict Y_0
c_{skip}	0	0	0	1
c_{out}	$(\sigma_{t\|0} + 10^{-3})^{-1}$	$(\sigma_{t\|0} + 10^{-3})^{-1}$	$(\sigma_{t\|0} + 10^{-3})^{-1}$	1
Loss	$\|\sigma_{t\|0} S^\top D_\theta + z\|_2^2$	$\|\sigma_{t\|0} D_\theta + Sz\|_2^2$	$\|\sigma_{t\|0} D_\theta + z\|_2^2$	$\|D_\theta - Y_0\|_2^2$
$K\nabla \log p_t$	KD_θ	D_θ	SD_θ	$-\Sigma_{t\|0}^{-1}(Y_t - \mathrm{e}^{-\frac{B(t)}{2}} D_\theta)$

Table F.1. Summary of different score parametrisations as well as the values for c_{skip} and c_{out} that we found to be optimal, based on the recommendation from Karras et al. (2022, Appendix B.6).

Choosing $\Lambda(t) = \Sigma_{t|0} = L_{t|0} L_{t|0}^\top$ we obtain

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|L_{t|0}^\top D_\theta(t, Y_t) + z\|_2^2\right] \tag{F.28}$$

$$= \mathbb{E}\left[\|\sigma_{t|0} S^\top D_\theta(t, Y_t) + z\|_2^2\right]. \tag{F.29}$$

Preconditioning by K Alternatively, one can train the neural network to approximate the preconditioned score $D_\theta \approx \mathbf{K}\nabla_{Y_t} \log p_t(Y_t|Y_0)$. The loss, weighted by $\Lambda = \sigma_{t|0}^2 \mathbf{I}$, is then given by

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|D_\theta(t, Y_t) + K\, L_{t|0}^{-\top} z\|_{\Lambda(t)}^2\right] \tag{F.30}$$

$$= \mathbb{E}\left[\|D_\theta(t, Y_t) + \sigma_{t|0}^{-1} Sz\|_{\Lambda(t)}^2\right] \tag{F.31}$$

$$= \mathbb{E}\left[\|\sigma_{t|0} D_\theta(t, Y_t) + Sz\|_2^2\right]. \tag{F.32}$$

Precondition by $S^\top$ A variation of the previous one, is to precondition the score by the transpose Cholesky of the limiting kernel gram matrix, such that $D_\theta \approx S^\top \nabla_{Y_t} \log p_t(Y_t|Y_0)$.

The loss, weighted by $\Lambda = \sigma_{t|0}^2 \mathbf{I}$, becomes

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|D_\theta(t, Y_t) + S^\top\, L_{t|0}^{-\top} z\|_{\Lambda(t)}^2\right] \tag{F.33}$$

$$= \mathbb{E}\left[\|D_\theta(t, Y_t) + \sigma_{t|0}^{-1} z\|_{\Lambda(t)}^2\right] \tag{F.34}$$

$$= \mathbb{E}\left[\|\sigma_{t|0} D_\theta(t, Y_t) + z\|_2^2\right]. \tag{F.35}$$

Predicting Y_0 Finally, an alternative strategy is to predict Y_0 from a noised version Y_t. In this case, the loss takes the simple form

$$\mathcal{L}(\theta) = \mathbb{E}\left[\|D_\theta(t, Y_t) - Y_0\|_2^2\right].$$

The score can be computed from the network's prediction following

$$\nabla \log p_t(Y_t|Y_0) = -\Sigma_{t|0}^{-1}(Y_t - \boldsymbol{\mu}_{t|0}) \tag{F.36}$$

$$= -\Sigma_{t|0}^{-1}(Y_t - \mathrm{e}^{-B(t)/2} Y_0) \tag{F.37}$$

$$\approx -\Sigma_{t|0}^{-1}\left(Y_t - \mathrm{e}^{-B(t)/2} D_\theta(t, Y_t)\right) \tag{F.38}$$

$$\tag{F.39}$$

Table F.1 summarises the different options for parametrising the score as well as the values for c_{skip} and c_{out} that we found to be optimal, based on the recommendation from Karras et al. (2022, Appendix B.6). In practice, we found the precondition by K parametrisation to produce the best results, but we refer to appendix F.4.1 for a more in-depth ablation study.

F.2.4. Exact (marginal) score in Gaussian setting. Interpolating between Gaussian processes $GP(m_0, \Sigma_0)$ and $GP(m, \mathrm{K})$

$$\begin{aligned} \mathrm{K}\nabla_{\bar{Y}_t} \log p_t(\boldsymbol{Y_t}) &= -\mathrm{K}\Sigma_t^{-1}(\boldsymbol{Y_t} - m_t) && \text{(F.40)} \\ &= -\mathrm{K}[\mathrm{K} + e^{-\int_{s=0}^{t} \beta_s ds}(\Sigma_0 - \mathrm{K})]^{-1}(\boldsymbol{Y_t} - m_t) && \text{(F.41)} \\ &= -\mathrm{K}(\mathrm{L}_t \mathrm{L}_t{}^{\top})^{-1}(\boldsymbol{Y_t} - m_t) && \text{(F.42)} \\ &= -\mathrm{K}\mathrm{L}_t{}^{\top -1}\mathrm{L}_t{}^{-1}(\boldsymbol{Y_t} - m_t) && \text{(F.43)} \\ & && \text{(F.44)} \end{aligned}$$

with $\Sigma_t = \mathrm{K} + e^{-\int_{s=0}^{t} \beta_s ds}(\Sigma_0 - \mathrm{K}) = \mathrm{L}_t \mathrm{L}_t{}^{\top}$ obtained via Cholesky decomposition.

F.2.5. Langevin dynamics. Under mild assumptions on $\nabla \log p_{T-t}$ (Durmus and Moulines, 2016) the following SDE

$$\mathrm{d}\boldsymbol{Y}_s = \tfrac{1}{2}\mathrm{K}\nabla \log p_{T-t}(\boldsymbol{Y}_s)\mathrm{d}s + \sqrt{\mathrm{K}}\mathrm{d}\boldsymbol{B}_s \tag{F.45}$$

admits a solution $(\boldsymbol{Y}_s)_{s\geq 0}$ whose law $\mathcal{L}(\boldsymbol{Y}_s)$ converges with geometric rate to p_{T-t} for any invertible matrix K.

F.2.6. Likelihood evaluation. Similarly to Song et al. (2020b), we can derive a deterministic process which has the same marginal density as the SDE (6.4), which is given by the following Ordinary Differential Equation (ODE)—referred as the probability flow ODE

$$\mathrm{d}\begin{pmatrix} \boldsymbol{Y}_t(x) \\ \log p_t(\boldsymbol{Y}_t(x)) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left\{m(x) - \boldsymbol{Y}_t(x) - \mathrm{K}(x,x)\nabla \log p_t(\boldsymbol{Y}_t(x))\right\}\beta_t \\ -\frac{1}{2}\operatorname{div}\left\{m(x) - \boldsymbol{Y}_t(x) - \mathrm{K}(x,x)\nabla \log p_t(\boldsymbol{Y}_t(x))\right\}\beta_t \end{pmatrix}\mathrm{d}t. \tag{F.46}$$

Once the score network is learnt, we can thus use it in conjunction with an ODE solver to compute the likelihood of the model.

F.3. Invariant neural diffusion processes

F.3.1. E(n)-equivariant kernels. A kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is equivariant if it satisfies the following constraints: (a) k is *stationary*, that is if for all $x, x' \in \mathbb{R}^n$

$$k(x, x') = k(x - x') \triangleq \tilde{k}(x - x') \tag{F.47}$$

and if (b) it satisfies the *angular constraint* for any $h \in H$

$$k(hx, hx') = \rho(h)k(x, x')\rho(h)^{\top}. \tag{F.48}$$

A trivial example of such an equivariant kernel is the diagonal kernel $k(x, x') = k_0(x, x')\mathrm{I}$ (Holderrieth et al., 2021b), with k_0 stationary. This kernel can be understood has having d independent Gaussian process uni-dimensional output, that is, there is no inter-dimensional correlation.

Less trivial examples, are the $\mathrm{E}(n)$ equivariant kernels proposed in Macêdo and Castro (2010). Namely curl-free and divergence-free kernels, allowing for instance to model electric or magnetic fields. Formally we have $k_{\text{curl}} = k_0 A$ and $k_{\text{div}} = k_0 B$ with k_0 stationary, e.g. squared exponential kernel $k_0(x, x') = \sigma^2 \exp\left(\frac{\|x-x'\|^2}{2l^2}\right)$, and A and B given by

$$A(x, x') = \mathrm{I} - \frac{(x - x')(x - x')^\top}{l^2} \tag{F.49}$$

$$B(x, x') = \frac{(x - x')(x - x')^\top}{l^2} + \left(n - 1 - \frac{\|x - x'\|^2}{l^2}\right)\mathrm{I}. \tag{F.50}$$

See Holderrieth et al. (Appendix C, 2021b) for a proof.

F.3.2. Proof of proposition 6.2. Below we give two proofs for the group invariance of the generative process, one via the probability flow ODE and one directly via Fokker-Planck.

Proof. The reverse probability flow associated with the forward SDE (6.4) with approximate score $\mathbf{s}_\theta(t, \cdot) \approx \nabla \log p_t$ is given by

$$\mathrm{d}\bar{Y}_t | x = \tfrac{1}{2}\left[-m(x) + \bar{Y}_t + \mathrm{K}(x, x)\mathbf{s}_\theta(T - t, x, \bar{Y}_t)\right]\mathrm{d}t \tag{F.51}$$

$$\triangleq b_{\text{ODE}}(t, x, \bar{Y}_t)\mathrm{d}t \tag{F.52}$$

This ODE induces a flow $\phi_t^b : X^n \times Y^n \to \mathrm{T}Y^n$ for a given integration time t, which is said to be G-equivariant if the vector field is $G-$equivariant itself, i.e. $b(t, g \cdot x, \rho(g)\bar{Y}_t) = \rho(g)b(t, x, \bar{Y}_t)$. We have that for any $g \in G$

$$b_{\text{ODE}}(t, g \cdot x, \rho(g)\bar{Y}_t) = \tfrac{1}{2}\left[-m(g \cdot x) + \rho(g)Y_t + \mathrm{K}(g \cdot x, g \cdot x)\ \mathbf{s}_\theta(t, g \cdot x, \rho(g)\bar{Y}_t)\right] \tag{F.53}$$

$$\overset{(1)}{=} \tfrac{1}{2}\left[-\rho(g)m(x) + \rho(g)Y_t + \rho(g)\mathrm{K}(x, x)\rho(g)^\top\ \mathbf{s}_\theta(t, g \cdot x, \rho(g)\bar{Y}_t)\right] \tag{F.54}$$

$$\overset{(2)}{=} \tfrac{1}{2}\left[-\rho(g)m(x) + \rho(g)Y_t + \rho(g)\mathrm{K}(x, x)\rho(g)^\top\rho(g)\ \mathbf{s}_\theta(t, x, \bar{Y}_t)\right] \tag{F.55}$$

$$\overset{(3)}{=} \tfrac{1}{2}\rho(g)\left[-m(x) + Y_t + \mathrm{K}(x, x)\ \mathbf{s}_\theta(t, x, \bar{Y}_t)\right] \tag{F.56}$$

$$= \rho(g)b_{\text{ODE}}(t, x, \bar{Y}_t) \tag{F.57}$$

with (1) from the G-invariant prior GP conditions on m and k, (2) assuming that the score network is G-equivariant and (3) assuming that $\rho(g) \in O(n)$. To prove the opposite direction, we can simply follow these computations backwards. Finally, we know that with a G-invariant probability measure p_{ref} and G-equivariant map ϕ, the pushforward probability measure $p_{\text{ref}}^{-1} \circ \phi$ is also G-invariant (Köhler et al., 2020; Papamakarios et al., 2019). Assuming a G-invariant prior GP, and a G-equivariant score network, we thus have that the generative model from section 6.2.3 defines marginals that are G-invariant. ∎

Proof. The reverse SDE associated of the forward SDE (6.4) with approximate score $\mathbf{s}_\theta(t,\cdot) \approx \nabla \log p_t$ is given by

$$\mathrm{d}\bar{\boldsymbol{Y}}_t|x = \left[-(m(x) - \bar{\boldsymbol{Y}}_t)/2 + \mathrm{K}(x,x)\mathbf{s}_\theta(T-t, x, \bar{\boldsymbol{Y}}_t)\right]\mathrm{d}t + \sqrt{\beta_t \mathrm{K}(x,x)}\mathrm{d}\boldsymbol{B}_t \tag{F.58}$$

$$\triangleq b_{\mathrm{SDE}}(t, x, \bar{\boldsymbol{Y}}_t)\mathrm{d}t + \Sigma^{1/2}(t,x)\ \mathrm{d}\boldsymbol{B}_t. \tag{F.59}$$

As for the probability flow drift b_{ODE}, we have that b_{SDE} is similarly G-equivariant, that is $b_{\mathrm{SDE}}(t, g\cdot x, \rho(g)\bar{\boldsymbol{Y}}_t) = \rho(g) b_{\mathrm{SDE}}(t, x, \bar{\boldsymbol{Y}}_t)$ for any $g \in G$. Additionally, we have that diffusion matrix is also G-equivariant as for any $g \in G$ we have $\Sigma(t, g\cdot x) = \beta_t \mathrm{K}(g\cdot x, g\cdot x) = \beta_t \rho(g)\mathrm{K}(x,x)\rho(g)^\top = \rho(g)\Sigma(t,x)\rho(g)^\top$ since K is the gram matrix of an G-equivariant kernel k.

Additionally assuming that b_{SDE} and Σ are bounded, Yim et al. (Proposition 3.6, 2023) says that the distribution of $\bar{\boldsymbol{Y}}_t$ is G-invariant, and in particular $\mathcal{L}(\bar{\boldsymbol{Y}}_0)$.

■

F.3.3. Equivariant posterior maps.

Invariant prior stochastic process implies an equivariant posterior map

THEOREM F.2. *Using the language of Weiler et al. (2023) our tensor fields are sections of an associated vector bundle $\mathcal{A}$ of a manifold M with a G structure. Let* Isom_{GM} *be the group of G-structure preserving isometries on M. The action of this group on a section of the bundle $f \in \Gamma(\mathcal{A})$ is given by*

$$\phi \triangleright f := \phi_{*,\mathcal{A}} \circ f \circ \phi^{-1}. \tag{F.60}$$

Let $f \sim P$, P a distribution over the space of section. Let $\phi \triangleright P$ be the law of of $\phi \triangleright f$. Let $\mu_x = \mathcal{L}(f(x)) = \pi_{x\#}P$, the law of f evaluated at a point, where π_x is the canonical projection operator onto the marginal at x, # the pushforward operator in the measure theory sense, $x \in M$ and y is in the fibre of the associated bundle. Let $\mu_x^{x',y} = \mathcal{L}(f(x)|f(x') = y') = \pi_x \mu^{x',y'} = \pi_{x\#}\mathcal{L}(f|f(x') = y')$, the conditional law of the process when given $f(x') = y'$.

Assume that the prior is invariant under the action of Isom_{GM}*, i.e. that*

$$\phi \triangleright \mu_x = \left(\phi_{*,\mathcal{A}}\right)_{\#}\mu_{\phi^{-1}(x)} = \mu_x \tag{F.61}$$

Then the conditional measures are equivariant, in the sense that

$$\phi \triangleright \mu_x^{x',y'} = \left(\phi_{*,\mathcal{A}}\right)_{\#}\mu_{\phi^{-1}(x)}^{x',y'} = \mu_x^{\phi^{-1}(x), \phi_{*,\mathcal{A}}(y)} = \mu_x^{\phi\triangleright(x',y')} \tag{F.62}$$

Proof. $\forall A, B$ test functions, $\phi \in \mathrm{Isom}_{GM}$,

$$\begin{aligned}
&\mathbb{E}[B(f(x'))A((\phi \triangleright f)(x))] \\
&= \mathbb{E}\left[B(f(x'))A\left(\phi_{*,\mathcal{A}} \circ f \circ \phi^{-1}(x)\right)\right] \\
&= \mathbb{E}\left[B(f(x'))\,\mathbb{E}\left[A\left(\phi_{*,\mathcal{A}}\left(F(\phi^{-1}(x))\right)\right) \,\middle|\, F(x')\right]\right] \\
&= \mathbb{E}\left[B(f(x')) \int A(y)\left(\phi_{*,\mathcal{A}}\right)_{\#}\mu^{x',f(x')}_{\phi^{-1}(x)}(\mathrm{d}y)\right] \\
&= \int B(y') \int A(y)\left(\phi_{*,\mathcal{A}}\right)_{\#}\mu^{x',f(x')}_{\phi^{-1}(x)}(\mathrm{d}y)\mu_{x'}(\mathrm{d}y') \\
&= \int B(y') \int A(y)\left(\phi \triangleright \mu^{x',f(x')}_{x}\right)(\mathrm{d}y)\mu_{x'}(\mathrm{d}y')
\end{aligned}$$

By invariance this quantity is also equal to

$$\begin{aligned}
&\mathbb{E}\left[B\left((\phi^{-1} \triangleright f)(x')\right)A((\phi^{-1} \triangleright \phi \triangleright f)(x))\right] \\
&= \mathbb{E}\left[B\left((\phi^{-1} \triangleright f)(x')\right)\,\mathbb{E}\left[A(f(x)) \,\middle|\, B\left((\phi^{-1} \triangleright f)(x')\right)\right]\right] \\
&= \mathbb{E}\left[B\left(\phi_{*,\mathcal{A}}(f(\phi^{-1}(x')))\right)\left[A(F(x)) \,\middle|\, \phi_{*,\mathcal{A}}(f(\phi^{-1}(x')))\right]\right] \\
&= \mathbb{E}\left[B\left(\tau^{-1}_{x',g}F(gx')\right) \int A(y)\mu_{x}^{\phi(x'),\phi^{-1}_{*,\mathcal{A}}(y)}\right](\mathrm{d}y) \\
&= \int B(y') \int A(y)\mu_{x}^{\phi\triangleright(x',y)}(\mathrm{d}y)\left(\phi^{-1}_{*,\mathcal{A}}\right)_{\#}\mu_{\phi(x')}(\mathrm{d}y') \\
&= \int B(y') \int A(y)\mu_{x}^{\phi\triangleright(x',y)}(\mathrm{d}y)\left(\phi^{-1} \triangleright \mu_{x'}\right)(\mathrm{d}y')
\end{aligned}$$

Hence

$$\left(\phi \triangleright \mu^{x',f(x')}_{x}\right)(\mathrm{d}y)\mu_{x'}(\mathrm{d}y') = \mu_{x}^{\phi\triangleright(x',y)}(\mathrm{d}y)\left(\phi^{-1} \triangleright \mu_{x'}\right)(\mathrm{d}y')$$

By the stated invariance $\phi^{-1} \triangleright \mu_{x'} = \mu_{x'}$, hence

$$\left(\phi \triangleright \mu^{x',f(x')}_{x}\right)(\mathrm{d}y) = \mu_{x}^{\phi\triangleright(x',y)}(\mathrm{d}y) \text{ a.e. } y'$$

So

$$\phi \triangleright \mu^{x',f(x')}_{x} = \mu_{x}^{\phi\triangleright(x',y)} \tag{F.63}$$

as desired. ∎

In this section, we show that:

(a) Algorithm 6.3 and algorithm F.3 `Repaint` from (Lugmayr et al., 2022) are equivalent in a specific setting.

(b) There exists a continuous limit (SDE) for both procedures. This SDE targets a probability density which *does not* correspond to $p(x_{t_0}|x_0^c)$.

(c) When $t_0 \to 0$ this probability measure converges to $p(x_0|x_0^c)$ which ensures the correctness of the proposed sampling scheme.

Algorithm F.3 RePaint (Lugmayr et al., 2022).

šeries Require: Score network $\mathbf{s}_\theta(t, \boldsymbol{x}, \boldsymbol{y})$, conditioning points $(\boldsymbol{x}^c, \boldsymbol{y}^c)$, query locations $\boldsymbol{x}^*$
$\bar{\boldsymbol{x}} = [\boldsymbol{x}^c, \boldsymbol{x}^*]$ ▷ Augmented inputs set
$\left[\boldsymbol{y}_T^c, \boldsymbol{y}_T^*\right] \sim \mathrm{N}(m(\bar{\boldsymbol{x}}), k(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}}))$ ▷ Sample initial noise
for $t \in \{T, T-\gamma, ..., \epsilon\}$ **do**
  $\tilde{\boldsymbol{y}}_t^* = \boldsymbol{y}_t^*$
  for $l \in \{1, \dots, L\}$ **do**
    $\boldsymbol{y}_t^c \sim \mathrm{N}(m_t(\boldsymbol{x}^c; \boldsymbol{y}^c), k_t(\boldsymbol{x}^c, \boldsymbol{x}^c; \boldsymbol{y}^c))$ ▷ Noise context outputs
    $Z \sim \mathrm{N}(0, \mathbf{I})$ ▷ Sample tangent noise
    $\left[\_, \tilde{\boldsymbol{y}}_{t-\gamma}^*\right] = \left[\boldsymbol{y}_t^c, \tilde{\boldsymbol{y}}_t^*\right] + \gamma \left\{-\frac{1}{2}\left(m(\bar{\boldsymbol{x}}) - \left[\boldsymbol{y}_t^c, \tilde{\boldsymbol{y}}_t^*\right]\right) + \mathrm{K}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})\mathbf{s}_\theta(t, \bar{\boldsymbol{x}}, \left[\boldsymbol{y}_t^c, \tilde{\boldsymbol{y}}_t^*\right])\right\} + \sqrt{\gamma}\mathrm{K}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{x}})^{1/2} Z$ ▷ Reverse step
    $Z \sim \mathrm{N}(0, \mathbf{I})$ ▷ Sample tangent noise
    $\tilde{\boldsymbol{y}}_t^* = \tilde{\boldsymbol{y}}_{t-\gamma}^* + \gamma \left\{\frac{1}{2}\left(m(\boldsymbol{x}^*) - \tilde{\boldsymbol{y}}_{t-\gamma}^*\right)\right\} + \sqrt{\gamma}\mathrm{K}(\boldsymbol{x}^*, \boldsymbol{x}^*)^{1/2} Z$ ▷ Forward step
  $\boldsymbol{y}_{t-\gamma}^* = \tilde{\boldsymbol{y}}_{t-\gamma}^*$
šeries return $\boldsymbol{y}_\epsilon^*$

We begin by recalling the conditional sampling algorithm we study in Algorithm 6.3 and Algorithm F.3.

First, we start by describing the RePaint algorithm (Lugmayr et al., 2022). We consider $(Z_k^0, Z_k^1, Z_k^2)_{k\in\mathbb{N}}$ a sequence of independent Gaussian random variable such that for any $k \in \mathbb{N}$, Z_k^1 and Z_k^2 are d-dimensional Gaussian random variables with zero mean and identity covariance matrix and Z_k^0 is a p-dimensional Gaussian random variable with zero mean and identity covariance matrix. We assume that the whole sequence to be inferred is of size d while the context is of size p. For simplicity, we only consider the Euclidean setting with $\mathrm{K} = \mathbf{I}$. The proofs can be adapted to cover the case $\mathrm{K} \neq \mathbf{I}$ without loss of generality.

Let us fix a time $t_0 \in \{0, T\}$. We consider the chain $(X_k)_{k\in\mathbb{N}}$ given by $X_0 \in \mathbb{R}^d$ and for any $k \in \mathbb{N}$, we define

$$X_{k+1/2} = \mathrm{e}^{\gamma} X_k + 2(\mathrm{e}^{\gamma} - 1)\nabla_{x_k} \log p_{t_0}([X_k, X_k^c]) + (\mathrm{e}^{2\gamma} - 1)^{1/2} Z_k^1, \tag{F.64}$$

where $X_k^c = \mathrm{e}^{-t_0} X_0^c + (1 - \mathrm{e}^{-2t_0})^{1/2} Z_k^0$. Finally, we consider

$$X_{k+1} = \mathrm{e}^{-\gamma} X_{k+1/2} + (1 - \mathrm{e}^{-2\gamma})^{1/2} Z_k^2. \tag{F.65}$$

Note that (F.64) corresponds to one step of *backward SDE* integration and (F.65) corresponds to one step of *forward SDE* integration. In both cases we have used the exponential integrator, see (De Bortoli, 2022) for instance. While we use the exponential integrator in the proofs for simplicity other integrators such as the classical Euler-Maruyama integration could have been used. Combining (F.64) and (F.65), we get that for any $k \in \mathbb{N}$ we have

$$X_{k+1} = X_k + 2(1 - \mathrm{e}^{-\gamma})\nabla_{x_k} \log p_{t_0}([X_k, X_k^c]) + (1 - \mathrm{e}^{-2\gamma})^{1/2}(Z_k^1 + Z_k^2). \tag{F.66}$$

Remarking that $(Z_k)_{k\in\mathbb{N}} = ((Z_k^1 + Z_k^3)/\sqrt{2})_{k\in\mathbb{N}}$ is a family of d-dimensional Gaussian random variables with zero mean and identity covariance matrix, we get that for any $k \in \mathbb{N}$

$$X_{k+1} = X_k + 2(1 - \mathrm{e}^{-\gamma})\nabla_{x_k} \log p_{t_0}([X_k, X_k^c]) + \sqrt{2}(1 - \mathrm{e}^{-2\gamma})^{1/2} Z_k, \tag{F.67}$$

where we recall that $X_k^c = \mathrm{e}^{-t_0} X_0^c + (1 - \mathrm{e}^{-2t_0})^{1/2} Z_k^0$. Note that the process (F.67) is another version of the `Repaint` algorithm (Lugmayr et al., 2022), where we have concatenated the denoising and noising procedure. With this formulation, it is clear that `Repaint` is equivalent to Algorithm 6.3. In what follows, we identify the limiting SDE of this process.

In what follows, we describe the limiting behaviour of (F.67) under mild assumptions on the target distribution. In what follows, for any $x_{t_0} \in \mathbb{R}^d$, we denote

$$b(x_{t_0}) = 2 \int_{\mathbb{R}^p} \nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x_{t_0}^c]) p_{t_0|0}(x_{t_0}^c | x_0^c) \mathrm{d}x_{t_0}^c. \tag{F.68}$$

We emphasize that $b/2 \neq \nabla_{x_{t_0}} \log p(\cdot | x_0^c)$. In particular, using Tweedie's identity, we have that for any $x_{t_0} \in \mathbb{R}^d$

$$\nabla \log p_{t_0}(x_{t_0} | x_0^c) = \int_{\mathbb{R}^p} \nabla_{x_{t_0}} \log p([x_{t_0}, x_{t_0}^c] | x_0^c) p(x_{t_0}^c | x_{t_0}, x_0^c) \mathrm{d}x_{t_0}^c. \tag{F.69}$$

We introduce the following assumption.

Assumption F.4. *There exist* $\mathtt{L}, C \geq 0$, $\mathtt{m} > 0$ *such that for any* $x_{t_0}^c, y_t^c \in \mathbb{R}^p$ *and* $x_{t_0}, y_t \in \mathbb{R}^d$

$$\left\| \nabla \log p_{t_0}([x_{t_0}, x_{t_0}^c]) - \nabla \log p_{t_0}([y_t, y_t^c]) \right\| \leq \mathtt{L}(\left\| x_{t_0} - y_t \right\| + \left\| x_{t_0}^c - y_t^c \right\|). \tag{F.70}$$

Assumption F.4 ensures that there exists a unique strong are studied in De Bortoli (2022). In the theoretical literature on diffusion models the Lipschitz assumption is classical, see

We denote $((\boldsymbol{X}_t^\gamma)_{t \geq 0})_{\gamma > 0}$ the family of processes such that for any $k \in \mathbb{N}$ and $\gamma > 0$, we have for any $t \in [k\gamma, (k+1)\gamma)$, $\boldsymbol{X}_t^\gamma = (1 - (t - k\gamma)/\gamma) \boldsymbol{X}_{k\gamma}^\gamma + (t - k\gamma)/\gamma \boldsymbol{X}_{(k+1)\gamma}^\gamma$ and

$$\boldsymbol{X}_{(k+1)\gamma}^\gamma = \boldsymbol{X}_{k\gamma}^\gamma + 2(1 - \mathrm{e}^{-\gamma}) \nabla_{\boldsymbol{X}_{k\gamma}^\gamma} \log p_{t_0}([\boldsymbol{X}_{k\gamma}^\gamma, \boldsymbol{X}_{k\gamma}^{c,n}]) + \sqrt{2}(1 - \mathrm{e}^{-2\gamma})^{1/2} \boldsymbol{Z}_{k\gamma}^\gamma, \tag{F.71}$$

where $(\boldsymbol{Z}_{k\gamma}^\gamma)_{k \in \mathbb{N}, \gamma > 0}$ is a family of independent d-dimensional Gaussian random variables with zero mean and identity covariance matrix and for any $k \in \mathbb{N}$, $\gamma > 0$, $\boldsymbol{X}_{k\gamma}^{c,\gamma} = \mathrm{e}^{-t_0} x_0^c + (1 - \mathrm{e}^{-2t_0})^{1/2} \boldsymbol{Z}_{k\gamma}^{0,\gamma}$, where $(\boldsymbol{Z}_{k\gamma}^{0,\gamma})_{k \in \mathbb{N}, \gamma > 0}$ is a family of independent p-dimensional Gaussian random variables with zero mean and identity covariance matrix. This is a *linear interpolation* of the `Repaint` algorithm in the form of (F.67).

Finally, we denote $(\boldsymbol{X}_t)_{t \geq 0}$ such that

$$\mathrm{d}\boldsymbol{X}_t = b(\boldsymbol{X}_t)\mathrm{d}t + 2\boldsymbol{B}_t, \qquad \boldsymbol{X}_0 = x_0. \tag{F.72}$$

We recall that b depends on t_0 but t_0 is *fixed* here. This means that we are at time t_0 in the diffusion and consider a *corrector* at this stage. The variable t does not corresponds to the backward evolution but to the forward evolution *in the corrector stage*. Under Assumption F.4, (F.72) admits a unique strong solution. The rest of the section is dedicated to the proof of the following result.

Theorem F.5. *Assume assumption F.4. Then* $\lim_{n \to \infty} \left(\boldsymbol{X}_t^{1/n} \right)_{t \geq 0} = (x \boldsymbol{X}_t)_{t \geq 0}$.

This result is an application of Stroock and Varadhan (2007, Theorem 11.2.3). It explains the *continuous* limit of the `Repaint` algorithm (Lugmayr et al., 2022).

In what follows, we verify that the assumptions of this result hold in our setting. For any $\gamma > 0$ and $x \in \mathbb{R}^d$, we define

$$b_\gamma(x) = (2/\gamma)[(1 - \mathrm{e}^{-\gamma}) \int_{\mathbb{R}^d} \nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x^c_{t_0}]) p_{t_0|0}(x^c_{t_0}|x^c_0) \mathrm{d}x^c_{t_0} \tag{F.73}$$

$$- (1/\gamma)\, \mathbb{E}\left[(\boldsymbol{X}^\gamma_{(k+1)\gamma} - \boldsymbol{X}^\gamma_{k\gamma}) \mathbf{1}_{\left\| X^\gamma_{(k+1)\gamma} - X^\gamma_{k\gamma} \right\| \geq 1} | \boldsymbol{X}_{k\gamma} = x \right], \tag{F.74}$$

$$\Sigma_\gamma(x) = (4/\gamma)(1 - \mathrm{e}^{-\gamma})^2 \int_{\mathbb{R}^d} \nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x^c_{t_0}])^{\otimes 2} p_{t_0|0}(x^c_{t_0}|x^c_0) \mathrm{d}x^c_{t_0} + (2/\gamma)(1 - \mathrm{e}^{-2\gamma}) \mathbf{I} \tag{F.75}$$

$$- (1/\gamma)\, \mathbb{E}\left[(\boldsymbol{X}^\gamma_{(k+1)\gamma} - \boldsymbol{X}^\gamma_{k\gamma})^{\otimes 2} \mathbf{1}_{\left\| X^\gamma_{(k+1)\gamma} - X^\gamma_{k\gamma} \right\| \geq 1} | \boldsymbol{X}_{k\gamma} = x \right]. \tag{F.76}$$

Note that for any $\gamma > 0$ and $x \in \mathbb{R}^d$, we have

$$b_\gamma(x) = \mathbb{E}\left[\mathbf{1}_{\left\| X^\gamma_{(k+1)\gamma} - X^\gamma_{k\gamma} \right\| \leq 1} (\boldsymbol{X}^\gamma_{(k+1)\gamma} - \boldsymbol{X}^\gamma_{k\gamma}) | \boldsymbol{X}^\gamma_{k\gamma} = x \right] \tag{F.77}$$

$$\Sigma_\gamma(x) = \mathbb{E}\left[\mathbf{1}_{\left\| X^\gamma_{(k+1)\gamma} - X^\gamma_{k\gamma} \right\| \leq 1} (\boldsymbol{X}^\gamma_{(k+1)\gamma} - \boldsymbol{X}^\gamma_{k\gamma})^{\otimes 2} | \boldsymbol{X}^\gamma_{k\gamma} = x \right] \tag{F.78}$$

$$\tag{F.79}$$

Lemma F.6. *Assume* Assumption F.4. *Then, we have that for any* $R, \varepsilon > 0$ *and* $\gamma \in (0, 1)$

$$\lim_{\gamma \to 0} \sup \left\{ \left\| \Sigma_\gamma(x) - \Sigma(x) \right\| \ : \ x \in \mathbb{R}^d, \ \|x\| \leq R \right\} = 0, \tag{F.80}$$

$$\lim_{\gamma \to 0} \sup \left\{ \left\| b_\gamma(x) - b(x) \right\| \ : \ x \in \mathbb{R}^d, \ \|x\| \leq R \right\} = 0, \tag{F.81}$$

$$\lim_{\gamma \to 0} (1/\gamma) \sup \left\{ \mathbb{P}\left(\left\| \boldsymbol{X}^\gamma_{(k+1)\gamma} - \boldsymbol{X}^\gamma_{k\gamma} \right\| \geq \varepsilon \, | \boldsymbol{X}_{k\gamma} = x \right) \ : \ x \in \mathbb{R}^d, \ \|x\| \leq R \right\} = 0. \tag{F.82}$$

Where we recall that for any $x \in \mathbb{R}^d$,

$$b(x) = 2 \int_{\mathbb{R}^p} \nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x^c_{t_0}]) p^x_{t|0}(x^c_{t_0}|x^c_0) \mathrm{d}x^c_{t_0}, \qquad \Sigma(x) = 4\mathbf{I}. \tag{F.83}$$

Proof. Let $R, \varepsilon > 0$ and $\gamma \in (0, 1)$. Using Assumption F.4, there exists $C > 0$ such that for any $x_{t_0} \in \mathbb{R}^d$ with $\left\| x_{t_0} \right\| \leq R$, we have $\left\| \nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x^c_{t_0}]) \right\| \leq C(1 + \left\| x^c_{t_0} \right\|)$. Since $p^c_{t_0|0}$ is Gaussian with zero mean and covariance matrix $(1 - \mathrm{e}^{-2t_0})\mathbf{I}$, we get that for any $p \in \mathbb{N}$, there exists $A_k \geq 0$ such that for any $x_{t_0} \in \mathbb{R}^d$ with $\left\| x_{t_0} \right\| \leq R$

$$\int_{\mathbb{R}^d} \left\| \nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x^c_{t_0}]) \right\|^p p^c_{t_0|0}(x^c_{t_0}|x^c_0) \mathrm{d}x^c_{t_0} \leq A_k(1 + \left\| x^c_0 \right\|^p). \tag{F.84}$$

Therefore, using this result and the fact that for any $s \geq 0$, $\mathrm{e}^{-s} \geq 1 - s$, we get that there exists $B_k \geq 0$ such that for any $k, p \in \mathbb{N}$ and for any $x_{t_0} \in \mathbb{R}^d$ with $\left\| x_{t_0} \right\| \leq R$

$$\mathbb{E}\left[\left\| \boldsymbol{X}_{(k+1)\gamma} - \boldsymbol{X}_{k\gamma} \right\|^p | \boldsymbol{X}_{k\gamma} = x \right] \leq B_k \gamma^{p/2} (1 + \left\| x^c_0 \right\|^p). \tag{F.85}$$

Therefore, combining this result and the Markov inequality, we get that for any $x_{t_0} \in \mathbb{R}^d$ with $\left\|x_{t_0}\right\| \leq R$ we have

$$\lim_{\gamma\to 0}(1/\gamma)\sup\left\{\mathbb{P}\left(\left\|\boldsymbol{X}^{\gamma}_{(k+1)\gamma} - \boldsymbol{X}^{\gamma}_{k\gamma}\right\| \geq \varepsilon \,|\boldsymbol{X}_{k\gamma} = x\right) \,:\, x \in \mathbb{R}^d,\ \|x\| \leq R\right\} = 0, \tag{F.86}$$

$$\lim_{\gamma\to 0}(1/\gamma)\left\|\mathbb{E}\left[(\boldsymbol{X}^{\gamma}_{(k+1)\gamma} - \boldsymbol{X}^{\gamma}_{k\gamma})\mathbf{1}_{\left\|\boldsymbol{X}^{\gamma}_{(k+1)\gamma} - \boldsymbol{X}^{\gamma}_{k\gamma}\right\|\geq 1}|\boldsymbol{X}_{k\gamma} = x\right]\right\| = 0, \tag{F.87}$$

$$\lim_{\gamma\to 0}(1/\gamma)\left\|\mathbb{E}\left[(\boldsymbol{X}^{\gamma}_{(k+1)\gamma} - \boldsymbol{X}^{\gamma}_{k\gamma})\mathbf{1}_{\left\|\boldsymbol{X}^{\gamma}_{(k+1)\gamma} - \boldsymbol{X}^{\gamma}_{k\gamma}\right\|\geq 1}|\boldsymbol{X}_{k\gamma} = x\right]\right\| = 0 \tag{F.88}$$

In addition, we have that for any $x_{t_0} \in \mathbb{R}^d$ with $R > 0$

$$|(2/\gamma)(1-\mathrm{e}^{-\gamma}) - 2|\left\|\int_{\mathbb{R}^d} \nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x^c_{t_0}])p_{t_0|0}(x^c_{t_0}|x^c_0)\mathrm{d}x^c_{t_0}\right\| \tag{F.89}$$

$$\leq A_1(1 + \left\|x^c_0\right\|)(2/\gamma)|\mathrm{e}^{-\gamma} - 1 + \gamma|. \tag{F.90}$$

We also have that for any $x_{t_0} \in \mathbb{R}^d$ with $R > 0$

$$(4/\gamma)|1-\mathrm{e}^{-\gamma}|^2\left\|\int_{\mathbb{R}^d} \nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x^c_{t_0}])^{\otimes 2}p_{t_0|0}(x^c_{t_0}|x^c_0)\mathrm{d}x^c_{t_0}\right\| \tag{F.91}$$

$$\leq A_2(1 + \left\|x^c_0\right\|^2)(4/\gamma)|1-\mathrm{e}^{-\gamma}|^2. \tag{F.92}$$

Combining this result, (F.86), the fact that $\lim_{\gamma\to 0}(4/\gamma)|1-\mathrm{e}^{-\gamma}|^2 = 0$ and $\lim_{\gamma\to 0}(2/\gamma)|\mathrm{e}^{-\gamma} - 1 +$ 0, we get that $\lim_{\gamma\to 0}\sup\left\{\left\|\Sigma_\gamma(x) - \Sigma(x)\right\| \,:\, x \in \mathbb{R}^d,\ \|x\| \leq R\right\} = 0$. Similarly, using (F.86), (F.89) and the fact that $\lim_{\gamma\to 0}(4/\gamma)|1-\mathrm{e}^{-\gamma}|^2 = 0$, we get that $\lim_{\gamma\to 0}\sup\left\{\left\|b_\gamma(x) - b($ 0. ■

We can now conclude the proof of Theorem F.5.

Proof. We have that $x \mapsto b(x)$ and $x \mapsto \Sigma(x)$ are continuous. Combining this result and Lemma F.6, we conclude the proof upon applying Stroock and Varadhan (2007, Theorem 11.2.3). ■

Theorem F.5 is a non-quantitative result which states what is the limit chain for the RePaint procedure. Note that if we do not resample, we get that

$$b^{\mathrm{cond}}(x) = 2\nabla_{x_{t_0}} \log p_{t_0}([x_{t_0}, x^c_{t_0}]), \qquad \Sigma(x) = 4\mathbf{I}. \tag{F.93}$$

Recalling (F.83), we get that (F.93) is an *amortised version* of b^{cond}. Similar convergence results can be derived in this case. Note that it is also possible to obtain quantitative discretization bounds between $(\boldsymbol{X}_t)_{t\geq 0}$ and $(\boldsymbol{X}^{1/n}_t)_{t\geq 0}$ under the ℓ^2 distance. These bounds are usually leveraged using the Girsanov theorem (Durmus and Moulines, 2017; Dalalyan, 2017). We leave the study of such bounds for future work.

We also remark that $b(x_{t_0})$ is *not* given by $\nabla \log p_{t_0}(x_{t_0}|x^c_0)$. Denoting U_{t_0} such that for any $x_{t_0} \in \mathbb{R}^d$

$$U_{t_0}(x_{t_0}) = -\int_{\mathbb{R}^p}(\log p_{t_0}(x_{t_0}|x^c_{t_0}))p_{t|0}(x^c_{t_0}|x^c_0)\mathrm{d}x^c_{t_0}, \tag{F.94}$$

we have that $\nabla U_{t_0}(x_{t_0}) = -b(x_{t_0})$, under mild integration assumptions. In addition, using Jensen's inequality, we have

$$\int_{\mathbb{R}^d} \exp[-U_{t_0}(x_{t_0})]\mathrm{d}x_{t_0} \leq \int_{\mathbb{R}^d}\int_{\mathbb{R}^p} p_{t_0}(x_{t_0}|x^c_{t_0}) p_{t|0}(x^c_{t_0}|x^c_0)\mathrm{d}x_{t_0}\mathrm{d}x^c_{t_0} \leq 1. \tag{F.95}$$

Hence, π_{t_0} with density proportional to $x \mapsto \exp[-U_{t_0}(x)]$ defines a valid probability measure.

We make the following assumption which allows us to control the ergodicity of the process $(\boldsymbol{X}_t)_{t\geq 0}$.

ASSUMPTION F.7. *There exist* $\mathtt{m} > 0$ *and* $C \geq 0$ *such that for any* $x_{t_0} \in \mathbb{R}^d$ *and* $x^c_{t_0} \in \mathbb{R}^p$

$$\langle \nabla_{x_t} \log p_{t_0}([x_t, x^c_t]), x_t\rangle \leq -\mathtt{m}\|x_t\|^2 + C(1 + \left\|x^c_t\right\|^2). \tag{F.96}$$

The following proposition ensures the ergodicity of the chain $(\boldsymbol{X}_t)_{t\geq 0}$. It is a direct application of Roberts and Tweedie (1996, Theorem 2.1).

PROPOSITION F.8. *Assume* Assumption F.4 *and* Assumption F.7. *Then,* π_{t_0} *is the unique invariant probability measure of* $(\boldsymbol{X}_t)_{t\geq 0}$ *and* $\lim_{t\to 0}\left\|\mathcal{L}(\boldsymbol{X}_t) - \pi_{t_0}\right\|_{\mathrm{TV}} = 0$, *where* $\mathcal{L}(\boldsymbol{X}_t)$ *is the distribution of* $\boldsymbol{X}_t$.

Finally, for any $t_0 > 0$, denoting π_{t_0} the probability measure with density U_{t_0} given for any $x_{t_0} \in \mathbb{R}^d$ by

$$U_{t_0}(x_{t_0}) = -\int_{\mathbb{R}^p} (\log p_{t_0}(x_{t_0}|x^c_{t_0})) p_{t|0}(x^c_{t_0}|x^c_0)\mathrm{d}x^c_{t_0}. \tag{F.97}$$

We show that the family of measures $(\pi_{t_0})_{t_0>0}$ approximates the posterior with density $x_0 \mapsto p(x_0|x^c_0)$ when t_0 is small enough.

PROPOSITION F.9. *Assume* Assumption F.4. *We have that* $\lim_{t_0\to 0} \pi_{t_0} = \pi_0$ *where* π_0 *admits a density with respect to. the Lebesgue measure given by* $x_0 \mapsto p(x_0|x^c_0)$.

Proof. This is a direct consequence of the fact that $p_{t|0}(\cdot|x^c_0) \to \delta_{x^c_0}$. ∎

This last results shows that even though we do not target $x_{t_0} \mapsto p_{t_0|0}(x_{t_0}|x^c_0)$ using this corrector term, we still target $p(x_0|x^c_0)$ as $t_0 \to 0$ which corresponds to the desired output of the algorithm.

F.4. EXPERIMENTAL DETAILS

Models, training and evaluation have been implemented in `Jax` (Bradbury et al., 2018). We used Python (Van Rossum and Drake Jr, 1995) for all programming, Hydra (Yadan, 2019), Numpy (Harris et al., 2020), Scipy (Virtanen et al., 2020), Matplotlib (Hunter, 2007), and Pandas (McKinney et al., 2010).

F.4.1. Regression 1d.

Data generation

We follow the same experimental setup as Bruinsma et al. (2020) to generate the 1d synthetic data. It consists of Gaussian (Squared Exponential (SE), Matérn($\frac{5}{2}$), Weakly Periodic) and non-Gaussian (Sawtooth and Mixture) sample paths, where Mixture is a combination of the other four datasets with equal weight. Figure 6.4 shows samples for each of these dataset. The Gaussian datasets are corrupted with observation noise with variance $\sigma^2 = 0.05^2$. The left column of figure 6.4 shows example sample paths for each of the 5 datasets.

The training data consists of 2^{14} sample paths while the test dataset has 2^12 paths. For each test path we sample the number of context points between 1 and 10, the number of target points are fixed to 50 for the GP datasets and 100 for the non-Gaussian datasets. The input range for the training and interpolation datasets is $[-2, 2]$ for both the context and target sets, while for the extrapolation task the context and target input points are drawn from $[2, 6]$.

Architecture. For all datasets, except Sawtooth, we use 5 bi-dimensional attention layers (Dutordoir et al., 2023) with 64 hidden dimensions and 8 output heads. For Sawtooth, we obtained better performance with a wider and shallower model consisting of 2 bi-dimensional attention layers with a hidden dimension of 128. In all experiment, we train the NDP-based models over 300 epochs using a batch size of 256. Furthermore, we use the Adam optimiser for training with the following learning rate schedule: linear warm-up for 10 epochs followed by a cosine decay until the end of training.

Ablation Limiting Kernels

The test log-likelihoods (TLLs) reported in appendix F.4.1 for the NDP models target a white limiting kernel and train to approximate the preconditioned score $K\nabla \log p_t$. Overall, we found this to be the best performing setting. Appendix F.4.1 shows an ablation study for different choices of limiting kernel and score parametrisation. We refer to table F.1 for a detailed derivation of the score parametrisations.

The dataset in the top row of the figure originates from a Squared Exponential (SE) GP with lengthscale $\ell = 0.25$. We compare the performance of three different limiting kernels: white (blue), a SE with a longer lengthscale $\ell = 1$ (orange), and a SE with a shorter lengthscale $\ell = 0.1$ (green). As the dataset is Gaussian, we have access to the true score. We observe that, across the different parametrisations, the white limiting kernel performance best. However, note that for the White kernel $K = I$ and thus the different parametrisations become identical. For non-white limiting kernels we see a reduction in performance for both the approximate and exact score. We attribute this to the additional complexity of learning a non-diagonal covariance.

In the bottom row of appendix F.4.1 we repeat the experiment for a dataset consisting of samples from the Periodic GP with lengthscale 0.5. We draw similar conclusions: the best performing limiting kernel, across the different parametrisations, is the White noise kernel.

Ablation Conditional Sampling

Next, we focus on the empirical performance of the different noising schemes in the conditional sampling, as discussed in figure 6.3. For this, we measure the the Kullback-Leibler (KL) divergence between two Gaussian distributions: the true GP-based conditional distribution, and an distribution created by drawing conditional sampling from the model and fitting a Gaussian to it using the empirical mean and covariance. We perform this test on the 1D squared exponential dataset (described above) as this gives us access to the true posterior. We use 2^{12} samples to estimate the empirical mean and covariance, and fix the number of context points to 3.

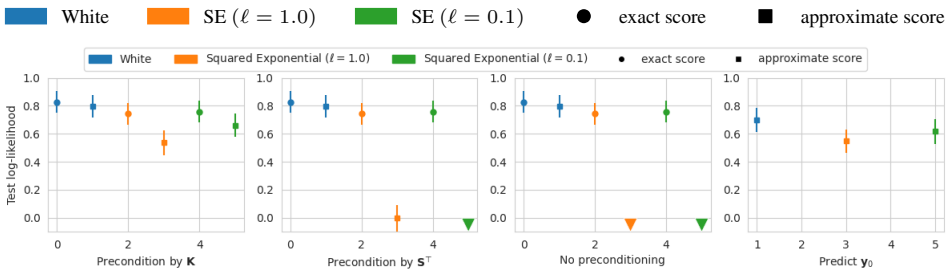
In figure F.2 we keep the total number of score evaluations fixed to 5000 and vary the number of steps in the inner (L) loop such that the number of outer steps is given by the ratio $5000/L$. From the figure, we observe that the particular choice of noising scheme is of less importance as long at least a couple (± 5) inner steps are taken. We further note that in this experiment we used the true score (available because of the Gaussianity of the dataset), which means that these results may differ if an approximate score network is used.

Table F.2. Mean test log-likelihood (TLL) ± 1 standard error estimated over 4096 test samples are reported. Statistically significant best non-GP model is in bold. The NP baselines (GNP, ConvCNP, ConvNP and ANP) are quoted from Bruinsma et al. (2020). '*' stands for a TLL below -10.

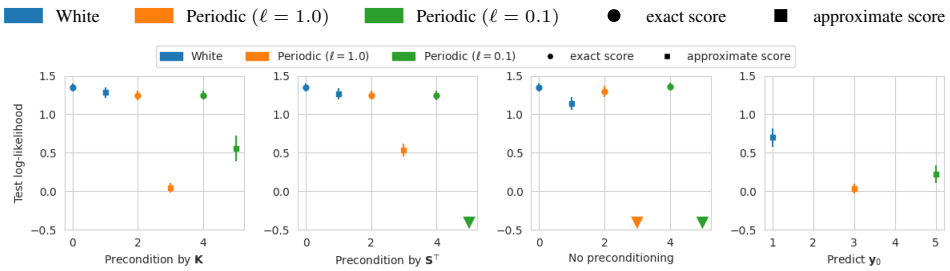
	SE	Matérn$-\frac{5}{2}$	Weakly Per.	Sawtooth	Mixture
Interpolation					
GP (optimum)	0.70±0.00	0.31±0.00	−0.32±0.00	n/a	n/a
$T(1)$−GeomNDP	**0.72**±0.03	**0.32**±0.03	**−0.38**±0.03	**3.39**±0.04	**0.64**±0.08
NDP*	**0.71**±0.03	**0.30**±0.03	**−0.37**±0.03	**3.39**±0.04	**0.64**±0.08
GNP	**0.70**±0.01	**0.30**±0.01	−0.47±0.01	0.42±0.01	0.10±0.02
ConvCNP	−0.80±0.01	−0.95±0.01	−1.20±0.01	0.55±0.02	−0.93±0.02
ConvNP	−0.46±0.01	−0.67±0.01	−1.02±0.01	1.20±0.01	−0.50±0.02
ANP	−0.61±0.01	−0.75±0.01	−1.19±0.01	0.34±0.01	−0.69±0.02
Generalisation					
GP (optimum)	0.70±0.00	0.31±0.00	−0.32±0.00	n/a	n/a
$T(1)$−GeomNDP	**0.70**±0.02	**0.31**±0.02	**−0.38**±0.03	**3.39**±0.03	**0.62**±0.02
NDP*	*	*	*	*	*
GNP	**0.69**±0.01	**0.30**±0.01	−0.47±0.01	0.42±0.01	0.10±0.02
ConvCNP	−0.81±0.01	−0.95±0.01	−1.20±0.01	0.53±0.02	−0.96±0.02
ConvNP	−0.46±0.01	−0.67±0.01	−1.02±0.01	1.19±0.01	−0.53±0.02
ANP	−1.42±0.01	−1.34±0.01	−1.33±0.00	−0.17±0.00	−1.24±0.01

F.4.2. Gaussian process vector fields.

Data We create synthetic datasets using samples from two-dimensional zero-mean Gaussian Processes with the following E(n)2-equivariant kernels: a diagonal Squared-Exponential (SE) kernel, a zero curl (Curl-free) kernel and a zero divergence (Div-free) kernel, as described in appendix F.3.1. We set the variance to $\sigma^2 = 1$ and the lengthscale to $\ell = \sqrt{5}$. We evaluate these Gaussian Processes on a disk grid, created via a 2D grid with 30×30 points regularly space on $[-10, 10]^2$



(a) Squared Exponential dataset with lengthscale $\ell = 0.25$



(b) Periodic dataset with lengthscale $\ell = 0.25$

Figure F.1. *Ablation study* targeting different limiting kernels and score parametrisations.

and keeping only the points inside the disk of radius 10. We create a training dataset of size 80×10^3. and a test dataset of size 10×10^3.

Models We compare two flavours of our model GeomNDP . One with a non-equivariant attention-based score network (Figure C.1, Dutordoir et al., 2023), referred as NDP*. Another one with a E(n)2-equivariant score architecture, based on steerable CNNs (Thomas et al., 2018; Weiler et al., 2018). We rely on the `e3nn` library (Geiger and Smidt, 2022) for implementation. A knn graph $\mathcal{E}$ is built with $k = 20$. The pairwise distances are first embed into $\mu(r_{ab})$ with a 'smooth_finite' basis of 50 elements via `e3nn.soft_one_hot_linspace`, and with a maximum radius of 2. The time is mapped via a sinusoidal embedding $\phi(t)$ (Vaswani et al., 2017). Then edge features are obtained as $e_{ab} = \Psi^{(e)}(\mu(r_{ab})||\phi(t))\ \forall (a,b) \in \mathcal{E}_k$ with $\Psi^{(e)}$ an MLP with 2 hidden layers of width 64. We use 5 `e3nn.FullyConnectedTensorProduct` layers with update given by $V_a^{k+1} = \sum_{b \in \mathcal{N}(a,\mathcal{E}_k)} V_a^k \otimes \left(\Psi^v(e_{ab}||V_a^k||V_b^k)\right) Y(\hat{r}_{ab})$ with Y spherical harmonics up to order 2m Ψ^v an MLP with 2 hidden layers of width 64 acting on invariant features, and node features V^k having irreps `12x0e + 12x0o + 4x1e + 4x1o`. Each layer has a gate non-linearity (Weiler et al., 2018).

We also evaluate two neural processes, a translation-equivariant ConvCNP (Gordon et al., 2020) with decoder architecture based on 2D convolutional layers (LeCun et al., 1998) and a C4 $\ltimes \mathbb{R}^2 \subset$ E(n)2-equivariant SteerCNP (Holderrieth et al., 2021b) with decoder architecture based on 2D steerable convolutions (Weiler et al., 2023). Specific details can be found in the accompanying codebase https://github.com/PeterHolderrieth/Steerable_CNPs of Holderrieth et al. (2021b).

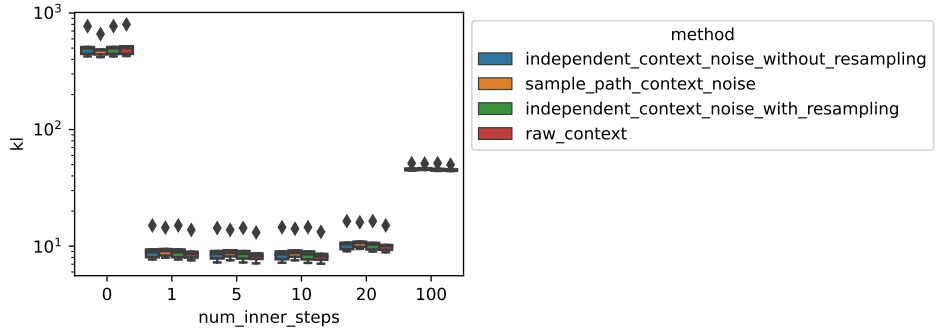


Figure F.2. Ablation noising schemes for conditional sampling.

Optimisation. Models are trained for $80k$ iterations, via (Kingma, 2014) with a learning rate of $5e-4$ and a batch size of 32. The neural diffusion processes are trained unconditionally, that is we feed GP samples evaluated on the full disk grid. Their weights are updated via with exponential moving average, with coefficient 0.99. The diffusion coefficient is weighted by $\beta: \ t \mapsto \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot t$, and $\beta_{\min} = 1e-4$, $\beta_{\max} = 15$.

As standard, the neural processes are trained by splitting the training batches into a context and evaluation set, similar to when evaluating the models. Models have been trained on A100-SXM-80GB GPUs.

Evaluation. We measure the predictive log-likelihood of the data process samples under the model on a held-out test dataset. The context sets are of size 25 and uniformly sampled from a disk grid of size 648, and the models are evaluated on the complementary of the grid. For neural diffusion processes, we estimate the likelihood by solving the associated probability flow ODE (F.46). The divergence is estimated with the Hutchinson estimator, with Rademacher noise, and 8 samples, whilst the ODE is solved with the 2nd order Heun solver, with 100 discretisation steps.

We also report the performance of the data-generating GP, and the same GP but with diagonal posterior covariance GP (DIAG.) .

F.4.3. Tropical cyclone trajectory prediction. Models.

Four models were evaluated.

The GP ($\mathbb{R} \to \mathbb{R}^2$) took the raw latitude-longitude data and normalised it. Using a 2-output RBF kernel with no covariance between the latitude and longitude and taking the cyclone time as input, placed a GP over the data. The hyperparameters of this kernel were optimised using a maximum likelihood grid search over the data. Note that this model places density outside the bounding box of $[-90, 90] \times [-180, 180]$ that defines the range of latitude and longitude, and so does not place a proper distribution on the space of paths on the sphere.

The STEREOGRAPHIC GP ($\mathbb{R} \to \mathbb{R}^2/\{0\}$) instead transformed the data under a stereographic projection centred at the north pole, and used the same GP and

optimisation as above. Since this model only places density on a set of measure zero that does not correspond to the sphere, it does induce a proper distribution on the space of paths on the sphere.

The NDP ($\mathbb{R} \to \mathbb{R}^2$) uses the same preprocessing as GP ($\mathbb{R} \to \mathbb{R}^2$) but uses a Neural Diffusion Process from (Dutordoir et al., 2023) to model the data. This has the same shortcomings as the GP ($\mathbb{R} \to \mathbb{R}^2$) in not placing a proper density on the space of paths on the sphere. The network used for the score function and the optimisation procedure is detailed below. A linear beta schedule was used with $\beta_0 = 1e-4$ and $\beta_1 = 10$. The reverse model was integrated back to $\epsilon = 5e-4$ for numerical stability. The reference measure was a white noise kernel with a variance 0.05. ODEs and SDEs were discretised with 1000 steps.

The GeomNDP ($\mathbb{R} \to \mathcal{S}^2$) works with the data projected into 3d space on the surface of the sphere. This projection makes no difference to the results of the model, but makes the computation of the manifold functions such as the exp map easier, and makes it easier to define a smooth score function on the sphere. This is done by outputting a vector for the score from the neural network in 3d space, and projecting it onto the tangent space of the sphere at the given point. For the necessity of this, see (De Bortoli et al., 2021). The network used for the score function and the optimisation procedure is detailed below. A linear beta schedule was used with $\beta_0 = 1e-4$ and $\beta_1 = 15$. The reverse model was integrated back to $\epsilon = 5e-4$ for numerical stability. The reference measure was a white noise kernel with a variance 0.05. ODEs and SDEs were discretised with 1000 steps.

Neural network. The network used to learn the score function for both NDP ($\mathbb{R} \to \mathbb{R}^2$) and GeomNDP ($\mathbb{R} \to \mathcal{S}^2$) is a bi-attention network from Dutordoir et al. (2023) with 5 layers, hidden size of 128 and 4 heads per layer. This results in 924k parameters.

Optimisation. NDP ($\mathbb{R} \to \mathbb{R}^2$) and GeomNDP ($\mathbb{R} \to \mathcal{S}^2$) were both optimised using (correctly implemented) Adam for 250k steps using a batch size of 1024 and global norm clipping of 1. Batches were drawn from the shuffled data and refreshed each time the dataset was exhausted. A learning rate schedule was used with 1000 warmup steps linearly from 1e-5 to 1e-3, and from there a cosine schedule decaying from 1e-3 to 1e-5. With even probability either the whole cyclone track was used in the batch, or 20 random points were sub-sampled to train the model better for the conditional sampling task.

Conditional sampling. The GP models used closed-form conditional sampling as described. Both diffusion-based models used the Langevin sampling scheme described in this work. 1000 outer steps were used with 25 inner steps. We use a $\psi = 1.0$ and $\lambda_0 = 2.5$. In addition at the end of the Langevin sampling, we run an additional 150 Langevin steps with $t = \epsilon$ as this visually improved performance.

Evaluation. For the model (conditional) log probabilities the GP models were computed in closed form. For the diffusion-based models, they were computed using the auxiliary likelihood ODE discretised over 1000 steps. The conditional probabilities were computed via the difference between the log-likelihood of the whole trajectory and the log-likelihood of the context set only. The mean squared errors were computed using the geodesic distance between 10 conditionally sampled

trajectories, described above.