Generative Models: Theory and Applications

Joe Benton
St Peter's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas 2023

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the intellectual contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification. My personal contributions are as detailed in the authorship forms at the end of each chapter. This dissertation is my own work except as specified in the text, acknowledgements and authorship forms.

Joe Benton
Michaelmas 2023

Acknowledgements

This thesis could not have existed without the support of my supervisors Arnaud Doucet and George Deligiannidis. Their formal supervision was of course essential, but far more valuable were their frank advice, their sustained encouragement, and their impeccable research instincts. I am incredibly grateful for their support to shape the PhD into the experience that I wanted it to be.

I have been very lucky to work with such an excellent group of collaborators in the Oxford Computational Statistics and Machine Learning group. Particular thanks must go to my coauthors: to Valentin for his early guidance and direction, to Kamélia for her thoughtfulness and razor-sharp wit, and to Yuyang and Andrew for their computational expertise and for negotiating with the cluster on my behalf.

The Department of Statistics and the StatML CDT of which I have been a part have provided a wonderful and stimulating research environment. I am grateful to the very large number of people who have worked to make that so. To my StatML colleagues and friends Alex, Angus, Max, Nic and Vik, thank you for making the whole journey such an enjoyable ride. To Anna and Sahra, thank you for welcoming us all into office G.05. Special mentions must also go to the Engineering and Physical Sciences Research Council, for their funding, and to Joanna, for her baking.

I am grateful also for the broader experience that I have been afforded during my time in Oxford. The PhD would have been a far less colourful endeavour without the people I met at EA Oxford, OUU, OUCCC, and in St Peter's College MCR. In addition, I am indebted to Redwood Research, for the opportunity to visit them in Berkeley, and to everyone whom I met there, for the horizon-expanding experience.

Finally, thank you to all of my friends, both old and new, who kept me smiling throughout. There are far too many of you to name individually, but in particular, to Anya, for listening, to Kitti, for understanding, to Jake, for encouraging, and to George, for never letting me lose sight of what was most important. And of course, to Mum, Dad, Sam and Daniel, for simply being there, always.

Abstract

Given some samples from a data distribution, the central aim of generative modeling is to generate more samples from approximately the same distribution. This framework has recently seen an explosion in popularity, with impressive applications in image generation, language modeling and protein synthesis. The striking success of these methods motivates two key questions: under what conditions do generative models provide accurate approximations to the underlying data distribution, and can we extend the range of scenarios in which they may be applied? This thesis considers these questions in the context of two classes of generative models: diffusion models and importance weighted autoencoders.

Diffusion models work by iteratively applying noise to the data distribution and then learning to remove this noise. They were originally introduced for real-valued data. However, for many potential applications our data is most naturally defined on another state space – perhaps a manifold, or a discrete space. We describe extensions of diffusion models to arbitrary state spaces, using generic Markov processes for noising, and show how such models can be effectively learned. We also provide a detailed study of a specific extension to discrete state spaces. Next, we investigate the approximation accuracy of diffusion models. We derive error bounds for flow matching – a generalization of diffusion models – and improve upon state-of-the-art bounds for diffusion models, using techniques inspired by stochastic localization.

Importance weighted autoencoders (IWAEs) work by learning a latent variable representation of the data, using importance sampling in the evidence lower bound to get a tighter variational objective. IWAEs suffer from several limitations, including posterior variance underestimation, poor signal-to-noise ratio during training, and weight collapse in the importance sampling ratios. We propose an extension of the IWAE – the VR-IWAE – that addresses the first two of these three issues. We then provide a detailed theoretical study of the third, showing that it persists even for the VR-IWAE. We provide empirical demonstrations of these phenomena on a range of simulated and real-world data.

Contents

1 Introduction and Literature Review

Contents

1.1 Motivation

Many of the most fundamental tasks which we may design algorithms to perform involve, at some basic level, generating data. For example, we may desire algorithms that can synthesize audio, write coherent text or draw images to match a description. However, for most interesting tasks, the distribution of plausible data is very complex and hard to specify explicitly – describing precisely what distinguishes a convincing image or piece of text is incredibly challenging, and initial attempts to build artificial intelligence systems that could perform these tasks to a human level mostly failed for this reason. Fortunately, for many types of data which we wish to generate, there are large quantities of existing data already available. This raises the possibility that we could leverage this existing data to learn something about the underlying

distribution, and then use what we have learned to inform our generation of new data.

Mathematically, this perspective is encapsulated in the core problem of generative modeling: given samples from a data distribution $p_{\text{data}}(\mathbf{x})$, can we generate additional synthetic samples coming from approximately the same distribution? More generally, we ask whether, given samples from a joint distribution $p_{\text{data}}(\boldsymbol{\xi}, \mathbf{x})$, we can generate approximate samples from the conditional distribution $p_{\text{data}}(\mathbf{x}|\boldsymbol{\xi})$. This allows us to, for example, generate images $\mathbf{x}$ conditioned on a given text prompt $\boldsymbol{\xi}$. Models that can perform this task are known as *generative models* and can be a hugely powerful technique for generating diverse types of conditional data.

In the last decade, the rise of deep learning has inspired an array of neural-network-based techniques for generative modeling. These include generative adversarial networks (Goodfellow et al., 2014), variational autoencoders (VAEs) (Kingma et al., 2014), normalizing flows (Rezende et al., 2015), autoregressive models (Oord et al., 2016b), diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b), and flow matching methods (Lipman et al., 2023; Albergo et al., 2023b). These models have lead to striking progress across a wide variety of domains, including image synthesis (Karras et al., 2020; Vahdat et al., 2020; Ramesh et al., 2022; Saharia et al., 2022), text generation (Brown et al., 2020; Anil et al., 2023), audio generation (Oord et al., 2016a; Popov et al., 2021; Le et al., 2023) and molecular structure generation (Ingraham et al., 2019; Trippe et al., 2023).

The incredible success of such methods raises two questions, which will be the central themes of this thesis:

- Can we understand the conditions under which these generative models provide accurate approximations to the underlying data distribution?
- Can we expand the range of scenarios to which these generative models may be applied?

While the space of possible generative modeling techniques is very large, in this thesis we choose to focus on a subset of existing techniques. In the first part of the thesis, we study diffusion modeling and its close cousin flow matching. In the second,

we turn our attention to variational autoencoders, and in particular an extension known as the importance weighted autoencoder (IWAE) (Burda et al., 2016).

Diffusion models were originally developed for data defined on $\mathbb{R}^d$ (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b). However, there are lots of forms of data which are not most naturally represented on $\mathbb{R}^d$. For example, text data is naturally discrete, while geospatial data typically takes value on the sphere S^2. While we could choose to embed these distributions in a real vector space, it may be preferable to develop diffusion models that respect the intrinsic structure of the data distribution. The development of such models is the subject of Chapters 2 and 3 of this thesis.

Second, we investigate how accurate the approximations provided by diffusion models are. Concretely, we study the rate at which diffusion models on $\mathbb{R}^d$ converge to the true data distribution. We hope that understanding these convergence rates provides insight into why diffusion models are empirically so successful. In addition, we hope that it sheds light on the conditions under which they will, or will not, provide an accurate approximation, and potentially motivates the design of more effective diffusion methods. This is the subject of Chapters 4 and 5.

Third, in Chapter 6 we turn to the VAE and its cousin the IWAE, which uses importance sampling in the variational objective to obtain a tighter variational bound. These methods have been observed to have several practical limitations which we seek to address. First, VAEs often suffer from posterior variance underestimation (Minka, 2005), meaning that they typically do not achieve good coverage of the data distribution. Previous work by Li et al. (2016) replaced the Kullback–Leibler (KL) divergence in the VAE objective with Rényi's alpha-divergence (Rényi, 1961) in an attempt to encourage mode-seeking behaviour; we ask if this insight can be generalized to the IWAE setting. Second, work has shown that although the IWAE provides a tighter variational bound, it suffers from a low signal-to-noise ratio (SNR) in practice (Rainforth et al., 2018). We study the effect that our alpha-divergence-based extension of the IWAE has on this problem. Thirdly, we suspect that the importance sampling in the IWAE bound should suffer from weight

collapse in high dimensions (Bengtsson et al., 2008), which may render it ineffective in practice; we aim to provide a quantitative analysis of this phenomenon and demonstrate its practical relevance.

In the next sections, we present an introduction to the main areas of study in this thesis: diffusion models, flow matching and variational autoencoders.

1.2 Diffusion Models

Diffusion models were first introduced in the discrete-time setting by Sohl-Dickstein et al. (2015), and subsequently came to prominence due to work by Ho et al. (2020). The first continuous-time diffusion models were introduced by Song et al. (2021b), who also made the connection with score matching (Song et al., 2019; Hyvärinen, 2005).

Diffusion models consist of a pair of stochastic processes. We start by defining a fixed *forward* or *noising* process $(X_t)_{t\in[0,T]}$, which is initialized according to the data distribution $X_0 \sim p_{\text{data}}$. We label the marginals of the forward process $q_t(\mathbf{x})$. Second, we construct a learned *reverse* or *generative* process $(Y_t)_{t\in[0,T]}$, with marginals denoted by $p_t(\mathbf{x})$. The basic idea is to pick the noising process $(X_t)_{t\in[0,T]}$ such that the marginals q_t converge to some easy-to-sample reference distribution π for large times, and to learn a generative process $(Y_t)_{t\in[0,T]}$ which approximates the reverse dynamics of the noising process. Then, we generate approximate samples from p_{data} by sampling $Y_0 \sim \pi$ and running the reverse process to produce a sample $Y_T \sim p_T$, which should approximate $q_0 = p_{\text{data}}$.

For continuous-time diffusion models on $\mathbb{R}^d$, we define the forward process via the stochastic differential equation (SDE)

$$\mathrm{d}X_t = b(X_t, t)\mathrm{d}t + \mathrm{d}B_t, \quad X_0 \sim p_{\text{data}},$$

for some function $b : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, where $(B_t)_{t\in[0,T]}$ is a standard Brownian motion (Song et al., 2021b). A typical choice is $b(\mathbf{x}, t) = -\frac{1}{2}\mathbf{x}$, in which case our forward process is an Ornstein–Uhlenbeck (OU) process, since the transition

densities $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ are analytically tractable for the OU process and the marginals converge exponentially to the standard Gaussian.

Under suitable regularity conditions on p_{data}, the reverse process satisfies the SDE

$$\mathrm{d}Y_t = \{-b(Y_t, T-t) + \nabla \log q_{T-t}(Y_t)\}\mathrm{d}t + \mathrm{d}B'_t,$$

where $(B'_t)_{t\in[0,T]}$ is another Brownian motion (Anderson, 1982; Cattiaux et al., 2022). Therefore, in order to learn the dynamics of the reverse process it suffices to approximate $\nabla \log q_t(\mathbf{x}_t)$. Typically, we learn a parametric approximation $s_\theta(\mathbf{x}_t, t)$, parameterized via a neural network with parameters $\theta \in \mathbb{R}^D$. We would like to fit θ by minimizing the L^2 objective function

$$\mathcal{I}_{\text{ESM}}(s_\theta) = \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)} \left[\|\nabla \log q_t(\mathbf{x}_t) - s_\theta(\mathbf{x}_t, t)\|^2\right].$$

This objective is not directly tractable, since it involves the marginals $q_t(\mathbf{x}_t)$. However, we can use techniques from score matching (Hyvärinen, 2005; Vincent, 2011) to derive the equivalent denoising score matching objective

$$\mathcal{I}_{\text{DSM}}(s_\theta) = \int_0^T \mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)} \left[\|\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) - s_\theta(\mathbf{x}_t, t)\|^2\right].$$

We fit our diffusion model by minimizing $\mathcal{I}_{\text{DSM}}$ using empirical estimates for the objective based on samples from the forward process and performing stochastic gradient descent on θ. Alternative equivalent objectives from the score matching literature, such as implicit score matching or sliced score matching can also be used for the objective function (Hyvärinen, 2005; Song et al., 2020). Although the use of $\mathcal{I}_{\text{DSM}}$ was initially motivated by Song et al. (2021b) by comparison to the score matching literature (Hyvärinen, 2005; Vincent, 2011), later works showed that this objective could also be derived in a principled fashion as a lower bound on the log-likelihood of the diffusion model (Huang et al., 2021).

Recently, several extensions of diffusion models to state spaces other than $\mathbb{R}^d$ have been proposed, including to discrete spaces (Austin et al., 2021; Hoogeboom et al., 2021) and Riemannian manifolds (De Bortoli et al., 2022; Huang et al., 2022). However, these extensions have typically been ad-hoc, with new training objectives needing to be proposed for each novel setting.

In recent years there have been substantial efforts to obtain theoretical guarantees on the performance of diffusion models, with most results taking the form of a bound on the distance between p_T and p_{data} under some distance measure, making some smoothness assumptions on the data distribution or forward process, and assuming some bound on the error of the score approximation. The first such results were derived by De Bortoli et al. (2021), Block et al. (2022), and De Bortoli (2022), but these all had exponential dependence in the relevant problem parameters. Subsequent works were able to obtain bounds that held under strong smoothness assumptions (e.g. a log-Sobolev inequality) (Lee et al., 2022; Yang et al., 2022), or provided non-quantitative bounds (Pidstrigach, 2022; Liu et al., 2022).

More recently, the first polynomial bounds on the convergence rates under reasonable smoothness assumptions and an L^2 bound on the error of the score matching approximation have been derived (Chen et al., 2023a; Chen et al., 2023c; Lee et al., 2023; Li et al., 2023; Conforti et al., 2023). The most prominent method for such bounds has been to use a version of Girsanov's theorem to bound the KL distance between the path measures of the forward and reverse processes (Chen et al., 2023a; Chen et al., 2023c). This gives a bound that can be expressed as the sum of terms corresponding to the L^2 error of the score approximation, the error from discretizing the reverse SDE and the convergence of the forward process. Methods mostly vary in their treatment of the error arising from the discretization of the reverse SDE.

1.3 Flow Matching

Diffusion models as described in the previous section can provide very high quality approximate samples from many data distributions. However, simulating the reverse process requires many steps and therefore many serial evaluations – often hundreds, or even thousands – of the neural network approximation, making diffusion models expensive to train and sample from (Kong et al., 2021; Zhang et al., 2023).

Various methods to speed up sampling from diffusion models have been proposed. These include using the probability flow ODE (Song et al., 2021b), a deterministic

processes with the same marginal distributions as the reverse SDE which is cheaper to simulate, using adaptive step sizes when simulating the reverse process (Jolicoeur-Martineau et al., 2021), and using denoising diffusion implicit models (Song et al., 2021a), which employ a Gaussian non-Markovian noising process to allow for faster deterministic sampling.

Flow matching, introduced independently by Lipman et al. (2023), Liu et al. (2023), and Albergo et al. (2023b) is another alternative to SDE-based diffusion models which allows for more efficient sampling. The core idea is to abstract away the particular choice of noising process and instead focus directly on the set of interpolating marginals $(p_t)_{t\in[0,T]}$. We first define a set of probability densities $(p_t)_{t\in[0,T]}$ such that $p_0 = \pi$ for some easy-to-sample reference distribution π and $p_T = p_{\text{data}}$. Then, we construct a deterministic process with marginals $(p_t)_{t\in[0,T]}$.

To define the marginals $(p_t)_{t\in[0,T]}$ we start by specifying an interpolating path between any pair of points $(\mathbf{x}_0, \mathbf{x}_T) \in \mathbb{R}^d \times \mathbb{R}^d$. We do this via an interpolant function $I(\mathbf{x}_0, \mathbf{x}_T, t) : \mathbb{R}^d \times \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, which may be either stochastic or deterministic, and which satisfies $I(\mathbf{x}_0, \mathbf{x}_T, 0) = \mathbf{x}_0$ and $I(\mathbf{x}_0, \mathbf{x}_T, T) = \mathbf{x}_T$ for all $\mathbf{x}_0, \mathbf{x}_T \in \mathbb{R}^d$. Then, we define the process $(X_t)_{t\in[0,T]}$ by taking $X_0 \sim \pi$, $X_T \sim p_{\text{data}}$ independently and setting $X_t = I(X_0, X_T, t)$ for each $t \in (0,T)$. We let $(p_t)_{t\in[0,T]}$ be the marginals of the process $(X_t)_{t\in[0,T]}$.

The process $(X_t)_{t\in[0,T]}$ is not a suitable choice of flow between π and p_{data} as it is not causal (i.e. the construction of X_t relies on knowledge of X_T). However, we may construct a suitable flow $(Z_t)_{t\in[0,T]}$ by defining the *expected velocity field* as

$$v^X(\mathbf{x}, t) = \mathbb{E}\left[\dot{X}_t \mid X_t = \mathbf{x}\right], \quad \text{for all } \mathbf{x} \in \text{supp}(X_t),$$

where $\dot{X}_t$ denote the time derivative of X_t, and letting $(Z_t)_{t\in[0,T]}$ be the solution to the ODE

$$\frac{\mathrm{d}Z_t}{\mathrm{d}t} = v^X(Z_t, t), \quad Z_0 \sim \pi. \tag{1.1}$$

This works because it can be shown that the process $(Z_t)_{t\in[0,T]}$ defined in this way has the same marginals $(p_t)_{t\in[0,T]}$ as the original process $(X_t)_{t\in[0,T]}$.

In order to simulate $(Z_t)_{t\in[0,T]}$, we learn a parametric approximation $v_\theta(\mathbf{x}, t)$ to $v^X(\mathbf{x}, t)$. We do this by minimizing the L^2 objective

$$\mathcal{L}(v_\theta) = \int_0^1 \mathbb{E}\left[\left\|v_\theta(X_t, t) - v^X(X_t, t)\right\|^2\right] \mathrm{d}t.$$

Although this is not tractable directly, we can rewrite it as

$$\mathcal{L}(v_\theta) = \int_0^1 \mathbb{E}\left[\left\|v_\theta(X_t, t) - \dot{X}_t\right\|^2\right] \mathrm{d}t + const,$$

where the constant is independent of v_θ and so irrelevant to the optimization. Once we have learned $v_\theta(\mathbf{x}, t)$ we can create approximate samples from p_{data} by drawing $Z_0 \sim \pi$, simulating the ODE (1.1) and noting that Z_T should be approximately distributed according to p_{data}.

For a certain stochastic Gaussian choice of interpolating function $I(\mathbf{x}_0, \mathbf{x}_T, t)$, the flow learned by flow matching reproduces the probability flow ODE sampling scheme for SDE-based diffusion models. However, flow matching can incorporate a much wider class of possible sample paths than standard diffusion models, allowing us to construct smoother interpolating paths and leading to more efficient training and sampling (Lipman et al., 2023; Liu et al., 2023). In addition, the flow matching training objective tends to be more stable and robust than previous score matching methods (Lipman et al., 2023).

Flow matching models can also be viewed as a specific instance of normalizing flows (Rezende et al., 2015; Chen et al., 2018), where the additional structure in the interpolant probability densities means that we can find equivalent training objectives which are much simpler to optimize. This makes training and sampling for flow matching much more efficient compared to previous normalizing flow methods (Grathwohl et al., 2019; Rozen et al., 2021; Ben-Hamu et al., 2022).

Bounds on the approximation error for deterministic sampling schemes such as flow matching are less well-developed than those for the stochastic sampling of diffusion models. Albergo et al. (2023b) bound the KL distance between the end-point marginals of two flow ODEs in terms of the Lipschitz constant of the velocity fields. However, their bound assumes a uniform-in-time Lipschitz constant,

which is unrealistic in many settings, such as if the data distribution is supported on a submanifold. Alternatively, Chen et al. (2023d) and Li et al. (2023) derive bounds on the error of the probability flow ODE for a diffusion model. However, Chen et al. (2023d) does not deal with learned approximations to the score, and neither deals with the more general flow matching setting. Other works attempting to bound the error of the flow matching procedure or probability flow ODE have required the injection of some stochasticity into the sampling procedure in order to smooth the flows (Albergo et al., 2023a; Chen et al., 2023b).

1.4 Variational Autoencoders

Variational autoencoders are another class of generative model based on latent variable models, introduced by Kingma et al. (2014) and Rezende et al. (2015). The basic idea is to learn to associate a latent representation $\mathbf{z}$ to each datapoint $\mathbf{x}$ and vice versa. If we enforce an easy-to-sample prior on $\mathbf{z}$, we can then generate synthetic datapoints by sampling $\mathbf{z}$ from this prior and calculating the associated datapoint $\mathbf{x}$.

Formally, VAEs consist of three parts: the *generative* or *decoder* network $p_\theta(\mathbf{x}|\mathbf{z})$, which gives a probabilistic mapping from the latent variable $\mathbf{z}$ to the datapoint $\mathbf{x}$, the *inference* or *encoder* network $q_\phi(\mathbf{z}|\mathbf{x})$, which gives a probabilistic mapping from the datapoint to the latent variable and which should invert p_θ, and the *prior* $p(\mathbf{z})$ on the latent variable. Together, the prior and the decoder combine to give the model likelihood $p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$. The encoder $q_\phi(\mathbf{z}|\mathbf{x})$ can be thought of as a parametric approximation to the posterior $p_\theta(\mathbf{z}|\mathbf{x})$, essentially performing amortized posterior inference.

In practice, the prior $p(\mathbf{z})$ is typically fixed, for example to be a standard Gaussian, while the decoder and encoder networks are typically parameterized via neural networks. A common choice of parameterization is to let $p_\theta(\mathbf{x}|\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{x})$ be Gaussian distributions with mean and variance given by the output of a neural network with parameters θ and ϕ respectively, i.e. $p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \lambda_\theta(\mathbf{z}), \sigma^2_\theta(\mathbf{x}))$ and $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_\phi(\mathbf{x}), \tau^2_\phi(\mathbf{x}))$ for some neural network functions λ_θ, σ_θ, μ_ϕ and τ_ϕ.

We would like to fit our VAE by maximizing the model log-likelihood $\log p_\theta(\mathbf{x})$. Unfortunately, this is intractable directly as it requires marginalizing over $\mathbf{z}$. However, we can lower bound the model log-likelihood in the following way,

$$\log p_\theta(\mathbf{x}) = \log \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] =: \ell(\theta,\phi,\mathbf{x}).$$

The quantity $\ell(\theta,\phi,\mathbf{x})$ is known as an evidence lower bound (ELBO), and can be used as a surrogate optimization objective. It can be shown that maximizing the ELBO is equivalent to minimizing the KL divergence $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))$.

We train our VAE by maximizing $\ell(\theta,\phi,\mathbf{x})$ over θ and ϕ simultaneously using stochastic gradient descent. We obtain unbiased estimates of the gradient with respect to θ using samples from the data distribution and empirically estimating the expectation. To obtain stable unbiased estimates of the gradient with respect to ϕ, we use the *reparameterization trick* (Kingma et al., 2014). This assumes that we can find a differentiable function $f_\phi(\mathbf{x},\boldsymbol{\varepsilon})$ of $\mathbf{x}$, ϕ and an auxiliary variable $\boldsymbol{\varepsilon}$ along with a distribution $q(\boldsymbol{\varepsilon})$ such that we can simulate $\mathbf{z} \sim q_\phi(\,\cdot\,|\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^d$ by drawing $\boldsymbol{\varepsilon} \sim q(\cdot)$ and setting $\mathbf{z} = f_\phi(\mathbf{x},\boldsymbol{\varepsilon})$. If this is the case, we can write

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})}\right] = \mathbb{E}_{q(\boldsymbol{\varepsilon})}\left[\nabla_\phi \log \frac{p_\theta(\mathbf{x}, f_\phi(\mathbf{x},\boldsymbol{\varepsilon}))}{q_\phi(f_\phi(\mathbf{x},\boldsymbol{\varepsilon})|\mathbf{x})}\right],$$

and then the RHS can be estimated using samples from the data distribution and an empirical estimate of the expectation.

One problem encountered by VAEs is that the ELBO objective heavily penalizes samples from the encoder which have low posterior probability under $p_\theta(\mathbf{z}|\mathbf{x})$, so $q_\theta(\mathbf{z}|\mathbf{x})$ tends to underestimate the true posterior variance (Minka, 2005). The importance weighted autoencoder was proposed by Burda et al. (2016) to address this problem. It modifies the ELBO objective used to train the VAE, replacing it with

$$\ell_N(\theta,\phi,\mathbf{x}) = \mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_N \sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{N}\sum_{i=1}^{N} \frac{p_\theta(\mathbf{x},\mathbf{z}_i)}{q_\phi(\mathbf{z}_i|\mathbf{x})}\right],$$

which can be viewed as an N-sample importance weighted estimate for the model log-likelihood. The hope is that this will be a decent lower bound if any of the samples $\mathbf{z}_i$ have reasonable probability mass under $p_\theta(\mathbf{z}|\mathbf{x})$.

The IWAE bound $\ell_N(\theta, \phi, \mathbf{x})$ is a lower bound on the model log-likelihood for any choice of N, and recovers the original ELBO for $N = 1$. In addition, the bound becomes tighter as N increases (Burda et al., 2016), and it can in fact be shown that the variational gap decays at rate $1/N$ when the latent dimension remains fixed (Maddison et al., 2017; Domke et al., 2018). We may hope that by providing a tighter variational approximation, the IWAE allows us to fit a more accurate model for the data distribution. However, Rainforth et al. (2018) show that although the IWAE bound is tighter, increasing the number of samples N causes the signal-to-noise ratio of gradient updates during training to deteriorate, limiting the usefulness of IWAEs in practice.

Other works have extended VAEs in a different direction, replacing the KL divergence used to construct the ELBO with another distance measure. Li et al. (2016) consider minimizing $\mathrm{D}_\alpha(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))$ instead of $\mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))$, where $\mathrm{D}_\alpha(q||p)$ denotes Rényi's alpha divergence between the probability distributions q and p, defined for $\alpha \in \mathbb{R} \setminus \{1\}$ by

$$\mathrm{D}_\alpha(q||p) = \frac{1}{\alpha - 1} \log \int q(\mathbf{z})^\alpha p(\mathbf{z})^{1-\alpha} \mathrm{d}\mathbf{z},$$

and extended to $\alpha = 1$ by continuity (when it recovers the KL divergence) (Rényi, 1961). It can be shown that minimizing $\mathrm{D}_\alpha(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x}))$ is equivalent to maximizing the objective

$$\ell^{(\alpha)}(\theta, \phi, \mathbf{x}) = \frac{1}{1 - \alpha} \log \left(\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[w_{\theta,\phi}(\mathbf{z}; \mathbf{x})^{1-\alpha} \right] \right),$$

where $w_{\theta,\phi}(\mathbf{z}; \mathbf{x}) = p_\theta(\mathbf{x}, \mathbf{z})/q_\phi(\mathbf{z}|\mathbf{x})$ denote the importance weights. We call $\ell^{(\alpha)}(\theta, \phi, \mathbf{x})$ the *variational Rényi (VR) bound*.

The variational Rényi bound is decreasing in α and is a lower bound on the model log-likelihood when $\alpha > 0$. Li et al. (2016) therefore suggest it as a suitable training objective for fitting our autoencoder. Unfortunately, the gradients of $\ell^{(\alpha)}(\theta, \phi, \mathbf{x})$ with respect to θ and ϕ are intractable, and Li et al. (2016) therefore rely on importance sampling approximations, leading to biased gradient estimates.

1.5 Thesis Outline

This dissertation follows the format of an integrated thesis. It is composed of five papers, all on the topics of understanding the theoretical properties of generative models and extending them to new settings. This section provides a brief overview of each chapter.

From Denoising Diffusions to Denoising Markov Models In Chapter 2, we extend the standard framework for score-based generative modeling using SDEs on $\mathbb{R}^d$ (Song et al., 2021b) to arbitrary state spaces. We replace the forward and reverse SDEs used on $\mathbb{R}^d$ with general Markov processes defined via generators $\mathcal{L}$ and $\mathcal{K}$ to obtain a generalization of a diffusion model which we term a *denoising Markov model* (DMM).

Following the methodology of Huang et al. (2021), we find an expression for the log-likelihood of our model, and then derive a lower bound for this model log-likelihood (an ELBO) (Theorem 1) using generalizations of the Fokker–Planck, Feynman–Kac and Girsanov theorems. We show that this lower bound can be maximized by finding the equivalent tractable training objective

$$\mathcal{I}_{\mathrm{ISM}}(\beta) = \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)} \left[\frac{\hat{\mathcal{L}}^* \beta(\mathbf{x}_t, t)}{\beta(\mathbf{x}_t, t)} + \hat{\mathcal{L}} \log \beta(\mathbf{x}_t, t) \right] \mathrm{d}t,$$

where $\beta(\mathbf{x}_t, t)$ parameterizes the reverse generator $\mathcal{K}$. Then, we may fit an arbitrary DMM by taking samples from the forward process, empirically estimating $\mathcal{I}_{ISM}(\beta)$ and minimizing it using stochastic gradient descent. This framework recovers several particular instances of diffusion models (Song et al., 2021b; De Bortoli et al., 2022) as special cases (Examples 5 and 7).

Next, we show that our training objective can be interpreted as a generalization of previous score matching objectives (Hyvärinen, 2005; Vincent, 2011). This gives a principled way to extend score matching to arbitrary state spaces, and several of the intuitions behind score matching carry over to our more general setting (Proposition 1). We also derive equivalent objectives (corresponding to implicit, explicit and denoising score matching) and show that previously studied

discrete-time diffusion models can be viewed as the natural first-order discretization of our framework (Theorem 3).

Finally, we show that DMMs are effective in practice. We demonstrate their performance on several tasks covering a range of classes of state space, including performing posterior inference for synthetic data on $\mathbb{R}^d$, image inpainting in discrete pixel space, pose estimation in $SO(3)$ and approximating distributions of measures on a finite state space.

A Continuous Time Framework for Discrete Denoising Models In Chapter 3, we study a particular extension of diffusion models to discrete state spaces. This can be viewed as a special application of the framework of the previous chapter.

We develop practical tools that allow for the efficient learning and sampling of the reverse process, which takes the form of a continuous-time Markov chain (CTMC) in discrete space. First, we propose a parameterization of the reverse generator $\hat{R}_t$ in terms of the forward generator R_t and reverse transition densities $q_{0|t}(\mathbf{x}_0|\mathbf{x}_t)$,

$$\hat{R}_t(\mathbf{x}, \tilde{\mathbf{x}}) = R_t(\mathbf{x}, \tilde{\mathbf{x}}) \sum_{\mathbf{x}_0 \in \mathcal{X}} \frac{q_{t|0}(\tilde{\mathbf{x}}|\mathbf{x}_0)}{q_{t|0}(\mathbf{x}|\mathbf{x}_0)} q_{0|t}(\mathbf{x}_0|\mathbf{x}), \quad \text{for } \mathbf{x} \neq \tilde{\mathbf{x}}.$$

Then, we learn an approximation $p^{\theta}_{0|t}(\mathbf{x}_0|\mathbf{x})$ to $q_{0|t}(\mathbf{x}_0|\mathbf{x})$, which provides a more stable target for learning. Second, we propose the use of tau-leaping, a method for simulating CMTCs from chemical physics, for efficient sampling of the reverse process (Gillespie, 2001). Tau-leaping works by fixing a time step, sampling all of the transitions that the CTMC would take within that time step, summing them all up, and then applying them as a single update at the end of the time step, thus avoiding having to simulate each transition individually. We demonstrate that these innovations lead to competitive performance on real-world image and audio generation tasks.

In addition, we provide a novel predictor-corrector scheme for our method, which increases the fidelity of generated samples. We also derive a bound on the TV error of the approximation learned by our model, showing that with a sufficiently good approximation to the reverse rate matrix and a sufficiently fine grid of time steps, we can approximate the original data distribution arbitrarily well.

Error Bounds for Flow Matching Methods In Chapter 4, we provide the first error bound for the general flow matching procedure with fully deterministic sampling which applies even when the data distribution does not have full support.

Our result comes in two parts. First, we use the Alekseev–Gröbner theorem to bound the difference between the true and approximate flow ODEs. This leads to a bound in 2-Wasserstein distance between the data distribution π_1 and our learned approximation $\hat{\pi}_1$, given in terms of the Lipschitz constant L_t of the approximate velocity field and the L^2 error ε of our flow approximation (Theorem 1),

$$W_2(\hat{\pi}_1, \pi_1) \leq \varepsilon \exp\Big\{ \int_0^1 L_t \, \mathrm{d}t \Big\}. \tag{1.2}$$

Second, we provide a bound on the Lipschitz constant of the true velocity field (Theorem 2). This bound holds under a certain regularity condition on the data distributions π_0 and π_1, specifically that if $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$ independently, then $\alpha_t X_0 + \beta_t X_1$ is λ-*regular* for each $t \in [0,1]$, where we say that an $\mathbb{R}^d$-valued random variable W is λ-regular if whenever we take $\tau \in (0, \infty)$ and $\xi \sim \mathcal{N}(0, \tau^2 I_d)$ independently of W and set $W' = W + \xi$, then we have $\|\mathrm{Cov}_{\xi|W'=\mathbf{x}}(\xi)\|_{\mathrm{op}} \leq \lambda\tau^2$ for all $\mathbf{x} \in \mathbb{R}^d$. (See Chapter 4 for more intuition on this regularity condition.) For suitable choices of interpolating paths, the bound we obtain on the Lipschitz constant is logarithmic in the problem parameters.

Combining this logarithmic dependence with the exponential in (1.2), we deduce that there is a class of functions $\mathcal{V}$ such that if we optimize the flow matching loss over all functions in $\mathcal{V}$, then the resulting flow matching procedure has an error which converges polynomially in the relevant problem parameters, subject to our smoothness assumption (Theorem 3). We also explain the particular relevance of our results in the case of the probability flow ODE, which can be viewed as a special case of flow matching.

Nearly d-Linear Convergence Bounds for Diffusion Models Via Stochastic Localization In Chapter 5, we derive the first convergence bounds for diffusion models which are linear in the dimension d of the data distribution (up to logarithmic

factors) without requiring smoothness assumptions on the data distribution, in the setting of early stopping. As in Chen et al. (2023a) and Chen et al. (2023c), we use a version of Girsanov's theorem to deduce that

$$\mathrm{KL}(Q||P^{q_T}) \leq \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q \left[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T-t_k)\|^2 \right] \mathrm{d}t,$$

where Q and P^{q_T} are the path measures of the true and approximate reverse diffusion process respectively. We then decompose the RHS into a term that is controlled by our assumption that the score approximation is L^2-accurate and a term which corresponds to the error arising from the discretization of the reverse SDE.

Controlling the error from discretization boils down to controlling the quantity $E_{s,t} := \mathbb{E}\left[\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)\|^2\right]$. We do this by decomposing it into several terms and applying Itô's lemma and Itô's isometry to deduce a differential inequality for $E_{s,t}$. Then, we use the equivalence between diffusion models and stochastic localization, as shown by Montanari (2023), to relate the coefficients of the differential inequality to quantities from stochastic localization.

The result is a bound on the convergence rate which is linear in the data dimension (up to logarithmic factors) without smoothness assumptions on the data distribution. It implies that diffusion models require at most $\tilde{O}(\frac{d \log^2(1/\delta)}{\varepsilon^2})$ steps to approximate a given data distribution, smoothed with Gaussian noise of variance δ, to within ε^2 in KL divergence.

Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders In Chapter 7, we address several limitations of the IWAE, in particular the problems of posterior variance underestimation, poor signal-to-noise ratio of gradient estimates, and weight collapse in the importance sampling term of the IWAE objective. Our main contribution is to introduce the VR-IWAE bound

$$\ell_N^{(\alpha)}(\theta, \phi, \mathbf{x}) = \frac{1}{1-\alpha} \mathbb{E}_{\mathbf{z}_1, \ldots, \mathbf{z}_N \sim q_\phi(\mathbf{z}|\mathbf{x})} \left[\log \left(\frac{1}{N} \sum_{i=1}^{N} w_{\theta,\phi}(\mathbf{z}; \mathbf{x})^{1-\alpha} \right) \right],$$

and to propose its use as an objective for training variational autoencoders. The VR-IWAE bound recovers the IWAE by setting $\alpha = 0$ and the VR bound as $N \to \infty$,

so can be thought of as a generalization unifying both methods and allowing us to interpolate between the different bounds and trade off between mode-seeking and mode-covering behaviour. In addition, the VR-IWAE recovers the ELBO as $\alpha \to 1$.

First, we show that $\ell_N^{(\alpha)}(\theta, \phi, \mathbf{x})$ is indeed a lower bound on the model log-likelihood for all $\alpha \in [0, 1)$, so is a suitable training objective. We also show how to derive reparameterized gradient estimates for $\ell_N^{(\alpha)}(\theta, \phi, \mathbf{x})$. We find that the resulting gradient descent scheme recovers the same gradient updates as the variational Rényi procedure proposed in Li et al. (2016). However, while Li et al. (2016) view their gradient estimates as biased in the variational Rényi setting, from our perspective they can be viewed as unbiased estimates of the gradients of the VR-IWAE bound.

Second, we show that for $\alpha \in (0, 1)$ the reparameterized gradient estimates of the VR-IWAE bound have better signal-to-noise properties than those of the IWAE bound. Concretely, we show that for $\alpha \in (0, 1)$ as $N \to \infty$ the SNR of the reparameterized estimates of $\nabla_\theta \ell_N^{(\alpha)}(\theta, \phi, \mathbf{x})$ and $\nabla_\phi \ell_N^{(\alpha)}(\theta, \phi, \mathbf{x})$ are both $O(\sqrt{N})$. In contrast, Rainforth et al. (2018) found that the SNR of the reparameterized estimates of $\nabla_\phi \ell_N(\theta, \phi, \mathbf{x})$ was $O(1/\sqrt{N})$, implying that for large N the ϕ-gradients of the IWAE have a much worse SNR than those of the VR-IWAE bound for $\alpha > 0$. We also derive doubly reparameterized gradient estimates for the VR-IWAE bound along the lines of Tucker et al. (2019) to improve the SNR in the general case.

Next, we turn our attention to investigating the tightness of the VR-IWAE bound. First, we generalize the analysis of Maddison et al. (2017) and Domke et al. (2018) when d is fixed and $N \to \infty$, and show that the variational gap of the VR-IWAE decreases at rate $O(1/N)$. However, we also indicate why we do not typically expect this assumption to be realistic – essentially, because in practice the latent dimension is usually large compared to the number of importance samples used in the IWAE bound.

We therefore study the behaviour of the VR-IWAE bound as both d and N go to infinity. We consider two assumptions on the distribution of the normalized importance weights: (i) $\log \overline{w}_{\theta,\phi}(\mathbf{z}; \mathbf{x})$ is normally distributed, or (ii) $\log \overline{w}_{\theta,\phi}(\mathbf{z}; \mathbf{x})$

can be expressed as the sum of d i.i.d. random variables (and therefore is approximately normally distributed by the central limit theorem). We motivate these assumptions by noting that if the encoder and decoder factorize over the latent dimensions, then $\log \overline{w}_{\theta,\phi}(\mathbf{z};\mathbf{x})$ can be expressed as the sum of d factors which we expect to be approximately independent and identically distributed. We expect this to be a reasonable approximation in high dimensions, and in Section 6 we demonstrate that this is indeed the case in practice.

We can decompose the variational gap into a term corresponding to the largest importance weight and a term corresponding to the contribution from all other importance weights. Under either assumption above, we find that the term corresponding to the largest important weight dominates, and we observe weight collapse in the importance sampling in the VR-IWAE, as described in Bengtsson et al. (2008). The key ingredients in our derivations are several concentration results and limit theorems for large deviations of Gaussian random variables and sums of i.i.d. random variables (Saulis et al., 1991).

Under assumption (i), our main conclusion is that as $d, N \to \infty$ the variational gap is asymptotically

$$\frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right),$$

so long as $\log N/d \to 0$. We see that in this regime, while increasing N does still decrease the variational gap, the impact is negligible compared to the dominant term $d\sigma^2/2$. Thus, unless N grows exponentially with d there is little to be gained over the ELBO from importance sampling in the variational objective. We also derive a similar conclusion under assumption (ii), so long as $\log N/d^{1/3} \to 0$.

Finally, we test the accuracy of our theoretical results in practice, on both simluated and real-world data. We find that while the results of Maddison et al. (2017) and Domke et al. (2018), which predict that the variational gap decays at a rate of $O(1/N)$, are reasonable for small values of d, when d becomes large our asymptotic results provide a much better fit to the observed variational gap.

2

From Denoising Diffusions to Denoising Markov Models

From Denoising Diffusions to Denoising Markov Models

Joe Benton†
Department of Statistics, University of Oxford, Oxford, UK
Yuyang Shi
Department of Statistics, University of Oxford, Oxford, UK
Valentin De Bortoli
ENS, Paris, France
George Deligiannidis
Department of Statistics, University of Oxford, Oxford, UK
Arnaud Doucet
Department of Statistics, University of Oxford, Oxford, UK

Summary. Denoising diffusions are state-of-the-art generative models exhibiting remarkable empirical performance. They work by diffusing the data distribution into a Gaussian distribution and then learning to reverse this noising process to obtain synthetic datapoints. The denoising diffusion relies on approximations of the logarithmic derivatives of the noised data densities using score matching. Such models can also be used to perform approximate posterior simulation when one can only sample from the prior and likelihood. We propose a unifying framework generalising this approach to a wide class of spaces and leading to an original extension of score matching. We illustrate the resulting models on various applications.

Keywords: denoising diffusions, generative models, posterior simulation, score matching, unifying framework

1. Introduction

Given a set of samples from an unknown distribution $p_{\text{data}}(\mathbf{x})$, generative modelling is the task of producing further synthetic samples coming from approximately the same distribution. Over the past decade, a variety of techniques have been developed to tackle this problem, including autoregressive models (Oord et al., 2016), generative adversarial networks (Goodfellow et al., 2014), variational autoencoders (Kingma and Welling, 2014) and normalising flows (Rezende and Mohamed, 2015). These methods have had significant success in generating perceptually realistic samples from complex data distributions, such as text and image data (Brown et al., 2020; Dhariwal and Nichol, 2021). A major motivation for the development of generative models is that they can be easily

†*Address for correspondence:* Joe Benton, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK. E-mail: benton@stats.ox.ac.uk

extended for Bayesian inference. In a typical setting, we make an observation $\boldsymbol{\xi}^*$ based on underlying datapoint $\mathbf{x}$, for example a category label or partial observation of $\mathbf{x}$, and want to sample from the posterior distribution $p_{\text{data}}(\mathbf{x}|\boldsymbol{\xi}^*)$. We achieve this by learning a conditional generative model for $\mathbf{x}$ given any observation $\boldsymbol{\xi}$ based on samples from $p_{\text{data}}(\mathbf{x}, \boldsymbol{\xi})$. This approach is particularly useful in high-dimensional scenarios where traditional sampling methods, such as Markov chain Monte Carlo (MCMC) methods or approximate Bayesian computation (ABC), are typically infeasible.

Recently, denoising diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021) have emerged as effective generative models for high-dimensional data. They work by incrementally adding noise to the data to transform the data distribution into an easy-to-sample reference distribution, and then learning to invert the noising process, which is achieved using score matching (Hyvärinen, 2005). Their use for inference has recently seen an explosion of applications, including text-to-speech generation (Popov et al., 2021), image inpainting and super-resolution (Song et al., 2021; Saharia et al., 2022) and protein structure modelling (Trippe et al., 2023).

Most of the current methodology, theory and applications of denoising diffusion models are for diffusion processes on $\mathbb{R}^d$. However, many distributions of interest are defined on different spaces. Recently, De Bortoli et al. (2022) and Huang et al. (2022) have extended continuous-time methods and the analogy with score matching from $\mathbb{R}^d$ to general Riemannian manifolds in order to model data with strong geometric prior. Several diffusion methods have also been developed for discrete data, such as text, music or graph structures (Austin et al., 2021; Hoogeboom et al., 2021; Campbell et al., 2022; Sun et al., 2023). Here though, the relationships to score matching, as well as between these various methods and the Euclidean diffusion case, are less clear. All these recent extensions have been somewhat ad hoc, with training objectives needing to be re-derived for each new application.

The main contribution of this paper is to provide a unifying framework for such models, which we call *denoising Markov models*, or DMMs. We demonstrate how to construct and train a DMM for data in any state space satisfying mild regularity conditions. This yields a principled procedure for using these models for unconditional generation and inference on a wider class of spaces than previously considered. Additionally this general framework leads to a principled extension of score matching to general spaces. Finally, we demonstrate the application of our framework on examples in continuous Euclidean space, discrete space, for Riemmanian manifolds and on the simplex.

2. Background

A denoising diffusion model is a generative model consisting of two stochastic processes. The *fixed* noising process takes a data point $\mathbf{x}_0$ drawn from a data distribution $q_0 := p_{\text{data}}$ on state space $\mathcal{X}$ and maps it stochastically to some $\mathbf{x}_T \in \mathcal{X}$. The *learned* generative process takes $\mathbf{x}_T \in \mathcal{X}$ drawn according to some initial distribution p_0 on $\mathcal{X}$ and maps it back stochastically to some $\mathbf{x}_0 \in \mathcal{X}$. Throughout, we denote the marginals of the noising and generative processes by $q_t(\mathbf{x})$ and $p_t(\mathbf{x})$ respectively for $t \in [0, T]$.

The basic idea is to pick a noising process so that $(q_t)_{t\geq 0}$ converges to some easy-to-sample-from distribution q_{ref}, which we then take to be p_0. We learn a generative

process which approximates the time-reversal of the noising process. Then, we can generate approximate samples from q_0 by sampling $\mathbf{x}_T \sim p_0$ and running the dynamics of the reverse process to produce a sample $\mathbf{x}_0 \sim p_T$, which should be close to q_0.

2.1. Continuous-time denoising diffusion models on $\mathbb{R}^d$

The framework for continuous-time diffusion models on $\mathbb{R}^d$ was first set out by Song et al. (2021). The "forward" noising process $(Y_t)_{t\in[0,T]}$ evolves according to the stochastic differential equation (SDE)

$$\mathrm{d}Y_t = b(Y_t, t)\mathrm{d}t + \mathrm{d}B_t, \qquad Y_0 = \mathbf{x}_0 \sim p_{\mathrm{data}}, \tag{1}$$

for some chosen function $b : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, and standard Brownian motion B. With this set-up, the time-reversed process $X_t = Y_{T-t}$ can be simulated by initialising $X_0 = \mathbf{x}_T \sim q_T$ and running the SDE

$$\mathrm{d}X_t = \{-b(X_t, T-t) + \nabla_{\mathbf{x}} \log q_{T-t}(X_t)\}\mathrm{d}t + \mathrm{d}\hat{B}_t, \tag{2}$$

where $q_t(\mathbf{x}_t)$ denotes the marginals of the forward process and $\hat{B}$ is another standard Brownian motion (Anderson, 1982). We typically choose our forward process to be an Ornstein–Uhlenbeck process, i.e. $b(\mathbf{x}, t) = -\mathbf{x}/2$, for which $q_T \approx q_{\mathrm{ref}} := \mathcal{N}(0, I_d)$, the standard Gaussian distribution on $\mathbb{R}^d$, for large T.

To simulate the reverse process, we must approximate $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$. We do this by fixing a parametric family of functions $s_\theta(\mathbf{x}, t)$, and then choosing the parameters θ to minimise the *denoising score matching* objective

$$\mathcal{I}_{\mathrm{DSM}}(\theta) = \frac{1}{2}\int_0^T \mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)}\left[||\nabla_{\mathbf{x}} \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) - s_\theta(\mathbf{x}_t, t)||^2\right] \mathrm{d}t, \tag{3}$$

where $q_{0,t}(\mathbf{x}_0, \mathbf{x}_t)$ and $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ denote the joint and conditional distributions of the SDE (1). The conditional is available in closed-form for the Ornstein–Uhlenbeck process. This is sensible since $\mathcal{I}_{\mathrm{DSM}}$ is minimised when $s_\theta(\mathbf{x}, t) = \nabla_{\mathbf{x}} \log q_t(\mathbf{x})$ for almost all $x \in \mathcal{X}$ and $t \in [0,T]$ (Song et al., 2021). If our score estimate were exact and $p_0 = q_T$, then we would have $p_t = q_{T-t}$ for all $t \in [0,T]$. In practice, we use a neural network to parameterise $s_\theta(\mathbf{x}, t)$ and use stochastic gradient descent to minimise $\mathcal{I}_{\mathrm{DSM}}(\theta)$.

Once we have a score estimate $s_\theta(\mathbf{x}, t)$, we compute approximate samples from the reverse process by running the approximate reverse process

$$\mathrm{d}X_t = \{-b(X_t, T-t) + s_\theta(X_t, T-t)\}\mathrm{d}t + \mathrm{d}\hat{B}_t \tag{4}$$

starting in $X_0 \sim p_0$ and setting $\mathbf{x}_0 = X_T$. In practice, we use suitable numerical integrators to simulate the approximate reverse process.

Alternatively, the objective $\mathcal{I}_{\mathrm{DSM}}$ can be derived from a lower bound on the model log-likelihood (also known as an Evidence Lower Bound, or ELBO) for $q_T(x)$, either using Girsanov's theorem and the chain rule for Kullback–Leibler divergences (Song et al., 2021), or by combining the Fokker–Planck equation and Feynman–Kac formula with Girsanov's theorem (Huang et al., 2021).

2.2. Diffusion models for inference

Denoising diffusions can also be used to sample approximately from a posterior $p_{\text{data}}(\mathbf{x}|\boldsymbol{\xi}^*)$ when we only have access to samples from the joint distribution $p_{\text{data}}(\mathbf{x}, \boldsymbol{\xi})$; see e.g. (Song et al., 2021). We first draw a sample $(\mathbf{x}_0, \boldsymbol{\xi}_0) \sim p_{\text{data}}$, set $Y_0 = \mathbf{x}_0$ and let $(Y_t)_{t\in[0,T]}$ evolve according to Equation (1). If we condition on $\boldsymbol{\xi}_0$, then the process Y has marginals $q_t(\mathbf{x}_t|\boldsymbol{\xi}_0) = \int q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)p_{\text{data}}(\mathbf{x}_0|\boldsymbol{\xi}_0)\mathrm{d}\mathbf{x}_0$, where $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ is the transition kernel of the forward diffusion in Equation (1). So, the time-reversed process $X_t = Y_{T-t}$ conditioned on $\boldsymbol{\xi}_0$ can be simulated by initialising $X_0 \sim q_T(\cdot|\boldsymbol{\xi}_0)$ and running the SDE

$$\mathrm{d}X_t = \{-b(X_t, T-t) + \nabla_{\mathbf{x}} \log q_{T-t}(X_t \mid \boldsymbol{\xi}_0)\}\mathrm{d}t + \mathrm{d}\hat{B}_t. \tag{5}$$

If we have $q_T(\cdot|\boldsymbol{\xi}) \approx q_{\text{ref}}$ for all $\boldsymbol{\xi}$ and an approximation $s_\theta(\mathbf{x}, \boldsymbol{\xi}, t)$ to $\nabla_{\mathbf{x}} \log q_t(\mathbf{x}|\boldsymbol{\xi})$, we can obtain approximate samples from $q_0(\cdot|\boldsymbol{\xi}^*) = p_{\text{data}}(\cdot|\boldsymbol{\xi}^*)$ for any given $\boldsymbol{\xi}^*$ by initialising $X_0 \sim p_0 := q_{\text{ref}}$, simulating the reverse dynamics in Equation (5) with $\nabla_{\mathbf{x}} \log q_{T-t}(X_t|\boldsymbol{\xi}_0)$ replaced by $s_\theta(X_t, \boldsymbol{\xi}^*, T-t)$, and setting $\mathbf{x}_0 = X_T$. To learn $s_\theta(\mathbf{x}, \boldsymbol{\xi}, t)$, we minimise

$$\mathcal{I}_{\text{DSM}}(\theta) = \frac{1}{2}\int_0^T \mathbb{E}_{q(\mathbf{x}_0, \mathbf{x}_t, \boldsymbol{\xi}_0)}\left[||\nabla_{\mathbf{x}} \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) - s_\theta(\mathbf{x}_t, \boldsymbol{\xi}_0, t)||^2\right] \mathrm{d}t,$$

where we denote $q(\mathbf{x}_0, \mathbf{x}_t, \boldsymbol{\xi}_0) = p_{\text{data}}(\mathbf{x}_0, \boldsymbol{\xi}_0)q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$. This objective is minimised when $s_\theta(\mathbf{x}, \boldsymbol{\xi}, t) = \nabla_{\mathbf{x}} \log q_t(\mathbf{x}|\boldsymbol{\xi})$ for almost all $x \in \mathcal{X}$ and $t \in [0, T]$ (Song et al., 2021).

2.3. Score matching

The objective $\mathcal{I}_{\text{DSM}}$ defined in Equation (3) can also be interpreted as a score matching objective. Score matching was introduced as a method for fitting unnormalised probability distributions defined on $\mathbb{R}^d$ by Hyvärinen (2005). It approximates a distribution $q_0(\mathbf{x})$ with a distribution of the form $p(\mathbf{x}; \theta) = q(\mathbf{x}; \theta)/Z(\theta)$ by minimising

$$\mathcal{J}(\theta) = \frac{1}{2}\mathbb{E}_{q_0(\mathbf{x})}\left[||\nabla_{\mathbf{x}} \log q_0(\mathbf{x}) - \nabla_{\mathbf{x}} \log q(\mathbf{x}; \theta)||^2\right],$$

known as an *explicit score matching* loss. This objective is intractable since it depends on $\nabla_{\mathbf{x}} \log q_0(\mathbf{x})$, but there are methods for rewriting it in an equivalent tractable form, including implicit and denoising score matching (Hyvärinen, 2005; Vincent, 2011). Equation (3), which corresponds to denoising score matching, can also be written in explicit, implicit or sliced score matching form (Huang et al., 2021).

3. A general framework for denoising Markov models

In this section, we set out a general framework for DMMs. First, we explain how to construct a DMM on an arbitrary state space with a forward noising process Y and backward generative process X. Second, we derive an expression for the model likelihood in terms of an expectation over an auxiliary process Z, defined in terms of X and running forward in time. Third, we derive an ELBO by using Girsanov's theorem to relate the expectation over Z to one over Y. Finally, we show how this ELBO can be used to get a tractable training objective. Our argument follows a similar structure to Huang et al. (2021), but we work in terms of generic Markov generators, rather than specific operators corresponding to diffusions on $\mathbb{R}^d$, and so require generalisations of the stochastic process results therein. For simplicity, we present the framework for unconditional generation and then explain how to adapt it for inference.

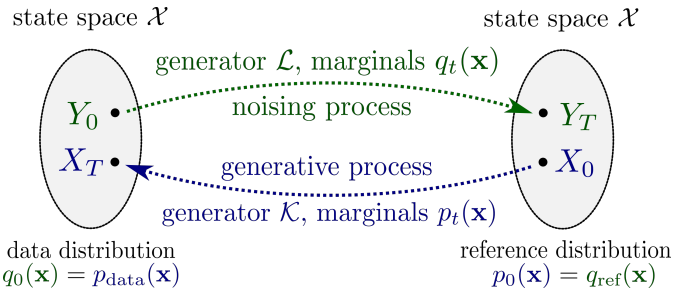

Fig. 1. Diagram of notation.

3.1. Notation and set-up

Our data is assumed to be distributed according to p_{data} on a state space $\mathcal{X}$. We assume only that $\mathcal{X}$ comes with some reference measure ν, with respect to which all probability densities will be defined, and satisfies some regularity conditions given in Appendix B.1. This includes $\mathbb{R}^d$, discrete spaces and Riemannian manifolds (with or without boundary).

Our DMM consists of a noising process $(Y_t)_{t\in[0,T]}$ and a generative process $(X_t)_{t\in[0,T]}$, which are Markov processes. We consider Y fixed and learn X to approximate the reverse of Y. Initially, we must fix a class of processes to which X and Y belong and within which we will optimise X. The particular class and parameterisation we choose will necessarily depend on $\mathcal{X}$, but a typical choice for $\mathcal{X} = \mathbb{R}^d$ would be a diffusion (see Example 1), while a typical choice when $\mathcal{X}$ is a finite discrete space may be a continuous-time Markov chain (CTMC) (see Example 2). Our notation is depicted in Fig. 1.

As X and Y are not necessarily time-homogeneous, it is helpful to define the extended processes $\overline{X}$ and $\overline{Y}$ by for example setting $X_t = X_T$ for $t \geq T$ and letting $\overline{X} = (X_t, t)_{t\geq 0}$. Then $\overline{X}$, $\overline{Y}$ are time-homogeneous Markov chains on the extended space $\mathcal{S} := \mathcal{X} \times [0, \infty)$.

In general, it is most convenient to define X and Y via the generators of $\overline{X}$ and $\overline{Y}$, which we denote by $\mathcal{K}$ and $\mathcal{L}$ respectively. Informally, the generator of a Markov process $\overline{W}$ with state space $\mathcal{S}$ is an operator $\mathcal{A}$ which acts on a subset $\mathcal{D}(\mathcal{A})$ of the space of functions $f : \mathcal{S} \to \mathbb{R}$ and satisfies $\mathcal{A}f = \lim_{s\to 0}(P_s f - f)/s$, where $(P_s)_{s\geq 0}$ is the transition semigroup associated to $\overline{W}$ and $P_s f(x) = \mathbb{E}\left[f(X_s)|X_0 = x\right]$. For a more formal definition, see Appendix A.1.

We denote the time marginals of the processes X, Y by $p_t(\mathbf{x})$, $q_t(\mathbf{x})$ respectively. We make some smoothness assumptions on p, in Appendix B.2, and assume that $\mathcal{K}$, $\mathcal{L}$ satisfy some regularity conditions, in Appendix B.3. Our assumptions hold for standard models in the literature (Euclidean diffusions, CTMCs and manifold diffusions; see Appendix F), plus some that are not covered previously, such as degenerate diffusions. For infinite dimensional spaces, the assumptions of Appendix B.1 may fail and more care is needed.

One consequence of our assumptions is that the operator $\mathcal{K}$ decomposes as $\mathcal{K} = \partial_t + \hat{\mathcal{K}}$, where $\hat{\mathcal{K}}$ operates only on the spatial variables of a function f. We can therefore view $\hat{\mathcal{K}}$ as an operator on functions from $\mathcal{X}$, rather than on functions from $\mathcal{S}$, and we denote by $\hat{\mathcal{K}}^*$ the adjoint of $\hat{\mathcal{K}}$ acting on functions on $\mathcal{X}$ (see Appendix A.2).

Example 1 (Euclidean Diffusion). *If X and Y are diffusions on $\mathbb{R}^d$ given by the SDEs $\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \mathrm{d}\hat{B}_t$ and $\mathrm{d}Y_t = b(Y_t, t)\mathrm{d}t + \mathrm{d}B_t$, where B and $\hat{B}$ are Brownian*

motions, then the corresponding generators are $\mathcal{K} = \partial_t + \mu \cdot \nabla + \frac{1}{2}\Delta$ *and* $\mathcal{L} = \partial_t + b \cdot \nabla + \frac{1}{2}\Delta$*, where* $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ *denotes the Laplacian. We then have* $\hat{\mathcal{K}}^* = -\mu \cdot \nabla - (\nabla \cdot \mu) + \frac{1}{2}\Delta$ *using integration by parts.*

Example 2 (Discrete Space CTMC). *If* X *and* Y *are CTMCs, then* $\mathcal{K} = \partial_t + A$ *and* $\mathcal{L} = \partial_t + B$*, where* A *and* B *are the time-dependent generator matrices of* X *and* Y*. In this case,* $\hat{\mathcal{K}}^* = A^T$*, the transpose of* A*.*

3.2. An expression for the model likelihood

We now derive an expression for the model likelihood $p_T(\mathbf{x})$. First, under our assumptions, a generalised form of the Fokker–Planck equation, stated precisely in Appendix C, implies that $\partial_t p = \hat{\mathcal{K}}^* p$ for ν-almost every $\mathbf{x} \in \mathcal{X}$. Typically, the adjoint operator $\hat{\mathcal{K}}^*$ resembles the generator of another process in the same class as X and Y. We formalise this idea by making the following assumption.

Assumption 1. *Let* $v(\mathbf{x}, t) = p_{T-t}(\mathbf{x})$*. Then we can write the equation* $\partial_t p = \hat{\mathcal{K}}^* p$ *in the form* $\mathcal{M}v + cv = 0$ *for some function* $c : \mathcal{S} \to \mathbb{R}$*, where* $\mathcal{M}$ *is the generator of another auxiliary Feller process* $\overline{Z} = (Z_t, t)_{t \geq 0}$ *on* $\mathcal{S}$*.*

Example 3 (Euclidean Diffusion). *For Euclidean diffusions, the Fokker–Planck equation can be written as* $\partial_t v = \mu \cdot \nabla v + (\nabla \cdot \mu)v - \frac{1}{2}\Delta v$*. Assumption 1 is satisfied with* $c = -(\nabla \cdot \mu)$ *and* $\mathcal{M} = \partial_t - \mu \cdot \nabla + \frac{1}{2}\Delta$*, noting that* $\mathcal{M}$ *is the generator of the diffusion process* Z *defined by* $\mathrm{d}Z_t = -\mu(Z_t, T - t)\mathrm{d}t + \mathrm{d}B'_t$*, where* B' *is a Brownian motion.*

Example 4 (Discrete Space CTMC). *In the CTMC case, if* $c_{\mathbf{x}} = \sum_{\mathbf{y} \in \mathcal{X}} A_{\mathbf{yx}}$*, and* $D_{\mathbf{xy}} = A_{\mathbf{yx}} - c_{\mathbf{x}} \mathbb{1}_{\mathbf{x}=\mathbf{y}}$*, then* $\mathcal{M} = \partial_t + D$ *is the generator of a CTMC and Assumption 1 is satisfied. Here* c *has a natural interpretation as a "discrete divergence".*

In general, we make two smoothness assumptions on c and v, given in Appendix B.4.

Given the Fokker–Planck equation and Assumption 1, we apply a generalised form of the Feynman–Kac Theorem (see Appendix C) to $\overline{Z}$ and v to get the following expression for the model likelihood, which generalises that of Huang et al. (2021):

$$p_T(\mathbf{x}) = v(\mathbf{x}, 0) = \mathbb{E}\left[p_0(Z_T) \exp\left\{ \int_0^T c(Z_s, s)\, \mathrm{d}s \right\} \;\middle|\; Z_0 = \mathbf{x} \right]. \tag{6}$$

This gives an expression in terms of an expectation over the auxiliary process Z. We next make this tractable by converting it into an expectation over Y.

3.3. Deriving a tractable lower bound on the model log-likelihood

We would like to train our model by finding a reverse process X which maximises the likelihood in Equation (6). Unfortunately this expression is intractable, but we can find a tractable lower bound for $\log p_T(\mathbf{x})$ which can then be used as a surrogate objective.

By taking logarithms in Equation (6) and applying Jensen's inequality, we get

$$\log p_T(\mathbf{x}) \geq \mathbb{E}_{\mathbb{Q}}\left[\log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} + \log p_0(Y_T) + \int_0^T c(Y_s, s)\, \mathrm{d}s \;\middle|\; Y_0 = \mathbf{x} \right] =: \mathcal{E}^{\infty} \tag{7}$$

where $\mathbb{P}$ and $\mathbb{Q}$ are the path measures of the processes $\overline{Z}$ and $\overline{Y}$ respectively and $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$ denotes the Radon–Nikodym derivative.

To write $\mathcal{E}^\infty$ in a tractable form we need to evaluate $\log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}$, which we do using a generalisation of Girsanov's theorem. To apply this result, we require that the generators of the auxiliary process and the noising process are related in the following way.

ASSUMPTION 2. *There is a bounded measurable function $\beta : \mathcal{S} \to (0, \infty)$ such that $\beta^{-1}\mathcal{M}f = \mathcal{L}(\beta^{-1}f) - f\mathcal{L}(\beta^{-1})$ for all $f : S \to \mathbb{R}$ such that $f \in \mathcal{D}(\mathcal{M})$ and $\beta^{-1}f \in \mathcal{D}(\mathcal{L})$.*

Since $\mathcal{M}$ is defined in terms of $\mathcal{K}$, we think of Assumption 2 as forcing a particular parameterisation of the generative process in terms of β. In general, not every generative process in the same class as $\mathcal{L}$ will have such a parameterisation. However, the true time-reversal of $\mathcal{L}$ can always be parameterised in this way with $\beta(\mathbf{x}, t) = p_t(\mathbf{x})$, so this parameterisation is sufficient to capture the optimal generative process. In addition, the objective in Theorem 1 below can often be interpreted and used for a much broader set of generative processes than those which satisfy Assumption 2.

Under Assumption 2, along with a further technical assumption given in Appendix B.5, we may apply a generalised form of Girsanov's Theorem (see Appendix C, and take $\alpha = \beta^{-1}$ in Theorem 6) and Dynkin's formula (see Appendix A.1) to get

$$\log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \int_0^T \left\{-\mathcal{L}\log\beta(Y_s, s) - \beta(Y_s, s)\mathcal{L}(\beta^{-1})(Y_s, s)\right\} \mathrm{d}s + \mathbb{Q}\text{-martingale}.$$

In addition, we get that $c = \beta\mathcal{L}(\beta^{-1}) - v^{-1}\beta\mathcal{L}(\beta^{-1}v)$ by combining Assumption 2 with $f = v$ and Assumption 1. This allows us to rewrite the ELBO from Equation (7) as

$$\mathcal{E}^\infty = \mathbb{E}_{\mathbb{Q}}\left[\log p_0(Y_T) - \int_0^T \left\{\frac{\mathcal{L}(\beta^{-1}v)(Y_s, s)}{\beta^{-1}(Y_s, s)v(Y_s, s)} + \mathcal{L}\log\beta(Y_s, s)\right\}\mathrm{d}s \;\middle|\; Y_0 = \mathbf{x}\right].$$

The final step required to get a tractable expression for $\mathcal{E}^\infty$ is to remove the function v from this expression. For this, we use the following lemma (see Appendix D).

LEMMA 1. *Let the generator $\mathcal{L}$ and the functions β and c be as above. Then, we have $v^{-1}\beta\mathcal{L}(\beta^{-1}v) + \mathcal{L}\log\beta = \beta^{-1}\hat{\mathcal{L}}^*\beta + \hat{\mathcal{L}}\log\beta$.*

THEOREM 1. *For DMMs as in Section 3.1–3.3, the log-likelihood is lower bounded by*

$$\mathcal{E}^\infty = \mathbb{E}_{\mathbb{Q}}\Big[\log p_0(Y_T)\Big|Y_0 = \mathbf{x}\Big] - \int_0^T \mathbb{E}_{\mathbb{Q}}\left[\frac{\hat{\mathcal{L}}^*\beta(Y_s, s)}{\beta(Y_s, s)} + \hat{\mathcal{L}}\log\beta(Y_s, s) \;\middle|\; Y_0 = \mathbf{x}\right]\mathrm{d}s. \quad (8)$$

This result extends the corresponding expression for $\mathbb{R}^d$ in Huang et al. (2021). We see the ELBO consists of a term representing the log-likelihood under the reference distribution and an implicit score matching term arising from the change in measure.

3.4. Finding suitable training objectives

Based on Theorem 1, we fit our generative model by maximising the expectation of $\mathcal{E}^\infty$ with respect to p_{data}. This is equivalent to minimising the objective

$$\mathcal{I}_{\text{ISM}}(\beta) = \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t, t)}{\beta(\mathbf{x}_t, t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t, t)\right]\mathrm{d}t, \quad (9)$$

which we call the *implicit score matching* objective, since it can be interpreted as an extension of implicit score matching from $\mathbb{R}^d$ (see Section 4 below for more intuition).

Since q_t and $\hat{\mathcal{L}}$ are determined by the noising process, which is known and assumed easy to sample from, $\mathcal{I}_{\mathrm{ISM}}(\beta)$ and its gradient with respect to β can be estimated in an unbiased fashion. Since β parameterises $\mathcal{M}$ via Assumption 2, and thus $\mathcal{K}$ through Assumption 1, minimising $\mathcal{I}_{\mathrm{ISM}}(\beta)$ over β is equivalent to learning the generative process.

We also have an equivalent *denoising score matching objective* (see Appendix E),

$$\mathcal{I}_{\mathrm{DSM}}(\beta) = \int_0^T \mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)}\left[\frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t,t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)\right]\mathrm{d}t. \tag{10}$$

Both objectives are minimised when $\beta(\mathbf{x},t) \propto q_t(\mathbf{x})$, as shown in Proposition 1. $\mathcal{I}_{\mathrm{DSM}}(\beta)$ can be interpreted as quantifying the difference between $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ and $\beta(\mathbf{x}_t,t)$ via the score matching operator $\Phi(f) = f^{-1}\mathcal{L}f - \mathcal{L}\log f$ introduced in Section 4 below. These objectives also generalise the following previously studied instances of diffusion models. For all derivations and remarks on the choice of parameterisation, see Appendix F.

Example 5 (Euclidean Diffusion). *In the setting of Example 1, Assumption 2 reduces to* $\nabla\log\beta = b + \mu$*, and we have* $f^{-1}\mathcal{L}f - \mathcal{L}\log f = \frac{1}{2}\left\|\nabla\log f\right\|^2$*. If we substitute* $s_\theta(\mathbf{x}_t,t) = \nabla\log\beta(\mathbf{x}_t,t)$*,* $\mathcal{I}_{\mathrm{DSM}}(\beta)$ *defined in Equation (10) reduces to Equation (3) and the reverse process is parameterised as in Equation (4). We thus recover the results of Song et al. (2021) and Huang et al. (2021).*

Example 6 (Discrete Space CTMC). *In the setting of Example 2, Assumption 2 reduces to* $A_{\mathbf{yx}} = \frac{\beta(\mathbf{x},t)}{\beta(\mathbf{y},t)}B_{\mathbf{xy}}$ *for all* $\mathbf{x} \neq \mathbf{y}$*. We may rewrite* $\mathcal{I}_{\mathrm{ISM}}$ *in terms of* A *to recover the objective of Campbell et al. (2022),*

$$\mathcal{I}_{\mathrm{ISM}}(A) = \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)}\Bigg[-A_{\mathbf{x}_t\mathbf{x}_t} - \sum_{\mathbf{y}\neq\mathbf{x}_t} B_{\mathbf{x}_t\mathbf{y}}\log A_{\mathbf{y}\mathbf{x}_t}\Bigg]\mathrm{d}t + \mathit{const}.$$

Example 7 (Riemannian Manifolds). *If* $\mathcal{X}$ *is a Riemannian manifold and we take* $\mathcal{K} = \partial_t + \mu\cdot\nabla + \frac{1}{2}\Delta$*,* $\mathcal{L} = \partial_t + b\cdot\nabla + \frac{1}{2}\Delta$ *where* Δ *is the Laplace–Beltrami operator associated to* $\mathcal{X}$*, and perform the reparameterisation* $s_\theta(\mathbf{x}_t,t) = \nabla\log\beta(\mathbf{x}_t,t)$*, then we recover the framework for training diffusion models on Riemannian manifolds given in De Bortoli et al. (2022) and Huang et al. (2022).*

3.5. Inference

To use DMMs for inference, we follow a similar procedure to Section 2.2. To noise a sample $(\mathbf{x}_0,\boldsymbol{\xi}_0) \sim p_{\mathrm{data}}$, we set $Y_0 = \mathbf{x}_0$ and let Y evolve according to $\mathcal{L}$. To generate $\mathbf{x}_0$ conditioned on an observation $\boldsymbol{\xi}^*$, we use a generative process $X^{\boldsymbol{\xi}^*}$ conditioned on $\boldsymbol{\xi}^*$. We parameterise $X^{\boldsymbol{\xi}^*}$ in terms of a function $\beta(\mathbf{x}_t,\boldsymbol{\xi}^*,t)$ which now takes $\boldsymbol{\xi}^*$ as an input.

We aim to learn $X^{\boldsymbol{\xi}^*}$ to approximate the time-reversal of Y conditioned on $\boldsymbol{\xi}^*$. The following extension of Theorem 1 (proved in Appendix D) gives us a way to do this.

THEOREM 2. *With the above set-up, minimising the objective*

$$\mathcal{I}_{\mathrm{DSM}}(\beta) = \int_0^T \mathbb{E}_{q(\mathbf{x}_0,\mathbf{x}_t,\boldsymbol{\xi}_0)}\left[\frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)\right]\mathrm{d}t$$

is equivalent to maximising a lower bound on the expected model log-likelihood.

Theorem 2 suggests that we may train conditional DMMs by maximising the objective $\mathcal{I}_{\mathrm{DSM}}(\beta)$ (or the equivalent $\mathcal{I}_{\mathrm{ISM}}(\beta)$ objective). Since $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ is known, we may do this by calculating an empirical estimate for $\mathcal{I}_{\mathrm{DSM}}(\beta)$ based on samples $(\mathbf{x}_0, \boldsymbol{\xi}_0)$ drawn from p_{data} and minimising over β. Then, we generate samples from $p_{\mathrm{data}}(\mathbf{x}_0|\boldsymbol{\xi}^*)$ by initialising $X_0^{\boldsymbol{\xi}^*} \sim p_0$, simulating the reverse process with generator $\mathcal{K}$ parameterised by $\beta = \beta(\cdot, \boldsymbol{\xi}^*, \cdot)$, and setting $\mathbf{x}_0 = X_T^{\boldsymbol{\xi}^*}$.

4. Score matching on general state-spaces

When X and Y are Euclidean diffusions, the objective $\mathcal{I}_{\mathrm{DSM}}(\beta)$ in Equation (10) becomes the score matching objective in Equation (3). Similarly, the objective $\mathcal{I}_{\mathrm{ISM}}(\beta)$ from Equation (9) reduces to the implicit score matching objective introduced by Hyvärinen (2005). This suggests we can view Equations (9) and (10) as generalisations of score matching objectives to arbitrary state spaces.

Given state space $\mathcal{X}$ on which we have a Markov process generator $\mathcal{L}$ and an unknown distribution $q_0(\mathbf{x})$ we wish to approximate, the corresponding *generalised implicit score matching* method learns an approximation $\varphi(\mathbf{x})$ to $q_0(\mathbf{x})$ by minimising

$$\mathcal{J}_{\mathrm{ISM}}(\varphi) = \mathbb{E}_{q_0(\mathbf{x})}\left[\frac{\hat{\mathcal{L}}^*\varphi(\mathbf{x})}{\varphi(\mathbf{x})} + \hat{\mathcal{L}}\log\varphi(\mathbf{x})\right].$$

We can show that $\mathcal{J}_{\mathrm{ISM}}$ is equivalent to the *generalised explicit score matching objective*

$$\mathcal{J}_{\mathrm{ESM}}(\varphi) = \mathbb{E}_{q_0(\mathbf{x})}\left[\frac{\mathcal{L}(q_0/\varphi)(\mathbf{x})}{(q_0(\mathbf{x})/\varphi(\mathbf{x}))} - \mathcal{L}\log(q_0/\varphi)(\mathbf{x})\right].$$

In addition, we define the corresponding *generalised denoising score matching* method, which learns an approximation $\varphi_\tau(\mathbf{x}_\tau)$ to the noised distribution $q_\tau(\mathbf{x}_\tau)$, formed by sampling $\mathbf{x}_0 \sim q_0(\cdot)$ and $\mathbf{x}_\tau \sim q_{\tau|0}(\,\cdot\,|\mathbf{x}_0)$, where $q_{\tau|0}$ is the transition probability associated to $\mathcal{L}$ run for time τ. It does this by minimising the objective

$$\mathcal{J}_{\mathrm{DSM}}(\varphi_\tau) = \mathbb{E}_{q_{0,\tau}(\mathbf{x}_0,\mathbf{x}_\tau)}\left[\frac{\mathcal{L}(q_{\tau|0}(\cdot|\mathbf{x}_0)/\varphi_\tau(\cdot))(\mathbf{x}_\tau)}{q_{\tau|0}(\mathbf{x}_\tau|\mathbf{x}_0)/\varphi_\tau(\mathbf{x}_\tau)} - \mathcal{L}\log(q_{\tau|0}(\cdot|\mathbf{x}_0)/\varphi_\tau(\cdot))(\mathbf{x}_\tau)\right].$$

$\mathcal{J}_{\mathrm{DSM}}$ is equivalent to both $\mathcal{J}_{\mathrm{ISM}}$ and $\mathcal{J}_{\mathrm{ESM}}$ when used to learn the smoothed distribution $q_\tau(\mathbf{x}_\tau)$ (see Appendix E). All three objectives extend the corresponding score matching objectives introduced for $\mathbb{R}^d$ by Hyvärinen (2005) and Vincent (2011). They also coincide with the extension of score matching for Riemannian manifolds of Mardia et al. (2016).

To illustrate further intuitions behind our objective functions, we define the *score matching operator* $\Phi(f) = f^{-1}\mathcal{L}f - \mathcal{L}\log f$. Note that the time component of Φ cancels,

so we can view it as an operator on $\mathcal{X}$. With this notation, the generalised explicit score matching objective becomes $\mathcal{J}_{\text{ESM}}(\varphi) = \mathbb{E}_{q_0(\mathbf{x})}[\Phi(q_0/\varphi)(\mathbf{x})]$. For Euclidean diffusions, $\Phi(f) = \frac{1}{2}||\nabla \log f||^2$ (see Example 5). In the general case, we view $\Phi(f)$ as measuring the magnitude of a logarithmic gradient of f. We interpret the objectives $\mathcal{J}_{\text{DSM}}$ and $\mathcal{J}_{\text{ESM}}$ as trying to fit φ to q_0 by minimising this logarithmic gradient of the ratio q_0/φ.

PROPOSITION 1. *Let Y be a Feller process with semigroup operators $(Q_t)_{t\geq 0}$, generator $\mathcal{L}$ and associated score matching operator Φ. Then:*

(a) $\Phi(f) \geq 0$ for all f in the domain of Φ, with equality if f is constant;
(b) for any probability measures π_1, π_2 on $\mathcal{X}$ and $t \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\pi_1 Q_t||\pi_2 Q_t) = -\mathbb{E}_{\pi_1 Q_t}\left[\Phi\left(\frac{\mathrm{d}(\pi_1 Q_t)}{\mathrm{d}(\pi_2 Q_t)}\right)\right],$$

where $\mathrm{KL}(\pi_1 Q_t||\pi_2 Q_t)$ denotes the Kullback–Leibler divergence between $\pi_1 Q_t$, $\pi_2 Q_t$.

Proposition 1(a) shows that Φ is always non-negative, so $\mathcal{J}_{\text{ESM}}$ is minimised if $\varphi(\mathbf{x}) \propto q_0(\mathbf{x})$. Thus minimising any of our generalised score matching objectives should typically correspond to learning an approximation to q_0. Note though that if Q_t is not ergodic and π_1, π_2 are different invariant distributions of Q_t then Proposition 1(b) implies that $\Phi(\mathrm{d}\pi_1/\mathrm{d}\pi_2) = 0$ π_1-a.e., even though $\mathrm{d}\pi_1/\mathrm{d}\pi_2$ is not constant. This suggests that generalised score matching may fail if the noising process is not ergodic. Proposition 1(b) was proved for score matching on $\mathbb{R}^d$ by Lyu (2009). It suggests we can interpret score matching as finding an approximation φ which minimises the decrease in KL divergence between q_0 and φ caused by adding an infinitesimal amount of noise to both according to $\mathcal{L}$.

Our generalised score matching methods give a principled way to extend score matching to fit unnormalised probability distributions on arbitrary spaces. Other extensions of score matching have been explored, including to arbitrary sub-domains of $\mathbb{R}^d$ (Yu et al., 2022), ratio matching (Hyvärinen, 2007) and marginalisation with generalised score matching (Lyu, 2009). However, these methods lack the generality of our framework and do not respect the intuition coming from $\mathbb{R}^d$ that Proposition 1(b) should hold. There are also many other density estimation methods that seek to learn ratios of density functions, including noise-contrastive estimation, which also approximates score matching under certain conditions (Gutmann and Hirayama, 2011).

5. Relationship to discrete time models

Denoising diffusion models were originally introduced in discrete time by Sohl-Dickstein et al. (2015). In this setting, the noising and generative processes are Markov chains $\mathbf{x}_{0:T} = (\mathbf{x}_{t_k})_{k=0}^N$ observed at a sequence of times $0 = t_0 < t_1 < \cdots < t_N = T$, with fixed forwards transition kernel $\tilde{q}(\mathbf{x}_{t_k}|\mathbf{x}_{t_{k-1}})$ and learned backwards kernel $\tilde{p}_\theta(\mathbf{x}_{t_{k-1}}|\mathbf{x}_{t_k})$. To fit discrete time diffusion models, Sohl-Dickstein et al. (2015) minimise the following Kullback–Leibler divergence with respect to θ:

$$\mathrm{KL}(\tilde{q}(\mathbf{x}_{0:T})||\tilde{p}_\theta(\mathbf{x}_{0:T})) = \sum_{k=1}^{N} \mathbb{E}_{\tilde{q}(\mathbf{x}_{t_{k-1}}, \mathbf{x}_{t_k})}\left[\log \frac{\tilde{q}(\mathbf{x}_{t_k}|\mathbf{x}_{t_{k-1}})}{\tilde{p}_\theta(\mathbf{x}_{t_{k-1}}|\mathbf{x}_{t_k})}\right] + \textit{const.} \tag{11}$$

Given any DMM with generators $\mathcal{K}, \mathcal{L}$ and marginals p_t, q_t as in Section 3, we define its *natural discretisation* to be the discrete-time model with $\tilde{q}(\mathbf{x}_{t_k}|\mathbf{x}_{t_{k-1}}) = q_{t_k|t_{k-1}}(\mathbf{x}_{t_k}|\mathbf{x}_{t_{k-1}})$ and $\tilde{p}_\theta(\mathbf{x}_{t_{k-1}}|\mathbf{x}_{t_k}) = p_{T-t_{k-1}|T-t_k}(\mathbf{x}_{t_{k-1}}|\mathbf{x}_{t_k})$. Then, the Kullback–Leibler divergence (11) for the natural discretisation can be viewed as a first-order approximation to $\mathcal{I}_{\mathrm{ISM}}$ for the continuous-time model.

LEMMA 2. *Suppose X, Y are fixed generative and noising processes with marginals p, q as in Section 3, and suppose that they are related as in Assumptions 1 and 2 for some sufficiently regular function β. Then for any $0 < s < t < T$ with $\gamma = t - s$,*

$$\gamma\, \mathbb{E}_{q_s(\mathbf{x}_s)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_s)}{\beta(\mathbf{x}_s)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_s)\right] = \mathbb{E}_{q_{s,t}(\mathbf{x}_s,\mathbf{x}_t)}\left[\log\frac{q_{t|s}(\mathbf{x}_t|\mathbf{x}_s)}{p_{T-s|T-t}(\mathbf{x}_s|\mathbf{x}_t)}\right] + o(\gamma).$$

Applying this lemma on each interval $[t_k, t_{k+1}]$, we get the following theorem.

THEOREM 3. *For any DMM, the objective (11) for its natural discretisation is equivalent to the natural discretisation of $\mathcal{I}_{\mathrm{ISM}}$ to first order in $\overline{\gamma} = \max_{k=0,\dots,N-1}|t_{k+1} - t_k|$.*

This theorem generalises to arbitrary state spaces a result of Ho et al. (2020), which demonstrated the equivalence of minimizing (11) and the score matching objective for Euclidean state spaces. For the proofs of Lemma 2 and Theorem 3, see Appendix H.

Lemma 2 also implies a general equivalence between one-step denoising autoencoders and score matching. Vincent (2011) discussed this equivalence for autoencoders using Gaussian noise in $\mathbb{R}^d$, but our methods allow us to extend this correspondence to arbitrary state spaces and noising processes. For more details, see Appendix I.

6. Experiments

We now present experiments demonstrating DMMs on several tasks and data spaces, for unconditional generation and conditional simulation. All details are in Appendix J.

6.1. Inference on $\mathbb{R}^d$ using diffusion processes

First, we use diffusion processes in $\mathbb{R}^d$ to perform approximate Bayesian inference for real-valued parameters. We consider $p_{\mathrm{data}}(\boldsymbol{\xi}|\mathbf{x}) = \prod_{i=1}^N p_{\mathrm{data}}(\xi_i|\mathbf{x})$, where $p_{\mathrm{data}}(\xi_i|\mathbf{x})$ is the g-and-k distribution with parameters $\mathbf{x} = (A, B, g, k)$ and $d = 4$, and we let $p_{\mathrm{data}}(\mathbf{x})$ be uniform on $[0, 10]^4$. The g-and-k distribution is a 4-parameter distribution in which A, B, g, k control the location, scale, skewness and kurtosis respectively.

We fix our noising process to be an Ornstein–Uhlenbeck process, and parameterise our reverse process as in Example 5, with $s_\theta(\mathbf{x}, \boldsymbol{\xi}, t)$ being given by a fully connected neural network. To train the model, we sample $(\mathbf{x}_0, \boldsymbol{\xi}_0) \sim p_{\mathrm{data}}$ and minimise the denoising score matching objective from Section 3.5 via stochastic gradient descent on θ.

To test our model, we first consider the case where there are a true set of underlying parameters $\mathbf{x}_{\mathrm{true}} = (3, 1, 2, 0.5)$. We generate an observation $\boldsymbol{\xi}_0 \sim p_{\mathrm{data}}(\boldsymbol{\xi}_0|\mathbf{x}_{\mathrm{true}})$ with $N = 250$, sample from the approximate posterior using our DMM and plot the result in Fig. 2. We compare our method with the semi-automatic ABC (SA-ABC) (Nunes and Prangle, 2015) and Wasserstein SMC (W-SMC) (Bernton et al., 2019) methodologies, as well as Sequential Neural Posterior, Likelihood and Ratio Estimation approaches (SNPE, SNLE and SNRE) (see e.g. Lueckmann et al. (2021)). We see in Fig. 2 that the

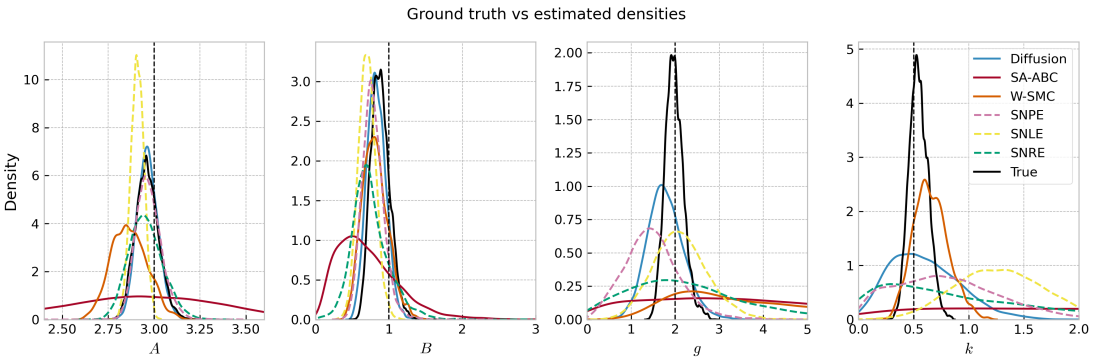

Fig. 2. Posterior kernel density estimates of samples generated using our DMM, SA-ABC, W-SMC, SNLE, SNPE and SNRE for the g-and-k distribution, with $\mathbf{x}_{\text{true}} = (3, 1, 2, 0.5)$ and $N = 250$.

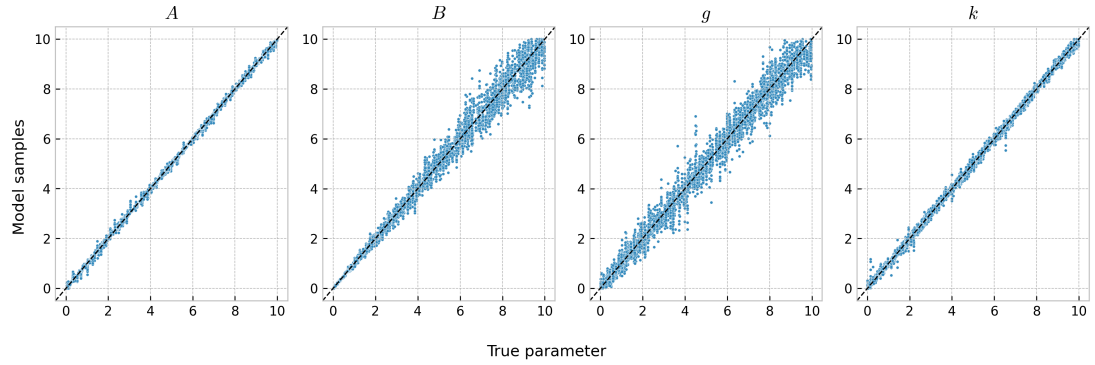

Fig. 3. Comparison of posterior samples $\mathbf{x}_0'$ from our DMM approximation to $p_{\text{data}}(\cdot|\boldsymbol{\xi}_0)$ and the true parameter value $\mathbf{x}_0$ for a range of $\mathbf{x}_0$ in the prior distribution, with $N = 10000$.

DMM achieves more accurate posterior estimation for all parameters, except the kurtosis parameter k for which W-SMC is more accurate. Among the other neural network-based approaches, SNPE appears most competitive on this task, but is less accurate than the DMM especially for parameters g and k. Additional experimental results comparing DMMs to other simulation-based inference methods can be found in (Sharrock et al., 2022; Geffner et al., 2023).

Next, we demonstrate that our model can perform inference for a range of observation values $\boldsymbol{\xi}^*$ simultaneously. We generate a series of 512 parameter values $\mathbf{x}_0$ drawn from $p_{\text{data}}(\mathbf{x}_0)$ and draw an observation $\boldsymbol{\xi}_0$ from $p_{\text{data}}(\boldsymbol{\xi}_0|\mathbf{x}_0)$ with $N = 10000$ for each $\mathbf{x}_0$. Then, we generate 8 samples $\mathbf{x}_0'$ from our approximation to the posterior $p_{\text{data}}(\mathbf{x}_0|\boldsymbol{\xi}_0)$ for each $\boldsymbol{\xi}_0$. We plot each component of the pairs $(\mathbf{x}_0, \mathbf{x}_0')$ in Fig. 3. We see our model is able to infer the original parameters across a range of parameter values.

6.2. Image inpainting and super-resolution using discrete-space CTMCs

Second, we demonstrate that our framework is applicable for large-scale Bayesian inverse problems, such as super-resolution and inpainting for images. For these problems, the prior $p_{\text{data}}(\mathbf{x})$ is the distribution of images. Most ABC techniques such as SA-ABC

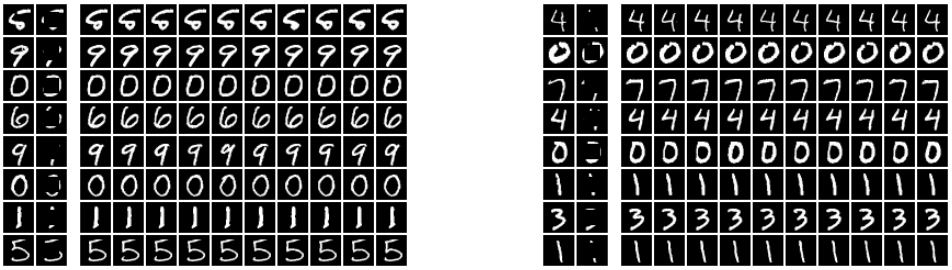

Fig. 4. Samples from the MNIST inpainting task. The first column in each set plots the ground truth images, and the second column has the centre 14×14 pixels missing.

and W-SMC are not applicable as they require an analytical expression for this prior, whereas DMMs do not rely on such an expression.

We consider performing image inpainting for MNIST digit images, where each image $\mathbf{x}_0$ has 28×28 pixels with values in $\{0, \dots, 255\}$, and the observed incomplete image $\boldsymbol{\xi}_0$ has the middle 14×14 pixels missing. Since our state space $\mathcal{X} = \{0, \dots, 255\}^{28 \times 28}$ is discrete, we use the set-up of Example 2 and let the generator of our noising process factor over pixel dimensions. We use the denoising parameterisation of the reverse process (see Appendix F.2) and train by minimising the form of the objective in Example 6.

To test our model, we plot the reconstructed image samples for a number of digits in Fig. 4. We observe that the samples we obtain are consistent with conditioning and appear to be realistic, but also display diversity in the shape of the strokes. In Appendix J.2, we also compare our method to a continuous state space approach.

In addition, we train a conditional discrete-space DMM to perform super-resolution on ImageNet images to demonstrate that this method provides perceptually high quality samples even in very high-dimensional scenarios. For details, see Appendix J.3.

6.3. Modelling distributions on $SO(3)$ using manifold diffusions

Thirdly, we demonstrate that DMMs can approximate distributions on manifolds using two tasks on $SO(3)$. Since $SO(3)$ is a Lie group and so a Riemannian manifold, we use the framework from Example 7. As our noising process, we use Brownian motion with generator $\mathcal{L} = \partial_t + \frac{1}{2}\Delta$. We can explicitly calculate the transition kernels $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ for this process, allowing us to use the denoising score matching objective. We parameterise this objective in terms of a neural network approximation $s_\theta(\mathbf{x}, t)$ of the score. This is in contrast to De Bortoli et al. (2022), in which the explicit transition kernels are not used for sampling the forward process or in the loss function, both of which require further approximations.

First we check that our DMM can learn simple mixtures of wrapped normal distributions $p_{\text{data}}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{N}^W(\mathbf{x}|\mu_m, \sigma_m^2)$, where $\mathcal{N}^W(\mathbf{x} \mid \mu_m, \sigma_m^2)$ is the wrapped normal distribution on $SO(3)$ with expectation μ_m and variance σ_m^2 (De Bortoli et al., 2022). We plot samples from our resulting DMM in Fig. 5. We see that our model provides a good fit to $p_{\text{data}}(\mathbf{x})$, covering all modes. In Appendix J.5, we provide additional results and show that we can also sample from the class conditional density $p_{\text{data}}(\mathbf{x}|m)$.

Second, we consider a more realistic pose estimation task on the SYMSOL dataset, which requires predicting the 3D orientation of various symmetric 3D solids based on

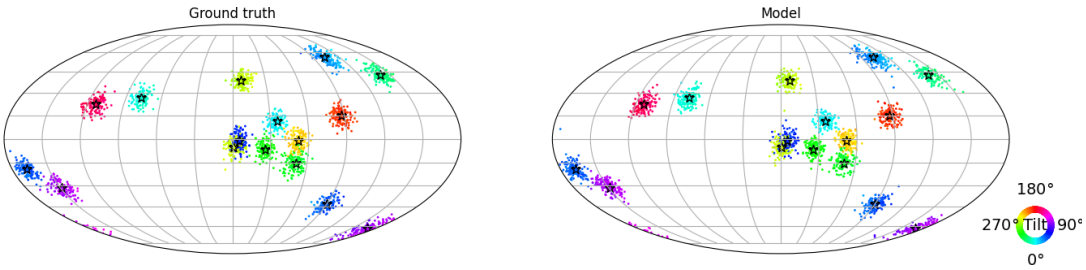

Fig. 5. Samples from the ground truth and our DMM approximation to the mixture of wrapped normal distributions. Each sample is denoted by a point, whose position represents the axis of rotation and whose colour represents the angle of rotation. Stars denote the true cluster means.

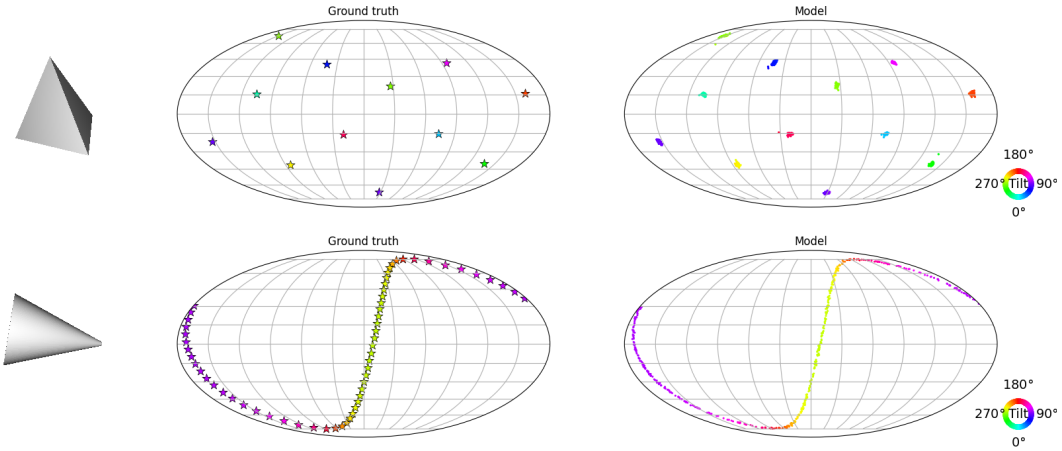

Fig. 6. Samples from the ground truth (plotted as stars, middle) and our pose estimation DMM (right) conditioned on 2D views of two shapes (left). The axis of rotation and rotation angle are represented by position and colour respectively.

2D views (Murphy et al., 2021). Due to the rotational symmetries, a key challenge is to predict all possible poses when only one possibility is presented in training. We use a conditional DMM where $\boldsymbol{\xi}$ is the 2D image view. Fig. 6 shows two sets of samples from our model conditioned on 2D images of two different solids. We see that our model learns to sample from the ground truth accurately and infer the full set of rotational symmetries for different views $\boldsymbol{\xi}$. For further experimental details and plots, see Appendix J.6.

6.4. Approximation of distributions over measures using Wright–Fisher diffusions

Finally, we present an example of learning to approximate a distribution over measures on a finite state space $E = \{1, \dots, N\}$. In this case $\mathcal{X} = \mathcal{P}(E)$, the space of measures on E. This is of particular interest in compositional data analysis (Greenacre, 2021). Elements of $\mathcal{X}$ can be parameterised by tuples of real numbers $\mathbf{p} = (p_1, \dots, p_N) \in [0,1]^N$ such that $\sum_{i=1}^N p_i = 1$. We could approximate the data distribution using a diffusion model on $\mathbb{R}^N$, but such a model would not reflect the fact that our distribution should be supported on a submanifold, the simplex. Using the standard setup for manifold diffusions as in Example 7 would not respect the boundary of the simplex. Other methods have been

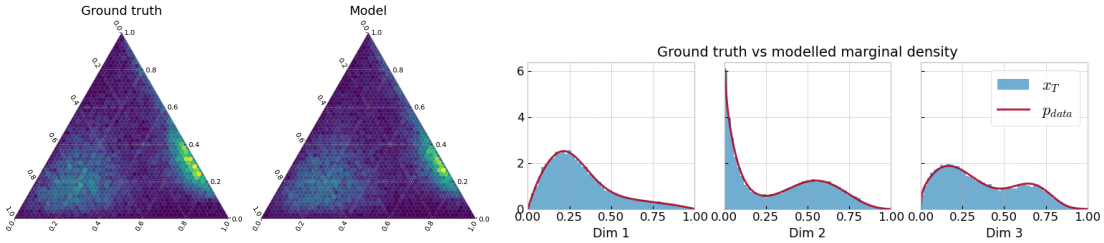

Fig. 7. Histograms of samples from our simplex DMM and the ground truth mixture of Dirichlet distributions for dimension $N = 3$, plotted over the whole space as a ternary plot (left) and over the marginals per dimension (right).

presented in the literature, but they rely on either reflected diffusions (Lou and Ermon, 2023) or on projections of the simplex (Richemond et al., 2022).

We therefore use Wright–Fisher diffusions, a process used in population genetics to model the evolution of allele frequencies, as our class of generative processes. A Wright–Fisher process has generator $\mathcal{L} = \partial_t + \frac{1}{2}\sum_{i,j=1}^{N} p_i\,(\delta_{ij} - p_j)\,\frac{\partial^2}{\partial p_i \partial p_j} + \sum_{i,j=1}^{N} q_{ij} p_i \frac{\partial}{\partial p_j}$, where $(q_{ij})_{i,j=1,\dots,N}$ is some matrix such that $\sum_{j=1}^{N} q_{ij} = 0$ for each $i = 1, \dots, N$. The process takes values in the space of measures on E, and so respects the structure of our data distribution (Ethier and Griffiths, 1993). For specific choices of q_{ij}, the process converges to a known invariant distribution and we can calculate the implicit score matching loss. For details of the theoretical setup, see Appendix F.4.

We evaluate the proposed method by modelling $p_{\text{data}}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} \text{Dirichlet}(\alpha_m)$, a mixture of Dirichlet distributions with parameters $\alpha_m \in \mathbb{R}^N$, for various values of N. Fig. 7 shows two visualisations of samples drawn from our DMM compared to ground truth samples in dimension $N = 3$. Our model is able to accurately approximate $p_{\text{data}}(\mathbf{x})$. For further evaluations and experimental details, see Appendix J.7.

7. Discussion

We have provided here a general framework which allows us to extend denoising diffusion models to general state-spaces. The resulting DMMs can be trained with principled objectives and used for inference, generalizing along the way score matching ideas. Their applicability and performance have been demonstrated on a range of problems. From a methodological point of view, the proposed framework is general enough to accommodate, for example, general noising processes, mixed continuous/discrete processes and some infinite-dimensional settings with finite representations (though our assumptions on the state space (see Appendix B.1) may fail to hold in the infinite-dimensional setting so more care is required).

However, we still lack a proper theoretical understanding of these models. Under realistic assumptions on the data distribution, De Bortoli (2023) and Chen et al. (2023) show that diffusion models on $\mathbb{R}^d$ can in theory learn essentially any distribution given a good enough score approximation and infinite data. However finite sample guarantees are currently absent. Moreover, p_{data} is typically an empirical measure as we only have access

to a finite set of datapoints, so q_t is a mixture of Gaussians for an Ornstein–Uhlenbeck noising diffusion and its score $\nabla \log q_t$ is thus available. If we were simulating samples using the exact time reversal of this diffusion, we would simply recover the empirical distribution. It is because we are approximating the time-reversal and in particular using an approximation of the scores that we are able to obtain novel samples. It is not yet clear why the approximation of the score using neural networks appears to provide perceptually realistic samples for many applications.

The effectiveness of such methods for inference, even in scenarios where standard MCMC or ABC techniques are not applicable (Sharrock et al., 2022; Geffner et al., 2023), may also be considered surprising. One perspective on the training process is that it involves the model constructing its own summary statistics that allow it to perform inference effectively on the training observations. It is not yet well understood why the summary statistics the model learns appear empirically effective, or what sorts of summary statistics our training procedure biases the model towards.

Overall, this contribution shows how the range of existing models relate to each other and may help applying DMMs in practice to a large variety of problems. However, our understanding of such models is still incomplete and deserves further attention.

Acknowledgments

Joe Benton was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1) and Yuyang Shi by the Huawei UK Fellowship Programme. Arnaud Doucet acknowledges support of the UK Dstl and EPSRC grant EP/R013616/1. This is part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative. He also acknowledges support from the EPSRC grants CoSines (EP/R034710/1) and Bayes4Health (EP/R018561/1).

References

Anderson, B. D. O. (1982). Reverse-time Diffusion Equation Models. *Stochastic Processes and their Applications 12*, 313–326.

Austin, J., D. D. Johnson, J. Ho, D. Tarlow, and R. van den Berg (2021). Structured Denoising Diffusion Models in Discrete State-Spaces. *NeurIPS*.

Bernton, E., P. E. Jacob, M. Gerber, and C. P. Robert (2019). Approximate Bayesian Computation with the Wasserstein Distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(2), 235–269.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language Models are Few-shot Learners. *NeurIPS*.

Campbell, A., J. Benton, V. De Bortoli, T. Rainforth, G. Deligiannidis, and A. Doucet (2022). A Continuous Time Framework for Discrete Denoising Models. *NeurIPS*.

Chen, S., S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang (2023). Sampling is as Easy as Learning the Score: Theory for Diffusion Models with Minimal Data Assumptions. *ICLR*.

De Bortoli, V. (2023). Convergence of Denoising Diffusion Models under the Manifold Hypothesis. *Transactions on Machine Learning Research*.

De Bortoli, V., E. Mathieu, M. Hutchinson, J. Thornton, Y. W. Teh, and A. Doucet (2022). Riemannian Score-Based Generative Modeling. *NeurIPS*.

Dhariwal, P. and A. Nichol (2021). Diffusion Models Beat GANs on Image Synthesis. *NeurIPS*.

Ethier, S. N. and R. C. Griffiths (1993). The Transition Function of a Fleming-Viot Process. *The Annals of Probability 21*, 1571–1590.

Geffner, T., G. Papamakarios, and A. Mnih (2023). Compositional Score Modeling for Simulation-based Inference. *arXiv preprint arXiv:2209.14249*.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative Adversarial Nets. *NeurIPS*.

Greenacre, M. (2021). Compositional Data Analysis. *Annual Review of Statistics and its Application 8*, 271–299.

Gutmann, M. U. and J.-i. Hirayama (2011). Bregman Divergence as General Framework to Estimate Unnormalized Statistical Models. *UAI*.

Ho, J., A. Jain, and P. Abbeel (2020). Denoising Diffusion Probabilistic Models. *NeurIPS*.

Hoogeboom, E., D. Nielsen, P. Jaini, P. Forré, and M. Welling (2021). Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions. *NeurIPS*.

Huang, C.-W., M. Aghajohari, A. J. Bose, P. Panangaden, and A. Courville (2022). Riemannian Diffusion Models. *NeurIPS*.

Huang, C.-W., J. H. Lim, and A. Courville (2021). A Variational Perspective on Diffusion-Based Generative Models and Score Matching. *NeurIPS*.

Hyvärinen, A. (2005). Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research 6*, 695–709.

Hyvärinen, A. (2007). Some Extensions of Score Matching. *Computational Statistics and Data Analysis 51*, 2499 – 2512.

Kingma, D. P. and M. Welling (2014). Auto-Encoding Variational Bayes. *ICLR*.

Lou, A. and S. Ermon (2023). Reflected Diffusion Models. *ICML*.

Lueckmann, J.-M., J. Boelts, D. S. Greenberg, P. J. Gonçalves, and J. H. Macke (2021). Benchmarking Simulation-Based Inference. *AISTATS*.

Lyu, S. (2009). Interpretation and Generalization of Score Matching. *UAI*.

Mardia, K. V., J. T. Kent, and A. K. Laha (2016). Score Matching Estimators for Directional Distributions. *arXiv preprint arXiv:1604.08470*.

Murphy, K. A., C. Esteves, V. Jampani, S. Ramalingam, and A. Makadia (2021). Implicit-PDF: Non-Parametric Representation of Probability Distributions on the Rotation Manifold. *ICML*.

Nunes, M. A. and D. Prangle (2015). abctools: An R Package for Tuning Approximate Bayesian Computation Analyses. *The R Journal 7*(2), 189–205.

Oord, A. v. d., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu (2016). WaveNet: A Generative Model for Raw Audio. *arXiv:1609.03499*.

Popov, V., I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov (2021). Grad-tts: A Diffusion Probabilistic Model for Text-to-speech. *ICML*.

Rezende, D. J. and S. Mohamed (2015). Variational Inference with Normalizing Flows. *ICML*.

Richemond, P. H., S. Dieleman, and A. Doucet (2022). Categorical SDEs with Simplex Diffusion. *arXiv preprint arXiv:2210.14784*.

Saharia, C., J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi (2022). Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–14.

Sharrock, L., J. Simons, S. Liu, and M. Beaumont (2022). Sequential Neural Score Estimation: Likelihood-Free Inference with Conditional Score Based Diffusion Models. *arXiv preprint arXiv:2210.04872*.

Sohl-Dickstein, J., E. A. Weiss, N. Maheswaranathan, and S. Ganguli (2015). Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *ICML*.

Song, Y., C. Durkan, I. Murray, and S. Ermon (2021). Maximum Likelihood Training of Score-Based Diffusion Models. *NeurIPS*.

Song, Y., J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole (2021). Score-Based Generative Modeling through Stochastic Differential Equations. *ICLR*.

Sun, H., L. Yu, B. Dai, D. Schuurmans, and H. Dai (2023). Score-based Continuous-time Discrete Diffusion Models. *ICLR*.

Trippe, B. L., J. Yim, D. Tischer, D. Baker, T. Broderick, R. Barzilay, and T. Jaakkola (2023). Diffusion Probabilistic Modeling of Protein Backbones in 3D for the Motif-scaffolding Problem. *ICLR*.

Vincent, P. (2011). A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation 23*, 1661–1674.

Yu, S., M. Drton, and A. Shojaie (2022). Generalized Score Matching for General Domains. *Information and Inference: A Journal of the IMA 11*(2), 739–780.

A. Background on Feller processes

We recall some basic definitions and properties associated with Feller processes which we use for the derivations in Section 3. Our principal source is Dong (2003).

A.1. Definition of a Feller process

Let S be a locally compact, separable metric space and let $C_0(S)$ denote the set of continuous functions $f : S \to \mathbb{R}$ such that for any $\epsilon > 0$ there exists a compact $K \subseteq S$ such that $|f(x)| < \epsilon$ for all $x \notin K$. Also, let $||f||$ denote the supremum norm on $C_0(S)$.

Definition 1 (Feller process). *A time-homogeneous Markov process $(X_t)_{t\geq 0}$ with state space S and associated transition semigroup $(P_t)_{t\geq 0}$ is a Feller process if:*

- *$P_t f \in C_0(S)$ for all $f \in C_0(S)$ and $t \geq 0$.*
- *$||P_t f|| \leq ||f||$ for all $f \in C_0(S)$.*
- *$P_t f(x) \to f(x)$ as $t \to 0$ for all $x \in S$ and $f \in C_0(S)$.*

Definition 2 (Generator of a Feller process). *Suppose X is a Feller process on S as above and f is a function in $C_0(S)$. If the limit*

$$\mathcal{A}f := \lim_{s\to 0} \frac{P_s f - f}{s}$$

exists in $C_0(S)$, we say that f is in the domain of the generator of X. We call the operator $\mathcal{A}$ defined in this way the generator of X and denote its domain by $\mathcal{D}(\mathcal{A})$.

In the main text, we are concerned with Feller processes $\overline{X}, \overline{Y}$ defined on the extended space $\mathcal{S} = \mathcal{X} \times [0, \infty)$ which are constructed by taking a time-inhomogeneous Markov process X on $\mathcal{X}$ and defining $\overline{X} = (X_t, t)_{t\geq 0}$. In this setting, we have the following variant of Dynkin's formula.

Lemma 3 (Dynkin's formula). *If $\overline{X} = (X_t, t)_{t\geq 0}$ is a Feller process on $\mathcal{S}$ with generator $\mathcal{A}$ and $f \in \mathcal{D}(\mathcal{A})$, then*

$$M_t^f = f(X_t, t) - f(X_0, 0) - \int_0^t \mathcal{A}f(X_s, s)\ \mathrm{d}s$$

is a martingale with respect to the natural filtration of $\overline{X}$.

Proof. See Theorem 27.20 in Dong (2003).

A.2. Adjoint of a generator

Given a state space S and a reference measure ν on S, we can define an inner product on $C_0(S)$ by letting

$$\langle f, h\rangle = \int_S fh\ \mathrm{d}\nu$$

for all $f, h \in C_0(S)$ such that the integral exists. This induces a Hilbert space structure on $C_0(S)$ and allows us to make the following definition, from Yosida (1965).

DEFINITION 3 (ADJOINT OF AN OPERATOR). *Given operator $\mathcal{A}$ with domain $\mathcal{D}(\mathcal{A})$ contained in $C_0(S)$, we define the adjoint operator $\mathcal{A}^*$ acting at function $f \in C_0(S)$ by*

$$\langle \mathcal{A}^* f, h \rangle = \langle f, \mathcal{A} h \rangle \qquad \text{for all } h \in \mathcal{D}(\mathcal{A}).$$

The domain $\mathcal{D}(\mathcal{A}^*)$ of $\mathcal{A}^*$ is the set of all functions f such that there exists some function $\mathcal{A}^* f$ for which the above holds.

B. Assumptions for Section 3

Here, we list the assumptions under which our derivations in Section 3 hold. Note that these assumptions can be verified in several relevant cases (see Appendix F).

B.1. Assumptions on the state space $\mathcal{X}$

ASSUMPTION 3. *The state space $\mathcal{X}$ is a locally compact, separable metric space. In addition, there exists a reference measure ν on $\mathcal{X}$ with respect to which all relevant probability distributions are absolutely continuous.*

B.2. Assumptions on the marginals p and q

ASSUMPTION 4. *We have $p_t \in \mathcal{D}(\hat{\mathcal{K}}^*)$ for each $t \in [0, T]$, where $\hat{\mathcal{K}}^*$ is the adjoint of the spatial part of the operator $\mathcal{K}$. In addition, p is differentiable with respect to t and $\partial_t p$ is bounded.*

B.3. Assumptions on the generators $\mathcal{K}$ and $\mathcal{L}$

ASSUMPTION 5. *$\overline{X}$ and $\overline{Y}$ are Feller processes with associated transition semigroups $(P_t)_{t\geq 0}$, $(Q_t)_{t\geq 0}$ and generators $\mathcal{K}, \mathcal{L}$ respectively.*

ASSUMPTION 6. *$\mathcal{K}$ decomposes as $\mathcal{K} = \partial_t + \hat{\mathcal{K}}$, where $\hat{\mathcal{K}} f$ is defined only in terms of the spatial arguments of f, so we may view it as an operator on (a subset of) $C_0(\mathcal{X})$.*

ASSUMPTION 7. *There exists a subset $\mathcal{D}_0 \subseteq \mathcal{D}(\hat{\mathcal{K}}) \cap L^2(\mathcal{X}, \nu)$ which is dense in $L^2(\mathcal{X}, \nu)$, satisfies $\hat{\mathcal{K}} h \in \mathcal{D}_0$ for all $h \in \mathcal{D}_0$ and such that every function in $\mathcal{D}_0$ is bounded and has compact support.*

B.4. Assumptions on $\mathcal{M}$ and c

ASSUMPTION 8. *The function $c : \mathcal{S} \to \mathbb{R}$ is bounded, and the function $v : \mathcal{S} \to \mathbb{R}$ is bounded, in $\mathcal{D}(\mathcal{M})$ and satisfies $\int_0^T \mathbb{E}\left[|\mathcal{M} v(Z_s, s)|^2\right] \mathrm{d}s < \infty$.*

B.5. Assumptions on β

ASSUMPTION 9. *The functions β^{-1}, $\beta^{-1} v$, $\log \beta$ and $\log v$ are in $\mathcal{D}(\mathcal{L})$, β^{-1} and $\beta \mathcal{L}(\beta^{-1})$ are both bounded, and $\beta \in \mathcal{D}(\hat{\mathcal{L}}^*)$.*

C. Stochastic process theory

We provide full statements of the general stochastic process results used in Section 3. For completeness, we also provide proofs of the given results adapted to our setting.

THEOREM 4 (FOKKER–PLANCK). *Let $(X_t)_{t\in[0,T]}$ be a Markov process with generator $\mathcal{K}$ and marginals p_t satisfying the assumptions in Appendix B. Then p satisfies the forward Kolmogorov equation $\partial_t p = \hat{\mathcal{K}}^* p$ for ν-almost every $\mathbf{x}$.*

PROOF. For any $h \in \mathcal{D}_0$, by Assumptions 4 and 7 we may write

$$\begin{aligned}\langle \partial_t p - \hat{\mathcal{K}}^* p, h\rangle &= \int_{\mathcal{X}} (\partial_t p)h - p(\hat{\mathcal{K}}h)\,\mathrm{d}\nu \\ &= \partial_t \mathbb{E}\left[h(X_t)\right] - \mathbb{E}\left[\hat{\mathcal{K}}h(X_t)\right].\end{aligned}$$

Applying Dynkin's formula to $f(\mathbf{x}, t) = h(\mathbf{x})$, taking expectations and using Fubini's theorem, we see that

$$\mathbb{E}\left[h(X_t)\right] - \mathbb{E}\left[h(X_0)\right] = \int_0^t \mathbb{E}\left[\hat{\mathcal{K}}h(X_s)\right]\,\mathrm{d}s.$$

Differentiating with respect to t, we deduce that $\langle \partial_t p - \hat{\mathcal{K}}^* p, h\rangle = 0$. Since this holds for all $h \in \mathcal{D}_0$ and $\mathcal{D}_0$ is dense in $L^2(\mathcal{X}, \nu)$, we conclude that $\partial_t p - \hat{\mathcal{K}}^* p = 0$ holds ν-a.e. as required.

THEOREM 5 (FEYNMAN–KAC). *Let $\overline{Z} = (Z_t, t)_{t\geq 0}$ be a Feller process on $\mathcal{S}$ with generator $\mathcal{M}$. Suppose that we are given functions $v, c : \mathcal{S} \to \mathbb{R}$ and $h : \mathcal{X} \to \mathbb{R}$ such that $\mathcal{M}, v, c$ solve equation $\mathcal{M}v + cv = 0$ (as in Assumption 1) with boundary condition $v(\cdot, T) = h(\cdot)$. Suppose also that Assumption 8 is satisfied. Then we have*

$$v(\mathbf{x}, \tau) = \mathbb{E}\left[h(Z_T)\exp\left\{\int_\tau^T c(Z_s, s)\,\mathrm{d}s\right\} \;\middle|\; Z_\tau = \mathbf{x}\right]$$

for all $0 \leq \tau \leq T$.

PROOF. This result is well-known in the case of Euclidean diffusion processes (Karatzas and Shreve, 1991). In the general case, the proof relies on the theory of semimartingales (see for example Métivier (1982)). Fix $\tau \in [0, T]$ and for all $t \in [\tau, T]$ define

$$S_t = v(Z_t, t)\exp\left\{\int_\tau^t c(Z_s, s)\mathrm{d}s\right\}$$

along with

$$V_t = v(Z_t, t), \qquad U_t = \exp\left\{\int_\tau^t c(Z_s, s)\mathrm{d}s\right\}.$$

Each of these processes is clearly a semimartingale, and so we may define $\mathrm{d}S_t$, $\mathrm{d}U_t$ and $\mathrm{d}V_t$ accordingly (Métivier, 1982). The following lemma will allow us to express $\mathrm{d}S_t$ in terms of $\mathrm{d}U_t$ and $\mathrm{d}V_t$.

LEMMA 4 (INTEGRATION BY PARTS FOR SEMIMARTINGALES). *If U and V are semimartingales and at least one is continuous then we have*

$$\mathrm{d}(U_t V_t) = U_{t-}\mathrm{d}V_t + U_{t-}\mathrm{d}V_t + \mathrm{d}[U,V]^c_t,$$

where $[\cdot,\cdot]^c_t$ denotes the quadratic covariation.

PROOF. This is Theorem 2.7.4(ii) of Pulido (2011), or follows from applying Theorem 27.1 of Métivier (1982) to the function $\varphi(U,V) = UV$.

Since $v \in \mathcal{D}(\mathcal{M})$ by Assumption 8, by Dynkin's formula we have that V is a semimartingale and we may decompose

$$\mathrm{d}V_t = \mathcal{M}v\mathrm{d}t + \mathrm{d}M^v_t$$

where M^v_t is a martingale. Also, since $c(x,t)$ is bounded by Assumption 8, U is a continuous, adapted, previsible process of finite variation and satisfies

$$\mathrm{d}U_t = c(Z_t,t)\exp\left\{\int_\tau^t c(Z_s,s)\,\mathrm{d}s\right\}\mathrm{d}t.$$

In addition, note that $\mathrm{d}[U,V]^c_t = 0$ since U is continuous and of finite variation. Therefore, by Lemma 4, we can calculate

$$\begin{aligned} S_t - S_\tau &= \int_\tau^t U_{s-}\mathrm{d}V_s + \int_\tau^t V_{s-}\mathrm{d}U_s + [U,V]^c_t \\ &= \int_\tau^t U_s\big\{\mathcal{M}v + cv\big\}\mathrm{d}s + \int_\tau^t U_s\,\mathrm{d}M^v_s \\ &= \int_\tau^t U_s\,\mathrm{d}M^v_s \end{aligned}$$

where we have used that $\mathcal{M}v + cv = 0$ in the last line. Therefore, S can be expressed as a stochastic integral with respect to the martingale M^v.

The conditions we have imposed through Assumption 8 on c and v imply that U is bounded and M^v is square-integrable. It follows, for example from Theorem 24.4.5 in (Métivier, 1982), that S is a local martingale and hence, since it is also bounded, a true martingale. We then have that

$$\begin{aligned} v(\mathbf{x},\tau) &= \mathbb{E}\left[S_\tau | Z_\tau = \mathbf{x}\right] = \mathbb{E}\left[S_T | Z_\tau = \mathbf{x}\right] \\ &= \mathbb{E}\left[h(Z_T)\exp\left\{\int_\tau^T c(Z_s,s)\mathrm{d}s\right\}\middle| Z_\tau = \mathbf{x}\right] \end{aligned}$$

as required.

THEOREM 6 (GIRSANOV). *Let $\overline{Y} = (Y_t,t)_{t\geq 0}$ and $\overline{Z} = (Z_t,t)_{t\geq 0}$ be Feller processes on $\mathcal{S}$ with generators $\mathcal{L}$, $\mathcal{M}$ and path measures $\mathbb{Q}$, $\mathbb{P}$ respectively, such that Y_0 and Z_0 have the same law. Suppose also that there exists a bounded, measurable function $\alpha : \mathcal{S} \to (0,\infty)$ in $\mathcal{D}(\mathcal{L})$ such that $\alpha^{-1}\mathcal{L}\alpha$ is bounded, and such that*

$$\alpha\mathcal{M}f = \mathcal{L}(f\alpha) - f\mathcal{L}\alpha \tag{12}$$

for all functions f such that $f \in \mathcal{D}(\mathcal{M})$ and $f\alpha \in \mathcal{D}(\mathcal{L})$. Then we have

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega) = \frac{\alpha(\omega_T, T)}{\alpha(\omega_0, 0)} \exp\Big\{ - \int_0^T \frac{\mathcal{L}\alpha(\omega_s, s)}{\alpha(\omega_s, s)} \, \mathrm{d}s \Big\}. \tag{13}$$

Proof. This essentially follows from the work of Palmowski and Rolski (2002). Using their terminology, their Proposition 3.2 implies α is a good function, so the RHS of Equation (13) is a martingale and we may define a measure $\tilde{\mathbb{P}}$ by

$$\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{Q}}(\omega) = \frac{\alpha(\omega_T, T)}{\alpha(\omega_0, 0)} \exp\Big\{ - \int_0^T \frac{\mathcal{L}\alpha(\omega_s, s)}{\alpha(\omega_s, s)} \, \mathrm{d}s \Big\}.$$

Under the measure $\tilde{\mathbb{P}}$, the canonical process $(\omega_t)_{t\in[0,T]}$ is still Markov. By the proof of their Theorem 4.2, we see that

$$\tilde{D}_t^f = f(Y_t, t) - \int_0^t \mathcal{M}f(Y_s, s)\mathrm{d}s$$

is a martingale for all sufficiently smooth functions f, implying that $\mathcal{M}$ is the generator of $(\omega_t)_{t\in[0,T]}$ under $\tilde{\mathbb{P}}$. It follows that $(\omega_t)_{t\in[0,T]}$ has the same law under $\tilde{\mathbb{P}}$ as $\overline{Z}$ does under $\mathbb{Q}$, which is sufficient to prove the result since $\overline{Y}$ and $\overline{Z}$ are Feller.

D. Proof from Section 3

We give the proofs of Lemma 1 and Theorem 2 from Section 3.

Lemma 1. *Let the generator $\mathcal{L}$ and the functions β and c be as above. Then, we have* $v^{-1}\beta\mathcal{L}(\beta^{-1}v) + \mathcal{L}\log\beta = \beta^{-1}\hat{\mathcal{L}}^*\beta + \hat{\mathcal{L}}\log\beta$.

Proof. Let us define $\hat{\mathcal{M}}$ to be the operator such that $\mathcal{M} = \hat{\mathcal{M}} + \partial_t$. Then, since $\hat{\mathcal{M}} + c = \mathcal{M} + c - \partial_t = \hat{\mathcal{K}}^*$, for any sufficiently rapidly decaying test function f we have

$$\langle \hat{\mathcal{M}}f, 1\rangle + \langle cf, 1\rangle = \langle \hat{\mathcal{K}}^* f, 1\rangle = \langle f, \hat{\mathcal{K}}1\rangle = 0,$$

so $\langle \hat{\mathcal{M}}f, 1\rangle = -\langle c, f\rangle$. Assumption 2, which states that $\beta^{-1}\mathcal{M}f = \mathcal{L}(\beta^{-1}f) - f\mathcal{L}(\beta^{-1})$ for all sufficiently rapidly decaying f, can be rearranged to $\hat{\mathcal{M}}f = \beta\hat{\mathcal{L}}(\beta^{-1}f) - \beta f\hat{\mathcal{L}}(\beta^{-1})$. So, it follows that

$$\begin{aligned}\langle c, f\rangle &= -\langle \beta\hat{\mathcal{L}}(\beta^{-1}f), 1\rangle + \langle \beta f\hat{\mathcal{L}}(\beta^{-1}), 1\rangle \\ &= -\langle f, \beta^{-1}\hat{\mathcal{L}}^*\beta\rangle + \langle f, \beta\hat{\mathcal{L}}(\beta^{-1})\rangle\end{aligned}$$

for any sufficiently rapidly decaying f. We conclude that $\beta^{-1}\hat{\mathcal{L}}^*\beta = \beta\hat{\mathcal{L}}(\beta^{-1}) - c$.

Next, using Assumption 2 with $f = v$ we can write

$$\begin{aligned}v^{-1}\beta\mathcal{L}(\beta^{-1}v) &= \beta\mathcal{L}(\beta^{-1}) + v^{-1}\mathcal{M}v \\ &= \beta\mathcal{L}(\beta^{-1}) - c \\ &= -\partial_t\log\beta + \beta\hat{\mathcal{L}}(\beta^{-1}) - c \\ &= -\partial_t\log\beta + \beta^{-1}\hat{\mathcal{L}}^*\beta.\end{aligned}$$

Finally, note that $-\mathcal{L}\log\beta + \hat{\mathcal{L}}\log\beta = -\partial_t \log\beta$. Combining this with the final line above, we get the desired result.

THEOREM 2. *With the above set-up, minimising the objective*

$$\mathcal{I}_{\mathrm{DSM}}(\beta) = \int_0^T \mathbb{E}_{q(\mathbf{x}_0,\mathbf{x}_t,\boldsymbol{\xi}_0)}\left[\frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)\right]\mathrm{d}t$$

is equivalent to maximising a lower bound on the expected model log-likelihood.

PROOF. Applying Theorem 1 to the generative process $X^{\boldsymbol{\xi}^*}$ conditioned on observation $\boldsymbol{\xi}^*$,

$$\log p_T(\mathbf{x}_0|\boldsymbol{\xi}^*) \geq \mathbb{E}_{\mathbb{Q}}\Big[\log p_0(Y_T)\Big|Y_0 = \mathbf{x}_0\Big] - \int_0^T \mathbb{E}_{\mathbb{Q}}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t,\boldsymbol{\xi}^*,t)}{\beta(\mathbf{x}_t,\boldsymbol{\xi}^*,t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t,\boldsymbol{\xi}^*,t)\;\middle|\; Y_0 = \mathbf{x}_0\right]\mathrm{d}t.$$

Replacing $\boldsymbol{\xi}^*$ by $\boldsymbol{\xi}_0$, letting $(\mathbf{x}_0,\boldsymbol{\xi}_0) \sim p_{\mathrm{data}}$ and taking expectations, we get

$$\mathbb{E}_{p_{\mathrm{data}}(\mathbf{x}_0,\boldsymbol{\xi}_0)}\left[\log p_T(\mathbf{x}_0|\boldsymbol{\xi}_0)\right] \geq \mathbb{E}_{q_T(\mathbf{x}_T)}\Big[\log p_0(\mathbf{x}_T)\Big] - \int_0^T \mathbb{E}_{q(\mathbf{x}_t,\boldsymbol{\xi}_0)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)}{\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)\right]\mathrm{d}t.$$

For any given $\boldsymbol{\xi}$, we have

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{x}_t|\boldsymbol{\xi})}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t,\boldsymbol{\xi},t)}{\beta(\mathbf{x}_t,\boldsymbol{\xi},t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t,\boldsymbol{\xi},t)\right]\\ &= \mathbb{E}_{q(\mathbf{x}_0,\mathbf{x}_t|\boldsymbol{\xi})}\left[\frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0,\boldsymbol{\xi})/\beta(\cdot,\boldsymbol{\xi},\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0,\boldsymbol{\xi})/\beta(\mathbf{x}_t,\boldsymbol{\xi},t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0,\boldsymbol{\xi})/\beta(\cdot,\boldsymbol{\xi},\cdot))(\mathbf{x}_t,t)\right] + const\end{aligned}$$

by the argument of Appendix E (see below), where the constant depends only on the dynamics of the forward process. Substituting $\boldsymbol{\xi}_0$ for $\boldsymbol{\xi}$ and taking expectations over $\boldsymbol{\xi}_0 \sim p_{\mathrm{data}}$, noting that $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0,\boldsymbol{\xi}_0) = q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$, we get

$$\begin{aligned}&\mathbb{E}_{q(\mathbf{x}_t,\boldsymbol{\xi}_0)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)}{\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)\right]\\ &= \mathbb{E}_{q(\mathbf{x}_0,\mathbf{x}_t,\boldsymbol{\xi}_0)}\left[\frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)\right] + const.\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{E}_{p_{\mathrm{data}}(\mathbf{x}_0,\boldsymbol{\xi}_0)}\left[\log p_T(\mathbf{x}_0|\boldsymbol{\xi}_0)\right] \geq \mathbb{E}_{q_T(\mathbf{x}_T)}\Big[\log p_0(\mathbf{x}_T)\Big]\\ &- \int_0^T \mathbb{E}_{q(\mathbf{x}_0,\mathbf{x}_t,\boldsymbol{\xi}_0)}\left[\frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)\right]\mathrm{d}t\\ &\qquad + const.\end{aligned}$$

The first term on the RHS and the constant are independent of the dynamics of the reverse process. Hence minimising

$$\mathcal{I}_{\mathrm{DSM}}(\beta) = \int_0^T \mathbb{E}_{q(\mathbf{x}_0,\mathbf{x}_t,\boldsymbol{\xi}_0)} \left[\frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t,\boldsymbol{\xi}_0,t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\boldsymbol{\xi}_0,\cdot))(\mathbf{x}_t,t) \right] \mathrm{d}t$$

is equivalent to maximising a lower bound on $\mathbb{E}_{p_{\mathrm{data}}(\mathbf{x}_0,\boldsymbol{\xi}_0)}\left[\log p_T(\mathbf{x}_0|\boldsymbol{\xi}_0)\right]$, which is the expected model log-likelihood.

E. Equivalence of generalised score matching objectives

First, we show that $\mathcal{I}_{\mathrm{ISM}}$ and $\mathcal{I}_{\mathrm{DSM}}$ are equivalent training objectives.

$$\begin{aligned}
&\mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)} \left[\frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t,t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\cdot))(\mathbf{x}_t,t) \right] \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} q_{0,t}(\mathbf{x}_0,\mathbf{x}_t) \left\{ \frac{\mathcal{L}(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)}{q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)/\beta(\mathbf{x}_t,t)} - \mathcal{L}\log(q_{\cdot|0}(\cdot|\mathbf{x}_0)/\beta(\cdot,\cdot))(\mathbf{x}_t,t) \right\} \mathrm{d}\nu(\mathbf{x}_0)\mathrm{d}\nu(\mathbf{x}_t) \\
&= \int_{\mathcal{X}} q_0(\mathbf{x}_0) \int_{\mathcal{X}} \left\{ \beta(\mathbf{x}_t,t)\hat{\mathcal{L}} \left(\frac{q_{t|0}(\cdot|\mathbf{x}_0)}{\beta(\cdot,t)} \right)(\mathbf{x}_t) - q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)\hat{\mathcal{L}}\log \left(\frac{q_{t|0}(\cdot|\mathbf{x}_0)}{\beta(\cdot,t)} \right)(\mathbf{x}_t) \right\} \mathrm{d}\nu(\mathbf{x}_t)\mathrm{d}\nu(\mathbf{x}_0) \\
&= \int_{\mathcal{X}} q_0(\mathbf{x}_0) \int_{\mathcal{X}} q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) \left\{ \frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t,t)}{\beta(\mathbf{x}_t,t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t,t) \right\} \mathrm{d}\nu(\mathbf{x}_t)\mathrm{d}\nu(\mathbf{x}_0) + const \\
&= \mathbb{E}_{q_t(\mathbf{x}_t)} \left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t)}{\beta(\mathbf{x}_t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t) \right] + const,
\end{aligned}$$

where the constants depend only on the dynamics of the forward process and so are fixed during training. Integrating from $t = 0$ to $t = T$, we conclude that $\mathcal{I}_{\mathrm{ISM}}$ and $\mathcal{I}_{\mathrm{DSM}}$ are equivalent.

There is also an *explicit score matching* form of the general DMM training objective as follows:

$$\mathcal{I}_{\mathrm{ESM}}(\beta) = \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)} \left[\frac{\mathcal{L}(q_{\cdot}(\cdot)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)}{q_t(\mathbf{x}_t)/\beta(\mathbf{x}_t,t)} - \mathcal{L}\log(q_{\cdot}(\cdot)/\beta(\cdot,\cdot))(\mathbf{x}_t,t) \right] \mathrm{d}t.$$

To see that this is equivalent to $\mathcal{I}_{\mathrm{ISM}}$ and $\mathcal{I}_{\mathrm{DSM}}$, observe

$$\begin{aligned}
&\mathbb{E}_{q_t(\mathbf{x}_t)}\left[\frac{\mathcal{L}(q_\cdot(\cdot)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)}{q_t(\mathbf{x}_t)/\beta(\mathbf{x}_t,t)} - \mathcal{L}\log(q_\cdot(\cdot)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)\right] \\
&= \int_{\mathcal{X}} q_t(\mathbf{x}_t)\left\{\frac{\mathcal{L}(q_\cdot(\cdot)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)}{q_t(\mathbf{x}_t)/\beta(\mathbf{x}_t,t)} - \mathcal{L}\log(q_\cdot(\cdot)/\beta(\cdot,\cdot))(\mathbf{x}_t,t)\right\}\mathrm{d}\nu(\mathbf{x}_t) \\
&= \int_{\mathcal{X}}\left\{\beta(\mathbf{x}_t,t)\hat{\mathcal{L}}\left(\frac{q_\cdot(\cdot)}{\beta(\cdot,\cdot)}\right) - q_t(\mathbf{x}_t)\hat{\mathcal{L}}\log\left(\frac{q_\cdot(\cdot)}{\beta(\cdot,\cdot)}\right)\right\}\mathrm{d}\nu(\mathbf{x}_t) \\
&= \int_{\mathcal{X}} q_t(\mathbf{x}_t)\left\{\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t,t)}{\beta(\mathbf{x}_t,t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t,t)\right\}\mathrm{d}\nu(\mathbf{x}_t) + const \\
&= \mathbb{E}_{q_t(\mathbf{x}_t)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_t)}{\beta(\mathbf{x}_t)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_t)\right] + const
\end{aligned}$$

and integrate from $t = 0$ to $t = T$.

F. Application to particular spaces

In this section, we show how our general framework can be applied in some particular cases of interest, namely to Euclidean diffusion processes, continuous-time Markov Chains on finite discrete state spaces, diffusions on Riemannian manifolds and the Wright–Fisher diffusion on the simplex.

A recurring theme we see in each example is that the default parameterisation given by our framework in terms of β is sub-optimal, either because we expect it to lead to numerical instabilities when optimising the training objective, or because it only captures a restricted subset of the class of reverse processes we are interested in. However, in each case it turns out to be possible to reparameterise the generative process in a way which captures a wider class of processes and lets us interpret the training objective on this wider class. This allows us to optimise our generative process over this wider class of processes. In addition this reparameterisation typically leads to a form of the objective that we expect to be more numerically stable in practice.

F.1. Real vector spaces

We show how our framework recovers the setup of Song et al. (2021), described in Section 2.1, in the case where $\mathcal{K}$ and $\mathcal{L}$ are the Euclidean diffusion processes given in Example 1. For convenience, we recall that X and Y satisfy the SDEs

$$\mathrm{d}X_t = \mu(X_t,t)\mathrm{d}t + \mathrm{d}\hat{B}_t, \qquad \mathrm{d}Y_t = b(Y_t,t)\mathrm{d}t + \mathrm{d}B_t, \tag{14}$$

respectively, and the corresponding generators are

$$\mathcal{K} = \partial_t + \mu\cdot\nabla + \frac{1}{2}\Delta, \qquad \mathcal{L} = \partial_t + b\cdot\nabla + \frac{1}{2}\Delta.$$

First, we check the assumptions made in Appendix B. If we let our reference measure ν be the Lebesgue measure, then Assumption 3 holds. Assumption 5 is satisfied whenever

b and μ are Lipschitz functions (Schilling and Partzsch, 2012, Corollaries 19.27 and 19.31), and Assumption 6 follows given the form of $\mathcal{K}$ above. For Assumption 7 we take $\mathcal{D}_0 = C_c^\infty(\mathbb{R}^d)$, the set of infinitely differentiable functions with compact support, and note that this is dense in $L^2(\mathcal{X}, \nu)$. Finally, we assume that the reverse process and p_0 are sufficiently regular that Assumptions 4, 8 and 9 hold.

Using integration by parts, we can calculate the adjoint of $\hat{\mathcal{K}}$. We have

$$\begin{aligned}\int f\hat{\mathcal{K}}h\mathrm{d}\nu &= \int f\left(\mu\cdot\nabla h + \frac{1}{2}\Delta h\right)\mathrm{d}\nu \\ &= -\int h\nabla\cdot(f\mu)\mathrm{d}\nu - \frac{1}{2}\int \nabla f\cdot\nabla h\mathrm{d}\nu \\ &= \int h\left(-\mu\cdot\nabla f - (\nabla\cdot\mu)f + \frac{1}{2}\Delta f\right)\mathrm{d}\nu,\end{aligned}$$

assuming f and h are sufficiently regular that all boundary terms are zero. Therefore,

$$\hat{\mathcal{K}}^* = -\mu\cdot\nabla - (\nabla\cdot\mu) + \frac{1}{2}\Delta.$$

We see that Assumption 1 holds if we let $c = -(\nabla\cdot\mu)$ and

$$\mathcal{M} = \partial_t - \mu\cdot\nabla + \frac{1}{2}\Delta,$$

noting that this is the generator of another diffusion process $\overline{Z}$ satisfying the SDE

$$\mathrm{d}Z_t = -\mu(Z_t, T-t)\mathrm{d}t + \mathrm{d}B'_t.$$

Given this form of $\mathcal{L}$ and $\mathcal{M}$, Assumption 2 then becomes

$$-\beta^{-1}\mu\cdot\nabla f + \frac{1}{2}\beta^{-1}\Delta f = b\cdot\nabla(\beta^{-1}f) + \frac{1}{2}\Delta(\beta^{-1}f) - fb\cdot\nabla(\beta^{-1}) - \frac{1}{2}f\Delta(\beta^{-1}),$$

which reduces to

$$\nabla\log\beta = \mu + b, \tag{15}$$

for some bounded measurable function β. This puts a restriction on the class of reverse processes $\mathcal{K}$ we may use; the condition that the drift μ must be expressible as $-b+\nabla\log\beta$ for some β is not automatically satisfied. However, the true time-reversal of the forward process will satisfy this property. In addition, we will show that we may reparameterise the training objective so that it can be interpreted for a broader class of reverse processes.

Assuming for the moment that Assumption 2 does hold, we can evaluate

$$\begin{aligned}\Phi(f) &= \frac{\mathcal{L}f}{f} - \mathcal{L}\log f \\ &= \frac{b\cdot\nabla f}{f} + \frac{1}{2}\frac{\Delta f}{f} - b\cdot\nabla\log f - \frac{1}{2}\Delta\log f \\ &= \frac{1}{2}\left\|\nabla\log f\right\|^2,\end{aligned}$$

and so the denoising score matching objective becomes

$$\mathcal{I}_{\mathrm{DSM}}(\beta) = \frac{1}{2}\int_0^T \mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)}\left[\left\|\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) - \nabla \log \beta(\mathbf{x}_t,t)\right\|^2\right] \mathrm{d}t. \tag{16}$$

Looking at Equations (15) and (16) suggests that it is more natural to parameterise the reverse process in terms of $s_\theta(\mathbf{x},t) = \nabla \log \beta(\mathbf{x},t)$ instead of $\beta(\mathbf{x},t)$. Making this substitution, the objective becomes

$$\mathcal{I}_{\mathrm{DSM}}(\theta) = \frac{1}{2}\int_0^T \mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)}\left[\left\|\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) - s_\theta(\mathbf{x}_t,t)\right\|^2\right] \mathrm{d}t,$$

recovering the objective of Song et al. (2021).

Parameterising in terms of $s_\theta(\mathbf{x},t)$ rather than $\beta(\mathbf{x},t)$ is preferable for a couple of reasons. First, $s_\theta(\mathbf{x},t)$ is targeting the score $\nabla \log q_t(\mathbf{x})$, while $\beta(\mathbf{x},t)$ is targeting $q_t(\mathbf{x})$, and we expect the former to typically be an easier target. Second, while Equation (16) only makes sense when the forward and backward processes are related via Assumption 2, the objective in Equation (3) is valid for any forward and backward diffusion processes as in Equation (14). Hence reparameterising allows us to capture a wider class of reverse processes in our optimisation.

F.2. Discrete state spaces

Next, we show how to apply our framework when X and Y are continuous-time Markov chains on a finite discrete state space as in Example 2. With a particular choice of parameterisation, we end up recovering the set-up of Campbell et al. (2022).

Recall that we start with $\mathcal{K} = \partial_t + A$ and $\mathcal{L} = \partial_t + B$, where A and B are the time-dependent generator matrices of X and Y respectively. From this it follows immediately that $\hat{\mathcal{K}}^* = A^T$. We will use the counting measure as our reference measure ν.

On a finite discrete space, all functions are bounded and have compact support, and $\mathcal{D}(\hat{\mathcal{K}}) = \mathcal{D}(\hat{\mathcal{L}}) = C_0(\mathcal{S})$ is the set of all functions on $\mathcal{X}$. Assumptions 3, 5, 6 and 7 follow immediately. In addition, we assume that the reverse process and p_0 are sufficiently regular that Assumptions 4, 8 and 9 always hold.

In order for Assumption 1 to hold, we need to find $\mathcal{M}$ and c such that $\mathcal{M}+c = \partial_t+\hat{\mathcal{K}}^*$ (viewed as operators). Since $\mathcal{M}$ should be the generator of another CTMC, we write $\mathcal{M} = \partial_t + D$ for some generator matrix D. We then require $D + c = A^T$, where c is viewed as a diagonal matrix and D must have zero row sums. This holds if and only if we take

$$c_{\mathbf{x}} = \sum_{\mathbf{y}\in\mathcal{X}} A_{\mathbf{yx}}, \qquad D_{\mathbf{xy}} = A_{\mathbf{yx}} - c_{\mathbf{x}}\mathbb{1}_{\mathbf{x}=\mathbf{y}}.$$

With this choice of $\mathcal{M}$, Assumption 2 becomes

$$\beta^{-1}(\mathbf{x},t)\sum_{\mathbf{z}\in\mathcal{X}} D_{\mathbf{xz}} f(\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{X}} B_{\mathbf{xz}}\, \beta^{-1}(\mathbf{z},t) f(\mathbf{z}) - f(\mathbf{x}) \sum_{\mathbf{z}\in\mathcal{X}} B_{\mathbf{xz}}\, \beta^{-1}(\mathbf{z},t)$$

for all $\mathbf{x} \in \mathcal{X}$. If we pick two distinct $\mathbf{x}, \mathbf{y}$ and set $f(\mathbf{z}) = \mathbb{1}_{\mathbf{z}=\mathbf{y}}$ in the above, we deduce

$$\beta^{-1}(\mathbf{x},t) D_{\mathbf{xy}} = \beta^{-1}(\mathbf{y},t) B_{\mathbf{xy}} \quad \text{for all } \mathbf{x} \neq \mathbf{y}.$$

Hence for Assumption 2 to hold, we require

$$A_{\mathbf{yx}} = \frac{\beta(\mathbf{x},t)}{\beta(\mathbf{y},t)} B_{\mathbf{xy}} \quad \text{for all } \mathbf{x} \neq \mathbf{y}. \tag{17}$$

An elementary check also shows that this condition is sufficient for Assumption 2 to hold for a given choice of β.

With this parameterisation, the implicit score matching objective becomes

$$\begin{aligned}\mathcal{I}_{\mathrm{ISM}}(\beta) &= \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)} \left[\frac{B^T \beta(\mathbf{x}_t)}{\beta(\mathbf{x}_t)} + B \log \beta(\mathbf{x}_t) \right] \mathrm{d}t \\ &= \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)} \left[\sum_{\mathbf{y} \in \mathcal{X}} \left\{ B_{\mathbf{y}\mathbf{x}_t} \frac{\beta(\mathbf{y},t)}{\beta(\mathbf{x}_t,t)} + B_{\mathbf{x}_t\mathbf{y}} \log \frac{\beta(\mathbf{y},t)}{\beta(\mathbf{x}_t,t)} \right\} \right] \mathrm{d}t.\end{aligned}$$

Unfortunately, fitting β directly using this objective is typically likely to perform poorly. This can be seen for a couple of reasons. Firstly, the optimal value of $\beta(\mathbf{x},t)$ is $q_t(\mathbf{x})$, and so learning $\beta(\mathbf{x},t)$ should be roughly as hard as targeting the marginals of the forward process directly. Secondly, the presence of β in the denominators can lead to numerical instabilities in regions where the forward process has low density.

Fortunately, we have at least a couple of methods for avoiding these problems available. The first is to find an equivalent formulation of the objective in terms of the generator of the reverse process, and then learn this generator using a denoising parameterisation. For $\mathbf{x} \neq \mathbf{y}$, we have

$$\begin{aligned} B_{\mathbf{xy}} \log \frac{\beta(\mathbf{y},t)}{\beta(\mathbf{x},t)} &= B_{\mathbf{xy}} \log \frac{B_{\mathbf{xy}}}{A_{\mathbf{yx}}} \\ &= -B_{\mathbf{xy}} \log A_{\mathbf{yx}} + const, \end{aligned}$$

where the constant depends only on the dynamics of the forward process, which are fixed. We can therefore write

$$\begin{aligned}\mathcal{I}_{\mathrm{ISM}}(A) &= \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)} \left[B_{\mathbf{x}_t\mathbf{x}_t} + \sum_{\mathbf{y} \neq \mathbf{x}_t} A_{\mathbf{x}_t\mathbf{y}} - \sum_{\mathbf{y} \neq \mathbf{x}_t} B_{\mathbf{x}_t\mathbf{y}} \log A_{\mathbf{y}\mathbf{x}_t} \right] \mathrm{d}t + const \\ &= \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)} \left[-A_{\mathbf{x}_t\mathbf{x}_t} - \sum_{\mathbf{y} \neq \mathbf{x}_t} B_{\mathbf{x}_t\mathbf{y}} \log A_{\mathbf{y}\mathbf{x}_t} \right] \mathrm{d}t + const,\end{aligned}$$

recovering the objective of Campbell et al. (2022). In addition, we can parameterise the reverse generator A via

$$A_{\mathbf{xy}}(\theta) = B_{\mathbf{yx}} \sum_{\mathbf{x}_0} \frac{q_{t|0}(\mathbf{y}|\mathbf{x}_0)}{q_{t|0}(\mathbf{x}|\mathbf{x}_0)} p_\theta^{(t)}(\mathbf{x}_0|\mathbf{x}_t) \quad \text{for } \mathbf{x} \neq \mathbf{y}, \tag{18}$$

where $p_\theta^{(t)}(\mathbf{x}_0|\mathbf{x}_t)$ is some learned estimate of the original datapoint $\mathbf{x}_0$ given the noised observation $\mathbf{x}_t$, and θ denotes the learnable parameters. This parameterisation should be

more stable, as it avoids potentially exploding denominators, and we expect predicting the original datapoint given the noised datapoint to be an easier goal than learning the marginals $q_t(\mathbf{x})$. See Campbell et al. (2022) for more details on this denoising parameterisation.

The second method is to reparameterise our objective in terms of the ratios $s_\theta(\mathbf{x}, \mathbf{y}, t) = \beta(\mathbf{y}, t)/\beta(\mathbf{x}, t)$. Doing this, the training objective becomes

$$\mathcal{I}_{\text{ISM}}(\theta) = \int_0^T \mathbb{E}_{q_t(\mathbf{x}_t)} \left[\sum_{\mathbf{y} \in \mathcal{X}} \left\{ B_{\mathbf{y}\mathbf{x}_t} s_\theta(\mathbf{x}_t, \mathbf{y}; t) - B_{\mathbf{x}_t \mathbf{y}} \log s_\theta(\mathbf{y}, \mathbf{x}_t; t) \right\} \right] \mathrm{d}t. \tag{19}$$

In addition, the generative process is now parameterised in terms of $s_\theta(\mathbf{x}, \mathbf{y}, t)$ via

$$A_{\mathbf{x}\mathbf{y}} = B_{\mathbf{y}\mathbf{x}} s_\theta(\mathbf{x}, \mathbf{y}; t) \quad \text{for } \mathbf{x} \neq \mathbf{y}. \tag{20}$$

Importantly, this objective matches the generalised objective from Section 3 when the noising and generative processes are related by Assumption 2, and is still minimised when $s_\theta(\mathbf{x}, \mathbf{y}; t) = q_t(\mathbf{y})/q_t(\mathbf{x})$.

This parameterisation is potentially beneficial for a couple of reasons. Firstly, by removing $\beta(\mathbf{x}, t)$ from the denominators, we expect that objective should be more numerically stable. Secondly, this parameterisation captures a wider class of potential reverse processes, since A is now given in terms of B via Equation (20), which is less restrictive than Equation (17).

As discussed further in Section 4, the integrand in Equation (19) may be viewed as a score matching objective for discrete state space. It shares certain similarities with ratio matching techniques (Hyvärinen, 2007), in particular targeting the ratios $\beta(\mathbf{y}, t)/\beta(\mathbf{x}, t)$. However, as far as we are aware this particular objective is not directly equivalent to any previously studied score matching objective in discrete state space (Hyvärinen, 2007; Lyu, 2009; Sohl-Dickstein et al., 2011).

F.3. Riemannian manifolds

Consider the case where $\mathcal{X}$ is a Riemannian manifold with metric tensor g and ν is the volume measure induced by g (so that Assumption 3 holds). A diffusion in $\mathcal{X}$ may be defined through its generator, so we let the noising and generative processes have generators

$$\mathcal{K} = \partial_t + \mu \cdot \nabla + \frac{1}{2}\Delta, \qquad \mathcal{L} = \partial_t + b \cdot \nabla + \frac{1}{2}\Delta.$$

respectively, where Δ is the Laplace-Beltrami operator defined in local coordinates by

$$\Delta f = \frac{1}{\sqrt{|g|}} \partial_i \big(\sqrt{|g|} g^{ij} \partial_j f \big)$$

and $|g|$ denotes the determinant of the metric tensor. For such processes, Assumption 5 is satisfied under mild regularity conditions on the manifold and the coefficients of the generators, as detailed by Molchanov (1968). As in the Euclidean diffusion case, Assumption 6 follows from the given form of $\mathcal{K}$, for Assumption 7 we may take $\mathcal{D}_0 = C_c^\infty(\mathcal{X})$ and note that this is dense in $L^2(\mathcal{X}, \nu)$ (Taylor, 2011, Section 4.4), and we

assume that the reverse process and p_0 are sufficiently regular that Assumptions 4, 8 and 9 hold.

To calculate the adjoint operator of $\hat{\mathcal{K}}$, we recall that the canonical volume element on $\mathcal{X}$ induced by g is given by

$$\mathrm{d}\omega = \sqrt{|g|}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n$$

and the divergence of a vector field $a : \mathcal{X} \to T\mathcal{X}$ on a Riemannian manifold is given by

$$\nabla \cdot a = \frac{1}{\sqrt{|g|}} \partial_i (a^i \sqrt{|g|}).$$

Then, using the generalised Stokes' Theorem, we have

$$\begin{aligned}
\langle f, \mu \cdot \nabla h \rangle &= \int_{\mathcal{X}} f \mu^i \, (\partial_i h) \sqrt{|g|}\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \\
&= -\int_{\mathcal{X}} h \, \partial_i (\mu^i f \sqrt{|g|})\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \\
&= \langle -(\nabla \cdot \mu) f - (\mu \cdot \nabla f), h \rangle,
\end{aligned}$$

where we assume f and h are sufficiently smooth that we may disregard boundary terms. In addition, we have

$$\begin{aligned}
\langle f, \Delta h \rangle &= \int_{\mathcal{X}} f \partial_i (\sqrt{|g|} g^{ij} \partial_j h)\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \\
&= -\int_{\mathcal{X}} \sqrt{|g|} g^{ij} (\partial_i f)(\partial_j h)\,\mathrm{d}x^1 \wedge \cdots \wedge \mathrm{d}x^n \\
&= \langle \Delta f, h \rangle.
\end{aligned}$$

We conclude that the adjoint operator is given by

$$\hat{\mathcal{K}}^* = -\mu \cdot \nabla - (\nabla \cdot \mu) + \frac{1}{2}\Delta.$$

Then, as in the Euclidean diffusion case we see that Assumption 1 holds if we let $c = -(\nabla \cdot \mu)$ and

$$\mathcal{M} = \partial_t - \mu \cdot \nabla + \frac{1}{2}\Delta,$$

noting that $\mathcal{M}$ is also the generator of a diffusion process Z on $\mathcal{X}$. We also find that Assumption 2 reduces to the condition $\nabla \log \beta = \mu + b$, as before.

Assuming this holds, we can evaluate

$$\begin{aligned}
\Phi(f) &= \frac{\mathcal{L} f}{f} - \mathcal{L} \log f \\
&= \frac{b \cdot \nabla f}{f} + \frac{1}{2}\frac{\Delta f}{f} - b \cdot \nabla \log f - \frac{1}{2}\Delta \log f \\
&= \frac{1}{2} \|\nabla \log f\|^2_{g(x)},
\end{aligned}$$

where $\|\cdot\|_{g(x)}$ denotes the norm on the tangent space $T_x\mathcal{X}$ induced by g.

Finally, as in the Euclidean diffusion we make a reparameterisation $s_\theta(\mathbf{x}, t) = \nabla \log \beta(\mathbf{x}, t)$ in order to sidestep Assumption 2 and provide an easier training target. The resulting denoising score matching objective is

$$\mathcal{I}_{\mathrm{DSM}}(\theta) = \frac{1}{2}\int_0^T \mathbb{E}_{q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)}\left[\left\|\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) - s_\theta(\mathbf{x}_t, t)\right\|^2_{g(\mathbf{x}_t)}\right] \mathrm{d}t,$$

which reproduces the result of De Bortoli et al. (2022) and Huang et al. (2022). Notably, we find that all the relevant formulae in the manifold case are essentially the same as in the Euclidean diffusion case, except for the inclusion of the metric tensor.

F.4. Wright–Fisher diffusions

Suppose we wish to approximate a distribution $p_{\mathrm{data}}(\cdot)$ over the space $\mathcal{X} = \mathcal{P}(E)$ of measures on a finite set $E = \{1, \ldots, N\}$. A natural class of stochastic processes on $\mathcal{X}$ are the Wright–Fisher diffusions, a model used in population genetics to describe the evolution of allele frequencies in a population over time (Ethier and Griffiths, 1993).

We can parameterise measures in $\mathcal{X}$ by tuples of real numbers $\mathbf{p} = (p_1, \ldots, p_N) \in [0,1]^N$ such that $\sum_{i=1}^N p_i = 1$. With this parameterisation, the Wright–Fisher diffusion has generator

$$\mathcal{L} = \partial_t + \frac{1}{2}\sum_{i,j=1}^N p_i\left(\delta_{ij} - p_j\right)\frac{\partial^2}{\partial p_i \partial p_j} + \sum_{i,j=1}^N q_{ij} p_i \frac{\partial}{\partial p_j},$$

with domain $\mathcal{D}(\mathcal{L}) = \{F|_{\mathcal{P}(E)} : F(p_1, \ldots, p_N) \in C^2(\mathbb{R}^N)\}$, where $(q_{ij})_{i,j=1,\ldots,N}$ is some matrix, potentially depending on $\mathbf{p}$ and t, such that $\sum_{j=1}^N q_{ij} = 0$ for each $i = 1, \ldots, N$.

If we take $q_{ij} = \frac{1}{2}\vartheta_j > 0$ for all $\mathbf{p} \in \mathcal{X}, t \in [0, T]$ and $i \neq j$, then this process is ergodic and its invariant distribution is Dirichlet(Θ), the Dirichlet distribution with parameters $\Theta = (\vartheta_1, \ldots, \vartheta_N)$ (Ethier and Griffiths, 1993). Moreover, the transition function of the process can be expressed as

$$P(t, \mathbf{p}, \cdot) = \sum_{n=0}^\infty d_n^\Theta(t) \sum_{\alpha \in (\mathbb{Z}_+^N) : |\alpha| = n} \binom{n}{\alpha} \prod_{i=1}^N p_i^{\alpha_i} \mathrm{Dirichlet}(\alpha + \Theta)(\cdot) \tag{21}$$

where $d_n^\Theta(t)$ are smooth functions of t given explicitly in Ethier and Griffiths (1993). It follows that if we take $\vartheta_j > 2$ for all j and we start the process in the interior of the simplex, then the process almost surely does not hit the boundary and the marginals of the forward process always vanish and have zero derivative at the boundary (since this holds for any Dirichlet distribution where all parameters are greater than 2).

Note that $\mathcal{X}$ is compact and hence locally compact and separable. Since we can view $\mathcal{X}$ as a subset of a linear subspace of $\mathbb{R}^N$, it also has a natural Lebesgue measure, which we take as the reference measure ν. Hence we satisfy Assumption 3.

We let our noising process have generator $\mathcal{L}$ as above and our generative process have generator

$$\mathcal{K} = \partial_t + \frac{1}{2}\sum_{i,j=1}^N p_i\left(\delta_{ij} - p_j\right)\frac{\partial^2}{\partial p_i \partial p_j} + \sum_{i,j=1}^N r_{ij} p_i \frac{\partial}{\partial p_j},$$

where $(r_{ij})_{i,j=1,\dots,N}$ is another matrix with zero row sums. The forward process Y is then Feller from Ethier and Kurtz (1993, Theorem 3.4). It follows that the extended forward process $\overline{Y}$ is also Feller and, since the process is pathwise continuous on a compact state space, this implies that the extended backward process $\overline{X}$ is also Feller, so Assumption 5 holds. Assumption 6 follows from the given form of $\mathcal{K}$, and for Assumption 7, we can take $\mathcal{D}_0 = \{F|_{\mathcal{P}(E)} : F(p_1, \dots, p_N) \in C^\infty(\mathbb{R}^N)\}$. As usual, we assume that the reverse process and p_0 are sufficiently regular that Assumptions 4, 8 and 9 hold.

In order to calculate the adjoint operator $\hat{\mathcal{K}}^*$, we require the following lemma, which is essentially a form of the integration by parts formula for the space $\mathcal{X}$.

LEMMA 5. *Suppose we have $F : \mathbb{R}^N \to \mathbb{R}^N$ such that for all $x \in \mathcal{X}$, $F(x) \cdot \mathbf{1} = 0$, where $\mathbf{1}$ is the unit vector in the $(1, \dots, 1)^T$ direction. In addition, suppose that $F(x) = 0$ for all $x \in \partial\mathcal{X}$. Then*

$$\int_{\mathcal{X}} \sum_{j=1}^{N} \frac{\partial F_j}{\partial p_j}(\mathbf{p}) \, \mathrm{d}\nu(\mathbf{p}) - \frac{1}{N} \int_{\mathcal{X}} \sum_{j,k=1}^{N} \frac{\partial F_k}{\partial p_j}(\mathbf{p}) \, \mathrm{d}\nu(\mathbf{p}) = 0.$$

PROOF. Since $F(x) \cdot \mathbf{1} = 0$ for all $x \in \mathcal{X}$, we can view F as a function from $\mathcal{X}$ to $T\mathcal{X}$, the tangent bundle of $\mathcal{X}$. Then, since $F(x) = 0$ for $x \in \partial\mathcal{X}$, by the generalised Stokes' theorem we have

$$\int_{\mathcal{X}} \nabla_{\mathcal{X}} \cdot F \, \mathrm{d}\nu = 0,$$

where $\nabla_{\mathcal{X}} \cdot F$ denotes the manifold divergence on $\mathcal{X}$. Finally, $\nabla_{\mathcal{X}} \cdot F = \nabla \cdot F - \mathbf{1} \cdot \nabla (F \cdot \mathbf{1})$, where ∇ is the standard gradient operator on $\mathbb{R}^N$, and so the result follows.

First, we need to calculate the adjoint of $\hat{\mathcal{K}}$. To deal with the first order term, we use Lemma 5 with $F_j = r_{ij} p_i f h$ for $i = 1, \dots, N$ in turn to get

$$\int_{\mathcal{X}} \sum_{j=1}^{N} \frac{\partial}{\partial p_j}(r_{ij} p_i f h) \, \mathrm{d}\nu(\mathbf{p}) - \frac{1}{N} \int_{\mathcal{X}} \sum_{j,k=1}^{N} \frac{\partial}{\partial p_j}(r_{ik} p_i f h) \, \mathrm{d}\nu(\mathbf{p}) = 0,$$

whenever $fh = 0$ on $\partial\mathcal{X}$. Since $\sum_{j=1}^{N} r_{ij} = 0$, the second term vanishes. Thus, summing over i we get

$$\int_{\mathcal{X}} \sum_{i,j=1}^{N} p_i f h \frac{\partial r_{ij}}{\partial p_j} \, \mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} \sum_{i=1}^{N} r_{ii} f h \, \mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} \sum_{i,j=1}^{N} r_{ij} p_i h \frac{\partial f}{\partial p_j} \, \mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} \sum_{i,j=1}^{N} r_{ij} p_i f \frac{\partial h}{\partial p_j} \, \mathrm{d}\nu(\mathbf{p}) = 0,$$

from which we deduce that

$$\int_{\mathcal{X}} f \left(\sum_{i,j=1}^{N} r_{ij} p_i \frac{\partial}{\partial p_j} \right) h \, \mathrm{d}\nu(\mathbf{p}) = - \int_{\mathcal{X}} h \left(\sum_{i,j=1}^{N} p_i \frac{\partial r_{ij}}{\partial p_j} + \sum_{i=1}^{N} r_{ii} + \sum_{i,j=1}^{N} r_{ij} p_i \frac{\partial}{\partial p_j} \right) f \, \mathrm{d}\nu(\mathbf{p}) \tag{22}$$

whenever $fh = 0$ on $\partial\mathcal{X}$.

To deal with the second order term, we use Lemma 5 with $F_j = p_i(\delta_{ij} - p_j)(f\partial_i h)$ for each $i = 1, \ldots, N$ in turn to get

$$\int_{\mathcal{X}} \sum_{j=1}^{N} \frac{\partial}{\partial p_j}(p_i(\delta_{ij} - p_j)(f\partial_i h))\, \mathrm{d}\nu(\mathbf{p}) - \frac{1}{N}\int_{\mathcal{X}} \sum_{k,j=1}^{N} \frac{\partial}{\partial p_j}(p_i(\delta_{ik} - p_k)(f\partial_i h))\, \mathrm{d}\nu(\mathbf{p}) = 0,$$

whenever $f\partial_i h = 0$ on $\partial\mathcal{X}$ for each $i = 1, \ldots, N$. Expanding the LHS, we get

$$\int_{\mathcal{X}} \sum_{j=1}^{N} (\delta_{ij} - p_i - \delta_{ij}p_i)(f\partial_i h)\mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} \sum_{j=1}^{N} p_i(\delta_{ij} - p_j)((\partial_j f)(\partial_i h) + f\partial_i\partial_j h)\mathrm{d}\nu(\mathbf{p})$$

$$- \frac{1}{N}\int_{\mathcal{X}} \sum_{j,k=1}^{N} (\delta_{ij}\delta_{ik} - \delta_{ij}p_k - \delta_{jk}p_i)(f\partial_i h)\mathrm{d}\nu(\mathbf{p}) - \frac{1}{N}\int_{\mathcal{X}} \sum_{j,k=1}^{N} p_i(\delta_{ik} - p_k)\partial_j(f\partial_i h)\mathrm{d}\nu(\mathbf{p}).$$

Now, the last term is zero since $\sum_{k=1}^{N} p_i(\delta_{ik} - p_k) = 0$. Simplifying and summing over i, we get

$$\int_{\mathcal{X}} \sum_{i=1}^{N} (1 - Np_i) f\partial_i h\, \mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} f \left(\sum_{i,j=1}^{N} p_i(\delta_{ij} - p_j) \frac{\partial^2}{\partial p_i \partial p_j} \right) h\, \mathrm{d}\nu(\mathbf{p})$$

$$+ \int_{\mathcal{X}} \sum_{i,j=1}^{N} p_i(\delta_{ij} - p_j)(\partial_j f)(\partial_i h)\, \mathrm{d}\nu(\mathbf{p}) = 0.$$

By symmetry, we may reverse the roles of f and h in this last equation and subtract the resulting equations to get

$$\int_{\mathcal{X}} \sum_{i=1}^{N} (1 - Np_i) f\partial_i h\, \mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} f \left(\sum_{i,j=1}^{N} p_i(\delta_{ij} - p_j) \frac{\partial^2}{\partial p_i \partial p_j} \right) h\, \mathrm{d}\nu(\mathbf{p})$$

$$= \int_{\mathcal{X}} \sum_{i=1}^{N} (1 - Np_i) h\partial_i f\, \mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} h \left(\sum_{i,j=1}^{N} p_i(\delta_{ij} - p_j) \frac{\partial^2}{\partial p_i \partial p_j} \right) f\, \mathrm{d}\nu(\mathbf{p}) \qquad (23)$$

whenever $f\nabla h = h\nabla f = 0$ on $\partial\mathcal{X}$. Finally, applying Lemma 5 with $F_i = fh(1 - Np_i)$, we get

$$\int_{\mathcal{X}} \sum_{j=1}^{N} \frac{\partial}{\partial p_j}(fh(1 - Np_j))\, \mathrm{d}\nu(\mathbf{p}) - \frac{1}{N}\int_{\mathcal{X}} \sum_{i,j=1}^{N} \frac{\partial}{\partial p_j}(fh(1 - Np_i))\, \mathrm{d}\nu(\mathbf{p}) = 0.$$

whenever $fh = 0$ on $\partial\mathcal{X}$. Expanding, we have

$$\int_{\mathcal{X}} \sum_{j=1}^{N} h(1 - Np_j)\partial_j f\, \mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} \sum_{j=1}^{N} f(1 - Np_j)\partial_j h\, \mathrm{d}\nu(\mathbf{p}) - N^2 \int_{\mathcal{X}} fh\mathrm{d}\nu(\mathbf{p})$$

$$- \frac{1}{N}\int_{\mathcal{X}} \sum_{i,j=1}^{N} (1 - Np_i)\partial_j(fh)\, \mathrm{d}\nu(\mathbf{p}) + N\int_{\mathcal{X}} fh\mathrm{d}\nu(\mathbf{p}) = 0,$$

which simplifies to

$$\int_{\mathcal{X}} \sum_{j=1}^{N} h(1-Np_j)\partial_j f \, \mathrm{d}\nu(\mathbf{p}) + \int_{\mathcal{X}} \sum_{j=1}^{N} f(1-Np_j)\partial_j h \, \mathrm{d}\nu(\mathbf{p}) - N(N-1)\int_{\mathcal{X}} fh\mathrm{d}\nu(\mathbf{p}) = 0. \tag{24}$$

Combining Equations (23) and (24), we see

$$\begin{aligned}\frac{1}{2}\int_{\mathcal{X}} f\left(\sum_{i,j=1}^{N} p_i(\delta_{ij}-p_j)\frac{\partial^2}{\partial p_i \partial p_j}\right) h \, \mathrm{d}\nu(\mathbf{p}) = \frac{1}{2}\int_{\mathcal{X}} h\left(\sum_{i,j=1}^{N} p_i(\delta_{ij}-p_j)\frac{\partial^2}{\partial p_i \partial p_j}\right) f \, \mathrm{d}\nu(\mathbf{p}) \\ + \int_{\mathcal{X}} h\left(\sum_{j=1}^{N}(1-Np_j)\frac{\partial}{\partial p_j}\right) f \, \mathrm{d}\nu(\mathbf{p}) \\ - \frac{N(N-1)}{2}\int_{\mathcal{X}} fh\mathrm{d}\nu(\mathbf{p}).\end{aligned} \tag{25}$$

Putting together Equations (22) and (25), we conclude that the operator

$$\begin{aligned}\mathcal{K}_0 = \frac{1}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij}-p_j\right)\frac{\partial^2}{\partial p_i \partial p_j} + \sum_{j=1}^{N}(1-Np_j)\frac{\partial}{\partial p_j} - \frac{N(N-1)}{2} \\ - \sum_{i,j=1}^{N} p_i\frac{\partial r_{ij}}{\partial p_j} - \sum_{i=1}^{N} r_{ii} - \sum_{i,j=1}^{N} r_{ij}p_i\frac{\partial}{\partial p_j}\end{aligned}$$

satisfies $\langle \mathcal{K}_0 f, h\rangle = \langle f, \mathcal{K}h\rangle$ for all functions $f, h \in \{F|_{\mathcal{P}(E)} : F(p_1,\ldots,p_N) \in C^2(\mathbb{R}^N)\}$ such that $fh = f\nabla h = h\nabla f = 0$ on $\partial\mathcal{X}$. We conclude that $\hat{\mathcal{K}}^* = \mathcal{K}_0$ and $h \in \mathcal{D}(\hat{\mathcal{K}}^*)$ for all h such that $h = \nabla h = 0$ on $\partial\mathcal{X}$. Therefore, we choose to define

$$\mathcal{M} = \partial_t + \frac{1}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij}-p_j\right)\frac{\partial^2}{\partial p_i \partial p_j} + \sum_{j=1}^{N}(1-Np_j)\frac{\partial}{\partial p_j} - \sum_{i,j=1}^{N} r_{ij}p_i\frac{\partial}{\partial p_j},$$

$$c = -\frac{N(N-1)}{2} - \sum_{i,j=1}^{N} p_i\frac{\partial r_{ij}}{\partial p_j} - \sum_{i=1}^{N} r_{ii}.$$

We see that Assumption 1 is satisfied, since v vanishes and has zero derivative on $\partial\mathcal{X}$ by our earlier remarks. Recalling that $q_{ij} = \frac{1}{2}\vartheta_j$ for $i \neq j$ and $\sum_{j=1}^{N} r_{ij} = 0$, if we let

$$u_{ij} = \vartheta_j + \frac{p_j}{p_i}(\vartheta_i - 1) - r_{ij}, \quad \text{for } i \neq j$$

and set $u_{ii} = -\sum_{j\neq i} u_{ij}$, then for each $j = 1, \dots, N$ we have

$$\begin{aligned}\sum_{i=1}^{N} u_{ij}p_i &= \sum_{i\neq j} u_{ij}p_i - \sum_{i\neq j} u_{ji}p_j \\ &= \sum_{i\neq j} (p_i\vartheta_j + p_j(\vartheta_i - 1)) - \sum_{i\neq j} (p_j\vartheta_i + p_i(\vartheta_j - 1)) - \sum_{i=1}^{N} r_{ij}p_i \\ &= 1 - Np_j - \sum_{i=1}^{N} r_{ij}p_i.\end{aligned}$$

We thus see that $\mathcal{M}$ is the generator of another Wright–Fisher process with transition matrix $(u_{ij})_{i,j=1,\dots,N}$. Hence Assumption 1 is satisfied. To check Assumption 2,

$$\begin{aligned}\beta\mathcal{L}(\beta^{-1}f) - \beta f\mathcal{L}(\beta^{-1}) &= \frac{\beta}{2}\sum_{i,j=1}^{N} p_i(\delta_{ij} - p_j)(\beta^{-1}(\partial_i\partial_j f) + 2(\partial_i f)(\partial_j\beta^{-1}) + f(\partial_i\partial_j\beta^{-1})) \\ &\quad + \beta\sum_{i,j=1}^{N} q_{ij}p_i(\beta^{-1}\partial_j f + f\partial_j\beta^{-1}) \\ &\quad - \frac{\beta f}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij} - p_j\right)\partial_i\partial_j\beta^{-1} - \beta f\sum_{i,j=1}^{N} q_{ij}p_i\partial_j\beta^{-1} \\ &= \frac{1}{2}\sum_{i,j=1}^{N} p_i(\delta_{ij} - p_j)(\partial_i\partial_j f) - \sum_{i,j=1}^{N} p_i(\delta_{ij} - p_j)(\partial_i f)(\partial_j\log\beta) \\ &\quad + \sum_{i,j=1}^{N} q_{ij}p_i(\partial_j f).\end{aligned}$$

Thus Assumption 2 holds if and only if

$$\sum_{i=1}^{N} u_{ij}p_i = -\sum_{i=1}^{N} p_i(\delta_{ij} - p_j)(\partial_i\log\beta) + \sum_{i=1}^{N} q_{ij}p_i$$

for each $j = 1, \dots, N$. This is satisfied if we take

$$r_{ij} = \vartheta_j + \frac{p_j}{p_i}(\vartheta_i - 1) - p_j\frac{\partial(\log\beta)}{\partial p_i} - q_{ij} \tag{26}$$

for $i \neq j$ and $r_{ii} = -\sum_{j\neq i} r_{ij}$. We choose this parameterisation since if we start the forward process in its invariant distribution and learn β so that the generative process is the exact time reversal of the forward process then $\beta(\mathbf{p}, t) \propto q_t(\mathbf{x}) \propto \prod_{i=1}^{N} p_i^{\vartheta_i - 1}$. In this case, Equation (26) reduces to $r_{ij} = \vartheta_j - q_{ij}$, so this parameterisation ensures that if we start the forward process in its invariant distribution and learn the reverse process perfectly then the transition matrix $(r_{ij})_{i,j=1,\dots,n}$ we learn is equal to the transition matrix $(q_{ij})_{i,j=1,\dots,n}$ of the forward process.

We can then calculate the score matching operator $\Phi(f) = f^{-1}\mathcal{L}f - \mathcal{L}\log f$,

$$\begin{aligned}\Phi(f) &= \frac{1}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij} - p_j\right) f^{-1}\frac{\partial^2 f}{\partial p_i \partial p_j} - \frac{1}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij} - p_j\right)\frac{\partial^2(\log f)}{\partial p_i \partial p_j} \\ &\qquad + \sum_{i,j=1}^{N} q_{ij}p_i f^{-1}\frac{\partial f}{\partial p_j} - \sum_{i,j=1}^{N} q_{ij}p_i\frac{\partial(\log f)}{\partial p_j} \\ &= \frac{1}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij} - p_j\right)\frac{\partial(\log f)}{\partial p_i}\frac{\partial(\log f)}{\partial p_j}.\end{aligned}$$

However, since we do not have access to the analytic forms of the transition kernel $q_{t|0}(\mathbf{p}_t|\mathbf{p}_0)$ for this model, we must fit β using the implicit score matching objective. We thus calculate

$$\begin{aligned}\frac{\hat{\mathcal{L}}^*\beta}{\beta} + \hat{\mathcal{L}}\log\beta &= \frac{1}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij} - p_j\right)\beta^{-1}\frac{\partial^2\beta}{\partial p_i \partial p_j} + \sum_{j=1}^{N}(1 - Np_j)\beta^{-1}\frac{\partial\beta}{\partial p_j} - \frac{N(N-1)}{2} \\ &\quad - \sum_{i,j=1}^{N} p_i\frac{\partial q_{ij}}{\partial p_j} - \sum_{i=1}^{N} q_{ii} - \sum_{i,j=1}^{N} q_{ij}p_i\beta^{-1}\frac{\partial\beta}{\partial p_j} \\ &\quad + \frac{1}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij} - p_j\right)\frac{\partial^2(\log\beta)}{\partial p_i \partial p_j} + \sum_{i,j=1}^{N} q_{ij}p_i\frac{\partial(\log\beta)}{\partial p_j} \\ &= \sum_{i,j=1}^{N} p_i\left(\delta_{ij} - p_j\right)\frac{\partial^2(\log\beta)}{\partial p_i \partial p_j} + \frac{1}{2}\sum_{i,j=1}^{N} p_i\left(\delta_{ij} - p_j\right)\frac{\partial(\log\beta)}{\partial p_i}\frac{(\partial\log\beta)}{\partial p_j} \\ &\quad + \sum_{j=1}^{N}(1 - Np_j)\frac{\partial(\log\beta)}{\partial p_j} + const,\end{aligned}$$

where we have discarded terms that do not depend on β. Noting that the loss and the reverse process only depend on β through $\partial(\log\beta)/\partial p_j$, we reparameterise in terms of $s^i_\theta(\mathbf{p}, t) = p_i\partial(\log\beta(\mathbf{p}, t))/\partial p_i$. (We include the extra factor of p_i for numerical stability reasons, since if we start in the stationary distribution then $p_i\partial(\log\beta(\mathbf{p}, t))/\partial p_i$ should be of constant scale.) Doing this, the implicit score matching objective becomes

$$\begin{aligned}\mathcal{I}_{\mathrm{ISM}}(\theta) = \int_0^T \mathbb{E}_{q_t(\mathbf{p}_t)}\Bigg[\sum_{i,j=1}^{N}\left(\delta_{ij} - p_j\right)\frac{\partial s^i_\theta(\mathbf{p}_t, t)}{\partial p_j} + \frac{1}{2}\sum_{i,j=1}^{N}(p_j^{-1}\delta_{ij} - 1)s^i_\theta(\mathbf{p}_t, t)s^j_\theta(\mathbf{p}_t, t) \\ + (1 - N)\sum_{j=1}^{N} s^j_\theta(\mathbf{p}_t, t)\Bigg]\mathrm{d}t, \quad (27)\end{aligned}$$

and the reverse process is parameterised as the Wright–Fisher diffusion with transition matrix $(r_{ij})_{i,j=1,\ldots,N}$ where

$$r_{ij}(\mathbf{p}, T - t) = \frac{1}{2}\vartheta_j + \frac{p_j}{p_i}(\theta_i - 1) - \frac{p_j}{p_i}s^i_\theta(\mathbf{p}, t), \quad i \neq j.$$

G. Proof of properties of the score matching operator

We give the proof of the properties of the score matching operator from Proposition 1.

PROPOSITION 1. *Let Y be a Feller process with semigroup operators $(Q_t)_{t\geq 0}$, generator $\mathcal{L}$ and associated score matching operator Φ. Then:*

(a) $\Phi(f) \geq 0$ for all f in the domain of Φ, with equality if f is constant;
(b) for any probability measures π_1, π_2 on $\mathcal{X}$ and $t \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\pi_1 Q_t||\pi_2 Q_t) = -\mathbb{E}_{\pi_1 Q_t}\left[\Phi\left(\frac{\mathrm{d}(\pi_1 Q_t)}{\mathrm{d}(\pi_2 Q_t)}\right)\right],$$

where $\mathrm{KL}(\pi_1 Q_t||\pi_2 Q_t)$ denotes the Kullback–Leibler divergence between $\pi_1 Q_t$, $\pi_2 Q_t$.

PROOF. Since log is a concave function, it follows that $\log(Q_t f) \geq Q_t(\log f)$ for all f in the domain of Φ with equality if f is constant. Hence

$$\frac{\log(Q_t f) - \log f}{t} \geq \frac{Q_t(\log f) - \log f}{t}$$

for all $t \geq 0$. Taking the limit $t \downarrow 0$, we deduce that $(\mathcal{L}f)/f \geq \mathcal{L}(\log f)$ which gives the first part of the lemma.

For the second part, we assume that $\pi_1 Q_t$ and $\pi_2 Q_t$ are absolutely continuous with respect to ν and let $\pi_{1,t}(\mathbf{x})$ and $\pi_{2,t}(\mathbf{x})$ respectively denote their densities. Then

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\pi_1 Q_t||\pi_2 Q_t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int \pi_{1,t}(\mathbf{x}) \log\left(\frac{\pi_{1,t}(\mathbf{x})}{\pi_{2,t}(\mathbf{x})}\right) \mathrm{d}\nu(\mathbf{x}) \\
&= \int \partial_t \pi_{1,t}(\mathbf{x}) \log\left(\frac{\pi_{1,t}(\mathbf{x})}{\pi_{2,t}(\mathbf{x})}\right) \mathrm{d}\nu(\mathbf{x}) + \int \pi_{1,t}(\mathbf{x}) \partial_t \log \pi_{1,t}(\mathbf{x}) \mathrm{d}\nu(\mathbf{x}) \\
&\quad - \int \pi_{1,t}(\mathbf{x}) \partial_t \log \pi_{2,t}(\mathbf{x}) \mathrm{d}\nu(\mathbf{x}) \\
&= \int \hat{\mathcal{L}}^* \pi_{1,t}(\mathbf{x}) \log\left(\frac{\pi_{1,t}(\mathbf{x})}{\pi_{2,t}(\mathbf{x})}\right) \mathrm{d}\nu(\mathbf{x}) + \int \hat{\mathcal{L}}^* \pi_{1,t}(\mathbf{x}) \mathrm{d}\nu(\mathbf{x}) \\
&\quad - \int \hat{\mathcal{L}}^* \pi_{2,t}(\mathbf{x}) \left(\frac{\pi_{1,t}(\mathbf{x})}{\pi_{2,t}(\mathbf{x})}\right) \mathrm{d}\nu(\mathbf{x})
\end{aligned}$$

using the Fokker–Planck equation on each term. Since $\langle \hat{\mathcal{L}}^* \pi_{1,t}, 1\rangle = \langle \pi_{1,t}, \hat{\mathcal{L}} 1\rangle = 0$, we may drop the second term and write

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\pi_1 Q_t||\pi_2 Q_t) &= \mathbb{E}_{\pi_{1,t}(\mathbf{x})}\left[\hat{\mathcal{L}} \log\left(\frac{\pi_{1,t}(\mathbf{x})}{\pi_{2,t}(\mathbf{x})}\right)\right] - \mathbb{E}_{\pi_{1,t}(\mathbf{x})}\left[\left(\frac{\pi_{2,t}(\mathbf{x})}{\pi_{1,t}(\mathbf{x})}\right) \hat{\mathcal{L}}\left(\frac{\pi_{1,t}(\mathbf{x})}{\pi_{2,t}(\mathbf{x})}\right)\right] \\
&= -\mathbb{E}_{\pi_{1,t}(\mathbf{x})}\left[\Phi\left(\frac{\pi_{1,t}(\mathbf{x})}{\pi_{2,t}(\mathbf{x})}\right)\right],
\end{aligned}$$

which is the desired result.

H. Discrete-time approximation proofs

In this section, we give the proofs of Lemma 2 and Theorem 3 from Section 5. In order to prove Lemma 2, we use a couple of lemmas which we present first.

LEMMA 6. *Given processes $\overline{X}$ and $\overline{Y}$ as in Section 3, define a process $\overline{W}$ by setting $\overline{W}_t = (X_{T-t}, t)$ and denote its generator by $\mathcal{N}$. Then we have*

$$v\mathcal{N}g = (\mathcal{M} + c)(vg)$$

for all sufficiently rapidly decaying functions g.

PROOF. First, we let $\overleftarrow{\mathcal{K}} = -\partial_t + \hat{\mathcal{N}}$ denote the generator of the time-reversal of $\overline{X}$. Then, the integration by parts formula of Cattiaux et al. (2023) implies that for all sufficiently rapidly decaying test functions f and g we have

$$\langle p_t f, \mathcal{K}g + \overleftarrow{\mathcal{K}}g\rangle + \langle p_t, \Gamma(f,g)\rangle = 0,$$

where $\Gamma(f,g) = \mathcal{K}(fg) - f\mathcal{K}g - g\mathcal{K}f$ denotes the carré du champ operator associated to $\mathcal{K}$. We deduce that

$$\begin{aligned}
\langle f, p_t\overleftarrow{\mathcal{K}}g\rangle &= -\langle p_t f, \mathcal{K}g\rangle - \langle p_t, \mathcal{K}(fg) - f\mathcal{K}g - g\mathcal{K}f\rangle \\
&= -\langle p_t f, \partial_t g\rangle - \langle \hat{\mathcal{K}}^* p_t, fg\rangle + \langle \hat{\mathcal{K}}^*(gp_t), f\rangle \\
&= -\langle p_t \partial_t g, f\rangle - \langle g\partial_t p_t, f\rangle + \langle \hat{\mathcal{K}}^*(gp_t), f\rangle \\
&= \langle \hat{\mathcal{K}}^*(gp_t) - \partial_t(p_t g), f\rangle
\end{aligned}$$

where in the third line we have used the Fokker–Planck equation. Since f was arbitrary, it follows that

$$p_t\overleftarrow{\mathcal{K}}g = \hat{\mathcal{K}}^*(gp_t) - \partial_t(p_t g).$$

Finally if we substitute $t \mapsto T - t$ in this final equation, we get

$$v\mathcal{N}g = (\hat{\mathcal{K}}^* + \partial_t)(vg),$$

which gives the desired result when combined with the definition of $\mathcal{M}$ and c from Assumption 1.

LEMMA 7. *Suppose $\beta : \mathcal{S} \to (0,\infty)$ is a function such that Assumption 2 holds. If we define $\zeta = \beta^{-1}v$, then for any function f decaying sufficiently rapidly, ζ satisfies*

$$\zeta\mathcal{N}f = \mathcal{L}(f\zeta) - f\mathcal{L}\zeta.$$

PROOF. For any sufficiently rapidly decaying f satisfying $f\zeta \in \mathcal{D}(\mathcal{L})$ and $vf \in \mathcal{D}(\mathcal{M})$, using Lemma 6 we have

$$\begin{aligned}
\mathcal{L}(f\zeta) - f\mathcal{L}\zeta &= \mathcal{L}(v\beta^{-1}f) - f\mathcal{L}(\beta^{-1}v) \\
&= \beta^{-1}\mathcal{M}(vf) - \beta^{-1}f\mathcal{M}v \\
&= \beta^{-1}v\mathcal{N}f - c\beta^{-1}vf + c\beta^{-1}vf \\
&= \zeta\mathcal{N}f.
\end{aligned}$$

Now we can give the proofs of Lemma 2 and Theorem 3.

LEMMA 2. *Suppose X, Y are fixed generative and noising processes with marginals p, q as in Section 3, and suppose that they are related as in Assumptions 1 and 2 for some sufficiently regular function β. Then for any $0 < s < t < T$ with $\gamma = t - s$,*

$$\gamma\, \mathbb{E}_{q_s(\mathbf{x}_s)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_s)}{\beta(\mathbf{x}_s)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_s)\right] = \mathbb{E}_{q_{s,t}(\mathbf{x}_s,\mathbf{x}_t)}\left[\log\frac{q_{t|s}(\mathbf{x}_t|\mathbf{x}_s)}{p_{T-s|T-t}(\mathbf{x}_s|\mathbf{x}_t)}\right] + o(\gamma).$$

PROOF. Let $\mathbb{P}^{\mathbf{x}_s}$ and $\mathbb{Q}^{\mathbf{x}_s}$ denote the path measures of $\overline{W}$ and $\overline{Y}$ respectively on the interval $[s,t]$ when we condition on the initial value $\mathbf{x}_s$. Assuming β is sufficiently regular so that ζ is bounded away from zero and infinity and $\zeta^{-1}\mathcal{L}\zeta$ is bounded and continuous in the time variable, by Girsanov's theorem and Lemma 7 we have

$$\frac{\mathrm{d}\mathbb{P}^{\mathbf{x}_s}}{\mathrm{d}\mathbb{Q}^{\mathbf{x}_s}}(\omega) = \frac{\zeta(\omega_t,t)}{\zeta(\omega_s,s)}\exp\left\{-\int_s^t \frac{\mathcal{L}\zeta(\omega_\tau,\tau)}{\zeta(\omega_\tau,\tau)}\mathrm{d}\tau\right\}.$$

Taking logarithms and writing $\gamma = t - s$, to first order in γ for any fixed path ω this becomes

$$\log\frac{\mathrm{d}\mathbb{P}^{\mathbf{x}_s}}{\mathrm{d}\mathbb{Q}^{\mathbf{x}_s}}(\omega) = \log\frac{\zeta(\omega_t,t)}{\zeta(\omega_s,s)} - \gamma\frac{\mathcal{L}\zeta(\omega_s,s)}{\zeta(\omega_s,s)} + o(\gamma).$$

Since the first order terms depend only on the value of the path at its endpoints (ω_s,ω_t), we conclude that

$$\log\frac{q_{t|s}(\mathbf{x}_t|\mathbf{x}_s)}{p_{T-t|T-s}(\mathbf{x}_t|\mathbf{x}_s)} = -\log\frac{\zeta(\mathbf{x}_t,t)}{\zeta(\mathbf{x}_s,s)} + \gamma\frac{\mathcal{L}\zeta(\mathbf{x}_s,s)}{\zeta(\mathbf{x}_s,s)} + o(\gamma).$$

It follows that

$$\log\frac{q_{t|s}(\mathbf{x}_t|\mathbf{x}_s)}{p_{T-s|T-t}(\mathbf{x}_s|\mathbf{x}_t)} = \log\frac{v(\mathbf{x}_t,t)}{v(\mathbf{x}_s,s)} - \log\frac{\zeta(\mathbf{x}_t,t)}{\zeta(\mathbf{x}_s,s)} + \gamma\frac{\mathcal{L}\zeta(\mathbf{x}_s,s)}{\zeta(\mathbf{x}_s,s)} + o(\gamma).$$

Taking expectations and using the definition of the generator as a stochastic derivative, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\left[\log\frac{q_{t|s}(\mathbf{x}_t|\mathbf{x}_s)}{p_{T-s|T-t}(\mathbf{x}_s|\mathbf{x}_t)}\right] &= \gamma\mathbb{E}_{q_s(\mathbf{x}_s)}\left[\frac{\mathcal{L}\zeta(\mathbf{x}_s,s)}{\zeta(\mathbf{x}_s,s)} - \mathcal{L}\log\zeta(\mathbf{x}_s,s) + \mathcal{L}\log v(\mathbf{x}_s,s)\right] + o(\gamma)\\ &= \gamma\mathbb{E}_{q_s(\mathbf{x}_s)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_s,s)}{\beta(\mathbf{x}_s,s)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_s,s)\right] + o(\gamma),\end{aligned}$$

where in the final line we have used Lemma 1.

THEOREM 3. *For any DMM, the objective (11) for its natural discretisation is equivalent to the natural discretisation of $\mathcal{I}_{\mathrm{ISM}}$ to first order in $\overline{\gamma} = \max_{k=0,\dots,N-1}|t_{k+1} - t_k|$.*

PROOF. Given time steps $0 = t_0 < t_1 < \cdots < t_N = T$, define $\gamma_k = t_{k+1} - t_k$ for $k = 0,\dots,N-1$ and set $\overline{\gamma} = \max_{k=0,\dots,N-1}\gamma_k$. Then the natural discretisation of the objective $\mathcal{I}_{\mathrm{ISM}}(\beta)$ is given by

$$\sum_{k=0}^{N-1}\gamma_k\mathbb{E}_{q_{t_k}(\mathbf{x}_{t_k})}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_{t_k})}{\beta(\mathbf{x}_{t_k})} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_{t_k})\right].$$

Using Lemma 2 with $s = t_{k-1}$ and $t = t_k$ for $k = 1, \ldots, N$, we get

$$\begin{aligned}
&\mathbb{E}_{\tilde{q}(\mathbf{x}_{t_{k-1}})}\left[\mathrm{KL}(\tilde{q}(\mathbf{x}_{t_{k-1}}|\mathbf{x}_{t_k})||\tilde{p}_\theta(\mathbf{x}_{t_{k-1}}|\mathbf{x}_{t_k}))\right] \\
&= \mathbb{E}_{\tilde{q}(\mathbf{x}_{t_{k-1}},\mathbf{x}_{t_k})}\left[\log \frac{\tilde{q}(\mathbf{x}_{t_k}|\mathbf{x}_{t_{k-1}})}{\tilde{p}_\theta(\mathbf{x}_{t_{k-1}}|\mathbf{x}_{t_k})}\right] + const \\
&= \gamma_{k-1}\, \mathbb{E}_{q_{t_{k-1}}(\mathbf{x}_{t_{k-1}})}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_{t_k})}{\beta(\mathbf{x}_{t_k})} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_{t_k})\right] + o(\gamma_{k-1}) + const.
\end{aligned}$$

Putting this together, we see that

$$\mathrm{KL}(\tilde{q}(\mathbf{x}_{0:T})||\tilde{p}_\theta(\mathbf{x}_{0:T})) = \sum_{k=0}^{N-1} \gamma_k \mathbb{E}_{q_{t_k}(\mathbf{x}_{t_k})}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_{t_k})}{\beta(\mathbf{x}_{t_k})} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_{t_k})\right] + o(\overline{\gamma}) + const,$$

so objective (11) is equivalent to the natural discretisation of $\mathcal{I}_{\mathrm{ISM}}$ to first order in $\overline{\gamma}$.

I. General equivalence between denoising autoencoders and score matching

A denoising autoencoder takes a datapoint $\mathbf{x}_0$ drawn from a data distribution q_0, noises it according to some density $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ and then tries to reconstruct $\mathbf{x}_0$ given the noised observation $\mathbf{x}_\tau$ (Vincent et al., 2008). Traditionally, $q_\tau(\mathbf{x}_\tau|\mathbf{x}_0)$ is taken to be Gaussian with mean $\mathbf{x}_0$ and some standard deviation σ and we make a point estimate $f_\theta(\mathbf{x}_\tau)$ for $\mathbf{x}_0$ given $\mathbf{x}_\tau$. The parameters θ are learned by minimising the MSE error

$$\mathcal{J}_{\mathrm{DAE}}(\theta) = \mathbb{E}_{q_{0,\tau}(\mathbf{x}_0,\mathbf{x}_\tau)}\left[\|f_\theta(\mathbf{x}_\tau) - \mathbf{x}_0\|^2\right].$$

For a general denoising autoencoder on state space $\mathcal{X}$, we allow a probabilistic reconstruction $p^{(\theta)}_{0|\tau}(\mathbf{x}_0|\mathbf{x}_\tau)$ of $\mathbf{x}_0$ depending on a set of parameters θ, rather than a point estimate. We fit θ by minimising the objective

$$\mathcal{J}_{\mathrm{DAE}}(\theta) = \mathbb{E}_{q_{0,\tau}(\mathbf{x}_0,\mathbf{x}_\tau)}\left[-\log p^{(\theta)}_{0|\tau}(\mathbf{x}_0|\mathbf{x}_\tau)\right].$$

Note that this reduces to the MSE objective in the case where $\mathcal{X} = \mathbb{R}^d$ and $p^{(\theta)}_{0|\tau}(\mathbf{x}_0|\mathbf{x}_\tau)$ is Gaussian with mean $f_\theta(\mathbf{x}_\tau)$.

Suppose now that we have a generalised denoising autodencoder where the noising distribution $q_{0,\tau}(\mathbf{x}_0, \mathbf{x}_\tau)$ is given by the endpoints of a Markov process on $\mathcal{X}$ with generator $\mathcal{L}$ and the denoising distribution $p^{(\theta)}_{0|\tau}(\mathbf{x}_0|\mathbf{x}_\tau)$ is given by the endpoints of a Markov process on $\mathcal{X}$ with generator $\mathcal{K}$. Suppose further that we parameterise the denoising process $\mathcal{K}$ via some function $\beta(\mathbf{x}, t)$ according to Assumptions 1 and 2 as in Section 3. Then Lemma 2 implies that $\mathcal{J}_{\mathrm{DAE}}$ is equivalent to first order to the objective

$$\mathcal{J}_{\mathrm{ISM}}(\beta) = \mathbb{E}_{q_\tau(\mathbf{x}_\tau)}\left[\frac{\hat{\mathcal{L}}^*\beta(\mathbf{x}_\tau,\tau)}{\beta(\mathbf{x}_\tau,\tau)} + \hat{\mathcal{L}}\log\beta(\mathbf{x}_\tau,\tau)\right],$$

or alternatively to the corresponding generalised denoising score matching objective as in Section 4.

This generalises the result of Vincent (2011), which demonstrated an equivalence between denoising autoencoders and denoising score matching in the case of Gaussian noise on $\mathbb{R}^d$. Indeed, we recover their result by considering the case where $q_{\tau|0}(\mathbf{x}_\tau|\mathbf{x}_0)$ and $p^{(\theta)}_{0|\tau}(\mathbf{x}_0|\mathbf{x}_\tau)$ are Gaussian, noting that these distributions are naturally induced as the distributions of the endpoints of diffusion processes.

Our work extends this equivalence between denoising autoencoders and generalised score matching as described in Section 4 to arbitrary state spaces and noising/denoising distributions, provided that the noising and denoising distributions can be viewed as the marginals at the endpoints of Markov processes with known generators.

J. Experimental details

We give the details of our experimental set-up and results from Section 6. Code for all of our experiments can be found at `github.com/yuyang-shi/generalized-diffusion`.

J.1. Inference on $\mathbb{R}^d$ using diffusion processes

The g-and-k distribution with parameters (A, B, g, k) is defined via its quantile function

$$F^{-1}(q|A, B, g, k) = A + B\left[1 + 0.8\tanh\left(\frac{gz(q)}{2}\right)\right]\left(1 + z(q)^2\right)^k z(q),$$

where $z(q)$ denotes the qth quantile of the standard Gaussian distribution, and we require $B > 0$ and $k > -0.5$. The parameters A, B, g, k control the location, scale, skewness and kurtosis of the distribution respectively (Prangle, 2020). The prior on the parameters is uniform on $[0, 10]^4$. For the diffusion model, we centre and rescale each parameter linearly to $[-1, 1]$ in our implementation, and transform back to $[0, 10]$ for reporting.

As our noising process, we use the Ornstein–Uhlenbeck process $\mathrm{d}Y_t = -\frac{1}{2}Y_t\mathrm{d}t + \mathrm{d}B_t$. This has generator $\mathcal{L} = \partial_t - \frac{1}{2}\mathbf{x}\cdot\nabla + \frac{1}{2}\Delta$ and transition densities $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ which are Gaussian and available analytically. We can sample from the forward process at time t by sampling $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$ and then $\mathbf{x}_t \sim q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$. In practice, we apply a time-rescaling to the noising process following Song et al. (2021), in order to apply less noise at small times and move more quickly to the reference distribution at large times, by considering

$$\mathrm{d}Y_t = -\frac{1}{2}\beta(t)Y_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}B_t.$$

The β schedule is set to be linear and monotonically increasing, i.e.

$$\beta(t) = \beta_{\min} + (\beta_{\max} - \beta_{\min})t. \tag{28}$$

We set $\beta_{\min} = 0.001$ and $\beta_{\max}$ is selected using a grid search from $2, 4, 6, 8, 10$.

The reverse process is parameterised in terms of a conditional score network $s_\theta(\mathbf{x}_t, \boldsymbol{\xi}, t)$ using multilayer perceptrons (MLPs). We first encode $\mathbf{x}$ and $\boldsymbol{\xi}$ into 128-dimensional encodings using two separate MLPs with 3 layers and 512 hidden units in each layer. We then concatenate the two encodings as well as the time t and pass through another MLP with 3 layers and 512 hidden units in each layer. The total number of neural

network parameters is approximately 1.9M. For $N = 250$, we take in $\boldsymbol{\xi}$ the full set of order statistics as inputs to our network, i.e. we sort the observation $\boldsymbol{\xi}$ and take all $n = 250$ values. For $N = 10000$, we take $n = 100$ evenly-spaced order statistics from our observation as inputs, following Fearnhead and Prangle (2012).

Since we have access to the analytic transition densities, we train using the denoising score matching objective $\mathcal{I}_{\mathrm{DSM}}(\theta)$. We use a total of 10^6 training samples $(\mathbf{x}_0, \boldsymbol{\xi}_0) \sim p_{\mathrm{data}}$ during training. We optimise the network using the Adam optimiser with batch size 512 and learning rate 0.0001 with a cosine annealing schedule for 2.5M iterations. For sampling, we use the Euler-Maruyama method with 1000 steps to simulate from the reverse SDE.

The ground truth posterior density is estimated with MCMC samples generated using the R package `gk` (Prangle, 2020). We compare our method with the semi-automatic ABC (SA-ABC) and Wasserstein SMC (W-SMC) methodologies using the R packages `abctools` (Nunes and Prangle, 2015) and `winference` (Bernton et al., 2019), as well as with Sequential Neural Posterior (Greenberg et al., 2019), Likelihood (Papamakarios et al., 2019) and Ratio Estimation (Durkan et al., 2020) approaches (SNPE, SNLE and SNRE) using the `sbi` Python package (Tejero-Cantero et al., 2020). All methods are set to use 10^6 data samples to generate 5000 posterior samples. We note that the default configurations offered by the `sbi` package for SNPE, SNLE and SNRE use comparatively smaller neural networks compared to our choice of score network $s_\theta(\mathbf{x}_t, \boldsymbol{\xi}, t)$ detailed above. We have correspondingly increased the size of the neural networks for the three methods to approximately the same number of parameters. We also use Neural Spline Flows (NSFs, Durkan et al. (2019)) for SNPE as it is reported to have superior performance (Lueckmann et al., 2021). Other settings are kept to the default values.

Compared to SA-ABC and W-SMC methodologies, neural-network based approaches including our DMM model require fitting a neural network and therefore are more computationally expensive at training time. However, our model is able to produce more accurate posterior estimates for fixed $\boldsymbol{\xi}_0$, and perform amortised inference across a range of parameter values using the same number of 10^6 data samples. Therefore, it is comparatively more data-efficient.

As well as the plots in the main text, we also provide a pair plot comparing the approximate posterior from our diffusion model to the ground truth joint distribution in Fig. 8. We see that our model provides results very close to the ground truth for the parameters A, B and g and can model the dependency between parameters, but gives a wider estimate in its reproduction of the posterior over k.

J.2. MNIST digit image inpainting using discrete-space CTMCs

Our implementation in discrete space closely follows that of Campbell et al. (2022), and we refer to their paper for further details. We denote our states as $\mathbf{x}_0 = (\mathbf{x}_0^1 \ldots, \mathbf{x}_0^D)$ and for our noising process we use a CTMC with generator matrix $B := B^{1:D}(\mathbf{x}^{1:D}, \mathbf{y}^{1:D})$ which factorises over the dimensions, so $B^{1:D}(\mathbf{x}^{1:D}, \mathbf{y}^{1:D}) = \sum_{i=1}^{D} \tilde{B}(\mathbf{x}^i, \mathbf{y}^i) \mathbb{1}_{\mathbf{x}^{1:D\setminus i} = \mathbf{y}^{1:D\setminus i}}$ for some rate matrix $\tilde{B}$ acting on a single dimension. Thus each pixel evolves independently as a CTMC on $\{0, \ldots, 255\}$ with rate matrix $\tilde{B}$. We use the Gaussian rate matrix of Campbell et al. (2022) for $\tilde{B}$, which respects the ordinal structure of our state space

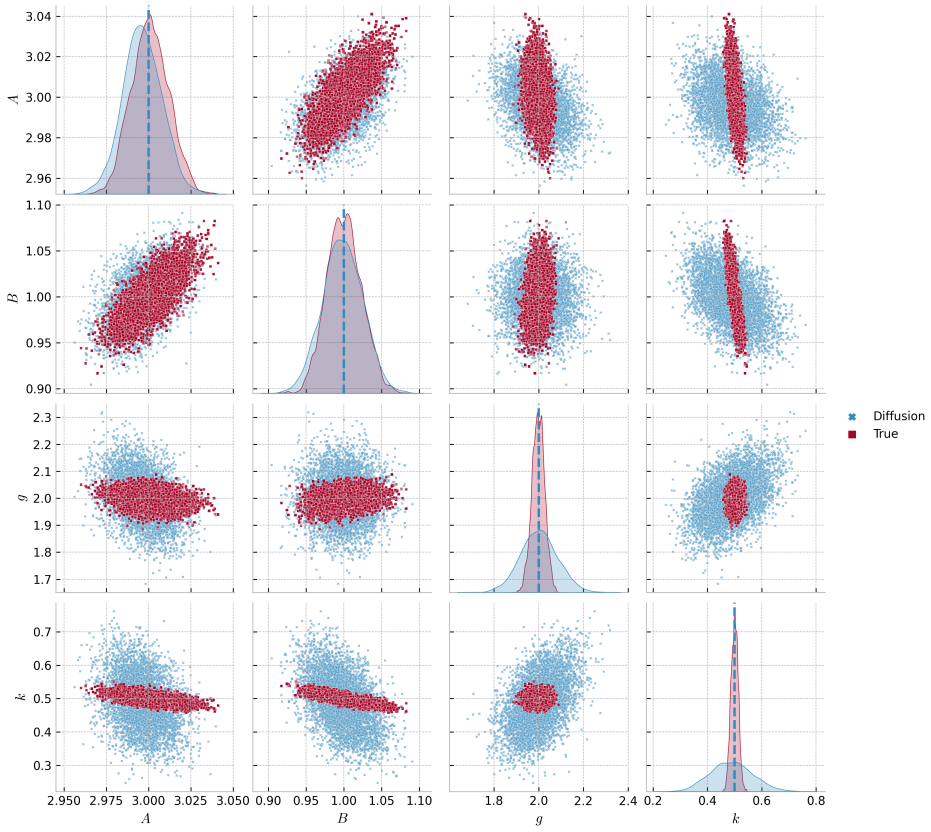

Fig. 8. Pair plots of the simulated posterior samples from the diffusion model and the ground truth distribution using MCMC for the g-and-k distribution example, with $\mathbf{x}_{\text{true}} = (3, 1, 2, 0.5)$ and $N = 10000$. The off-diagonal plots are the pairwise scatter plots between each component of $\mathbf{x}$, and the diagonal plots reproduce each parameter's marginal kernel density estimate.

and has a discretised Gaussian as its invariant distribution. The transition probabilities for this forward process can be calculated analytically efficiently by diagonalising the matrix and using matrix exponentials. This allows us to sample directly from the forward process at time t.

Since we have access to the forward transition probabilities, we use the denoising parameterisation of the reverse process in terms of $p_\theta^{(t)}(\mathbf{x}_0|\mathbf{x}_t)$ given in Equation (18), which we expect to lead to more stable training. We parameterise $p_\theta^{(t)}(\mathbf{x}_0|\mathbf{x}_t, \boldsymbol{\xi})$ using a convolutional U-net (Ho et al., 2020), taking as inputs both $\mathbf{x}_t$ and $\boldsymbol{\xi}$ (concatenated in the channel dimension), as well as a sinusoidal embedding of the time t. The total number of neural network parameters is approximately 6.1M. The output of the network is defined as the mean and log scale of a logistic distribution for each pixel. The logistic distribution is then discretised into bins $\{0, \ldots, 255\}$, and $p_\theta^{(t)}(\mathbf{x}_0|\mathbf{x}_t, \boldsymbol{\xi})$ is defined as the product of the discretised logistic distributions across dimensions.

We used the MNIST dataset (LeCun et al., 2010) which consists of images of hand-

Table 1. PSNR and SSIM scores for MNIST 14x14 inpainting using discrete-space and continuous-space DMMs. Higher values denote better performance.

	Discrete-space	Continuous-space (raw)	Continuous-space (rounded)
PSNR	16.63	16.72	**16.75**
SSIM	**0.757**	0.706	0.723

written digits. To train our model, we minimise the objective given in Example 6. For optimisation, we use the Adam optimiser with batch size 128 and learning rate 0.0002 for 1M iterations. In order to simulate the reverse process efficiently, we use a tau-leaping approximation with 1000 steps (for more details see Campbell et al. (2022)).

We compare our method to a continuous state space approach, as used for example in Song et al. (2021) and presented in Appendix F.1. We first normalize the data to range $[-1, 1]$, and then learn a continuous-space diffusion model with an Ornstein—Uhlenbeck noising process. All training configurations are kept the same as the discrete-space DMM. We report the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) for both methods in Table 1. PSNR and SSIM are two image quality metrics which measure the similarity between the generated posterior image and the ground truth. PSNR measures the pixel-by-pixel difference between two images and is a direct transformation of the mean squared error (MSE), whereas SSIM is a structural and more perceptional metric based on luminance, contrast and structure. For the continuous-space diffusion model, we report values for both the raw output samples (rescaled back to original scale), as well as with a further rounding step to the nearest integer in $\{0, \dots, 255\}$. The discrete-space and continuous-space models appear to achieve comparable results, with the discrete-space model having a slightly worse PSNR score, but slightly better SSIM score, suggesting comparable perceptual quality.

J.3. Large-scale image super-resolution using discrete-space CTMCs

We perform an additional experiment using discrete-space DMMs for a large-scale image inverse problem on the ImageNet dataset (Russakovsky et al., 2015). We train a DMM using CTMC noising and generative processes to perform 4-fold image super-resolution.

Each input image has 64×64 pixels and three RGB colour channels, and we aim to output images at the higher resolution of 256×256 pixels which are consistent with the input images. Our state space $\mathcal{X} = \{0, \dots, 255\}^{3 \times 256 \times 256}$.

The noising process, reverse process parameterisation, and neural network design are the same as in Section J.2, but we use a larger neural network for this task. As the starting point of our network optimisation, we utilise the pretrained network weights for continuous diffusions by Dhariwal and Nichol (2021), but we retrain the network for our discrete-space DMM using the objective in Example 6. The total number of neural network parameters is approximately 311.8M. We train the network using the Adam optimiser with batch size 4 and learning rate 2×10^{-5} for an additional 200000 iterations. For sampling, we use tau-leaping with 1000 steps.

We plot the simulated super-resolution samples in Fig. 9 for a number of low-resolution images generated from the ImageNet validation dataset. As shown in the images, the discrete diffusion model outputs different super-resolution samples that are realistic to

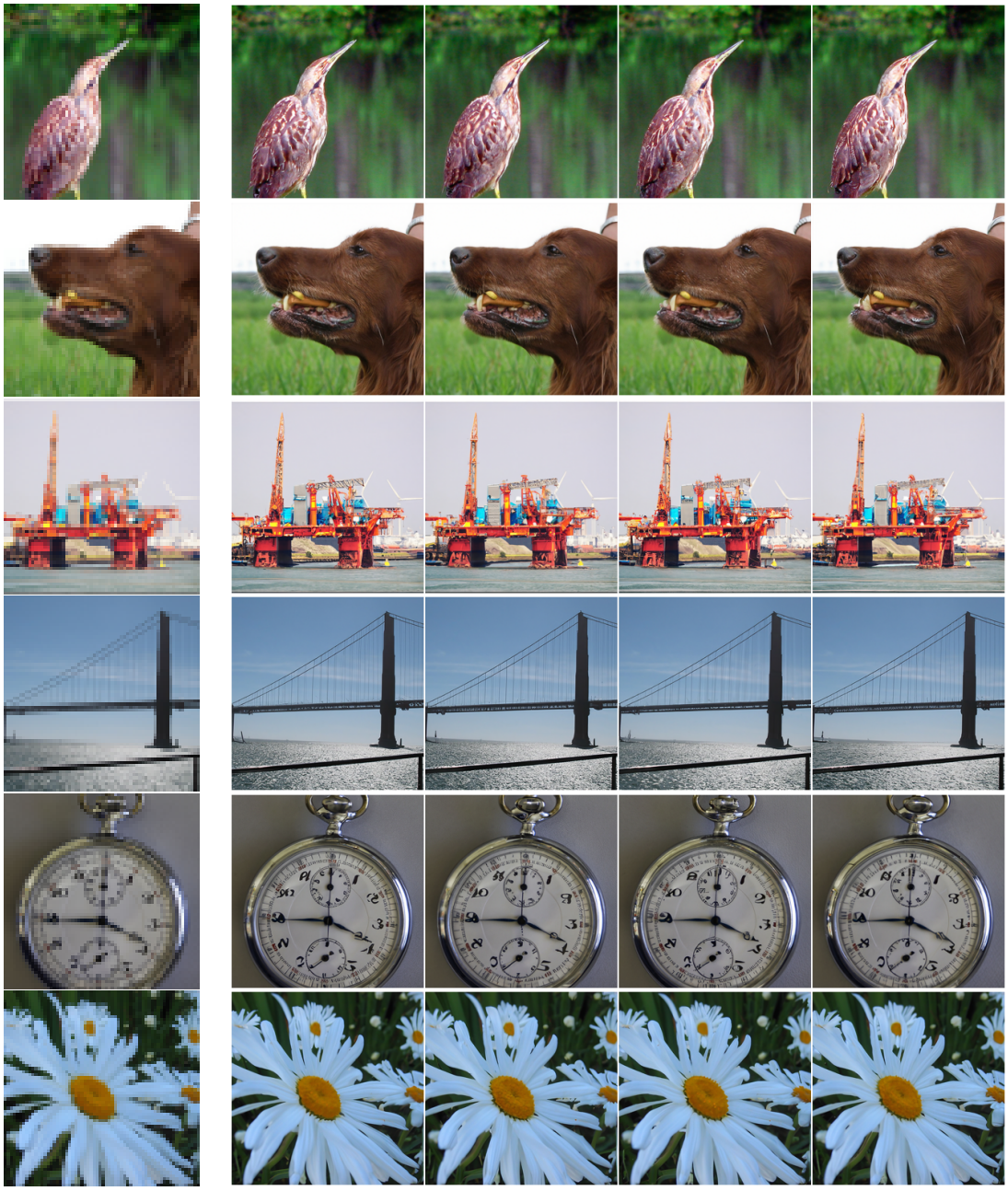

Fig. 9. Image super-resolution results ($64 \times 64 \rightarrow 256 \times 256$) on the ImageNet dataset using our DMM. The first column is the input image and remaining columns are samples from the DMM.

the eye, and coherent with the low-resolution images, demonstrating that DMMs can continue to provide high-quality posterior samples even in very high-dimensional scenarios situations where the prior $p_{\text{data}}(\mathbf{x})$ is unavailable and standard ABC or MCMC techniques are not available.

J.4. Modelling distributions on $SO(3)$ using manifold diffusions

Recall that our noising process on $SO(3)$ is Brownian motion with generator $\mathcal{L} = \partial_t + \frac{1}{2}\Delta$. Since $SO(3)$ is compact, this converges to the uniform measure for large times; see e.g. De Bortoli et al. (2022). For this process, the transition probabilities can be explicitly written as

$$q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) \propto \sum_{\ell=0}^{\infty}(2\ell+1)e^{-\ell(\ell+1)t/2}\frac{\sin\left(\left(\ell+\frac{1}{2}\right)\alpha\right)}{\sin(\alpha/2)}, \tag{29}$$

where $\alpha = \arccos\left[2^{-1}(\text{Tr}(\mathbf{x}_0^T\mathbf{x}_t)-1)\right]$ is the angle between $\mathbf{x}_t$ and $\mathbf{x}_0$, and $\mathbf{x}_t, \mathbf{x}_0 \in SO(3)$ are in matrix form. For completeness, we provide the derivation of this result below in Section J.4.1.

Given this expression, to sample from $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$, we follow Leach et al. (2022) and first sample the rotation axis v uniformly from the sphere $S^2 \subset \mathbb{R}^3$. Then, we sample the rotation angle $\alpha \in [0,\pi]$ using inverse transform sampling from the distribution

$$f_t(\alpha) = \frac{1-\cos(\alpha)}{\pi}\sum_{\ell=0}^{\infty}(2\ell+1)e^{-\ell(\ell+1)t/2}\frac{\sin\left(\left(\ell+\frac{1}{2}\right)\alpha\right)}{\sin(\alpha/2)},$$

where the normalising factor $(1-\cos(\alpha))/\pi$ is the measure on rotation angles induced by the uniform measure on $SO(3)$. For larger t, we find that the above series converges quickly and evaluating summation terms up to $l = 5$ gives an accurate approximation. For $t < 1$, the above series converges slowly, and so we use the approximation

$$f_t(\alpha) \approx \frac{1-\cos(\alpha)}{2\sqrt{\pi}\sin(\alpha/2)}\left(\frac{t}{2}\right)^{-\frac{3}{2}} e^{\frac{t}{8}-\frac{\alpha^2}{2t}}\left[\alpha - e^{-\frac{2\pi^2}{t}}\left((\alpha-2\pi)e^{\frac{2\pi\alpha}{t}} + (\alpha+2\pi)e^{-\frac{2\pi\alpha}{t}}\right)\right]$$

from Leach et al. (2022) instead. From the angle α and the axis $v = (x, y, z)$, we define the skew symmetric matrix V associated to v to be

$$V = \begin{pmatrix} 0 & z & -y \\ -z & 0 & x \\ y & -x & 0 \end{pmatrix}$$

and calculate the corresponding rotation matrix using Rodrigues' formula

$$R = I + \sin(\alpha)V + (1-\cos(\alpha))V^2.$$

Finally, we set $\mathbf{x}_t = R\mathbf{x}_0$. In this way, we can directly sample from the noising process at time t.

The reverse process is generated by $\mathcal{K} = \partial_t + s_\theta(\mathbf{x},t)\cdot\nabla + \frac{1}{2}\Delta$ by Example 7, and the score network is parameterised as $s_\theta(\mathbf{x},t) = \sum_{i=1}^3 s_\theta^i(\mathbf{x},t)E_i(\mathbf{x})$, using a basis $\{E_i\}_{i=1}^3$ of the tangent bundle.

We use the denoising score matching objective $\mathcal{I}_{\mathrm{DSM}}(\theta)$ to learn θ (see Section F.3). To compute the score $\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$, we use automatic differentiation on Equation (29), where $\mathbf{x}_t, \mathbf{x}_0 \in \mathbb{R}^{3\times 3}$ are represented in matrix form, followed by projection to the tangent space at $\mathbf{x}_t$. For small times, we find this can be numerically unstable, and so we use Varadhan's approximation

$$\lim_{t\to 0}\ t\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = \exp_{\mathbf{x}_t}^{-1}(\mathbf{x}_0)$$

for the heat kernel $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)$ at small times instead (De Bortoli et al., 2022).

Once we have learned the score network, we generate approximate samples from the reverse process using the Geodesic Random Walk method of De Bortoli et al. (2022), which corresponds to performing an Euler-Maruyama discretisation, taking Gaussian steps in the tangent space and then projecting back to the manifold using the exponential map.

J.4.1. Derivation of analytic transition probabilities

First, we calculate the metric tensor using the quaternion chart on $SO(3)$, where the unit quaternion $w + x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ represents a rotation by an angle $\alpha = 2\cos^{-1}(w)$ about the axis (x, y, z), and we consider the coordinates (x, y, z) to be our local chart. If $r = w + x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$, we find the metric at r by considering two small displacements $r + \mathrm{d}r$ and $r + \mathrm{d}r'$, rotating r back to the identity, and then using the fact that near the identity the metric is given by $4\mathrm{d}x^2 + 4\mathrm{d}y^2 + 4\mathrm{d}z^2$ (where the scaling is chosen to correspond to the definition of the exponential map used by De Bortoli et al. (2022) and Leach et al. (2022)). Writing

$$\begin{aligned} r + \mathrm{d}r &= (w + \mathrm{d}w) + (x + \mathrm{d}x)\mathbf{i} + (y + \mathrm{d}y)\mathbf{j} + (z + \mathrm{d}z)\mathbf{k}, \\ r + \mathrm{d}r' &= (w + \mathrm{d}w') + (x + \mathrm{d}x')\mathbf{i} + (y + \mathrm{d}y')\mathbf{j} + (z + \mathrm{d}z')\mathbf{k}, \end{aligned}$$

where we have $w\mathrm{d}w + x\mathrm{d}x + y\mathrm{d}y + z\mathrm{d}z = 0$ and $w\mathrm{d}w' + x\mathrm{d}x' + y\mathrm{d}y' + z\mathrm{d}z' = 0$, and noting that composition of rotations corresponds to multiplication in the quaternion algebra, we have

$$\begin{aligned} r^{-1}(r + \mathrm{d}r) &= (w - x\mathbf{i} - y\mathbf{j} - z\mathbf{k})\left((w + \mathrm{d}w) + (x + \mathrm{d}x)\mathbf{i} + (y + \mathrm{d}y)\mathbf{j} + (z + \mathrm{d}z)\mathbf{k}\right) \\ &= 1 + (-x\mathrm{d}w + w\mathrm{d}x - y\mathrm{d}z + z\mathrm{d}y)\,\mathbf{i} + (-y\mathrm{d}w + w\mathrm{d}y - z\mathrm{d}x + x\mathrm{d}z)\,\mathbf{j} \\ &\quad + (-z\mathrm{d}w + w\mathrm{d}z - x\mathrm{d}y + y\mathrm{d}x)\,\mathbf{k} \end{aligned}$$

and similarly for $r^{-1}(r + \mathrm{d}r')$. Therefore, the metric is expressed by

$$4\left\{\left(w + \frac{x^2}{w}\right)\mathrm{d}x + \left(-y + \frac{xz}{w}\right)\mathrm{d}z + \left(z + \frac{xy}{w}\right)\mathrm{d}y\right\}^2 + \textit{cyclic terms}.$$

Multiplying out, collecting like terms and inspecting the coefficients of $\mathrm{d}x^2$, $\mathrm{d}x\mathrm{d}y$ etc., we see that

$$g_{ij} = \frac{4}{w^2}\begin{pmatrix} w^2 + x^2 & xy & xz \\ xy & w^2 + y^2 & yz \\ xz & yz & w^2 + z^2 \end{pmatrix}$$

and we can calculate $|g| = 1/w^2$. Inverting the metric, we get

$$g^{ij} = \frac{1}{4}\begin{pmatrix}(1-x^2) & -xy & -xz\\ -xy & (1-y^2) & -yz\\ -xz & -yz & (1-z^2)\end{pmatrix}.$$

Now, we want to switch to using w as a coordinate, and to find expressions for Δf where $f(w)$ is a function only of w. To this end, we have

$$\nabla f = \frac{\partial f}{\partial w}\mathrm{d}w = -\frac{1}{w}\frac{\partial f}{\partial w}\left(x\mathrm{d}x + y\mathrm{d}y + z\mathrm{d}z\right),$$
$$g^{ij}(\nabla_f)_j = -\frac{1}{4w}\frac{\partial f}{\partial w}\begin{pmatrix}(1-x^2) & -xy & -xz\\ -xy & (1-y^2) & -yz\\ -xz & -yz & (1-z^2)\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = -\frac{w}{4}\frac{\partial f}{\partial w}\begin{pmatrix}x\\ y\\ z\end{pmatrix},$$

so

$$\Delta f = w\,\partial_i\left(\frac{1}{w}g^{ij}(\nabla f)_j\right) = -\frac{3w}{4}\frac{\partial f}{\partial w} + \frac{1-w^2}{4}\frac{\partial^2 f}{\partial w^2}.$$

If we make the substitution $w = \cos(\alpha/2)$, where α is the angle of the corresponding rotation, then $\mathrm{d}w = -\frac{1}{2}\sin(\alpha/2)\mathrm{d}\alpha$, and we get

$$\Delta f = \cot(\alpha/2)\frac{\partial f}{\partial \alpha} + \frac{\partial^2 f}{\partial \alpha^2}.$$

To find the transition probabilities, we must solve the Fokker–Planck equation

$$\frac{\partial q}{\partial t} = \frac{1}{2}\Delta q$$

on $SO(3)$, subject to the initial condition of a delta mass at I. By symmetry, we know the solution will be rotationally symmetric, so we can write the solution as $q(\alpha, t)$. Now, we look for separable solutions of the form $q(\alpha, t) = T(t)A(\alpha)$. We see that we must have

$$\frac{1}{T}\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{1}{2A}\left(\cot(\alpha/2)\frac{\mathrm{d}A}{\mathrm{d}\alpha} + \frac{\mathrm{d}^2 A}{\mathrm{d}\alpha^2}\right).$$

Separating the two equations, we see that we require

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{1}{2}\lambda T, \qquad \cot(\alpha/2)\frac{\mathrm{d}A}{\mathrm{d}\alpha} + \frac{\mathrm{d}^2 A}{\mathrm{d}\alpha^2} = \lambda A,$$

for some fixed λ. The first equation has solution $T(t) = e^{\lambda t/2}$, while a solution to the second is given by

$$A(\alpha) = \frac{\sin\left(\left(\mu + \frac{1}{2}\right)\alpha\right)}{\sin(\alpha/2)},$$

where μ satisfies $-\mu(\mu+1) = \lambda$. In addition, the boundary conditions force μ to be an integer. Combining these expressions, we see that the solution is of the form

$$q(\alpha, t) = \sum_{\ell=0}^{\infty}\beta_\ell e^{-\ell(\ell+1)t/2}\frac{\sin\left(\left(\ell + \frac{1}{2}\right)\alpha\right)}{\sin(\alpha/2)}$$

for some coefficients β_ℓ. Finally, we have the initial condition that $q(\alpha, 0) = 0$ for $\alpha > 0$ and $\int_{SO(3)} q(\mathbf{x}, 0) f(\mathbf{x}) \mathrm{d}\nu(\mathbf{x}) = f(I)$ where ν is the uniform probability measure on $SO(3)$. Up to a scaling factor, this is satisfied if and only if $\beta_\ell \propto (2\ell + 1)$. Putting this all together, we obtain Equation (29).

J.5. Mixture of wrapped normal distributions on $SO(3)$

We consider modelling a mixture of wrapped normal distributions on $SO(3)$. The wrapped normal distribution $\mathcal{N}^W(\mathbf{x} \mid \mu, \sigma^2)$ with mean μ and variance σ^2 is defined here as the transformed distribution via sampling $\mathbf{w} \sim \mathcal{N}(\mathbf{w} \mid 0, \sigma^2)$, where $\mathbf{w} \in \mathbb{R}^{3\times 3}$, from the standard normal distribution with variance σ^2, projecting $\mathbf{w}$ onto the tangent space via $\mathbf{v} = \frac{\mathbf{w} - \mathbf{w}^T}{2}$, then applying the exponential map $\mathbf{x} = \exp_\mu(\mathbf{v})$ at μ. While we could apply standard parametric learning methods which involve learning of $\{\mu_m, \sigma_m\}$ directly, we do not rely on the specific form of the data distribution p_{data}, which allows us to model different distributions flexibly. We consider modelling of a mixture of wrapped normal distributions with $M = 16$ mixtures.

We apply a time-rescaling for the noising process, which is given by $\mathcal{L} = \partial_t + \frac{1}{2}\beta(t)\Delta$ with the linear β schedule given in Equation (28). Then, the reverse process is generated by $\mathcal{K} = \partial_t + \beta(t) s_\theta(\mathbf{x}, t) \cdot \nabla + \frac{1}{2}\beta(t)\Delta$. We use an MLP with 5 layers and 512 hidden units in each layer to output a vector of dimension 3 parameterising $\{s^i_\theta(\mathbf{x}, t)\}^3_{i=1}$. We train the network using the Adam optimiser with batch size 512 and learning rate 0.0002 with a cosine annealing schedule for 100000 iterations.

We learn both the unconditional distribution $p_{\text{data}}(\mathbf{x})$ and the conditional distribution $p_{\text{data}}(\mathbf{x}|m)$ when conditioned on the cluster member m. In the conditional case, we learn a conditional score model $s_\theta(\mathbf{x}, m, t)$ under the same settings.

Fig. 10 shows the results from our conditional model for $p_{\text{data}}(\mathbf{x}|m)$, where we compare the unwrapped distributions in the tangent space between the ground truth normal distribution and the modelled distribution of mixture member $m = 1$, and plot a representative sample from our conditional model. We see that our model targets the correct mixture accurately. Our visualisations of distributions on $SO(3)$ are adapted from Murphy et al. (2021).

We compare our method to the method of De Bortoli et al. (2022), in which the denoising diffusion model for this task is trained by simulating the forward process using the Geodesic Random Walk and using the DSM loss with Varadhan's approximation, rather than using the analytic transition densities given in Appendix J.4.1 as we do. We compare the two methods using the learned models' test-set log-likelihood, calculated using the probability flow ODE as in De Bortoli et al. (2022), as well as the average time per training iteration. Our results are shown in Table 2. We see that both methods achieve comparable log-likelihoods, but our method is about 15% more efficient during training since having the analytic transition densities means that we can simulate the forward noising process in a single step.

J.6. Pose estimation on the SYMSOL dataset

We give details for the pose estimation task on the SYMSOL dataset. We use a similar network design for the conditional score $s_\theta(\mathbf{x}_t, \boldsymbol{\xi}, t)$ as Murphy et al. (2021), composed of

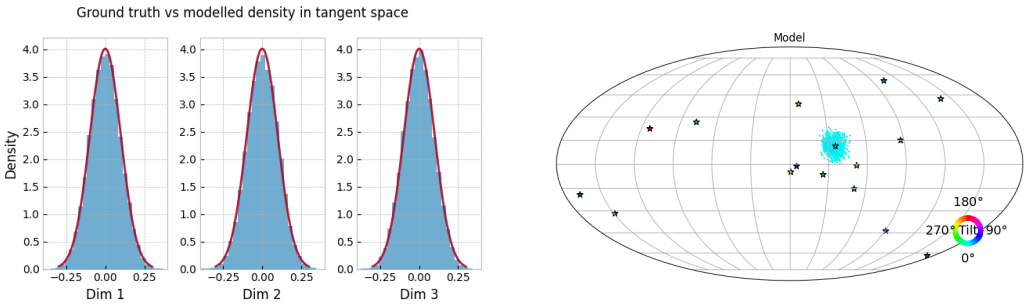

Fig. 10. (Left) Histogram of samples from our model conditioned on the mixture member $m = 1$ compared to the ground truth normal density, represented in the tangent space of $SO(3)$. (Right) Conditional samples from the model for $m = 1$. The axis of rotation and rotation angle are represented by position and colour respectively.

Table 2. Test set log-likelihood and time per training iteration for denoising models on $SO(3)$. Mean and standard deviation reported over 5 seeds.

	$M = 16$	$M = 32$	$M = 64$	Time per iteration (ms)
De Bortoli et al. (2022)	0.864±0.026	0.174±0.025	-0.516±0.016	55.18±2.783
Analytic (ours)	0.872±0.026	0.175±0.025	-0.515±0.016	47.23±2.134

a vision recognition model for processing the input images $\boldsymbol{\xi}$, and an MLP for outputting the score. For the vision recognition model, we utilise pretrained ResNet-50 backbone without the final fully-connected classification layer, which outputs a 2048-dimensional embedding. We next get sinusoidal positional embeddings of $\mathbf{x}_t$ and t, use linear layers to transform all embeddings into 256 dimensions and take the summed embedding. This also allows efficient computations of embeddings with a single $\boldsymbol{\xi}$ and multiple values of $(\mathbf{x}_t, t)$ as the computationally expensive forward pass through the vision recognition model only needs to be taken once. Thus, we simulate a small number of $(\mathbf{x}_t, t)$ pairs given each pair $(\mathbf{x}_0, \boldsymbol{\xi})$ at each step for more efficient training. We finally pass the embedding into an MLP with 3 layers and 256 hidden units in each layer.

Compared to the Implicit-PDF methodology by Murphy et al. (2021), which maintains a grid on $SO(3)$ and approximates the density pointwise, our DMM model directly learns a sampling method and does not require maintaining a grid. Therefore, our method is more general and not specific to $SO(3)$. For our implementation, we modify their network structure to take in the time t, and output the score parameterisation of dimension 3 as opposed to the unnormalised log density of dimension 1. We optimise the network using the Adam optimiser with batch size 128 and learning rate 0.0001 with a cosine annealing schedule for 100000 iterations.

We include further visualisations of the generated samples when conditioned on 2D views of different shapes in Fig. 11. As shown in the plots, the samples generated using DMM are all close to the ground truth and cover all modes of the class of rotational symmetries.

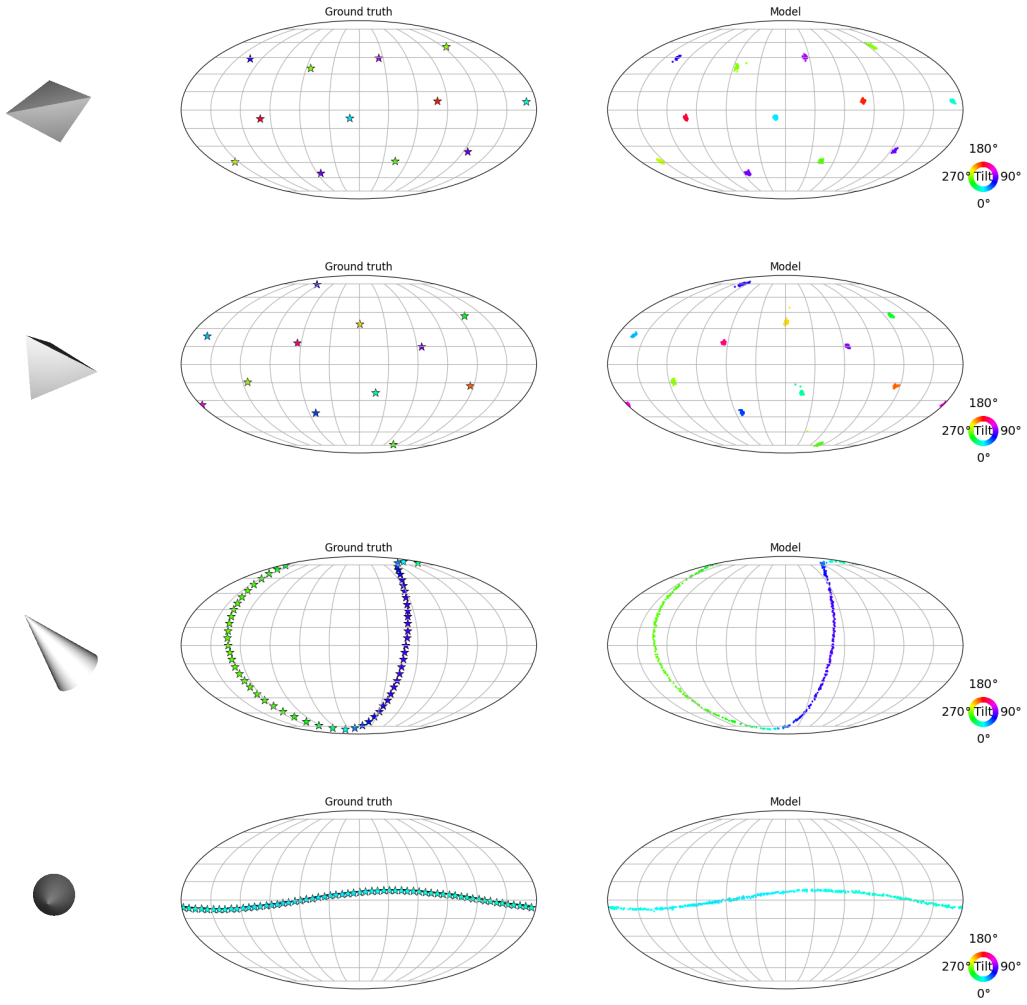

Fig. 11. Samples from the ground truth (plotted as stars, middle) and our pose estimation DMM (right) conditioned on 2D views of shapes (left). The axis of rotation and rotation angle are represented by position and colour respectively.

Table 3. True test data log-likelihood compared to the DMM model ELBO as given by (8) using the ISM loss for the mixture of Dirichlet example. Mean and standard deviation reported over 5 seeds.

Dimension of simplex	$N = 3$	$N = 5$	$N = 10$	$N = 20$
Data	1.321±0.340	4.122±0.242	15.288±0.389	45.914±0.694
Model	1.158±0.160	4.017±0.208	15.061±0.428	45.494±0.698

J.7. Approximation of distributions over measures using Wright–Fisher diffusions

Finally, we evaluate the Wright–Fisher diffusion framework from Appendix F.4 for modeling distributions over measures on a finite state space. We test our framework by attempting to model mixtures of Dirichlet distributions $p_{\text{data}}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} \text{Dirichlet}(\alpha_m)$ with parameters $\alpha_m \in \mathbb{R}^N$. We consider $M = 4$ mixtures and vary the number of dimensions N of the simplex.

As in Appendix F.4, we use a Wright–Fisher diffusion with $q_{ij} = \vartheta_j$ for all $i \neq j$ as our noising process, and set $\vartheta_j = 3$ for all $j = 1, \dots, N$. We also apply a time rescaling to the forward process as in Equation (28). We set $\beta_{\min} = 0.001$ and $\beta_{\max}$ is selected using a grid search from $0.5, 1, 2$. We simulate the forward diffusion process using the exact simulation algorithm of Jenkins and Spanò (2017), which exploits the eigenfunction decomposition of the Wright–Fisher process transition function given in Equation (21) and works by sampling from the ancestral process $A^{\Theta}_{\infty}(t)$ whose distribution is determined by the functions $\{d^{\Theta}_n(t) : n = 0, 1, \dots\}$. For very small times t, we also use a normal approximation for simulating $A^{\Theta}_{\infty}(t)$. For more details, we refer the reader to Jenkins and Spanò (2017).

We learn the score network with the parameterisation $s^i_\theta(\mathbf{p}, t) = p_i \partial(\log \beta(\mathbf{p}, t))/\partial p_i$ using the implicit score matching loss (27). We parameterise $s_\theta(\mathbf{p}, t)$ using an MLP with 4 layers and 512 hidden units in each layer to output a vector of dimension N. We train the network using the Adam optimiser with batch size 128 and learning rate 0.0001 with a cosine annealing schedule for 100000 iterations.

We visualise the results of this experiment in Fig. 7 for a 3-dimensional example. As can be seen, the DMM model is able to learn the ground truth distribution very accurately. We also report in Table 3 the ground truth log-likelihood of the data distribution $p_{\text{data}}(\mathbf{x})$ and the ELBO of the DMM model given by (8) using the ISM loss, as the number of dimensions N increases. We observe that the model's ELBO is consistently close to the true data log-likelihood, which demonstrates the scalability of the DMM model.

References

Cattiaux, P., G. Conforti, I. Gentil, and C. Léonard (2023). Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques 59*(4), 1844–1881.

Dong, R. (2003). Feller Processes and Semigroups. *Lecture notes, UC Berkeley, https://www.stat.berkeley.edu/~pitman/s205s03/lecture27.pdf*.

Durkan, C., A. Bekasov, I. Murray, and G. Papamakarios (2019). Neural Spline Flows. *NeurIPS*.

Durkan, C., I. Murray, and G. Papamakarios (2020). On Contrastive Learning for Likelihood-free Inference. *ICML*.

Ethier, S. N. and T. G. Kurtz (1993). Fleming–Viot Processes in Population Genetics. *SIAM Journal on Control and Optimization 31*, 345–386.

Fearnhead, P. and D. Prangle (2012). Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic Approximate Bayesian Computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(3), 419–474.

Greenberg, D. S., M. Nonnenmacher, and J. H. Macke (2019). Automatic Posterior Transformation for Likelihood-Free Inference. *ICML*.

Jenkins, P. A. and D. Spanò (2017). Exact Simulation of the Wright–Fisher Diffusion. *The Annals of Applied Probability 27*(3).

Karatzas, I. and S. E. Shreve (1991). *Brownian Motion and Stochastic Calculus*. Springer Science & Business Media.

Leach, A., S. M. Schmon, M. T. Degiacomi, and C. G. Willcocks (2022). Denoising Diffusion Probabilistic Models on SO(3) for Rotational Alignment. *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*.

LeCun, Y., C. Cortes, and C. Burges (2010). MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*.

Molchanov, S. A. (1968). Strong Feller Property of Diffusion Processes on Smooth Manifolds. *Theory of Probability & Its Applications 13*, 471–475.

Métivier, M. (1982). *Semimartingales*. De Gruyter.

Palmowski, Z. and T. Rolski (2002). A Technique for Exponential Change of Measure for Markov Processes. *Bernoulli 8*, 767–785.

Papamakarios, G., D. C. Sterratt, and I. Murray (2019). Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows. *AISTATS*.

Prangle, D. (2020). gk: An R Package for the g-and-k and Generalised g-and-h Distributions. *The R Journal 12*(1), 7–20.

Pulido, S. (2011). Semimartingales and stochastic integration. *Lecture Notes, CMU, https://www.andrew.cmu.edu/user/calmost/pdfs/21-882-int_lec.pdf*.

Russakovsky, O., 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 (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*(3), 211–252.

Schilling, R. L. and L. Partzsch (2012). *Brownian Motion: An Introduction to Stochastic Processes*. De Gruyter.

Sohl-Dickstein, J., P. B. Battaglino, and M. R. Deweese (2011). New Method for Parameter Estimation in Probabilistic Models: Minimum Probability Flow. *Physical Review Letters 107*.

Taylor, M. E. (2011). *Partial Differential Equations I: Basic Theory*. Springer.

Tejero-Cantero, A., J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves, D. S. Greenberg, and J. H. Macke (2020). sbi: A Toolkit for Simulation-based Inference. *Journal of Open Source Software 5*(52), 2505.

Vincent, P., H. Larochelle, Y. Bengio, and P. A. Manzagol (2008). Extracting and Composing Robust Features with Denoising Autoencoders. *ICML*.

Yosida, K. (1965). *Functional Analysis*. Springer Science & Business Media.

Statement of Authorship

Title of paper:	From Denoising Diffusions to Denoising Markov Models
Publication status:	Published
Publication details:	Joe Benton, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, Arnaud Doucet. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

Student Confirmation

Student name:	Joe Benton
Contribution to the paper:	The inspiration for the paper came out of my discussions with Valentin De Bortoli, George Deligiannidis and Arnaud Doucet. I provided all of the theoretical results in the paper (Sections 2–5), as well as writing the paper. The experiments (Section 6) were implemented by Yuyang Shi with my guidance.

Signature: J Beaton Date: 29/10/2023

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: George Deligiannidis, Associate Professor of Statistics

Supervisor comments: None

Signature: Date: 31/10/2023

3

A Continuous Time Framework for Discrete Denoising Models

A Continuous Time Framework for Discrete Denoising Models

Andrew Campbell[1] **Joe Benton**[1] **Valentin De Bortoli**[2]

Tom Rainforth[1] **George Deligiannidis**[1] **Arnaud Doucet**[1]

[1]Department of Statistics, University of Oxford, UK [2]CNRS ENS Ulm, Paris, France
{campbell, benton, rainforth, deligian, doucet}@stats.ox.ac.uk
valentin.debortoli@gmail.com

Abstract

We provide the first complete continuous time framework for denoising diffusion models of discrete data. This is achieved by formulating the forward noising process and corresponding reverse time generative process as Continuous Time Markov Chains (CTMCs). The model can be efficiently trained using a continuous time version of the ELBO. We simulate the high dimensional CTMC using techniques developed in chemical physics and exploit our continuous time framework to derive high performance samplers that we show can outperform discrete time methods for discrete data. The continuous time treatment also enables us to derive a novel theoretical result bounding the error between the generated sample distribution and the true data distribution.

1 Introduction

Diffusion/score-based/denoising models [1, 2, 3, 4] are a popular class of generative models that achieve state-of-the-art sample quality with good coverage of the data distribution [5] all whilst using a stable, non-adversarial, simple to implement training objective. The general framework is to define a forward noising process that takes in data and gradually corrupts it until the data distribution is transformed into a simple distribution that is easy to sample. The model then learns to reverse this process by learning the logarithmic gradient of the noised marginal distributions known as the score.

Most previous work on denoising models operates on a continuous state space. However, there are many problems for which the data we would like to model is discrete. This occurs, for example, in text, segmentation maps, categorical features, discrete latent spaces, and the direct 8-bit representation of images. Previous work has tried to realize the benefits of the denoising framework on discrete data problems, with promising initial results [6, 7, 8, 9, 10, 11, 12, 13].

All of these previous approaches train and sample the model in discrete *time*. Unfortunately, working in discrete time has notable drawbacks. It generally forces the user to pick a partition of the process at training time and the model only learns to denoise at these fixed time points. Due to the fixed partition, we are then limited to a simple ancestral sampling strategy. In continuous time, the model instead learns to denoise for any arbitrary time point in the process. This complete specification of the reverse process enables much greater flexibility in defining the reverse sampling scheme. For example, in continuous state spaces, continuous time samplers that greatly reduce the sampling time have been devised [14, 15, 16, 17] as well as ones that improve sample quality [4, 18]. The continuous time interpretation has also enabled the derivation of interesting theoretical properties such as error bounds [19] in continuous state spaces.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

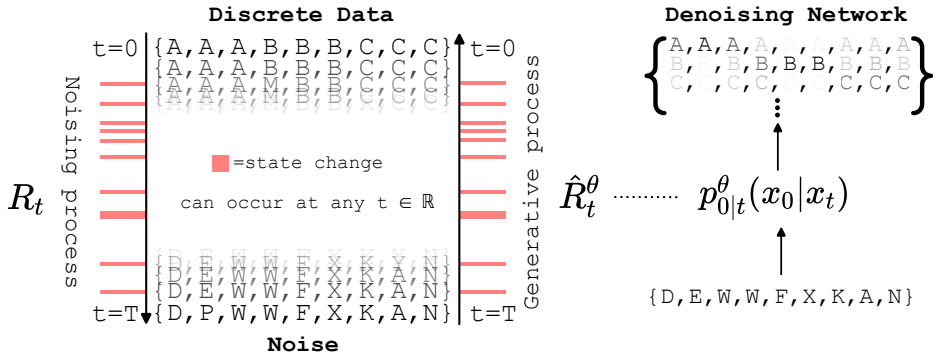

Figure 1: The forward noising process corrupts data according to R_t, the rate of corruption events at time t. The noising process' time reversal gives the generative process which is defined through $\hat{R}_t^\theta$, the rate of generative events at time t. $\hat{R}_t^\theta$ is parameterized through the denoising network, $p_{0|t}^\theta(x_0|x_t)$, which outputs categorical probabilities over clean x_0 values conditioned on a noisy x_t.

To allow these benefits to be exploited for discrete state spaces as well, we formulate a continuous time framework for discrete denoising models. Specifically, our contributions are as follows. We formulate the forward noising process as a Continuous Time Markov Chain (CTMC) and identify the generative CTMC that is the time-reversal of this process. We then bound the log likelihood of the generated data distribution, giving a continuous time equivalent of the ELBO that can be used for efficient training of a parametric approximation to the true generative reverse process. To efficiently simulate the parametric reverse process, we leverage tau-leaping [20] and propose a novel predictor-corrector type scheme that can be used to improve simulation accuracy. The continuous time framework allows us to derive a bound on the error between the true data distribution and the samples generated from the approximate reverse process simulated with tau-leaping. Finally, we demonstrate our proposed method on the generative modeling of images from the CIFAR-10 dataset and monophonic music sequences. Notably, we find our tau-leaping with predictor-corrector sampler can provide higher quality CIFAR10 samples than previous discrete time discrete state approaches, further closing the performance gap between when images are modeled as discrete data or as continuous data.

Proofs for all propositions and theorems are given in the Appendix.

2 Background on Discrete Denoising Models

In the discrete time, discrete state space case, we aim to model discrete data $x_0 \in \mathcal{X}$ with finite cardinality $S = |\mathcal{X}|$. We assume $x_0 \sim p_{\text{data}}(x_0)$ for some discrete data distribution $p_{\text{data}}(x_0)$. We define a forward noising process that transforms $p_{\text{data}}(x_0)$ to some distribution $q_K(x_K)$ that closely approximates an easy to sample distribution $p_{\text{ref}}(x_K)$. This is done by defining forward kernels $q_{k+1|k}(x_{k+1}|x_k)$ that all admit p_{ref} as a stationary distribution and mix reasonably quickly. For example, one can use a simple uniform kernel [6, 8], $q_{k+1|k}(x_{k+1}|x_k) = \delta_{x_{k+1},x_k}(1-\beta) + (1 - \delta_{x_{k+1},x_k})\beta/(S-1)$ where δ is a Kronecker delta. The corresponding p_{ref} is the uniform distribution over all states. Other choices include: an absorbing state kernel—where for each state there is a small probability that it transitions to some absorbing state—or a discretized Gaussian kernel—where only transitions to nearby states have significant probability (valid for spaces with ordinal structure) [8].

After defining $q_{k+1|k}$, we have a forward joint decomposition as follows

$$q_{0:K}(x_{0:K}) = p_{\text{data}}(x_0) \prod_{k=0}^{K-1} q_{k+1|k}(x_{k+1}|x_k).$$

The joint distribution $q_{0:K}(x_{0:K})$ also admits a reverse decomposition:

$$q_{0:K}(x_{0:K}) = q_K(x_K) \prod_{k=0}^{K-1} q_{k|k+1}(x_k|x_{k+1}) \text{ where } q_{k|k+1}(x_k|x_{k+1}) = \frac{q_{k+1|k}(x_{k+1}|x_k) q_k(x_k)}{q_{k+1}(x_{k+1})}.$$

Here $q_k(x_k)$ denotes the marginal of $q_{0:K}(x_{0:K})$ at time k. If one had access to $q_{k|k+1}$ and could sample q_K exactly, then samples from $p_{\text{data}}(x_0)$ could be produced by first sampling $x_K \sim q_K(\cdot)$ and then ancestrally sampling the reverse kernels, i.e. $x_k \sim q_{k|k+1}(\cdot|x_{k+1})$.

However, in practice, $q_{k|k+1}$ is intractable and needs to be approximated with a parametric reverse kernel, $p^\theta_{k|k+1}$. This kernel is commonly defined through the analytic $q_{k|k+1,0}$ distribution and a parametric 'denoising' model $p^\theta_{0|k+1}$ [6, 8],

$$\begin{aligned} p^\theta_{k|k+1}(x_k|x_{k+1}) &\triangleq \textstyle\sum_{x_0} q_{k|k+1,0}(x_k|x_{k+1},x_0)p^\theta_{0|k+1}(x_0|x_{k+1}) \\ &= q_{k+1|k}(x_{k+1}|x_k)\textstyle\sum_{x_0}\frac{q_{k|0}(x_k|x_0)}{q_{k+1|0}(x_{k+1}|x_0)}p^\theta_{0|k+1}(x_0|x_{k+1}). \end{aligned} \tag{1}$$

Though $q_K(x_K)$ is also intractable, for large K we can reliably approximate it with $p_{\text{ref}}(x_K)$. Note that the faster the transitions mix, the more accurate this approximation becomes. Approximate samples from $p_{\text{data}}(x_0)$ can then be obtained by sampling the generative joint distribution

$$p^\theta_{0:K}(x_{0:K}) = p_{\text{ref}}(x_K)\textstyle\prod_{k=0}^{K-1} p^\theta_{k|k+1}(x_k|x_{k+1}),$$

where θ is trained through minimizing the negative discrete time (DT) ELBO which is an upper bound on the negative model log-likelihood

$$\mathbb{E}_{p_{\text{data}}(x_0)}\left[-\log p^\theta_0(x_0)\right] \le \mathbb{E}_{q_{0:K}(x_{0:K})}\left[-\log\tfrac{p^\theta_{0:K}(x_{0:K})}{q_{1:K|0}(x_{1:K}|x_0)}\right] = \mathcal{L}_{\text{DT}}(\theta).$$

It was shown in [1] that $\mathcal{L}_{\text{DT}}$ can be re-written as

$$\begin{aligned} \mathcal{L}_{\text{DT}}(\theta) = \mathbb{E}_{p_{\text{data}}(x_0)}\Big[&\text{KL}(q_{K|0}(x_K|x_0)||p_{\text{ref}}(x_K)) - \mathbb{E}_{q_{1|0}(x_1|x_0)}\left[\log p^\theta_{0|1}(x_0|x_1)\right] \\ &+ \textstyle\sum_{k=1}^{K-1}\mathbb{E}_{q_{k+1|0}(x_{k+1}|x_0)}\left[\text{KL}(q_{k|k+1,0}(x_k|x_{k+1},x_0)||p^\theta_{k|k+1}(x_k|x_{k+1}))\right]\Big] \end{aligned}$$

where KL is the Kullback–Leibler divergence. The forward kernels $q_{k+1|k}$ are chosen such that $q_{k|0}(x_k|x_0)$ can be computed efficiently in a time independent of k. With this, θ can be efficiently trained by taking a random selection of terms from $\mathcal{L}_{\text{DT}}$ in each minibatch and performing a stochastic gradient step.

3 Continuous Time Framework

3.1 Forward process and its time reversal

Our method is built upon a continuous time process from $t = 0$ to $t = T$. State transitions can occur at any time during this process as opposed to the discrete time case where transitions only occur when one of the finite number of transition kernels is applied (see Figure 1). This process is known as a Continuous Time Markov Chain (CTMC), we provide a short overview of CTMCs in Appendix A for completeness. Giving an intuitive introduction here, we can define a CTMC through an initial distribution q_0 and a transition rate matrix $R_t \in \mathbb{R}^{S\times S}$. If the current state is $\tilde{x}$, then the transition rate matrix entry $R_t(\tilde{x}, x)$ is the instantaneous rate (occurrences per unit time) at which state $\tilde{x}$ transitions to state x. Loosely speaking, the next state in the process will likely be one for which $R_t(\tilde{x}, x)$ is high, and furthermore, the higher the rate is, the less time it will take for this transition to occur.

It turns out that the transition rate, R_t, also defines the infinitesimal transition probability for the process between the two time points $t - \Delta t$ and t

$$q_{t|t-\Delta t}(x|\tilde{x}) = \delta_{x,\tilde{x}} + R_t(\tilde{x}, x)\Delta t + o(\Delta t),$$

where $o(\Delta t)$ represents terms that tend to zero at a faster rate than Δt. Comparing to the discrete time case, we see that R_t assumes an analogous role to the discrete time forward kernel $q_{k+1|k}$ in how we define the forward process. Therefore, just as in discrete time, we design R_t such that: i) the forward process mixes quickly towards an easy to sample (stationary) distribution, p_{ref}, (e.g. uniform), ii) we can analytically obtain $q_{t|0}(x_t|x_0)$ distributions to enable efficient training (see Section 4.1 for how this is done). We initialize the forward CTMC at $q_0(x_0) = p_{\text{data}}(x_0)$ at time $t = 0$. We denote the marginal at time $t = T$ as $q_T(x_T)$, which should be close to $p_{\text{ref}}(x_T)$.

We now consider the time reversal of the forward process, which will take us from the marginal $q_T(x_T)$ back to the data distribution $p_{\text{data}}(x_0)$ through a reverse transition rate matrix, $\hat{R}_t \in \mathbb{R}^{S\times S}$:

$$q_{t|t+\Delta t}(\tilde{x}|x) = \delta_{\tilde{x},x} + \hat{R}_t(x, \tilde{x})\Delta t + o(\Delta t).$$

In discrete time, one uses Bayes rule to go from $q_{k+1|k}$ to $q_{k|k+1}$. We can use similar ideas to calculate $\hat{R}_t$ from R_t as per the following result.

Proposition 1. *For a forward in time CTMC, $\{x_t\}_{t\in[0,T]}$, with rate matrix R_t, initial distribution $p_{\text{data}}(x_0)$ and terminal distribution $q_T(x_T)$, there exists a CTMC with initial distribution $q_T(x_T)$ at $t = T$, terminal distribution $p_{\text{data}}(x_0)$ at $t = 0$ and transition rate matrix $\hat{R}_t$ that runs backwards in time and is almost everywhere equivalent to the time reversal of the forward CTMC, $\{x_t\}_{t\in[T,0]}$. Furthermore, $\hat{R}_t$ is related to R_t by the following expression*

$$\hat{R}_t(x, \tilde{x}) = R_t(\tilde{x}, x) \sum_{x_0} \frac{q_{t|0}(\tilde{x}|x_0)}{q_{t|0}(x|x_0)} q_{0|t}(x_0|x) \quad \textit{for} \quad x \neq \tilde{x},$$

where $q_{t|0}(x|x_0)$ are the conditional marginals of the forward process and $q_{0|t}(x_0|x) = q_{t|0}(x|x_0)p_{\text{data}}(x_0)/q_t(x)$ with $q_t(x)$ being the marginal of the forward process at time t. When $x = \tilde{x}$, $\hat{R}_t(x, x) = -\sum_{x'\neq x} \hat{R}_t(x, x')$ because the rows must sum to zero (see Appendix A).

Unfortunately, $\hat{R}_t$ is intractable due to the intractability of $q_t(x)$ and thus of $q_{0|t}(x_0|x)$. Therefore, we consider an approximation $\hat{R}^\theta_t$ of $\hat{R}_t$ by approximating $q_{0|t}(x_0|x)$ with a parametric denoising model, $p^\theta_{0|t}(x_0|x)$:

$$\hat{R}^\theta_t(x, \tilde{x}) = R_t(\tilde{x}, x) \sum_{x_0} \frac{q_{t|0}(\tilde{x}|x_0)}{q_{t|0}(x|x_0)} p^\theta_{0|t}(x_0|x) \quad \text{for} \quad x \neq \tilde{x}$$

and $\hat{R}^\theta_t(x, x) = -\sum_{x'\neq x} \hat{R}^\theta_t(x, x')$ as before. As a further analogy to the discrete time case, notice that when $x \neq \tilde{x}$, $\hat{R}^\theta_t$ has the same form as the discrete time parametric reverse kernel, $p^\theta_{k|k+1}$ defined in eq (1) but with the forward kernel, $q_{k+1|k}$, replaced by the forward rate, R_t.

3.2 Continuous Time ELBO

In discrete time, θ is trained by minimizing the discrete time negative ELBO, $\mathcal{L}_{\text{DT}}$, formed from the forward and reverse processes. We mirror this approach in continuous time by minimizing the corresponding continuous time (CT) negative ELBO, $\mathcal{L}_{\text{CT}}$, as derived below.

Proposition 2. *For the reverse in time CTMC with initial distribution $p_{ref}(x_T)$, terminal distribution $p^\theta_0(x_0)$, and reverse rate $\hat{R}^\theta_t$, an upper bound on the negative model log-likelihood, $\mathbb{E}_{p_{\text{data}}(x_0)}[-\log p^\theta_0(x_0)]$, is given by*

$$\mathcal{L}_{\text{CT}}(\theta) = T\, \mathbb{E}_{t\sim\mathcal{U}(0,T)q_t(x)r_t(\tilde{x}|x)} \left[\left\{ \sum_{x'\neq x} \hat{R}^\theta_t(x, x') \right\} - \mathcal{Z}^t(x) \log \left(\hat{R}^\theta_t(\tilde{x}, x) \right) \right] + C,$$

where C is a constant independent of θ and

$$\mathcal{Z}^t(x) = \textstyle\sum_{x'\neq x} R_t(x, x') \qquad\qquad r_t(\tilde{x}|x) = (1 - \delta_{\tilde{x},x})R_t(x, \tilde{x})/\mathcal{Z}^t(x).$$

Here $r_t(\tilde{x}|x)$ gives the probability of transitioning from x to $\tilde{x}$, given that we know a transition occurs at time t. We can optimize this objective efficiently with stochastic gradient descent. For a gradient update, we sample a batch of datapoints from $p_{\text{data}}(x_0)$, noise each datapoint using a random time, $t \sim \mathcal{U}(0, T)$, $x \sim q_{t|0}(x|x_0)$ and finally sample an auxiliary $\tilde{x}$ from $r_t(\tilde{x}|x)$ for each x. Intuitively, $(x, \tilde{x})$ are a pair of states following the forward in time noising process. Minimizing the second term in $\mathcal{L}_{\text{CT}}$ maximizes the reverse rate for this pair, but going in the backwards direction, $\tilde{x}$ to x. This is how $\hat{R}^\theta_t$ learns to reverse the noising process. Intuition on the first term and a direct comparison to $\mathcal{L}_{\text{DT}}$ is given in Appendix C.1.

The first argument of $\hat{R}^\theta_t$ is input into $p^\theta_{0|t}$ so we naively require two network forward passes on x and $\tilde{x}$ to evaluate the objective. We can avoid this by approximating the $q_t(x)$ sample in the first term with $\tilde{x}$ meaning we need only evaluate the network once on $\tilde{x}$. The approximation is valid because, as we show in Appendix C.4, $\tilde{x}$ is approximately distributed according to $q_{t+\delta t}$ for δt very small.

4 Efficient Forward and Backward Sampling

4.1 Choice of Forward Process

The transition rate matrix R_t needs to be chosen such that the forward process: i) mixes quickly towards p_{ref}, and ii) the $q_{t|0}(x|x_0)$ distributions can be analytically obtained. The Kolmogorov

differential equation for the CTMC needs to be integrated to obtain $q_{t|0}(x|x_0)$. This can be done analytically when R_t and $R_{t'}$ commute for all t, t', see Appendix E. An easy way to meet this condition is to let $R_t = \beta(t)R_b$ where $R_b \in \mathbb{R}^{S\times S}$ is a user-specified time independent base rate matrix and $\beta(t) \in \mathbb{R}$ is a time dependent scalar. We then obtain the analytic expression

$$q_{t|0}(x = j|x_0 = i) = \left(Q\text{exp}\left[\Lambda \int_0^t \beta(s)ds\right] Q^{-1}\right)_{ij}$$

where $R_b = Q\Lambda Q^{-1}$ is the eigendecomposition of matrix R_b and exp$[\cdot]$ the element-wise exponential.

Our choice of β schedule is guided by [3, 4], $\beta(t) = ab^t \log(b)$. The hyperparameters a and b are selected such that $q_T(x) \approx p_{\text{ref}}(x)$ at the terminal time $t = T$ while having a steady speed of 'information corruption' which ensures that $\hat{R}_t$ does not vary quickly in a short span of time.

We experiment with a variety of R_b matrices, for example, a uniform rate, $R_b = \mathbb{1}\mathbb{1}^T - S\text{Id}$, where $\mathbb{1}\mathbb{1}^T$ is a matrix of ones and Id is the identity. For problems with a heavy spatial bias, e.g. images, we can instead use a forward rate that only encourages transitions to nearby states; details and the links to the corresponding discrete time processes can be found in Appendix E.

4.2 Factorizing Over Dimensions

Our aim is to model data that is D dimensional, with each dimension taking one value from S possibilities. We now slightly redefine notation and say $\boldsymbol{x}^{1:D} \in \mathcal{X}^D$, $|\mathcal{X}| = S$. In this setting, calculating transition probabilities naively would require calculating S^D rate values corresponding to each of the possible next states. This is intractable for any reasonably sized S and D. We avoid this problem simply by factorizing the forward process such that each dimension propagates independently. Since this is a continuous time process and each dimension's forward process is independent of the others, the probability two or more dimensions transition at exactly the same time is zero. Therefore, overall in the full dimensional forward CTMC, each transition only ever involves a change in exactly one dimension. For the time reversal CTMC, it will also be true that exactly one dimension changes in each transition. This makes computation tractable because of the S^D rate values, only $D \times (S-1) + 1$ are non-zero - those corresponding to transitions where exactly one dimension changes plus the no change transition. Finally, we note that even though dimensions propagate independently in the forward direction, they are not independent in the reverse direction because the starting points for each dimension's forward process are not independent for non factorized p_{data}. The following proposition shows the exact forms for the forward and reverse rates in this case.

Proposition 3. *If the forward process factorizes as $q_{t|s}(\boldsymbol{x}_t^{1:D}|\boldsymbol{x}_s^{1:D}) = \prod_{d=1}^D q_{t|s}(x_t^d|x_s^d)$, $t > s$, then the forward and reverse rates are of the form*

$$R_t^{1:D}(\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}^{1:D}) = \textstyle\sum_{d=1}^D R_t^d(\tilde{x}^d, x^d)\delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}},$$

$$\hat{R}_t^{1:D}(\boldsymbol{x}^{1:D}, \tilde{\boldsymbol{x}}^{1:D}) = \textstyle\sum_{d=1}^D R_t^d(\tilde{x}^d, x^d)\delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}} \sum_{x_0^d} q_{0|t}(x_0^d|\boldsymbol{x}^{1:D}) \frac{q_{t|0}(\tilde{x}^d|x_0^d)}{q_{t|0}(x^d|x_0^d)},$$

where $R_t^d \in \mathbb{R}^{S\times S}$ and $\delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}}$ is 1 when all dimensions except for d are equal.

To find $\hat{R}_t^{\theta\,1:D}$ we simply replace $q_{0|t}(x_0^d|\boldsymbol{x}^{1:D})$ with $p_{0|t}^\theta(x_0^d|\boldsymbol{x}^{1:D})$ which is easily modeled with a neural network that outputs conditionally independent state probabilities in each dimension. In Appendix C.3 we derive the form of $\mathcal{L}_{\text{CT}}$ when we use this factorized form for $R_t^{1:D}$ and $\hat{R}_t^{\theta\,1:D}$.

4.3 Simulating the Generative Reverse Process with Tau-Leaping

The parametric generative reverse process is a CTMC with rate matrix $\hat{R}_t^{\theta\,1:D}$. Simulating this process from distribution $p_{\text{ref}}(\boldsymbol{x}_T^{1:D})$ at time $t = T$ back to $t = 0$ will produce approximate samples from $p_{\text{data}}(\boldsymbol{x}_0^{1:D})$. The process could be simulated exactly using Gillespie's Algorithm [21, 22, 23] which alternates between i) sampling a holding time to remain in the current state and ii) sampling a new state according to the current rate matrix, $\hat{R}_t^{\theta\,1:D}$ (see Appendix F). This is inefficient for large D because we would need to step through each transition individually and so only one dimension would change for each simulation step.

Instead, we use tau-leaping [20, 23], a very popular approximate simulation method developed in chemical physics. Rather than step back through time one transition to the next, tau-leaping leaps

from t to $t-\tau$ and applies all transitions that occurred in $[t-\tau, t]$ simultaneously. To make a leap, we assume $\hat{R}_t^{\theta\,1:D}$ and $\boldsymbol{x}_t^{1:D}$ remain constant in $[t-\tau, t]$. As we propagate from t to $t-\tau$, we count all of the transitions that occur, but hold off on actually applying them until we reach $t-\tau$, such that $\boldsymbol{x}_t^{1:D}$ remains constant in $[t-\tau, t]$. Assuming $\hat{R}_t^{\theta\,1:D}$ and $\boldsymbol{x}_t^{1:D}$ remain constant, the number of times a transition from $\boldsymbol{x}_t^{1:D}$ to $\tilde{\boldsymbol{x}}^{1:D}$ occurs in $[t-\tau, t]$ is Poisson distributed with mean $\tau \hat{R}_t^{\theta\,1:D}(\boldsymbol{x}_t^{1:D}, \tilde{\boldsymbol{x}}^{1:D})$. Once we reach $t-\tau$, we apply all transitions that occurred simultaneously i.e. $\boldsymbol{x}_{t-\tau}^{1:D} = \boldsymbol{x}_t^{1:D} + \sum_i P_i(\tilde{\boldsymbol{x}}_i^{1:D} - \boldsymbol{x}_t^{1:D})$ where P_i is a Poisson random variable with mean $\tau \hat{R}_t^{\theta\,1:D}(\boldsymbol{x}_t^{1:D}, \tilde{\boldsymbol{x}}_i^{1:D})$. Note the sum assumes a mapping from $\mathcal{X}$ to $\mathbb{Z}$.

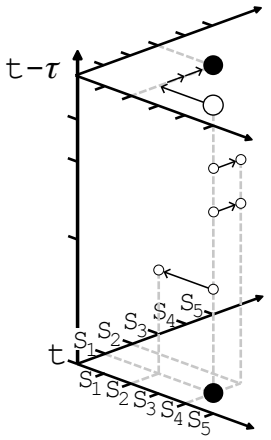

Figure 2: 3D visualization of one tau-leaping step from $x_t^{1:2} = \{S_4, S_1\}$ to $x_{t-\tau}^{1:2} = \{S_2, S_3\}$. Here, $D = 2$, $|\mathcal{X}| = 5$, $P_{12} = 1$, $P_{22} = 2$, all other $P_{ds} = 0$.

Using our knowledge of $\hat{R}_t^{\theta\,1:D}$, we can further unpack this update. Namely, $\hat{R}_t^{\theta\,1:D}(\boldsymbol{x}_t^{1:D}, \tilde{\boldsymbol{x}}^{1:D})$ can only be non-zero when $\tilde{\boldsymbol{x}}^{1:D}$ has a different value to $\boldsymbol{x}_t^{1:D}$ in exactly one dimension (rates for multi-dimensional changes are zero). Explicitly summing over these options we get $\boldsymbol{x}_{t-\tau}^{1:D} = \boldsymbol{x}_t^{1:D} + \sum_{d=1}^{D} \sum_{s=1\setminus x_t^d}^{S} P_{ds}(s - x_t^d)\boldsymbol{e}^d$ where $\boldsymbol{e}^d$ is a one-hot vector with a 1 at dimension d and P_{ds} is a Poisson random variable with mean $\tau \hat{R}_t^{\theta\,1:D}(\boldsymbol{x}_t^{1:D}, \boldsymbol{x}_t^{1:D} + (s - x_t^d)\boldsymbol{e}^d)$. Since multiple P_{ds} can be non-zero, we see that tau-leaping allows $\boldsymbol{x}_t^{1:D}$ to change in multiple dimensions in a single step. Figure 2 visualizes this idea. During the $[t-\tau, t]$ interval, one jump occurs in dimension 1 and two jumps occur in dimension 2. These are all applied simultaneously once we reach $t-\tau$. When our discrete data has ordinal structure (e.g. Section 6.2) our mapping to $\mathbb{Z}$ is not arbitrary and making multiple jumps within the same dimension ($\sum_{s=1\setminus x_t^d}^{S} P_{ds} > 1$) is meaningful. In the non-ordinal/categorical case (e.g. Section 6.3) the mapping to $\mathbb{Z}$ is arbitrary and so, although taking simultaneous jumps in different dimensions is meaningful, taking multiple jumps within the same dimension is not. For this type of data, we reject changes to x_t^d for any d for which $\sum_{s=1\setminus x_t^d}^{S} P_{ds} > 1$. In practice, the rejection rate is very small when $R_t^{1:D}$ is suitable for categorical data (e.g. uniform), see Appendix H.3. In Section 4.5, our error bound accounts for this low probability of rejection and also the low probability of an out of bounds jump that we observe in practice in the ordinal case.

The tau-leaping approximation improves with smaller τ, recovering exact simulation in the limit as $\tau \to 0$. Exact simulation is similar to an autoregressive model in that only one dimension changes per step. Increasing τ and thus the average number of dimensions changing per step gives us a natural way to modulate the 'autoregressiveness' of the model and trade sample quality with compute (Figure 4 right). We refer to our method of using tau-leaping to simulate the reverse CTMC as τLDR (tau-leaping denoising reversal) which we formalize in Algorithm 1 in Appendix F.

We note that theoretically, one could approximate $\hat{R}_t^{\theta\,1:D}$ as constant in the interval $[t-\tau, t]$, and construct a transition probability matrix by solving the forward Kolmogorov equation with the matrix exponential $P_{t-\tau|t} \approx \exp(\tau \hat{R}_t^{\theta\,1:D})$. However, for the learned $\hat{R}_t^{\theta\,1:D} \in \mathbb{R}^{S^D \times S^D}$ matrix, it is intractable to compute this matrix exponential so we use tau-leaping for sampling instead.

4.4 Predictor-Corrector

During approximate reverse sampling, we aim for the marginal distribution of samples at time t to be close to $q_t(x_t)$ (the marginal at time t of the true CTMC). The continuous time framework allows us to exploit additional information to more accurately follow the reverse progression of marginals, $\{q_t(x_t)\}_{t\in[T,0]}$ and improve sample quality. Namely, after a tau-leaping 'predictor' step using rate $\hat{R}_t^{\theta}$, we can apply 'corrector' steps with rate R_t^c which has $q_t(x_t)$ as its stationary distribution. The corrector steps bring the distribution of samples at time t closer to the desired $q_t(x_t)$ marginal. R_t^c is easy to calculate as stated below

Proposition 4. *For a forward CTMC with marginals $\{q_t(x_t)\}_{t\in[0,T]}$, forward rate, R_t, and corresponding reverse CTMC with rate $\hat{R}_t$, the rate $R_t^c = R_t + \hat{R}_t$ has $q_t(x_t)$ as its stationary distribution.*

In practice, we approximate R_t^c by replacing $\hat{R}_t$ with $\hat{R}_t^\theta$. This is directly analogous to Predictor-Corrector samplers in continuous state spaces [4] that predict by integrating the reverse SDE and correct with score-based Markov chain Monte Carlo steps, see Appendix F.2 for further discussion.

4.5 Error Bound

Our continuous time framework also allows us to provide a novel theoretical bound on the error between the true data distribution and the sample distribution generated via tau-leaping (without predictor-corrector steps), in terms of the error in our approximation of the reverse rate and the mixing of the forward noising process.

We assume we have a time-homogeneous rate matrix R_t on $\mathcal{X}$, from which we construct the factorized rate matrix $R_t^{1:D}$ on $\mathcal{X}^D$ by setting $R_t^d = R_t$ for each d. Note that by rescaling time by a factor of $\beta(t)$ we can transform our choice of rate from Section 4.1 to be time-homogeneous. We will denote $|R| = \sup_{t\in[0,T],x\in\mathcal{X}} |R_t(x,x)|$, and let t_{mix} be the (1/4)-mixing time of the CTMC with rate R_t (see [24, Chapter 4.5]).

Theorem 1. *For any $D \geq 1$ and distribution p_{data} on $\mathcal{X}^D$, let $\{x_t\}_{t\in[0,T]}$ be a CTMC starting in p_{data} with rate matrix $R_t^{1:D}$ as above. Suppose that $\hat{R}_t^{\theta\,1:D}$ is an approximation to the reverse rate matrix and let $(y_k)_{k=0,1,\ldots,N}$ be a tau-leaping approximation to the reverse dynamics with maximum step size τ. Suppose further that there is some constant $M > 0$ independent of D such that*

$$\sum_{y\neq x} \left| \hat{R}_t^{1:D}(x,y) - \hat{R}_t^{\theta\,1:D}(x,y) \right| \leq M \tag{2}$$

for all $t \in [0,T]$. Then under the assumptions in Appendix B.5, there are constants $C_1, C_2 > 0$ depending on $\mathcal{X}$ and R_t but not D such that, if $\mathcal{L}(y_0)$ denotes the law of y_0, we have the total variation bound

$$||\mathcal{L}(y_0) - p_{\text{data}}||_{\text{TV}} \leq 3MT + \left\{ \left(|R|SDC_1\right)^2 + \tfrac{1}{2}C_2(M + C_1SD|R|) \right\} \tau T + 2\exp\left\{ -\frac{T\log^2 2}{t_{\text{mix}}\log 4D} \right\}$$

The first term of the above bound captures the error introduced by our approximation of the reverse rate $\hat{R}_t^{1:D}$ with $\hat{R}_t^{\theta\,1:D}$. The second term reflects the error introduced by the tau-leaping approximation, and is linear in both T and τ, showing that as we take our tau-leaping steps to be arbitrarily small, the error introduced by tau-leaping goes to zero. The final term describes the mixing of the forward chain, and captures the error introduced since p_{ref} and q_T are not exactly equal.

We choose to make the dependence of the bound on the dimension D explicit, since we are specifically interested in applying tau-leaping to high dimensional problems where we make transitions in different dimensions simultaneously in a single time step. The bound grows at worst quadratically in the dimension, versus e.g. exponentially. The bound is therefore useful in showing us that we do not need to make τ impractically small in high dimensions. Other than gaining these intuitions, we do not expect the bound to be particularly tight in practice and further it would not be practical to compute because of the difficulty in finding M, C_1 and C_2.

The assumptions listed in Appendix B.5 hold approximately for tau-leaping in practice when we use spatially biased rates for ordinal data such that jump sizes are small or uniform rates for non-ordinal data such that the dimensional rejection rate is small. These assumptions could be weakened, however, Theorem 1 would become much more involved, obscuring the intuition and structure of the problem.

5 Related Work

The application of denoising models to discrete data was first described in [1] using a binomial diffusion process for a binary dataset. Each reverse kernel $p_{k|k+1}^\theta$ was directly parameterized without using a denoising model $p_{0|k}^\theta$. In [25] an approach for discrete categorical data was suggested using a uniform forward noising kernel, $q_{k+1|k}$, and a reverse kernel parameterized through a denoising model, though no experiments were performed with the approach. Experiments on text and segmentation maps were then performed with a similar model in [6]. Other forward kernels were introduced in [8] that are more appropriate for certain data types such as the spatially biased Gaussian kernel. [9, 13] apply the approach to discrete latent space modeling using uniform and absorbing state forward

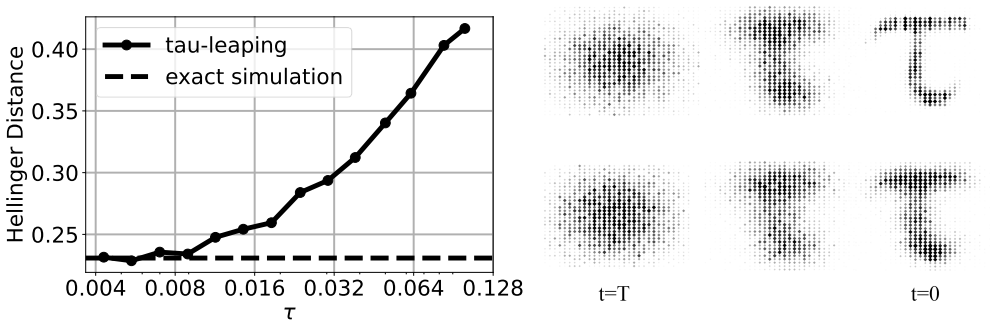

Figure 3: **Left:** Hellinger distance between the true training distribution and generated sample distributions with exact simulation or tau-leaping. With τ small, we simulate the reverse CTMC with the same fidelity as the exact simulation. **Top Right:** Histograms of the marginals during the reverse generative process simulated using tau-leaping with $\tau = 0.004$. Darker and larger diamonds represent increased density. **Bottom Right:** The same for $\tau = 0.1$, note the reduced sample quality.

kernels. Whilst a link to continuous time for the forward process is mentioned in [8], all of these approaches train and sample in discrete time. We show in Appendix G that this involves making an implicit approximation for multi-dimensional data. We extend this line of work by training and sampling in continuous time.

Other works also operate in discrete space but less rigidly follow the diffusion framework. A corruption process tailored to text is proposed in [12], whereby token deletion and insertion is also incorporated. [26] also focus on text, creating a generative reverse chain that repeatedly applies the same denoising kernel. The corruption distribution is also defined through the same denoising kernel to reduce distribution shift between training and sampling. In [7], a more standard masking based forward process is used but the reversal is interpreted from an order agnostic autoregressive perspective. They also describe how their model can be interpreted as the reversal of a continuous time absorbing state diffusion but do not utilize this perspective in training or sampling. [27] propose a denoising type framework that can be used on binary data where the forward and reverse process share the same transition kernel. Finally, in [11], the discrete latent space of a VQVAE is modeled by quantizing an underlying continuous state space diffusion with probabilistic quantization functions.

6 Experiments

6.1 Demonstrative Example

We first verify the method can accurately produce samples from the entire support of the data distribution and that tau-leaping can accurately simulate the reverse CTMC. To do this, we create a dataset formed of 2d samples of a state space of 32 arranged such that the histogram of the training dataset forms a 'τ' shape. We train a denoising model using the $\mathcal{L}_{\text{CT}}$ objective with $p^{\theta}_{0|t}$ parameterized through a residual MLP (full details in Appendix H.1). We then sample the parameterized reverse process using an exact method (up to needing to numerically integrate the reverse rate) and tau-leaping. Figure 3 top-right shows the marginals during reverse simulation with $\tau = 0.004$ and we indeed produce samples from the entire support of p_{data}. Furthermore, we find that with sufficiently small τ, we can match the fidelity of exact simulation of the reverse CTMC (Figure 3 left). The value of τ dictates the number of network evaluations in the reverse process according to NFE $= T/\tau$. In all experiments we use $T = 1$. Exact simulation results in a non zero Hellinger distance between the generated and training distributions because of imperfections in the learned $\hat{R}^{\theta}_t$ model.

6.2 Image Modeling

We now demonstrate that our continuous time framework gives us improved generative modeling performance versus operating in discrete time. We show this on the CIFAR-10 image dataset. Images are typically stored as discrete data, each pixel channel taking one value from 256 possibilities. Continuous state space methods have to somehow get around this fact by, for example, adding a discretization function at the end of the generative process [3] or adding uniform noise to the data.

Table 1: Sample quality metrics and model likelihoods for diffusion methods modeling CIFAR10 in discrete state space. Diffusion methods modeling CIFAR10 in continuous space are included for reference. The Inception Score (IS) and Fréchet Inception Distance (FID) are calculated using 50000 generated samples with respect to the training dataset as is standard practice. The ELBO values are reported on the test set in bits per dimension.

	Method	IS (↑)	FID (↓)	ELBO (↑)
Discrete state	D3PM Absorbing [8]	6.78	30.97	-4.40
	D3PM Gauss [8]	8.56	7.34	$\mathbf{-3.44}$
	τLDR-0 (ours)	8.74	8.10	-3.59
	τLDR-10 (ours)	**9.49**	**3.74**	-3.59
Continuous state	DDPM [3]	9.46	3.17	-3.75
	NCSN [4]	9.89	2.20	-

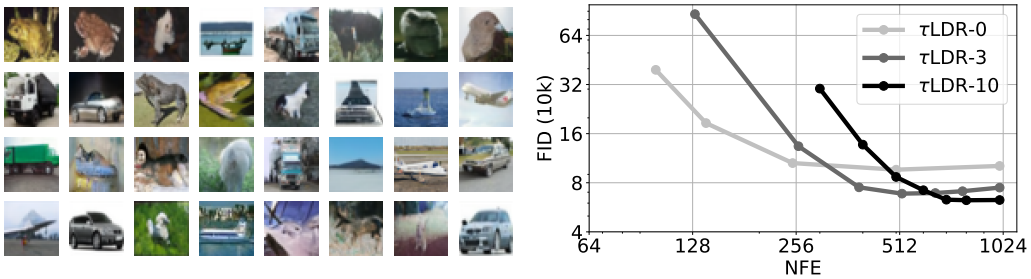

Figure 4: **Left:** Unconditional CIFAR10 samples from our τLDR-10 model **Right:** FID scores for the generated CIFAR10 samples versus number of $p^\theta_{0|t}$ evaluations during sampling (variation induced by varying τ). Calculated with 10k samples, hence the discrepancy with Table 1 [28].

Here, we model the images directly in discrete space. We parameterize $p^\theta_{0|t}$ using the standard U-net architecture [3] with the modifications for discrete state space suggested by [8]. We use a spatially biased rate matrix and train with an augmented $\mathcal{L}_{\text{CT}}$ loss including direct $p^\theta_{0|t}$ supervision, full experimental details are in Appendix H.2.

Figure 4 left shows randomly generated unconditional CIFAR10 samples from the model and we report sample quality metrics in Table 1. We see that our method (τLDR-0) with 0 corrector steps has better Inception Score but worse FID than the D3PM discrete time method. However, our τLDR-10 method with 10 corrector steps per predictor step at the end of the reverse sampling process ($t < 0.1T$) greatly improves sample quality, beating the discrete time method in both metrics and further closes the performance gap with methods modeling images as continuous data. The derivation of the corrector rate which gave us this improved performance required our continuous time framework. D3PM achieves the highest ELBO but we note that this does not correlate well with sample quality. In Table 1, τ was adjusted such that both τLDR-0 and τLDR-10 used 1000 $p^\theta_{0|t}$ evaluations in the reverse sampling procedure. We show how FID score varies with number of $p^\theta_{0|t}$ evaluations for τLDR-$\{0, 3, 10\}$ in Figure 4 right. The optimum number of corrector steps depends on the sampling budget, with lower numbers of corrector steps being optimal for tighter budgets. This is due to the increased τ required to maintain a fixed budget when we use a larger number of corrector steps.

6.3 Monophonic Music

In this experiment, we demonstrate our continuous time model improves generation quality on non-ordinal/categorical discrete data. We model songs from the Lakh pianoroll dataset [29, 30]. We select all monophonic sequences from the dataset such that at each of the 256 time steps either one from 128 notes is played or it is a rest. Therefore, our data has state space size $S = 129$ and dimension $D = 256$. We scramble the ordering of the state space when mapping to $\mathbb{Z}$ to destroy any ordinal structure. We parameterize $p^\theta_{0|t}$ with a transformer architecture [31] and train using a conditional form of $\mathcal{L}_{\text{CT}}$ targeting the conditional distribution of the final 14 bars (224 time steps) given the first 2 bars of the song. We use a uniform forward rate matrix, R_t, full experimental details

Table 2: Metrics comparing generated conditional samples and ground truth completions. We compute these over the test set showing mean±std with respect to 5 samples for each test song.

Model	Hellinger Distance	Proportion of Outliers
τLDR-0 Birth/Death	0.3928 ± 0.0010	0.1316 ± 0.0012
τLDR-0 Uniform	0.3765 ± 0.0013	0.1106 ± 0.0010
τLDR-2 Uniform	$\mathbf{0.3762 \pm 0.0015}$	$\mathbf{0.1091 \pm 0.0014}$
D3PM Uniform [8]	0.3839 ± 0.0002	0.1137 ± 0.0010

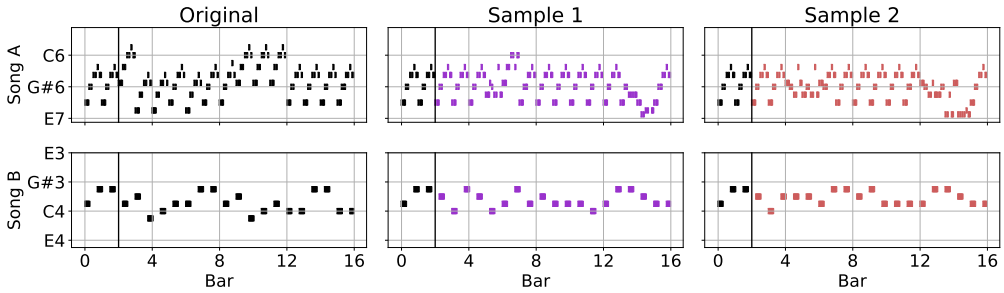

Figure 5: Conditional completions of an unseen music sequence. The conditioning 2 bars are shown to the left of the black line. More examples and audio recordings are linked in Appendix H.3.

are given in Appendix H.3. Conditional completions of unseen test songs are shown in Figure 5. The model is able to faithfully complete the piece in the same style as the conditioning bars.

We quantify sample quality in Table 2. We use two metrics: the Hellinger distance between the histograms of generated and ground truth notes and the proportion of outlier notes in the generations but not in the ground truth. Using our method, we compare between a birth/death and uniform forward rate matrix R_t. The birth/death rate is only non-zero for adjacent states whereas the uniform rate allows transitions between arbitrary states which is more appropriate for the categorical case thus giving improved sample quality. Adding 2 corrector steps per predictor step further improves sample quality. We also compare to the discrete time method D3PM [8] with its most suitable corruption process for categorical data. We find it performs worse than our continuous time method.

7 Discussion

We have presented a continuous time framework for discrete denoising models. We showed how to efficiently sample the generative process with tau-leaping and provided a bound on the error of the generated samples. On discrete data problems, we found our predictor-corrector sampler improved sample quality versus discrete time methods. Regarding limitations, our model requires many model evaluations to produce a sample. Our work has opened the door to applying the work improving sampling speed on continuous data [14, 15, 16, 17, 32] to discrete data problems too. Modeling performance on images is also slightly behind continuous state space models, we hope this gap is further closed with bespoke discrete state architectures and corruption process tuning. Finally, we note that the ELBO values for the discrete time model on CIFAR10 are better than for our method. In this work, we focused on sample quality rather than using our model to give data likelihoods e.g. for compression downstream tasks.

Acknowledgements

Andrew Campbell and Joe Benton acknowledge support from the EPSRC CDT in Modern Statistics and Statistical Machine Learning (EP/S023151/1). Arnaud Doucet is partly supported by the EPSRC grant EP/R034710/1. He also acknowledges support of the UK Defence Science and Technology Laboratory (DSTL) and EPSRC under grant EP/R013616/1. This is part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative. This project made use of time on Tier 2 HPC facility JADE2, funded by EPSRC (EP/T022205/1).

References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, 2015.

[2] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 2019.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[4] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

[6] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 2021.

[7] Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. *International Conference on Learning Representations*, 2022.

[8] Jacob Austin, Daniel Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 2021.

[9] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 2021.

[10] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. *International Conference on Machine Learning*, 2022.

[11] Max Cohen, Guillaume Quispe, Sylvain Le Corff, Charles Ollion, and Eric Moulines. Diffusion bridges vector quantized variational autoencoders. *International Conference on Machine Learning*, 2022.

[12] Daniel D Johnson, Jacob Austin, Rianne van den Berg, and Daniel Tarlow. Beyond in-place corruption: Insertion and deletion in denoising probabilistic models. *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (INNF+)*, 2021.

[13] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[14] Alexia Jolicoeur-Martineau, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080*, 2021.

[15] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

[16] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *International Conference on Learning Representations*, 2022.

[17] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[18] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped Langevin diffusion. *International Conference on Learning Representations*, 2022.

[19] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 2021.

[20] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[21] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[22] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[23] Darren J Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC, 2018.

[24] David Levin, Yuval Peres, and Elizabeth Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021.

[26] Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. Step-unrolled denoising autoencoders for text generation. *International Conference on Learning Representations*, 2022.

[27] Anirudh Goyal, Nan Rosemary Ke, Surya Ganguli, and Yoshua Bengio. Variational walkback: Learning a transition operator as a stochastic recurrent net. *Advances in Neural Information Processing Systems*, 2017.

[28] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[29] Colin Raffel. *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. PhD thesis, Columbia University, 2016.

[30] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. *AAAI Conference on Artificial Intelligence*, 2018.

[31] Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *International Society for Music Information Retrieval*, 2021.

[32] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *International Conference on Learning Representations*, 2022.

[33] David F Anderson. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The Journal of Chemical Physics*, 127(21):214107, 2007.

[34] Chie Furusawa, Shinya Kitaoka, Michael Li, and Yuri Odagiri. Generative probabilistic image colorization. *arXiv preprint arXiv:2109.14518*, 2021.

[35] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *AAAI Conference on Artificial Intelligence*, 2018.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015.

[38] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *International Conference on Learning Representations*, 2017.

[39] Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. Pixelsnail: An improved autoregressive generative model. *International Conference on Machine Learning*, 2018.

[40] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

[41] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. `https://github.com/mseitzer/pytorch-fid`, August 2020. Version 0.2.1.

Appendix

Contents

The appendix is organized as follows. In Section A, we provide a short introduction to Continuous Time Markov Chains, including the relevant results we use in this work. Proofs for all the Propositions and Theorems from the main text are in Section B. We then describe in Section C some additional intuitions and forms of our proposed objective, $\mathcal{L}_{\text{CT}}$. In Section D, we describe how an additional direct denoising model supervision term can be added to the objective to improve empirical performance. Details for how we define the forward process in our model can be found in Section E. Section F describes in more detail how CTMCs can be simulated and includes the algorithmic description of tau-leaping. We argue in Section G that operating in discrete time forces an implicit assumption when using a factorized forward process on multi-dimensional data. Full experimental details for all investigations can be found in Section H as well as additional plots and results from our models. Finally, in Section I, we consider the social impacts of our research.

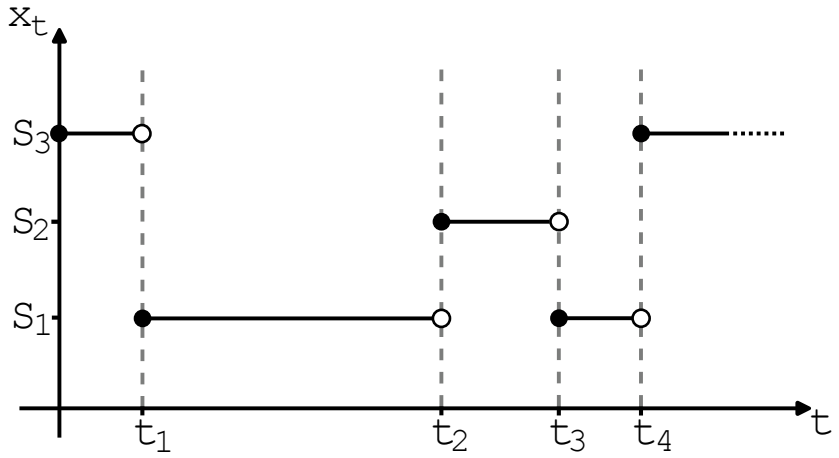

Figure 6: Schematic representation of a 1-dimensional CTMC with 3 states.

A Primer on Continuous Time Markov Chains

A Continuous Time Markov Chain (CTMC) is a right continuous stochastic process $\{x_t\}_{t\in[0,T]}$ satisfying the Markov property, with x_t taking values in a discrete state space $\mathcal{X}$. Since the CTMC is Markov, future behaviour of the process depends only on the current state and not the history. A schematic representation of a CTMC path is shown in Figure 6. The process repeatedly transitions from one state to another after having waited in the previous state for a randomly determined amount of time.

A CTMC can be completely characterised by its jumps and holding times. Specifically, the time between each jump or *holding time* is exponentially distributed with mean $\nu(x)$ where x is the state in which the process is holding. The next state that is jumped to is drawn from a jump probability distribution $r(\tilde{x}|x)$. The holding and jumping procedure is then repeated.

There is an equivalent definition involving the transition rate matrix, $R \in \mathbb{R}^{S\times S}$, that we use in the main paper. The transition rate matrix is defined as

$$R(\tilde{x}, x) = \lim_{\Delta t\to 0} \frac{q_{t|t-\Delta t}(x|\tilde{x}) - \delta_{x,\tilde{x}}}{\Delta t} \tag{3}$$

where $R(\tilde{x}, x)$ is the $(\tilde{x}, x)$ element of the transition rate matrix and $q_{t|t-\Delta t}(x|\tilde{x})$ is the infinitesimal transition probability of being in state x at time t given that the process was in state $\tilde{x}$ at time $t - \Delta t$. Conversely, the CTMC can itself be defined through this infinitesimal transition probability

$$q_{t|t-\Delta t}(x|\tilde{x}) = \delta_{x,\tilde{x}} + R(\tilde{x}, x)\Delta t + o(\Delta t) \tag{4}$$

where $o(\Delta t)$ represents terms that tend to zero at a faster rate than Δt. From this definition of the transition rate matrix, we can infer the following properties:

$$R(\tilde{x}, x) \geq 0 \quad \text{for} \quad \tilde{x} \neq x, \qquad R(x, x) \leq 0, \qquad R(x, x) = -\textstyle\sum_{x'\neq x} R(x, x') \tag{5}$$

$R(\tilde{x}, x)$ is the rate at which probability mass moves from state $\tilde{x}$ to x. $R(x, x)$ is the total rate at which probability mass moves out of state x and is thus negative.

In the time-homogeneous case, R has simple relations to the jump and holding time definitions.

$$\nu(x) = -\frac{1}{R(x, x)} \qquad r(\tilde{x}|x) = (1 - \delta_{\tilde{x},x})\frac{R(x, \tilde{x})}{-R(x, x)}$$

In the time-inhomogeneous case, our transition rate matrix will now depend on time, R_t, and these simple relations to the jump and holding time definition do not hold. However, R_t will still follow equations (3), (4) and (5).

The CTMC transition probabilities satisfy the Kolmogorov forward and backward equations. For $t > s$,

$$\text{Kolmogorov forward equation} \quad \partial_t q_{t|s}(x|\tilde{x}) = \sum_y q_{t|s}(y|\tilde{x}) R_t(y, x)$$

$$\text{Kolmogorov backward equation} \quad \partial_s q_{t|s}(x|\tilde{x}) = -\sum_y R_s(\tilde{x}, y) q_{t|s}(x|y)$$

The Kolmogorov forward equation also gives us a differential equation for the marginals of the CTMC.

$$\partial_t q_t(x) = \sum_y q_t(y) R_t(y, x).$$

Exponential and Poisson Random Variables In the time homogeneous case, holding times are exponentially distributed with mean $\nu(x) = -1/R(x, x)$. The tau-leaping algorithm relies on the fact that the number of events in interval $[0, t]$ is Poisson distributed with mean $\frac{1}{\nu}t$ when the inter-event times are exponentially distributed with mean ν.

B Proofs

B.1 Proof of Proposition 1

Proof. We recall that a process $\{x_t\}_{t\in[0,T]}$ taking values in $\mathcal{X}$ is called a CTMC if it is right-continuous and satisfies the Markov property. Denote $\{y_t\}_{t\in[0,T]} = \{x_{T-t}\}_{t\in[0,T]}$ except at the jump times of the forward process τ_n with $n \in \mathbb{N}$, where $y_{T-\tau_n} = x_{\tau_n}^- = \lim_{t\leq\tau_n, t\to\tau_n} x_t$. Hence, $\{y_t\}_{t\in[0,T]}$ is almost surely equal to $\{x_{T-t}\}_{t\in[0,T]}$ and is right-continuous. Since the Markov property is symmetric, we get that $\{y_t\}_{t\in[0,T]}$ is a CTMC. We now compute its transition matrix. Let $x, \tilde{x} \in \mathcal{X}$ with $x \neq \tilde{x}$, using the Kolmogorov forward equation, we have

$$\partial_t p_{t|s}(\tilde{x}|x) = \textstyle\sum_{y\in\mathcal{X}} p_{t|s}(y|x) \hat{R}_{T-t}(y, \tilde{x}) \ ,$$

where $\{p_{t|s},\ s, t \in [0, T],\ t > s\}$ is the transition probability system associated with $\{y_t\}_{t\in[0,T]}$ and $\{\hat{R}_{T-t}\}_{t\in[0,T]}$ is the transition rate matrix associated with $\{y_t\}_{t\in[0,T]}$. Note that

$$\begin{aligned}
p_{t|s}(x = j|\tilde{x} = i) &= \mathbb{P}(y_t = j \mid y_s = i) \\
&= \mathbb{P}(x_{T-t} = j \mid x_{T-s} = i) \\
&= \mathbb{P}(x_{T-s} = i|x_{T-t} = j) \frac{\mathbb{P}(x_{T-t} = j)}{\mathbb{P}(x_{T-s} = i)} \\
&= q_{T-s|T-t}(\tilde{x} = i|x = j) \frac{q_{T-t}(x = j)}{q_{T-s}(\tilde{x} = i)}
\end{aligned}$$

where $\{q_{t|s}, s, t \in [0, T], t > s\}$ is the transition probability system associated with $\{x_t\}_{t\in[0,T]}$ and $\{q_t, t \in [0, T]\}$ are the marginals of $\{x_t\}_{t\in[0,T]}$. Now, writing the backward Kolmogorov equation for $\{x_t\}_{t\in[0,T]}$

$$\partial_s q_{t|s}(\tilde{x}|x) = -\textstyle\sum_{y\in\mathcal{X}} R_s(x, y) q_{t|s}(\tilde{x}|y)$$

Re-labeling the time indices we obtain,

$$\begin{aligned}
\partial_{T-t} q_{T-s|T-t}(\tilde{x}|x) &= -\textstyle\sum_{y\in\mathcal{X}} R_{T-t}(x, y) q_{T-s|T-t}(\tilde{x}|y) \\
\partial_t q_{T-s|T-t}(\tilde{x}|x) &= \textstyle\sum_{y\in\mathcal{X}} R_{T-t}(x, y) q_{T-s|T-t}(\tilde{x}|y)
\end{aligned}$$

Letting $s \to t$ and using that $\lim_{s\to t} q_{T-s|T-t}(x|\tilde{x}) = 0$, we get that

$$\begin{aligned}
\hat{R}_{T-t}(x,\tilde{x}) &= \lim_{s\to t} \partial_t p_{t|s}(\tilde{x}|x) \\
&= \lim_{s\to t} \partial_t \left(q_{T-s|T-t}(x|\tilde{x}) \frac{q_{T-t}(\tilde{x})}{q_{T-s}(x)} \right) \\
&= \lim_{s\to t} \left[\partial_t \left(q_{T-s|T-t}(x|\tilde{x}) \right) \frac{q_{T-t}(\tilde{x})}{q_{T-s}(x)} + q_{T-s|T-t}(x|\tilde{x}) \frac{\partial_t q_{T-t}(\tilde{x})}{q_{T-s}(x)} \right] \\
&= \lim_{s\to t} \partial_t \left(q_{T-s|T-t}(x|\tilde{x}) \right) \frac{q_{T-t}(\tilde{x})}{q_{T-s}(x)} \\
&= R_{T-t}(\tilde{x},x) \frac{q_{T-t}(\tilde{x})}{q_{T-t}(x)}
\end{aligned}$$

Re-labeling the time-indices on the rate matrices, we obtain

$$\hat{R}_t(x,\tilde{x}) = R_t(\tilde{x},x) \frac{q_t(\tilde{x})}{q_t(x)}$$

Now we write the marginal ratio $\frac{q_t(\tilde{x})}{q_t(x)}$ in a different form

$$\begin{aligned}
\frac{q_t(\tilde{x})}{q_t(x)} &= \sum_{x_0} \frac{p_{\text{data}}(x_0)}{q_t(x)} q_{t|0}(\tilde{x}|x_0) \\
&= \sum_{x_0} \frac{q_{0|t}(x_0|x)}{q_{t|0}(x|x_0)} q_{t|0}(\tilde{x}|x_0).
\end{aligned}$$

Substituting in this form for the marginal ratio concludes the proof.

□

B.2 Proof of Proposition 2

In this section, we detail two proofs for Proposition 2. The first is a formal proof using results from stochastic processes. We then provide a second informal proof for the same result to gain intuition into the $\mathcal{L}_{\text{CT}}$ objective that only relies on elementary results from CTMCs.

Proof 1 - Stochastic Processes

Proof. Let us write $\mathbb{Q}$ for the path measure of the forward CTMC with rate matrix R_t, $\hat{\mathbb{Q}}$ for the path measure of its exact time reversal and $\mathbb{P}^\theta$ for the path measure of the approximate reverse process with rate matrix $\hat{R}^\theta_t$. Also, we use superscripts to notate conditioning on the starting point, for example $\mathbb{Q}^{x_0}$ denotes the path measure of the forward process conditioned to start in x_0.

With this notation, we have

$$\begin{aligned}
-\log p_0^\theta(x_0) &= -\log \int p_{\text{ref}}(\mathrm{d}x_T) \int_{\{\hat{W}_T = x_0\}} \mathbb{P}^{\theta,x_T}(\mathrm{d}w) \\
&= -\log \int q_{T|0}(\mathrm{d}x_T) \int_{\{\hat{W}_T = x_0\}} \hat{\mathbb{Q}}^{x_T}(\mathrm{d}w) \frac{\mathrm{d}p_{\text{ref}}}{\mathrm{d}q_{T|0}}(x_T) \frac{\mathrm{d}\mathbb{P}^{\theta,x_T}}{\mathrm{d}\hat{\mathbb{Q}}^{x_T}}(w) \\
&= -\log \int q_{T|0}(\mathrm{d}x_T) \int \hat{\mathbb{Q}}(\mathrm{d}w|\hat{W}_0 = x_T, \hat{W}_T = x_0) \frac{\mathrm{d}p_{\text{ref}}}{\mathrm{d}q_{T|0}}(x_T) \frac{\mathrm{d}\mathbb{P}^{\theta,x_T}}{\mathrm{d}\hat{\mathbb{Q}}^{x_T}}(w) \mathbb{Q}^{x_T}\{\hat{W}_0 = x_0\} \\
&\leq \int q_{T|0}(\mathrm{d}x_T) \int \hat{\mathbb{Q}}(\mathrm{d}w|\hat{W}_0 = x_T, \hat{W}_T = x_0) \left\{ -\log \frac{\mathrm{d}\mathbb{P}^{\theta,x_T}}{\mathrm{d}\hat{\mathbb{Q}}^{x_T}}(w) \right\} + C,
\end{aligned}$$

where $\mathbb{P}^\theta, \hat{\mathbb{Q}}$ run in the reverse time direction. Writing $\hat{W}_s$ for a reverse path and integrating wrt $p_{\text{data}}(\mathrm{d}x_0)$ we have

$$\int p_{\text{data}}(x_0)[-\log p_0^\theta(x_0)] \leq \int p_{\text{data}}(x_0) \int q_{T|0}(\mathrm{d}x_T) \int \hat{\mathbb{Q}}(\mathrm{d}\hat{W}|\hat{W}_0 = x_T, \hat{W}_T = x_0)$$
$$\times \left\{ \int_{s=0}^{T} \hat{R}^\theta_{T-s}(\hat{W}_s)\mathrm{d}s - \sum_{s:\hat{W}_{s-}\neq\hat{W}_s} \log \mathbb{P}^\theta_{T-s}(\hat{W}_s|\hat{W}_{s-}) R^\theta_{T-s}(\hat{W}_{s-}) \right\} + C,$$

where $\hat{R}^\theta_t(x)$ is shorthand for $-\hat{R}^\theta_t(x,x)$.

When $x_0 \sim p_{\text{data}}, x_T \sim q_{T|0}(\cdot|x_0), \hat{W} \sim \hat{\mathbb{Q}}(\mathrm{d}W|\hat{W}_0 = x_T, \hat{W}_T = x_0)$, the reverse path is distributed according to $p_{\text{data}}(\mathrm{d}x_0)\mathbb{Q}_{x_0}(\mathrm{d}W)$ and therefore $(\hat{W}_{s-}, \hat{W}_s)$ is distributed like $(W_{T-s}, W_{(T-s)-})$ and thus we have

$$\int p_{\text{data}}(x_0)[-\log p_0^\theta(x_0)]$$
$$\leq \int p_{\text{data}}(x_0)\mathbb{Q}_{x_0}(\mathrm{d}W) \left\{ \int_{s=0}^{T} \hat{R}^\theta_{T-s}(W_{(T-s)-})\mathrm{d}s - \sum_{s:W_{(T-s)-}\neq W_{T-s}} \log \mathbb{P}^\theta_{T-s}(W_{(T-s)-}|W_{T-s}) \hat{R}^\theta_{T-s}(W_{T-s}) \right\} + C$$

Using Dynkin's lemma and the fact that $\mathbb{P}^\theta_t(x|y)\hat{R}^\theta_t(y) = \hat{R}^\theta_t(y,x)$ we can re-expresss this final line as

$$= \iint_{s=0}^{T} q_{T-s}(\mathrm{d}x) \left\{ \sum_{z\neq x} \hat{R}^\theta_{T-s}(x,z) - \sum_{z\neq x} R_{T-s}(x,z) \frac{\sum_{y\neq x} R_{T-s}(x,y)}{\sum_{z\neq x} R_{T-s}(x,z)} \log \hat{R}^\theta_{T-s}(y,x) \right\}$$
$$= \iint_{s=0}^{T} q_{T-s}(\mathrm{d}x) r_{T-s}(\mathrm{d}y|x) \left\{ \sum_{z\neq x} \hat{R}^\theta_{T-s}(x,z) - \sum_{z\neq x} R_{T-s}(x,z) \log \hat{R}^\theta_{T-s}(y,x) \right\}$$
$$= \iint_{s=0}^{T} q_s(\mathrm{d}x) r_s(\mathrm{d}y|x) \left\{ \sum_{z\neq x} \hat{R}^\theta_s(x,z) - \sum_{z\neq x} R_s(x,z) \log \hat{R}^\theta_s(y,x) \right\}$$

which rearranges to give the continuous time ELBO in the form of Proposition 2.

□

Proof 2 - Limit of Discrete Time ELBO

Proof. Consider a partitioning of $[0,T]$, $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k < t_{k+1} < \cdots < t_{K-1} < t_K = T$. Let $t_k - t_{k-1} = \Delta t$ for all k. In subscripts we use k as a shorthand for t_k when this does not cause confusion. Considering a CTMC with this time partitioning converts the problem into a discrete time Markov Chain with forward transition kernel, $q_{k+1|k}(x_{k+1}|x_k)$ and parameterized reverse kernel, $p^\theta_{k|k+1}(x_k|x_{k+1})$. Therefore, we can write the negative ELBO in its discrete time form, $\mathcal{L}_{\text{DT}}$

$$\mathcal{L}_{\text{DT}}(\theta) = \mathbb{E}_{p_{\text{data}}(x_0)} \Big[\mathrm{KL}(q_{K|0}(x_K|x_0)||p_{\text{ref}}(x_K)) - \mathbb{E}_{q_{1|0}(x_1|x_0)} \left[\log p^\theta_{0|1}(x_0|x_1) \right]$$
$$+ \sum_{k=1}^{K-1} \mathbb{E}_{q_{k+1|0}(x_{k+1}|x_0)} \left[\mathrm{KL}(q_{k|k+1,0}(x_k|x_{k+1}, x_0)||p^\theta_{k|k+1}(x_k|x_{k+1})) \right] \Big]$$

In the following, we will write the transition kernels in terms of the CTMC rate matrices and take the limit as $\Delta t \to 0$ to obtain a continuous time negative ELBO.

First, consider one item from the inner sum of $\mathcal{L}_{\text{DT}}$

$$L_k = \mathbb{E}_{p_{\text{data}}(x_0)q_{k+1|0}(x_{k+1}|x_0)} \left[\mathrm{KL}(q_{k|k+1,0}(x_k|x_{k+1}, x_0)||p^\theta_{k|k+1}(x_k|x_{k+1})) \right]$$
$$= -\mathbb{E}_{p_{\text{data}}(x_0)q_{k+1|0}(x_{k+1}|x_0)q_{k|k+1,0}(x_k|x_{k+1},x_0)} \left[\log p^\theta_{k|k+1}(x_k|x_{k+1}) \right] + C$$
$$= -\mathbb{E}_{q_k(x_k)q_{k+1|k}(x_{k+1}|x_k)} \left[\log p^\theta_{k|k+1}(x_k|x_{k+1}) \right] + C$$

where we have absorbed terms that do not depend on θ into C. We now write $p^\theta_{k|k+1}(x_k|x_{k+1})$ in terms of $\hat{R}^\theta_k$.

$$p^\theta_{k|k+1}(x_k|x_{k+1}) = \delta_{x_k,x_{k+1}} + \hat{R}^\theta_k(x_{k+1},x_k)\Delta t + o(\Delta t)$$

$$\begin{aligned}
\log p^\theta_{k|k+1}(x_k|x_{k+1}) =& \log\left(\delta_{x_k,x_{k+1}} + \hat{R}^\theta_k(x_{k+1},x_k)\Delta t + o(\Delta t)\right) \\
=& \delta_{x_k,x_{k+1}} \log\left(1 + \hat{R}^\theta_k(x_k,x_k)\Delta t + o(\Delta t)\right) \\
&+ (1-\delta_{x_k,x_{k+1}}) \log\left(\hat{R}^\theta_k(x_{k+1},x_k)\Delta t + o(\Delta t)\right) \\
=& \delta_{x_k,x_{k+1}} \left(\hat{R}^\theta_k(x_k,x_k)\Delta t + o(\Delta t)\right) \\
&+ (1-\delta_{x_k,x_{k+1}}) \log\left(\hat{R}^\theta_k(x_{k+1},x_k)\Delta t + o(\Delta t)\right) \qquad (6)
\end{aligned}$$

where on the last line we have used the series expansion for $\log(1+z) = z - \frac{z^2}{2} + o(z^2)$ valid for $|z| \leq 1, z \neq -1$. For any finite $R^\theta_k(x_k,x_k)$, Δt can be taken small enough such that the series expansion holds. We now substitute this form for $\log p^\theta_{k|k+1}$ into L_k and further write the expectation over $q_{k+1|k}(x_{k+1}|x_k) = \delta_{x_k,x_{k+1}} + R_k(x_k,x_{k+1})\Delta t + o(\Delta t)$ as an explicit sum.

$$\begin{aligned}
L_k = -\mathbb{E}_{q_k(x_k)}\Bigg[\sum_{x_{k+1}}\Bigg\{&\left[\delta_{x_k,x_{k+1}} + R_k(x_k,x_{k+1})\Delta t + o(\Delta t)\right]\times \\
&\Big[\delta_{x_k,x_{k+1}}\left(\hat{R}^\theta_k(x_k,x_k)\Delta t + o(\Delta t)\right) \\
&+ \left(1-\delta_{x_k,x_{k+1}}\right)\log\left(\hat{R}^\theta_k(x_{k+1},x_k)\Delta t + o(\Delta t)\right)\Big]\Bigg\}\Bigg] + C
\end{aligned}$$

$$\begin{aligned}
L_k = -\mathbb{E}_{q_k(x_k)}\Bigg[\sum_{x_{k+1}}\Bigg\{&\delta_{x_k,x_{k+1}}\hat{R}^\theta_k(x_k,x_k)\Delta t \\
&+ (1-\delta_{x_k,x_{k+1}})R_k(x_k,x_{k+1})\Delta t\times \\
&\log\left(\hat{R}^\theta_k(x_{k+1},x_k)\Delta t + o(\Delta t)\right) + o(\Delta t)\Bigg\}\Bigg] + C
\end{aligned}$$

We can isolate $\hat{R}^\theta_k$ within the $\log$ through the following re-arrangement

$$\begin{aligned}
&\Delta t \log\left(\hat{R}^\theta_k(x_{k+1},x_k)\Delta t + o(\Delta t)\right) \\
&= \Delta t \log \Delta t + \Delta t \log\left(\hat{R}^\theta_k(x_{k+1},x_k) + o(1)\right) \\
&= \Delta t \log \Delta t + \Delta t \log\left(1 + o(1)\right) + \Delta t \log\left(\hat{R}^\theta_k(x_{k+1},x_k)\right)
\end{aligned}$$

where the first two terms are independent of θ and tend to 0 as $\Delta t \to 0$. Note that we assume $\hat{R}^\theta_k(x_{k+1},x_k) > 0$ for $x_{k+1} \neq x_k$ pairs which have $R_k(x_k,x_{k+1}) > 0$. This assumption is valid because, for $x_{k+1} \neq x_k$, we have

$$\hat{R}^\theta_k(x_{k+1},x_k) = R_k(x_k,x_{k+1})\sum_{x_0}\frac{q_{k|0}(x_k|x_0)}{q_{k|0}(x_{k+1}|x_0)}p^\theta_{0|k}(x_0|x_{k+1})$$

and we assume $p^\theta_{0|k}(x_0|x_{k+1}) > 0$ which is valid when we parameterize $p^\theta_{0|k}$ with a softmax output. We assume an irreducible Markov chain, hence $q_{k|0} > 0$ for $t_k > 0$.

With this re-arrangement, and absorbing constant terms into C, we obtain

$$L_k = -\mathbb{E}_{q_k(x_k)}\Bigg[\sum_{x_{k+1}}\Bigg\{\delta_{x_k,x_{k+1}}\hat{R}_k^\theta(x_k,x_k)\Delta t \\ + (1-\delta_{x_k,x_{k+1}})R_k(x_k,x_{k+1})\Delta t\log\left(\hat{R}_k^\theta(x_{k+1},x_k)\right) \\ + o(\Delta t)\Bigg\}\Bigg] + C$$

$$L_k = -\mathbb{E}_{q_k(x_k)}\left[\hat{R}_k^\theta(x_k,x_k)\Delta t + \sum_{x_{k+1}\neq x_k} R_k(x_k,x_{k+1})\Delta t\log\hat{R}_k^\theta(x_{k+1},x_k) + o(\Delta t)\right]$$

The second term can be re-written so that it is more efficient to approximate with Monte Carlo. Currently the denoising model $p_{0|k}^\theta$ has to be evaluated for each term in the sum $\sum_{x_{k+1}\neq x_k}$ which would require multiple forward passes of the neural network. We can instead create a new probability distribution to sample from as follows. Define

$$r_k(x_{k+1}|x_k) = (1-\delta_{x_k,x_{k+1}})\frac{R_k(x_k,x_{k+1})}{\mathcal{Z}^k(x_k)}$$

where

$$\mathcal{Z}^k(x_k) = \sum_{x'_{k+1}\neq x_k} R_k(x_k,x'_{k+1})$$

So we now have

$$L_k = -\mathbb{E}_{q_k(x_k)r_k(x_{k+1}|x_k)}\left[\hat{R}_k^\theta(x_k,x_k)\Delta t + \mathcal{Z}^k(x_k)\Delta t\log\hat{R}_k^\theta(x_{k+1},x_k) + o(\Delta t)\right]$$

Examining the other terms in $\mathcal{L}_{\text{DT}}$ we have $\mathbb{E}_{p_{\text{data}}(x_0)}\left[\text{KL}(q_{K|0}(x_K|x_0)||p_{\text{ref}}(x_K))\right]$ which does not depend on θ and $\mathbb{E}_{q_{1|0}(x_1|x_0)}\left[\log p_{0|1}^\theta(x_0|x_1)\right]$ which we expand here

$$\begin{aligned}
&\mathbb{E}_{q_{1|0}(x_1|x_0)}\left[\log p_{0|1}^\theta(x_0|x_1)\right] \\
&\quad= \sum_{x_1}\{\delta_{x_1,x_0} + \Delta t R_1(x_0,x_1) + o(\Delta t)\}\log p_{0|1}^\theta(x_0|x_1) \\
&\quad= \log p_{0|1}^\theta(x_0|x_0) + \Delta t\sum_{x_1} R_1(x_0,x_1)\log p_{0|1}^\theta(x_0|x_1) + o(\Delta t) \\
&\quad= \Delta t\hat{R}_1^\theta(x_0,x_0) + \Delta t\sum_{x_1} R_1(x_0,x_1)\log p_{0|1}^\theta(x_0|x_1) + o(\Delta t)
\end{aligned}$$

where on the final line we have used eq 6. In summary,

$$\begin{aligned}
\mathcal{L}_{\text{DT}} =& \Delta t\mathbb{E}_{p_{\text{data}}(x_0)q_{1|0}(x_1|x_0)}\left[-\hat{R}_1^\theta(x_0,x_0) + \sum_{x_1} R_1(x_0,x_1)\log p_{0|1}^\theta(x_0|x_1)\right] \\
&- \Delta t\sum_{k=1}^{K-1}\mathbb{E}_{q_k(x_k)r_k(x_{k+1}|x_k)}\left[\hat{R}_k^\theta(x_k,x_k) + \mathcal{Z}^k(x_k)\log\hat{R}_k^\theta(x_{k+1},x_k)\right] \\
&+ o(\Delta t) + C
\end{aligned}$$

We now take the limit of $\mathcal{L}_{\text{DT}}$ as $\Delta t\to 0$ and $K\to\infty$.

$$\lim_{\Delta t\to 0}\mathcal{L}_{\text{DT}} = \mathcal{L}_{\text{CT}} = -\int_0^T \mathbb{E}_{q_t(x)r_t(\tilde{x}|x)}\left[\hat{R}_t^\theta(x,x) + \mathcal{Z}^t(x)\log\left(\hat{R}_t^\theta(\tilde{x},x)\right)\right]dt + C$$

We can estimate the integral with Monte Carlo if we consider it to be an expectation with respect to a uniform distribution over times $(0,T)$. We also write $\hat{R}_t^\theta(x,x)$ explicitly as the negative off diagonal row sum to obtain

$$\mathcal{L}_{\text{CT}}(\theta) = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)q_t(x)r_t(\tilde{x}|x)}\left[\left\{\sum_{x'\neq x}\hat{R}_t^\theta(x,x')\right\} - \mathcal{Z}^t(x)\log\left(\hat{R}_t^\theta(\tilde{x},x)\right)\right] + C.$$

□

B.3 Proof of Proposition 3

Proof. We assume $q_{t|s}(\boldsymbol{x}_t^{1:D}|\boldsymbol{x}_s^{1:D})$ factorizes as $\prod_{d=1}^D q_{t|s}(x_t^d|x_s^d)$ where $q_{t|s}(x_t^d|x_s^d)$, $d = 1,\dots,D$ are the transition probabilities for independent singular dimensional CTMCs each with forward rate $R_t^d(\tilde{x}^d, x^d)$. In the following, we will drop time subscripts on x arguments. To find the correspondence between $R_t^{1:D}$ and R_t^d, we use the Kolmogorov forward equation

$$\partial_t q_{t|s}(\boldsymbol{x}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}) = \sum_{\boldsymbol{y}^{1:D}} q_{t|s}(\boldsymbol{y}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}) R_t^{1:D}(\boldsymbol{y}^{1:D}, \boldsymbol{x}^{1:D})$$

Substitute in our factorized form for $q_{t|s}$ into the LHS

$$\begin{aligned}
\partial_t q_{t|s}(\boldsymbol{x}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}) &= \partial_t \left\{ \prod_{d=1}^D q_{t|s}(x^d|\tilde{x}^d) \right\} \\
&= \sum_{d=1}^D q_{t|s}(\boldsymbol{x}^{1:D\setminus d}|\tilde{\boldsymbol{x}}^{1:D\setminus d}) \partial_t q_{t|s}(x^d|\tilde{x}^d) \\
&= \sum_{d=1}^D q_{t|s}(\boldsymbol{x}^{1:D\setminus d}|\tilde{\boldsymbol{x}}^{1:D\setminus d}) \sum_{y^d} q_{t|s}(y^d|\tilde{x}^d) R_t^d(y^d, x^d) \\
&= \sum_{d=1}^D \sum_{\boldsymbol{y}^{1:D}} q_{t|s}(\boldsymbol{x}^{1:D\setminus d}|\tilde{\boldsymbol{x}}^{1:D\setminus d}) q_{t|s}(y^d|\tilde{x}^d) R_t^d(y^d, x^d) \delta_{\boldsymbol{x}^{1:D\setminus d}, \boldsymbol{y}^{1:D\setminus d}} \\
&= \sum_{d=1}^D \sum_{\boldsymbol{y}^{1:D}} q_{t|s}(\boldsymbol{y}^{1:D\setminus d}|\tilde{\boldsymbol{x}}^{1:D\setminus d}) q_{t|s}(y^d|\tilde{x}^d) R_t^d(y^d, x^d) \delta_{\boldsymbol{x}^{1:D\setminus d}, \boldsymbol{y}^{1:D\setminus d}} \\
&= \sum_{\boldsymbol{y}^{1:D}} q_{t|s}(\boldsymbol{y}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}) \sum_{d=1}^D R_t^d(y^d, x^d) \delta_{\boldsymbol{x}^{1:D\setminus d}, \boldsymbol{y}^{1:D\setminus d}}
\end{aligned}$$

We therefore obtain

$$\sum_{\boldsymbol{y}^{1:D}} q_{t|s}(\boldsymbol{y}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}) R_t^{1:D}(\boldsymbol{y}^{1:D}, \boldsymbol{x}^{1:D}) = \sum_{\boldsymbol{y}^{1:D}} q_{t|s}(\boldsymbol{y}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}) \sum_{d=1}^D R_t^d(y^d, x^d) \delta_{\boldsymbol{x}^{1:D\setminus d}, \boldsymbol{y}^{1:D\setminus d}}$$

This must be true for all possible factorizable forward process transitions, $q_{t|s}$, including $q_{t|s}(\boldsymbol{y}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}) = \delta_{\boldsymbol{y}^{1:D}, \tilde{\boldsymbol{x}}^{1:D}}$. This choice gives us our forward rate relation

$$R_t^{1:D}(\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}^{1:D}) = \sum_{d=1}^D R_t^d(\tilde{x}^d, x^d) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}}$$

Substituting this into our expression for the reverse rate from Proposition 1 we obtain

$$\begin{aligned}
\hat{R}_t^{1:D}(\boldsymbol{x}^{1:D}, \tilde{\boldsymbol{x}}^{1:D}) &= \sum_{\boldsymbol{x}_0^{1:D}} \sum_{d=1}^D R_t^d(\tilde{x}^d, x^d) \frac{q_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}_0^{1:D})}{q_t(\boldsymbol{x}^{1:D}|\boldsymbol{x}_0^{1:D})} q_{0|t}(\boldsymbol{x}_0^{1:D}|\boldsymbol{x}^{1:D}) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}} \\
&= \sum_{\boldsymbol{x}_0^{1:D}} \sum_{d=1}^D R_t^d(\tilde{x}^d, x^d) \frac{q_{t|0}(\tilde{x}^d|x_0^d)}{q_{t|0}(x^d|x_0^d)} q_{0|t}(\boldsymbol{x}_0^{1:D}|\boldsymbol{x}^{1:D}) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}} \\
&= \sum_{d=1}^D R_t^d(\tilde{x}^d, x^d) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}} \sum_{x_0^d} q_{0|t}(x_0^d|\boldsymbol{x}^{1:D}) \frac{q_{t|0}(\tilde{x}^d|x_0^d)}{q_{t|0}(x^d|x_0^d)} \sum_{\boldsymbol{x}_0^{1:D\setminus d}} q_{0|t}(\boldsymbol{x}_0^{1:D\setminus d}|x_0^d, \boldsymbol{x}^{1:D}) \\
&= \sum_{d=1}^D R_t^d(\tilde{x}^d, x^d) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}} \sum_{x_0^d} q_{0|t}(x_0^d|\boldsymbol{x}^{1:D}) \frac{q_{t|0}(\tilde{x}^d|x_0^d)}{q_{t|0}(x^d|x_0^d)}
\end{aligned}$$

□

B.4 Proof of Proposition 4

Proof. By the Kolmogorov forward equation applied to the forwards process, we have

$$\partial_t q_t(x_t) = \sum_y R_t(y, x_t) q_t(y)$$

In addition, applying the Kolmogorov forward equation to the reverse process, which has the same marginals as the forward but time-reversed, we get

$$-\partial_t q_t(x_t) = \sum_y \hat{R}_t(y, x_t) q_t(y)$$

Summing these two equations gives

$$\sum_y \left\{ R_t(y, x_t) + \hat{R}_t(y, x_t) \right\} q_t(y) = 0$$

Therefore, by comparison with the Kolmogorov equation, $R_t + \hat{R}_t$ is the rate matrix of a CTMC with invariant distribution q_t. □

B.5 Proof of Theorem 1

In this section, we derive a bound on the error of our tau-leaping diffusion model. Because the tau-leaping approximation is only interesting in the case where multiple jumps are made along different dimensions in a single step, we choose to make the dependence of our bound on the dimension of our model explicit, rather than simply considering the case of fixed D and $\tau \to 0$.

Recall from the main text that we have a time-homogeneous rate matrix R_t on $\mathcal{X}$, from which we construct the factorised rate matrix $R_t^{1:D}$ on $\mathcal{X}^D$ by setting $R_t^d = R_t$ for each d, and will denote $|R| = \sup_{t \in [0,T], x \in \mathcal{X}} |R_t(x, x)|$, and let t_{mix} be the (1/4)-mixing time of the CTMC with rate R_t. We also define addition on the state space $\mathcal{X}^D$ using a mapping from $\mathcal{X}$ to $\mathbb{Z}$ as in Section 4.3 and component-wise addition.

Theorem 1. *For any $D \geq 1$ and distribution p_{data} on $\mathcal{X}^D$, let $\{x_t\}_{t \in [0,T]}$ be a CTMC starting in p_{data} with rate matrix $R_t^{1:D}$ as above. Suppose that $\hat{R}_t^{\theta\, 1:D}$ is an approximation to the reverse rate matrix and let $(y_k)_{k=0,1,\ldots,N}$ be a tau-leaping approximation to the reverse dynamics with maximum step size τ. Suppose further that there is some constant $M > 0$ independent of D such that*

$$\sum_{y \neq x} \left| \hat{R}_t^{1:D}(x, y) - \hat{R}_t^{\theta\, 1:D}(x, y) \right| \leq M$$

for all $t \in [0, T]$. Then under the assumptions listed below, there are constants $C_1, C_2 > 0$ depending on $\mathcal{X}$ and R_t but not D such that, if $\mathcal{L}(y_0)$ denotes the law of y_0, we have the total variation bound

$$||\mathcal{L}(y_0) - p_{\text{data}}||_{\text{TV}} \leq 3MT + \left\{ \left(|R| S D C_1 \right)^2 + \tfrac{1}{2} C_2 (M + C_1 S D |R|) \right\} \tau T + 2 \exp\left\{ - \frac{T \log^2 2}{t_{\text{mix}} \log 4D} \right\}$$

The above theorem holds under the following assumptions, where we write $x \sim y$ for $x, y \in S^D$ if they differ in at most one coordinate.

Assumption 1. *The data distribution p_{data} is strictly positive.*

Assumption 2. *There exists a constant $C_1 > 0$, depending on S and R_t but not D, such that for all $t \in [0, T]$ and $x, y \in S^D$ such that $x \sim y$, we have*

$$\frac{q_t(x)}{q_t(y)} \leq C_1.$$

Assumption 3. *There exists a constant $C_2 > 0$, depending on S and R_t but not D, such that for all $t \in [0, T]$ and all $x, y \in S^D$ such that $x \sim y$, we have*

$$\sum_z \left| \hat{R}_t(x, x + z) - \hat{R}_t(y, y + z) \right| \leq C_2.$$

If instead we were to allow C_1 and C_2 to depend on the dimension D, then Assumptions 2 and 3 follow trivially from Assumption 1 and the finiteness of the state space. However, we choose the stronger formulation above in order to make explicit the dependence of the error bound on the dimension, as previously explained.

As remarked in the main text, in most cases of practical interest (including the two examples explored in Section 6), Assumption 3 holds only approximately. However, we still expect the bound in Assumption 3 to hold whenever x, y are in addition chosen such that the tau-leaping approximation of the reverse process makes a jump between them with reasonably high probability. For example, in the case where our data is ordinal, we expect that for any $x \sim y$ jumps from x to y are only common when x is close to y, and thus $\hat{R}_t(x, x+z)$ and $\hat{R}_t(y, y+z)$ should be reasonably close whenever a jump from x to y occurs. Under a weaker assumption of this form, the proof of Theorem 1 can be adapted to work along similar lines, at the cost of a significant increase in technicality. We therefore choose to focus on the simpler case where Assumption 3 holds as it illustrates the key ideas.

In order to prove Theorem 1 we will require the following lemmas.

Proposition 5. *Let $(x_t)_{t\in[0,T]}$ and $(y_t)_{t\in[0,T]}$ be continuous time Markov chains on a finite state space S with generators G_t and H_t respectively which are both bounded and continuous in t. Let the Markov kernels associated to X and Y be K and L respectively. Then for any probability distribution ν on S we have*

$$||\nu K - \nu L||_{\text{TV}} \leq \int_0^T \sup_{x\in S} \Big\{ \sum_{y\neq x} |G_t(x,y) - H_t(x,y)| \Big\} \, \mathrm{d}t$$

Proof. We define a coupling of $(x_t)_{t\in[0,T]}$ and $(y_t)_{t\in[0,T]}$ as follows, based on the construction in Chapter 20.1 of [24]. First take $Z \sim \nu$ and set $x_0 = y_0 = Z$. Also define the variables $\tilde{x}_0 = \tilde{y}_0 = Z$.

Next, fix λ such that $|G_t(x,x)|, |H_t(x,x)| \leq \lambda$ for all $x \in S, t \in [0,T]$, let $(N_s)_{1\leq s\leq T}$ be a Poisson process on $[0,T]$ of rate λ, and set $N_0 = 0$. We write $N = N_T$, and $S_1, S_2, \ldots, S_N$ for the arrival times and set $S_{n+1} = T$. We construct x_t and y_t for $t > 0$ inductively as follows. For $t \in [0, S_1)$ let $x_t = y_t = x_0$. Let $1 \leq j \leq N$. Given $(x_r : r < S_j)$, $(y_r : r < S_j)$, and $\tilde{x}_j, \tilde{y}_j$, define the following probability measures

$$\rho_j(\tilde{x}_j, w) := \begin{cases} G_{S_j}(\tilde{x}_j, w)/\lambda, & w \neq \tilde{x}_j \\ 1 - G_{S_j}(\tilde{x}_j, w)/\lambda, & w = \tilde{x}_j, \end{cases}$$
$$\rho'_j(\tilde{y}_j, w) := \begin{cases} H_{S_j}(\tilde{y}_j, w)/\lambda, & w \neq \tilde{y}_j \\ 1 - H_{S_j}(\tilde{y}_j, w)/\lambda, & w = \tilde{y}_j. \end{cases}$$

Sample $(\tilde{x}_{j+1}, \tilde{y}_{j+1})$ from a maximal coupling of (ρ_j, ρ'_j) and for $t \in [S_j, S_{j+1})$ set $x_t = \tilde{x}_{j+1}$, $y_t = \tilde{y}_{j+1}$. Finally set $x_T = x_{S_N}$ and $y_T = y_{S_N}$.

Now, observe that $(x_t, y_t)_{t\in[0,T]}$ defined in this way is a coupling of the given Markov chains. Moreover,

$$\begin{aligned} ||\nu K - \nu L||_{\text{TV}} &\leq \mathbb{P}(x_T \neq y_T) \\ &= \mathbb{E}\left[\sum_{j=1}^{N} \mathbb{I}\left\{x_s = y_s,\, s < S_j\right\} \mathbb{I}\left\{x_{S_j} \neq y_{S_j}\right\} \right] \\ &= \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \sum_{j=0}^{n} \mathbb{E}\left[\mathbb{I}\left\{x_s = y_s,\, s < S_j\right\} \mathbb{I}\left\{x_{S_j} \neq y_{S_j}\right\} \right] \end{aligned}$$

and using the fact that jumps are coupled maximally

$$
\begin{aligned}
&= \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \sum_{j=0}^{n} \mathbb{E}\Big[\mathbb{I}\{x_s = y_s,\, s < S_j\} \times \|\rho_j(X_{S_{j-1}}, \cdot) - \tilde{\rho}_j(X_{S_{j-1}}, \cdot)\|_{\mathrm{TV}}\Big] \\
&= \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \sum_{j=0}^{n} \mathbb{E}\Big[\mathbb{I}\{x_s = y_s,\, s < S_j\} \frac{1}{\lambda} \sum_z |G_{S_j}(x_{S_{j-1}}, z) - H_{S_j}(x_{S_{j-1}}, z)|\Big] \\
&= \frac{1}{\lambda} \mathbb{E}\Big[\sum_{s: x_s \neq x_{s-}} \sum_z |G_s(x_{s-}, z) - H_s(x_{s-}, z)|\Big] \\
&= \frac{1}{\lambda} \int_{s=0}^{T} \mathbb{E}\Big[\lambda \sum_z |G_s(x_{s-}, z) - H_s(x_{s-}, z)|\Big] \\
&= \int_{s=0}^{T} \mathbb{E}\Big[\sum_z |G_s(x_{s-}, z) - H_s(x_{s-}, z)|\Big] \mathrm{d}s
\end{aligned}
$$

as required. □

Proposition 6. *For all $t \in [0, T]$ and $x, y \in \mathcal{X}^D$ such that $x \sim y$, we have*

$$|\partial_t \hat{R}_t(x, y)| \leq 2|R|^2 SDC_1^2$$

Moreover, it follows that $\hat{R}_t$ is bounded and continuous in t.

Proof. Omitting the superscripts for brevity where the notation is clear, we have

$$
\begin{aligned}
\left|\partial_t \hat{R}_t^{1:D}(x^{1:D}, y^{1:D})\right| &= \left|R_t(y, x) \partial_t \left\{\frac{q_t(y)}{q_t(x)}\right\}\right| \\
&= \left|R_t(y, x) \left\{\frac{q_t(y)}{q_t(x)} \frac{\sum_z R_t(z, y) q_t(z)}{q_t(y)} - \frac{q_t(y)}{q_t(x)} \frac{\sum_z R_t(z, x) q_t(z)}{q_t(x)}\right\}\right| \\
&\leq 2|R|^2 SDC_1^2
\end{aligned}
$$

where the second line follows from Kolmogorov's forward equation and the final inequality follows from Assumption 2 plus the fact that $R_t(z, x)$ (resp. $R_t(z, y)$) is only non-zero when $x \sim z$ (resp. $y \sim z$), and there are at most $|S||D|$ values of x (resp. y) for which this holds. □

We now give the proof of Theorem 1.

Proof of Theorem 1. Let us label the time steps used in tau-leaping by $0 = t_0 < t_1 < \cdots < t_N = T$, denote $\tau_k = t_k - t_{k-1}$, and denote the target stationary distribution by $\pi^D(x^{1:D}) = \prod_{d=1}^{D} \pi(x^d)$, where π is the invariant distribution of the single-dimensional transition matrix R_t^1.

Also, let $\mathcal{R}_k^{\theta,(\tau)}$ be the Markov kernel corresponding to applying the tau-leaping approximation with rate matrix $\hat{R}_{t_k}^{\theta}$ to move from t_k to t_{k-1}, and denote $\mathcal{R}^{\theta,(\tau)} = \mathcal{R}_N^{\theta,(\tau)} \mathcal{R}_{N-1}^{\theta,(\tau)} \ldots \mathcal{R}_1^{\theta,(\tau)}$ so that $\mathcal{R}^{\theta,(\tau)}$ expresses the full dynamics of the tau-leaping process and we have $\mathcal{L}(\hat{y}_0) = \pi^D \mathcal{R}^{\theta,(\tau)}$.

Then, as in [19] we can decompose

$$||\pi^D \mathcal{R}^{\theta,(\tau)} - p_d||_{\mathrm{TV}} \leq ||\pi^D \mathcal{R}^{\theta,(\tau)} - \pi^D (\mathbb{P}^R)_{T|0}||_{\mathrm{TV}} + ||\pi^D - q_T||_{\mathrm{TV}}$$

where $\mathbb{P}^R$ is the path measure of the exact reverse process.

We deal with the second term first. Let t_{mix} be the (1/4)-mixing time of the single-dimension CTMC with rate matrix R_t^1, i.e.

$$t_{\mathrm{mix}} = \inf\left\{t \geq 0 : \sup_{x_0^1 \in S} ||q_{t|0}(\,\cdot\,|x_0^1) - \pi||_{\mathrm{TV}} \leq \frac{1}{4}\right\}$$

It then follows from

$$||q_{t|0}(\,\cdot\,|x_0^{1:D}) - \pi^D||_{\text{TV}} \leq \sum_{d=1}^{D} ||q_{t|0}(\,\cdot\,|x_0^d) - \pi||_{\text{TV}}$$

that t^D_{mix}, the (1/4)-mixing time of the full CTMC with rate matrix $R_t^{1:D}$, satisfies the inequality $t^D_{mix} \leq \{1 + \lceil \log_2 D \rceil\} t_{\text{mix}}$. If we view $(x_{mt^D_{mix}})_{m\in\mathbb{N}}$ as a discrete-time Markov chain, then standard results on Markov chain mixing (see, for example, Chapter 4.5 of [24]) show that

$$||q_{mt^D_{mix}|0}(\,\cdot\,|x_0^{1:D}) - \pi^D||_{\text{TV}} \leq 2^{-m}$$

It then follows that for any $T \geq 0$ we have

$$||\pi^D - q_T||_{\text{TV}} \leq 2\exp\left\{-\frac{T\log 2}{t^D_{mix}}\right\} \leq 2\exp\left\{-\frac{T\log^2 2}{t_{\text{mix}}\log 4D}\right\}$$

completing the bound on the second term.

To bound the first term, we define $\mathcal{P}_k = (\mathbb{P}^R)_{T-t_{k-1}|T-t_k}$ and decompose it as

$$\begin{aligned}
||\pi\mathcal{R}^{\theta,(\tau)} - \pi(\mathbb{P}^R)_{T|0}||_{\text{TV}} &\leq \sup_\nu ||\nu\mathcal{R}_N^{\theta,(\tau)} \dots \mathcal{R}_1^{\theta,(\tau)} - \nu\mathcal{P}_N \dots \mathcal{P}_1||_{\text{TV}} \\
&\leq \sup_\nu ||\nu\mathcal{R}_N^{\theta,(\tau)}\mathcal{R}_{N-1}^{\theta,(\tau)} \dots \mathcal{R}_1^{\theta,(\tau)} - \nu\mathcal{R}_N^{\theta,(\tau)}\mathcal{P}_{N-1} \dots \mathcal{P}_1||_{\text{TV}} \\
&\quad + \sup_\nu ||\nu\mathcal{R}_N^{\theta,(\tau)}\mathcal{P}_{N-1} \dots \mathcal{P}_1 - \nu\mathcal{P}_N\mathcal{P}_{N-1} \dots \mathcal{P}_1||_{\text{TV}} \\
&\leq \sup_\nu ||\nu\mathcal{R}_{N-1}^{\theta,(\tau)} \dots \mathcal{R}_1^{\theta,(\tau)} - \nu\mathcal{P}_{N-1} \dots \mathcal{P}_1||_{\text{TV}} + \sup_\nu ||\nu\mathcal{R}_N^{\theta,(\tau)} - \nu\mathcal{P}_N||_{\text{TV}} \\
&\leq \sum_{k=1}^{N} \sup_\nu ||\nu\mathcal{R}_k^{\theta,(\tau)} - \nu\mathcal{P}_k||_{\text{TV}}
\end{aligned}$$

by proceeding inductively. So it suffices to find bounds on the total variation distance accumulated on each interval $[t_{k-1}, t_k]$.

Let $\mathcal{R}_k^\theta$ be the Markov kernel corresponding to running the chain from t_k to t_{k-1} with constant rate matrix $\hat{R}^\theta_{t_k}$. Since by Proposition 6 the reverse rate matrix $\hat{R}_t$ is bounded and continuous in t, using Proposition 5 we made deduce that for any distribution ν on S we have

$$\begin{aligned}
||\nu\mathcal{P}_k - \nu\mathcal{R}_k^\theta||_{\text{TV}} &\leq \int_{t_{k-1}}^{t_k} \sup_{x\in S}\Big\{\sum_{y\neq x} \big|\hat{R}_t(x,y) - \hat{R}^\theta_{t_k}(x,y)\big|\Big\}\,\mathrm{d}t \\
&\leq \int_{t_{k-1}}^{t_k} \sup_{x\in S}\Big\{\sum_{y\neq x} \big|\hat{R}_t(x,y) - \hat{R}_{t_k}(x,y)\big|\Big\}\,\mathrm{d}t \\
&\quad + \int_{t_{k-1}}^{t_k} \sup_{x\in S}\Big\{\sum_{y\neq x} \big|\hat{R}_{t_k}(x,y) - \hat{R}^\theta_{t_k}(x,y)\big|\Big\}\,\mathrm{d}t
\end{aligned}$$

The first half of this expression can be bounded using the Mean Value Theorem, according to

$$\begin{aligned}
\int_{t_{k-1}}^{t_k} \sup_{x\in S}\Big\{\sum_{y\neq x} \big|\hat{R}_t(x,y) - \hat{R}_{t_k}(x,y)\big|\Big\}\,\mathrm{d}t &\leq \int_{t_{k-1}}^{t_k} |t - t_k| \cdot 2|R|^2 S^2 D^2 C_1^2\,\mathrm{d}t \\
&\leq \big(|R|SDC_1\tau_k\big)^2
\end{aligned}$$

where in the first line we have used that the summand is only non-zero when $y \sim x$, and there are at most $|S||D|$ values of y for which this holds. The second term can be bounded using condition (2), to get

$$\int_{t_{k-1}}^{t_k} \sup_{x\in S}\Big\{\sum_{y\neq x} \big|\hat{R}_{t_k}(x,y) - \hat{R}^\theta_{t_k}(x,y)\big|\Big\}\,\mathrm{d}t \leq M\tau_k$$

Combining these two expressions, we get a bound on $||\nu\mathcal{P}_k - \nu\mathcal{R}_k^\theta||_{\text{TV}}$.

$$||\nu\mathcal{P}_k - \nu\mathcal{R}_k^\theta||_{\text{TV}} \leq \left(|R|SDC_1\tau_k\right)^2 + M\tau_k$$

It remains to bound $||\nu\mathcal{R}_k^\theta - \nu\mathcal{R}_k^{\theta,(\tau)}||_{\text{TV}}$. Note that performing tau-leaping with rate matrix $\hat{R}_{t_k}^\theta$ starting in x_{t_k} is equivalent to running a continuous time Markov chain from time t_k to t_{k-1} with constant rate matrix $\hat{R}_{t_k}^{\theta,(\tau)}$ given by

$$\hat{R}_{t_k}^{\theta,(\tau)}(x, y) = \hat{R}_{t_k}^\theta(x_{t_k}, y - x + x_{t_k})$$

(followed potentially by a clamping operation to keep us within $\mathcal{X}^D$). By an analogous argument to the proof of Proposition 5,

$$||\delta_{x_{t_k}}\mathcal{R}_k^\theta - \delta_{x_{t_k}}\mathcal{R}_k^{\theta,(\tau)}||_{\text{TV}} \leq \int_{t_{k-1}}^{t_k} \mathbb{E}\Big[\sum_{y \neq x_t} |\hat{R}_{t_k}^\theta(x_t, y) - \hat{R}_{t_k}^\theta(x_{t_k}, y - x_t + x_{t_k})|\Big]\,\mathrm{d}t$$

where the expectation is taken over $(x_t)_{t\in[t_{k-1},t_k]}$ distributed according to the exact CTMC with rate matrix $\hat{R}_{t_k}^\theta$. (Note we have disregarded the clamping operation, since this can only decrease the resulting total variation distance.)

We may rewrite this bound in terms of the exact reverse process using condition (2) to get

$$||\delta_{x_{t_k}}\mathcal{R}_k^\theta - \delta_{x_{t_k}}\mathcal{R}_k^{\theta,(\tau)}||_{\text{TV}} \leq \int_{t_{k-1}}^{t_k} \mathbb{E}\Big[2M + \sum_{y \neq x_t} |\hat{R}_{t_k}(x_t, y) - \hat{R}_{t_k}(x_{t_k}, y - x_t + x_{t_k})|\Big]\,\mathrm{d}t$$

Let J_t be the number of jumps that (x_t) makes between t_k and t, and label the times of these jumps as $s_1, \ldots, s_j$ where $t \leq s_1 \leq \cdots \leq s_j \leq t_k$ and $j = J_t$ for convenience. Then by Assumption 3, we have

$$\begin{aligned}\sum_{y \neq x_t} |\hat{R}_{t_k}(x_t, y) - \hat{R}_{t_k}(x_{t_k}, y - x_t + x_{t_k})| &\leq \sum_z |\hat{R}_{t_k}(x_t, x_t + z) - \hat{R}_{t_k}(x_{s_1}, x_{s_1} + z)| + \ldots \\ &\quad + \sum_z |\hat{R}_{t_k}(x_{s_j}, x_{s_j} + z) - \hat{R}_{t_k}(x_{t_k}, x_{t_k} + z)| \\ &\leq C_2 J_t\end{aligned}$$

where we have made the substitution $z = y - x_{t_k}$. We conclude that

$$\begin{aligned}||\delta_{x_{t_k}}\mathcal{R}_k^\theta - \delta_{x_{t_k}}\mathcal{R}_k^{\theta,(\tau)}||_{\text{TV}} &\leq \int_{t_{k-1}}^{t_k} \mathbb{E}\left[2M + C_2 J_t\right]\,\mathrm{d}t \\ &\leq 2M|t_k - t_{k-1}| + C_2 \int_{t_{k-1}}^{t_k} |t_k - t| \cdot \sup_x |\hat{R}_{t_k}^\theta(x, x)|\,\mathrm{d}t \\ &\leq 2M\gamma_k + \frac{1}{2}C_2|\hat{R}_{t_k}^\theta|\tau_k^2 \\ &\leq 2M\tau_k + \frac{1}{2}C_2(M + C_1SD|R|)\tau_k^2\end{aligned}$$

where to bound $\mathbb{E}[J_t]$ we have observed that jumps of (x_t) occur at a rate bounded above by $\sup_x |\hat{R}_{t_k}^\theta(x, x)|$, and in the last line we have used the condition (2) and Assumption 2. Since the above holds for any choice of x_{t_k}, it follows that

$$\sup_\nu ||\nu\mathcal{R}_k^\theta - \nu\mathcal{R}_k^{\theta,(\tau)}||_{\text{TV}} \leq 2M\tau_k + \frac{1}{2}C_2(M + C_1SD|R|)\tau_k^2$$

Summing over k and putting all our bounds together, we get

$$||\mathcal{L}(y_0) - p_{\text{data}}||_{\text{TV}} \leq 3MT + \left\{\left(|R|SDC_1\right)^2 + \tfrac{1}{2}C_2(M + C_1SD|R|)\right\}\tau T + 2\exp\left\{-\tfrac{T\log^2 2}{t_{\text{mix}}\log 4D}\right\}$$

as required. □

C Continuous Time ELBO Details

C.1 Comparison with the Discrete Time ELBO

It is easiest to gain intuition on the $\mathcal{L}_{\text{CT}}$ objective by comparing it to its discrete time counterpart, $\mathcal{L}_{\text{DT}}$, and examining the way in which $\mathcal{L}_{\text{DT}}$ in the limit becomes $\mathcal{L}_{\text{CT}}$ when we take the time step size to be very small. We repeat the definition of $\mathcal{L}_{\text{CT}}$ here for convenience

$$\mathcal{L}_{\text{CT}}(\theta) = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)q_t(x)r_t(\tilde{x}|x)}\left[\left\{\sum_{x'\neq x}\hat{R}_t^\theta(x,x')\right\} - \mathcal{Z}^t(x)\log\left(\hat{R}_t^\theta(\tilde{x},x)\right)\right] + C.$$

Recall that a single term from the KL sum in $\mathcal{L}_{\text{DT}}$ up to an additive constant independent of θ is

$$-\mathbb{E}_{q_k(x_k)q_{k+1|k}(x_{k+1}|x_k)}\left[\log p^\theta_{k|k+1}(x_k|x_{k+1})\right].$$

Minimizing this term is to sample (x_k, x_{k+1}) from the forward dynamics and then maximize the assigned model probability for the pairing in the reverse direction. A similar idea can be used to understand $\mathcal{L}_{\text{CT}}$. First, we write $\log p^\theta_{k|k+1}(x_k|x_{k+1})$ in terms of $\hat{R}^\theta_k$ as

$$\begin{aligned}\log p^\theta_{k|k+1}(x_k|x_{k+1}) =& \delta_{x_k,x_{k+1}}\left(\hat{R}^\theta_k(x_k,x_k)\Delta t + o(\Delta t)\right)\\ &+ (1-\delta_{x_k,x_{k+1}})\log\left(\hat{R}^\theta_k(x_{k+1},x_k)\Delta t + o(\Delta t)\right)\end{aligned}$$

where we have separated the cases when $x_k = x_{k+1}$ and when $x_k \neq x_{k+1}$ (see the proof of $\mathcal{L}_{\text{CT}}$ for the full details). The first term will become the $\sum_{x'\neq x}\hat{R}^\theta_t(x,x')$ term in $\mathcal{L}_{\text{CT}}$ whilst the second term will become the $\mathcal{Z}^t(x)\log\left(\hat{R}^\theta_t(\tilde{x},x)\right)$ term. Now, when we minimize $\mathcal{L}_{\text{CT}}$, we are sampling $(x,\tilde{x})$ from the forward process and then maximizing the assigned model probability for the pairing in the reverse direction, just as in $\mathcal{L}_{\text{DT}}$. The slight extra complexity comes from the fact we are considering the case when $x_k = x_{k+1}$ and the case when $x_k \neq x_{k+1}$ separately. When $x_k = x_{k+1}$, this corresponds to the first term in $\mathcal{L}_{\text{CT}}$ which we can see is minimizing the reverse rate out of x which is exactly maximizing the model probability for no transition to occur. When $x_k \neq x_{k+1}$, this corresponds to the second term in $\mathcal{L}_{\text{CT}}$, which is maximizing the reverse rate from $\tilde{x}$ to x which in turn maximizes the model probability for the $\tilde{x}$ to x transition to occur.

C.2 Conditional Form

For the conditional form of $\mathcal{L}_{\text{CT}}$, denoted as $\bar{\mathcal{L}}_{\text{CT}}$, we instead upper bound the negative conditional model log-likelihood, $\mathbb{E}_{p_{\text{data}}(x_0,y)}[-\log p^\theta_0(x_0|y)]$ where y is our conditioner. $\bar{\mathcal{L}}_{\text{CT}}$ has the following form

$$\bar{\mathcal{L}}_{\text{CT}}(\theta) = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)p_{\text{data}}(x_0,y)q_{t|0}(x|x_0)r_t(\tilde{x}|x)}\left[\left\{\sum_{x'\neq x}\hat{R}^\theta_t(x,x'|y)\right\} - \mathcal{Z}^t(x)\log\left(\hat{R}^\theta_t(\tilde{x},x|y)\right)\right] + C,$$

where

$$\begin{aligned}\hat{R}^\theta_t(x,\tilde{x}|y) =& R_t(\tilde{x},x)\sum_{x_0}\frac{q_{t|0}(\tilde{x}|x_0)}{q_{t|0}(x|x_0)}p^\theta_{0|t}(x_0|x,y) \quad \text{for} \quad x\neq\tilde{x}.\\ =& -\sum_{x'\neq x}\hat{R}^\theta_t(x,x'|y) \quad \text{for} \quad x=\tilde{x}\end{aligned}$$

This follows easily from considering the conditional form of the discret time ELBO, $\bar{\mathcal{L}}_{\text{DT}}$ and using the same arguments as before to go from discrete time to continuous time.

$$\mathbb{E}_{p_{\text{data}}(x_0,y)}[-\log p^\theta_0(x_0|y)] \leq \mathbb{E}_{p_{\text{data}}(x_0,y)q_{1:K|0}(x_{1:K}|x_0)}\left[-\log\frac{p^\theta_{0:K}(x_{0:K}|y)}{q_{1:K|0}(x_{1:K}|x_0)}\right] = \bar{\mathcal{L}}_{\text{DT}}$$

C.3 Continuous Time ELBO with Factorization Assumptions

In the following Proposition, we show the form of $\mathcal{L}_{\text{CT}}$ when we use a factorized forward process. We note that in the proof we rearrange the sampling distribution from $p_{\text{data}}(\boldsymbol{x}_0^{1:D})q_{t|0}(\boldsymbol{x}^{1:D}|\boldsymbol{x}_0^{1:D})r_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}^{1:D})$ to $p_{\text{data}}(\boldsymbol{x}_0^{1:D})\psi_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}_0^{1:D})\phi_t(\boldsymbol{x}^{1:D}|\tilde{\boldsymbol{x}}^{1:D},\boldsymbol{x}_0^{1:D})$. This is not strictly necessary but it allows us to analytically sum over the intermediate $\boldsymbol{x}^{1:D}$ variable which greatly reduces the variance of the resulting objective.

Proposition 7. *The $\mathcal{L}_{\text{CT}}$ objective when we substitute in the factorized forms for the forward and reverse process given in Proposition 3 is*

$$\mathcal{L}_{\text{CT}} = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)p_{\text{data}}(\boldsymbol{x}_0^{1:D})q_{t|0}(\boldsymbol{x}^{1:D}|\boldsymbol{x}_0^{1:D})}\left[\sum_{d=1}^{D}\sum_{x'^d\neq x^d}\hat{R}_t^{\theta\,d}(\boldsymbol{x}^{1:D},x'^d)\right]$$
$$- T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)p_{\text{data}}(\boldsymbol{x}_0^{1:D})\psi_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}_0^{1:D})}\left[\sum_{d=1}^{D}\sum_{x^d\neq\tilde{x}^d}\phi_t(x^d|\tilde{\boldsymbol{x}}^{1:D},\boldsymbol{x}_0^{1:D})\mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D/d}\circ x^d)\log\left(\hat{R}_t^{\theta\,d}(\tilde{\boldsymbol{x}}^{1:D},x^d)\right)\right]$$
$$+ C$$

with

$$\hat{R}_t^{\theta\,d}(\boldsymbol{x}^{1:D},\tilde{x}^d) = R_t^d(\tilde{x}^d,x^d)\sum_{x_0^d}p_{0|t}^{\theta}(x_0^d|\boldsymbol{x}^{1:D})\frac{q_{t|0}(\tilde{x}^d|x_0^d)}{q_{t|0}(x^d|x_0^d)}$$

$$\mathcal{Z}^t(\boldsymbol{x}^{1:D}) = \sum_{d=1}^{D}\sum_{\tilde{x}^d\neq x^d}R_t^d(x^d,\tilde{x}^d)$$

$$\phi_t(x^d|\tilde{\boldsymbol{x}}^{1:D},\boldsymbol{x}_0^{1:D}) = \frac{R_t^d(x^d,\tilde{x}^d)q_{t|0}(\tilde{\boldsymbol{x}}^{1:D\setminus d}\circ x^d|\boldsymbol{x}_0^{1:D})}{\mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D\setminus d}\circ x^d)\sum_{d'=1}^{D}\sum_{x'^{d'}\neq\tilde{x}^{d'}}\frac{R_t^{d'}(x'^{d'},\tilde{x}^{d'})}{\mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D\setminus d'}\circ x'^{d'})}q_{t|0}(\tilde{\boldsymbol{x}}^{1:D\setminus d'}\circ x'^{d'}|\boldsymbol{x}_0^{1:D})}$$

where $\circ$ represents the concatenation of a $D-1$ dimensional vector, $\boldsymbol{x}^{1:D\setminus d}$ with a scalar x^d, such that the resultant D dimensional vector has x^d at its d^{th} dimension. $\psi_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}_0^{1:D})$ is defined as the marginal of the forward noising process joint, $\int q_{t|0}(\boldsymbol{x}^{1:D}|\boldsymbol{x}_0^{1:D})r_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}^{1:D})d\boldsymbol{x}^{1:D}$.

Proof. We first re-write the general form of $\mathcal{L}_{\text{CT}}$ here

$$\mathcal{L}_{\text{CT}}(\theta) = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)q_t(x)r_t(\tilde{x}|x)}\left[\left\{\sum_{x'\neq x}\hat{R}_t^{\theta}(x,x')\right\} - \mathcal{Z}^t(x)\log\left(\hat{R}_t^{\theta}(\tilde{x},x)\right)\right] + C$$

where

$$\mathcal{Z}^t(x) = \textstyle\sum_{x'\neq x}R_t(x,x') \qquad\qquad r_t(\tilde{x}|x) = (1-\delta_{\tilde{x},x})R_t(x,\tilde{x})/\mathcal{Z}^t(x).$$

With a factorized forward process, $\hat{R}_t^{\theta}$ becomes

$$\hat{R}_t^{\theta\,1:D}(\boldsymbol{x}^{1:D},\tilde{\boldsymbol{x}}^{1:D}) = \sum_{d=1}^{D}\hat{R}_t^{\theta\,d}(\boldsymbol{x}^{1:D},\tilde{x}^d)\delta_{\boldsymbol{x}^{1:D\setminus d},\tilde{\boldsymbol{x}}^{1:D\setminus d}}$$

where

$$\hat{R}_t^{\theta\,d}(\boldsymbol{x}^{1:D},\tilde{x}^d) = R_t^d(\tilde{x}^d,x^d)\sum_{x_0^d}p_{0|t}^{\theta}(x_0^d|\boldsymbol{x}^{1:D})\frac{q_{t|0}(\tilde{x}^d|x_0^d)}{q_{t|0}(x^d|x_0^d)}$$

Substituting this form for $\hat{R}_t^{\theta\,1:D}$ into the first term in $\mathcal{L}_{\text{CT}}$ we get

$$\sum_{\boldsymbol{x}'^{1:D}\neq\boldsymbol{x}^{1:D}}\sum_{d=1}^{D}\hat{R}_t^{\theta\,d}(\boldsymbol{x}^{1:D},x'^d)\delta_{\boldsymbol{x}^{1:D\setminus d},\boldsymbol{x}'^{1:D\setminus d}}$$
$$= \sum_{d=1}^{D}\sum_{x'^d}\hat{R}_t^{\theta\,d}(\boldsymbol{x}^{1:D},x'^d)\sum_{\boldsymbol{x}'^{1:D\setminus d}}\delta_{\boldsymbol{x}^{1:D\setminus d},\boldsymbol{x}'^{1:D\setminus d}}(1-\delta_{\boldsymbol{x}'^{1:D},\boldsymbol{x}^{1:D}})$$
$$= \sum_{d=1}^{D}\sum_{x'^d\neq x^d}\hat{R}_t^{\theta\,d}(\boldsymbol{x}^{1:D},x'^d)$$

Now we tackle the second term in $\mathcal{L}_{\text{CT}}$. We first re-arrange the distribution over which we take the expectation:

$$p_{\text{data}}(\boldsymbol{x}_0^{1:D})q_{t|0}(\boldsymbol{x}^{1:D}|\boldsymbol{x}_0^{1:D})r_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}^{1:D}) = p_{\text{data}}(\boldsymbol{x}_0^{1:D})\psi_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}_0^{1:D})\phi_t(\boldsymbol{x}^{1:D}|\tilde{\boldsymbol{x}}^{1:D},\boldsymbol{x}_0^{1:D})$$

We have,

$$\begin{aligned}\phi_t(\boldsymbol{x}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}_0^{1:D}) &\propto q_{t|0}(\boldsymbol{x}^{1:D}|\boldsymbol{x}_0^{1:D}) r_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}^{1:D}) \\ &= q_{t|0}(\boldsymbol{x}^{1:D}|\boldsymbol{x}_0^{1:D})(1 - \delta_{\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}^{1:D}}) \frac{\sum_{d=1}^D R_t^d(x^d, \tilde{x}^d) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}}}{\mathcal{Z}^t(\boldsymbol{x}^{1:D})} \\ &= \sum_{d=1}^D \frac{R_t^d(x^d, \tilde{x}^d)}{\mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D\setminus d} \circ x^d)} q_{t|0}(\tilde{\boldsymbol{x}}^{1:D\setminus d} \circ x^d | \boldsymbol{x}_0^{1:D}) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}} (1 - \delta_{\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}^{1:D}})\end{aligned}$$

To find the normalization constant, we can sum the proportional term over $\boldsymbol{x}^{1:D}$

$$\begin{aligned}&\sum_{\boldsymbol{x}^{1:D}} \sum_{d=1}^D \frac{R_t^d(x^d, \tilde{x}^d)}{\mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D\setminus d} \circ x^d)} q_{t|0}(\tilde{\boldsymbol{x}}^{1:D\setminus d} \circ x^d | \boldsymbol{x}_0^{1:D}) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}} (1 - \delta_{\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}^{1:D}}) \\ &= \sum_{d=1}^D \sum_{x^d \neq \tilde{x}^d} \frac{R_t^d(x^d, \tilde{x}^d)}{\mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D\setminus d} \circ x^d)} q_{t|0}(\tilde{\boldsymbol{x}}^{1:D\setminus d} \circ x^d | \boldsymbol{x}_0^{1:D})\end{aligned}$$

Therefore,

$$\phi_t(\boldsymbol{x}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}_0^{1:D}) = (1 - \delta_{\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}^{1:D}}) \sum_{d=1}^D \phi_t(x^d|\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}_0^{1:D}) \delta_{\boldsymbol{x}^{1:D\setminus d}, \tilde{\boldsymbol{x}}^{1:D\setminus d}}$$

where

$$\phi_t(x^d|\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}_0^{1:D}) = \frac{R_t^d(x^d, \tilde{x}^d) q_{t|0}(\tilde{\boldsymbol{x}}^{1:D\setminus d} \circ x^d | \boldsymbol{x}_0^{1:D})}{\mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D\setminus d} \circ x^d) \sum_{d'=1}^D \sum_{x'^{d'} \neq \tilde{x}^{d'}} \frac{R_t^{d'}(x'^{d'}, \tilde{x}^{d'})}{\mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D\setminus d'} \circ x'^{d'})} q_{t|0}(\tilde{\boldsymbol{x}}^{1:D\setminus d'} \circ x'^{d'} | \boldsymbol{x}_0^{1:D})}$$

Now we write the second term as

$$\begin{aligned}&T\, \mathbb{E}_{t\sim\mathcal{U}(0,T) p_{\text{data}}(\boldsymbol{x}_0^{1:D}) \psi_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}_0^{1:D})} \left[- \sum_{\boldsymbol{x}^{1:D}} \phi_t(\boldsymbol{x}^{1:D}|\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}_0^{1:D}) \mathcal{Z}^t(\boldsymbol{x}^{1:D}) \log \hat{R}_t^{\theta\, 1:D}(\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}^{1:D}) \right] \\ &= -T\, \mathbb{E}_{t\sim\mathcal{U}(0,T) p_{\text{data}}(\boldsymbol{x}_0^{1:D}) \psi_t(\tilde{\boldsymbol{x}}^{1:D}|\boldsymbol{x}_0^{1:D})} \left[\sum_{d=1}^D \sum_{x^d \neq \tilde{x}^d} \phi_t(x^d|\tilde{\boldsymbol{x}}^{1:D}, \boldsymbol{x}_0^{1:D}) \mathcal{Z}^t(\tilde{\boldsymbol{x}}^{1:D/d} \circ x^d) \log \left(\hat{R}_t^{\theta\, d}(\tilde{\boldsymbol{x}}^{1:D}, x^d) \right) \right]\end{aligned}$$

□

C.4 One Forward Pass

To evaluate the $\mathcal{L}_{\text{CT}}$ objective, we naively need to perform two forward passes of the denoising network: $p_{0|t}^\theta(x_0|x)$ to calculate $\hat{R}_t^\theta(x, x')$ and $p_{0|t}^\theta(x_0|\tilde{x})$ to calculate $\hat{R}_t^\theta(\tilde{x}, x)$. This is wasteful because $\tilde{x}$ is created from x by applying a single forward transition which on multi-dimensional problems means $\tilde{x}$ differs from x in only a single dimension. To exploit the fact that $\tilde{x}$ and x are very similar, we approximate the sample $x \sim q_t(x)$ with the sample $\tilde{x} \sim \sum_x q_t(x) r_t(\tilde{x}|x)$. This gives the more efficient objective,

$$\mathcal{L}_{\text{eCT}}(\theta) = T\, \mathbb{E}_{t\sim\mathcal{U}(0,T) q_t(x) r_t(\tilde{x}|x)} \left[\left\{ \sum_{x' \neq \tilde{x}} \hat{R}_t^\theta(\tilde{x}, x') \right\} - \mathcal{Z}^t(x) \log \left(\hat{R}_t^\theta(\tilde{x}, x) \right) \right] + C$$

Table 3: Metrics on the monophonic music dataset comparing training with the efficient $\mathcal{L}_{\text{eCT}}$ objective vs the original $\mathcal{L}_{\text{CT}}$ objective. We compute these over the test set showing mean$\pm$std with respect to 5 samples for each test song.

Model	Hellinger Distance	Proportion of Outliers
τLDR-0 Uniform $\mathcal{L}_{\text{eCT}}$	0.3765 ± 0.0013	0.1106 ± 0.0010
τLDR-0 Uniform $\mathcal{L}_{\text{CT}}$	0.3797 ± 0.0009	0.1128 ± 0.0007

The approximation is valid because $q_t(x)$ and $\sum_x q_t(x) r_t(\tilde{x}|x)$ are very similar distributions, as we now show.

$$\begin{aligned}
\sum_x q_t(x) r_t(\tilde{x}|x) &= \sum_x q_t(x)(1 - \delta_{x,\tilde{x}}) \frac{R_t(x,\tilde{x})}{\sum_{x' \neq x} R_t(x,x')} \\
&\propto -q_t(\tilde{x}) R_t(\tilde{x},\tilde{x}) + \sum_x q_t(x) R_t(x,\tilde{x}) \\
&= q_t(\tilde{x}) \sum_{x' \neq \tilde{x}} R_t(\tilde{x},x') + \partial_t q_t(\tilde{x}) \\
&\propto q_t(\tilde{x}) + \frac{1}{\sum_{x' \neq x} R_t(\tilde{x},x')} \partial_t q_t(\tilde{x}) \\
&= q_t(\tilde{x}) + \delta t \, \partial_t q_t(\tilde{x})
\end{aligned}$$

where on the third line we have used the Kolmogorov forward equation and defined $\delta_t = 1/\sum_{x' \neq x} R_t(\tilde{x},x')$. The distribution $\sum_x q_t(x) r_t(\tilde{x}|x)$ is therefore $q_{t+\delta t}(\tilde{x})$ approximated using a first-order Taylor expansion around $q_t(\tilde{x})$. We notice that δt is the average time to the next transition at time t. δt can be calculated for the practical settings we consider, its varies between $2\times10^{-6}T$ and $2\times10^{-8}T$ in the image modelling task and is $1\times10^{-3}T$ in the monophonic music task.

We perform an ablation experiment comparing between training with $\mathcal{L}_{\text{eCT}}$ and $\mathcal{L}_{\text{CT}}$ on the monophonic music dataset, the results are shown in Table 3. We find that we gain a small boost in performance when using the more efficient $\mathcal{L}_{\text{eCT}}$ objective alongside the improved efficiency. We hypothesize that this is because of a slight reduction in variance for the $\mathcal{L}_{\text{eCT}}$ objective due to increased negative correlation between the two terms in the objective when $\tilde{x}$ is shared between them.

D Direct Denoising Model Supervision

Following [8], we can introduce direct $p^\theta_{0|t}$ supervision into the optimization objective which has been found empirically to improve performance. We first contextualize the change by expressing $\mathcal{L}_{\text{CT}}$ with the dependence on $p^\theta_{0|t}$ made explicit.

$$\begin{aligned}
\mathcal{L}_{\text{CT}} = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T) q_t(x) r_t(\tilde{x}|x)} \Big[\Big\{ \sum_{x' \neq x} R_t(x',x) \sum_{x_0} \frac{q_{t|0}(x'|x_0)}{q_{t|0}(x|x_0)} p^\theta_{0|t}(x_0|x) \Big\} \\
- \mathcal{Z}^t(x) \log \Big(R_t(x,\tilde{x}) \sum_{x_0} \frac{q_{t|0}(x|x_0)}{q_{t|0}(\tilde{x}|x_0)} p^\theta_{0|t}(x_0|\tilde{x}) \Big) \Big] + C
\end{aligned}$$

The signal for $p^\theta_{0|t}(x_0|x)$ comes through a sum over x_0 weighted by the ratio $\frac{q_{t|0}(x|x_0)}{q_{t|0}(\tilde{x}|x_0)}$. We can also provide a direct denoising signal by predicting the clean datapoint x_0 from the corrupted version x and using the negative log-likelihood loss.

$$L_{ll}(\theta) = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T) p_{\text{data}}(x_0) q_{t|0}(x|x_0)} \Big[-\log p^\theta_{0|t}(x_0|x) \Big]$$

Proposition 8. *The true denoising distribution, $q_{0|t}$, minimizes L_{ll}*

Proof.

$$T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)q_t(x)}\left[\mathrm{KL}\left(q_{0|t}(x_0|x)\,||\,p^\theta_{0|t}(x_0|x)\right)\right]$$
$$= T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)p_{\text{data}}(x_0)q_{t|0}(x|x_0)}\left[-\log p^\theta_{0|t}(x_0|x)\right] + C$$

where C is a constant independent of θ. Therefore, minimizing L_{ll} is equivalent to minimizing the KL divergence between $q_{0|t}$ and $p^\theta_{0|t}$, which is minimized when $p^\theta_{0|t} = q_{0|t}$. □

If we obtain the true denoising distribution, $p^\theta_{0|t} = q_{0|t}$, then we will have the true reverse rate, $\hat{R}_t$. [8] find that optimizing with an objective combining L_{ll} and $\mathcal{L}_{\text{DT}}$ performs best, which we can also do in continuous time

$$\min_\theta \quad \mathcal{L}_{\text{CT}}(\theta) + \lambda L_{ll}(\theta)$$

where λ is a hyperparameter. In [8], it was found that training with L_{ll} alone resulted in poorer performance than when the ELBO was included in the loss. We provide a theoretical hypothesis as to why this may be the case here. We show that minimizing L_{ll} is equivalent to minimizing an upper bound on the negative ELBO in discrete time and thus by training with L_{ll} we are simply minimizing a looser bound on the negative model log-likelihood than if we were to use the negative ELBO directly.

Proposition 9. *Minimizing the sum of negative log-likelihoods*

$$\sum_{k=0}^{K-1}\mathbb{E}_{p_{\text{data}}(x_0)q_{k+1|0}(x_{k+1}|x_0)}\left[-\log p^\theta_{0|k+1}(x_0|x_{k+1})\right]$$

is equivalent to minimizing an upper bound on the negative ELBO.

Proof.

$$\mathcal{L}_{\text{DT}}(\theta) = \mathbb{E}_{p_{\text{data}}(x_0)}\Big[\mathrm{KL}(q_{K|0}(x_K|x_0)||p_{\text{ref}}(x_K)) - \mathbb{E}_{q_{1|0}(x_1|x_0)}\left[\log p^\theta_{0|1}(x_0|x_1)\right]$$
$$+ \textstyle\sum_{k=1}^{K-1}\mathbb{E}_{q_{k+1|0}(x_{k+1}|x_0)}\left[\mathrm{KL}(q_{k|k+1,0}(x_k|x_{k+1},x_0)||p^\theta_{k|k+1}(x_k|x_{k+1}))\right]\Big]$$

Consider one term from the sum

$$L_k = \mathbb{E}_{p_{\text{data}}(x_0)q_{k+1|0}(x_{k+1}|x_0)}\left[\mathrm{KL}(q_{k|k+1,0}(x_k|x_{k+1},x_0)||p^\theta_{k|k+1}(x_k|x_{k+1})\right]$$
$$= \mathbb{E}_{q_{k+1}(x_{k+1})q_{k|k+1}(x_k|x_{k+1})}\left[-\log p^\theta_{k|k+1}(x_k|x_{k+1})\right]$$
$$+ \mathbb{E}_{p_{\text{data}}(x_0)q_{k+1|0}(x_{k+1}|x_0)q_{k|k+1,0}(x_k|x_{k+1},x_0)}\left[\log q_{k|k+1,0}(x_k|x_{k+1},x_0)\right]$$

Now,

$$\mathbb{E}_{q_{k+1}(x_{k+1})q_{k|k+1}(x_k|x_{k+1})}\left[-\log p^\theta_{k|k+1}(x_k|x_{k+1})\right]$$
$$= \mathbb{E}_{q_{k+1}(x_{k+1})q_{k|k+1}(x_k|x_{k+1})}\left[-\log\sum_{\tilde{x}_0} q(x_k|\tilde{x}_0,x_{k+1})p^\theta_{0|k+1}(\tilde{x}_0|x_{k+1})\right]$$
$$= \mathbb{E}_{q_{k+1}(x_{k+1})q_{k|k+1}(x_k|x_{k+1})}\left[-\log\sum_{\tilde{x}_0}\frac{q_{0|k}(\tilde{x}_0|x_k)q_{k|k+1}(x_k|x_{k+1})}{q_{0|k+1}(\tilde{x}_0|x_{k+1})}p^\theta_{0|k+1}(\tilde{x}_0|x_{k+1})\right]$$
$$\leq \mathbb{E}_{q_{k+1}(x_{k+1})q_{k|k+1}(x_k|x_{k+1})q_{0|k}(\tilde{x}_0|x_k)}\left[-\log\frac{q_{k|k+1}(x_k|x_{k+1})}{q_{0|k+1}(\tilde{x}_0|x_{k+1})}p^\theta_{0|k+1}(\tilde{x}_0|x_{k+1})\right]$$
$$= \mathbb{E}_{p_{\text{data}}(x_0)q_{k+1|0}(x_{k+1}|x_0)}\left[-\log p^\theta_{0|k+1}(x_0|x_{k+1})\right]$$
$$+ \mathbb{E}_{p_{\text{data}}(x_0)q_{k|0}(x_k|x_0)q_{k+1|k}(x_{k+1}|x_k)}\left[-\log\frac{q_{k|k+1}(x_k|x_{k+1})}{q_{0|k+1}(x_0|x_{k+1})}\right]$$

Therefore,

$$\begin{aligned}\mathcal{L}_{\text{DT}} \leq \sum_{k=0}^{K-1} \left\{ \mathbb{E}_{p_{\text{data}}(x_0) q_{k+1|0}(x_{k+1}|x_0)} \left[-\log p^{\theta}_{0|k+1}(x_0|x_{k+1}) \right] \right\} \\ + \mathbb{E}_{p_{\text{data}}(x_0)} \left[\text{KL}(q_{K|0}(x_K|x_0)||p_K(x_K)) \right] \\ + \sum_{k=1}^{K-1} \left\{ \mathbb{E}_{p_{\text{data}}(x_0) q_{k+1|0}(x_{k+1}|x_0) q_{k|k+1,0}(x_k|x_{k+1},x_0)} \left[\log q_{k|k+1,0}(x_k|x_{k+1}, x_0) \right] \right. \\ \left. + \mathbb{E}_{p_{\text{data}}(x_0) q_{k|0}(x_k|x_0) q_{k+1|k}(x_{k+1}|x_k)} \left[-\log \frac{q_{k|k+1}(x_k|x_{k+1})}{q_{0|k+1}(x_0|x_{k+1})} \right] \right\}\end{aligned}$$

We can see that only the first term depends on θ. □

E Choice of Forward Process

We need to choose the structure of R_t such that we can analytically obtain $q_{t|0}$ marginals to enable efficient training.

Proposition 10. *If R_t and $R_{t'}$ commute for all t, t' then $q_{t|0}(x = j|x_0 = i) = \left(exp\left[\int_0^t R_s ds\right]\right)_{ij}$ where exp here is understood to be the matrix exponential function.*

Proof. Let $(P_t)_{ij} = q_{t|0}(x = j|x_0 = i)$. We show that $P_t = \exp\left(\int_0^t R_s ds\right)$ is a solution to the Kolmogorov forward equation, which in matrix form reads, $\partial_t P_t = P_t R_t$. Writing the matrix exponential in sum form

$$\begin{aligned}P_t &= \sum_{k=0}^{\infty} \frac{1}{k!} \left(\int_0^t R_s ds \right)^k \\ &= \text{Id} + \int_0^t R_s ds + \frac{1}{2!} \left(\int_0^t R_s ds \right)^2 + \frac{1}{3!} \left(\int_0^t R_s ds \right)^3 + \dots\end{aligned}$$

Now, differentiating and using the fact that R_t, $R_{t'}$ commute.

$$\begin{aligned}\partial_t P_t &= R_t + \int_0^t R_s ds R_t + \frac{1}{2!} \left(\int_0^t R_s ds \right)^2 R_t + \dots \\ &= \left\{ \sum_{k=0}^{\infty} \frac{1}{k!} \left(\int_0^t R_s ds \right)^k \right\} R_t \\ &= P_t R_t\end{aligned}$$

□

As stated in the main text, we achieve the commutative property by selecting $R_t = \beta(t) R_b$ where $\beta(t)$ is a time dependent scalar and R_b is a constant base matrix. We can utilize the eigendecomposition

of $R_b = Q\Lambda Q^{-1}$ to efficiently calculate P_t,

$$\begin{aligned}
P_t &= \exp\left(\int_0^t \beta(s) R_b ds\right) \\
&= \sum_{k=0}^{\infty} \frac{1}{k!}\left(\int_0^t \beta(s) R_b ds\right)^k \\
&= \sum_{k=0}^{\infty} \frac{1}{k!}\left(Q\Lambda Q^{-1}\int_0^t \beta(s) ds\right)^k \\
&= \sum_{k=0}^{\infty} \frac{1}{k!} Q\left(\Lambda\int_0^t \beta(s) ds\right)^k Q^{-1} \\
&= Q\left\{\sum_{k=0}^{\infty} \frac{1}{k!}\left(\Lambda\int_0^t \beta(s) ds\right)^k\right\} Q^{-1} \\
&= Q\exp\left[\Lambda\int_0^t \beta(s) ds\right] Q^{-1}
\end{aligned}$$

Since Λ is a diagonal matrix, the matrix exponential coincides with the element wise exponential making the final expression tractable to compute. We choose $\beta(t) = ab^t \log b$ because this makes the integral which dictates the variance of $q_{t|0}$ have a simple form $\int_0^t \beta(s)ds = ab^t - a$.

For categorical problems, we found a uniform rate matrix works well, $R_t = \beta \mathbb{1}\mathbb{1}^T - \beta S\mathrm{Id}$. This is directly analogous to the discrete time uniform transition matrix: $P = \alpha \mathbb{1}\mathbb{1}^T + (1 - S\alpha)\mathrm{Id}$ with α depending on the time discretization used. Indeed, if one calculates the corresponding discrete transition matrix for the uniform R_t rate through the matrix exponential, the uniform transition matrix is obtained. Another categorical corruption process is the absorbing state process. In discrete time, the transition matrix is given by $P = \alpha \mathbb{1}\mathbf{e}_*^T + (1-\alpha)\mathrm{Id}$ where $\mathbf{e}_*$ is the one-hot encoding of the absorbing state. The corresponding absorbing state continuous time process has transition rate matrix: $R_t = \beta \mathbf{1}\mathbf{e}_*^T - \beta\mathrm{Id}$. The correspondence for more complex transition matrices e.g. the Discretized Gaussian matrix in [8] is much harder to find analytically especially if the time inhomogeneous case is considered. For datasets with an ordinal structure, we construct a new rate matrix that maintains a bias towards nearby states using a similar approach as that taken by [8] to construct the Discretized Gaussian matrix.

We construct this matrix by first picking a desired stationary distribution, p_{ref}, and then filling in matrix entries such that we encourage transitions to nearby states whilst keeping p_{ref} as our stationary distribution. Specifically, we let p_{ref} be a discretized Gaussian over the state space, i.e.

$$p_{\text{ref}}(x) \propto \exp\left[-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right]$$

To find a condition on the rate such that this is the case, recall the Kolmogorov differential equation for the marginals

$$\partial_t q_t(x) = \sum_{\tilde{x}} q_t(\tilde{x}) R_b(\tilde{x}, x)$$

Now, consider a rate that is in detailed balance with p_{ref}

$$p_{\text{ref}}(\tilde{x}) R_b(\tilde{x}, x) = p_{\text{ref}}(x) R_b(x, \tilde{x})$$

Substituting this rate into the Kolmogorov equation, we see that p_{ref} is the stationary distribution

$$\begin{aligned}
\partial_t p_{\text{ref}}(x) &= \sum_{\tilde{x}} p_{\text{ref}}(\tilde{x}) R_b(\tilde{x}, x) \\
&= \sum_{\tilde{x}} p_{\text{ref}}(x) R_b(x, \tilde{x}) \\
&= p_{\text{ref}}(x) \sum_{\tilde{x}} R_b(x, \tilde{x}) \\
&= 0
\end{aligned}$$

where the last line follows from the fact that the row sum of a rate matrix is zero. Note that any $R_t = \beta(t)R_b$ will also have this stationary distribution as the multiplication by $\beta(t)$ can be seen as just a scaling of the time axis. From the detailed balance equation, we gain a condition on R_b such that our desired p_{ref} is the stationary distribution

$$\frac{R_b(\tilde{x}, x)}{R_b(x, \tilde{x})} = \frac{p_{\text{ref}}(x)}{p_{\text{ref}}(\tilde{x})} = \exp\left[\frac{(\tilde{x} - \mu_0)^2}{2\sigma_0^2} - \frac{(x - \mu_0)^2}{2\sigma_0^2}\right]$$

This gives constraints on diagonal elements within R_b but does not fully define the entire matrix. To do this, we first make the assumption that μ is selected to be at the center of the state space. Then we set off diagonal terms to the right of the diagonal in the top half of the rate matrix and off diagonal terms to the left of the diagonal in the bottom half to be 1. Finally, progressing in from the top and bottom of the rate matrix we make definitions of rate matrix values that have not already been defined by the detailed balance condition. For clarity, we provide a pictorial representation of this scheme for an 8×8 rate matrix below

$$\begin{bmatrix} \cdot & 1 & \square & \square & \square & \square & \square & \square \\ \triangle & \cdot & 1 & \square & \square & \square & \square & \triangle \\ \triangle & \triangle & \cdot & 1 & \square & \square & \triangle & \triangle \\ \triangle & \triangle & \triangle & \cdot & 1 & \triangle & \triangle & \triangle \\ \triangle & \triangle & \triangle & \triangle & \cdot & \triangle & \triangle & \triangle \\ \triangle & \triangle & \triangle & \square & 1 & \cdot & \triangle & \triangle \\ \triangle & \triangle & \square & \square & \square & 1 & \cdot & \triangle \\ \triangle & \square & \square & \square & \square & \square & 1 & \cdot \end{bmatrix}$$

where $\square$ represents a value we will define, $\triangle$ represents a value that is defined relative to another entry through the detailed balance condition and $\cdot$ is a diagonal entry that is equal to the negative off diagonal row sum. We could define $\square$ values to be 0 to gain a sparse rate matrix, however, we found in early experiments that allowing transitions to further away states greatly reduces the mixing time and gives better performance. We define $\square$ in each row similarly, by setting it equal to $\exp[-i^2/\sigma_r^2]$ where i is the distance away from the '1' value in that row and σ_r is a hyperparameter defining the length scale in state space of a typical transition. This biases our forward process to make transitions between nearby states, at a length scale of σ_r.

F CTMC Simulation

F.1 Exact CTMC and Tau-Leaping

In this section, we first describe exact CTMC simulation before giving an algorithmic description of tau-leaping.

When a CTMC has a time-homogeneous rate matrix, we can use Gillespie's Algorithm [21, 22, 23] to exactly simulate it. This algorithm is based on the jump chain/holding time definition of the CTMC. It repeats the following two steps:

- Draw a holding time from an exponential distribution with mean $-1/R(x, x)$ and wait in the current state x for that amount of time.
- Sample the next state from $r(\tilde{x}|x) = (1 - \delta_{x,\tilde{x}})\frac{R(x,\tilde{x})}{\sum_{x' \neq x} R(x,x')}$

This Algorithm can be adjusted for the case when we have a time-inhomogeneous rate matrix using the modified next reaction method [33]. However, both algorithms still step through each transition in the CTMC individually and are thus unsuitable in our case because only one dimension would change for each simulation step making it very computationally expensive to produce a sample. Instead we use tau-leaping that allows multiple dimensions to change in a single simulation step. We detail this method in Algorithm 1.

F.2 Predictor-Corrector Discussion

In this section we compare predictor-corrector sampling schemes as applied to continuous state spaces and discrete state spaces.

Algorithm 1: Generative Reverse Process Simulation with Tau-Leaping

$t \leftarrow T$
$\boldsymbol{x}_t^{1:D} \sim p_{\text{ref}}(\boldsymbol{x}_T^{1:D})$
while $t > 0$ **do**
  Compute $p_{0|t}^{\theta}(x_0^d|\boldsymbol{x}_t^{1:D}), d = 1, \ldots, D$ with one forward pass of the denoising network
  for $d = 1, \ldots, D$ **do**
    for $s = 1, \ldots, S \backslash x_t^d$ **do**
      $\hat{R}_t^{\theta\, d}(\boldsymbol{x}_t^{1:D}, s) \leftarrow R_t^d(s, x_t^d) \sum_{x_0^d} p_{0|t}^{\theta}(x_0^d|\boldsymbol{x}_t^{1:D}) \frac{q_{t|0}(s|x_0^d)}{q_{t|0}(x_t^d|x_0^d)}$
      $P_{ds} \leftarrow \text{Poisson}\Big(\tau \hat{R}_t^{\theta\, d}(\boldsymbol{x}_t^{1:D}, s)\Big)$
    end
  end
  for $d = 1, \ldots, D$ **do**
    if *data is categorical AND* $\sum_{s=1}^{S} P_{ds} > 1$ **then**
      $x_{t-\tau}^d \leftarrow x_t^d$ `// reject change`
    else
      $x_{t-\tau}^d \leftarrow x_t^d + \sum_{s=1}^{S} P_{ds} \times (s - x_t^d)$
    end
  end
  $\boldsymbol{x}_{t-\tau}^{1:D} \leftarrow \text{Clamp}(\boldsymbol{x}_{t-\tau}^{1:D}, \text{min} = 1, \text{max} = S)$
  $t \leftarrow t - \tau$
end

The predictor-corrector scheme in continuous state spaces was introduced in [4]. It consists of alternating between a predictor step and a corrector step:

$$\text{Predictor} \quad \boldsymbol{x}_i \leftarrow \boldsymbol{x}_{i+1} + \gamma_i s_\theta(\boldsymbol{x}_{i+1}, i+1) + \sqrt{\gamma_i} z, \quad z \sim \mathcal{N}(0, I)$$
$$\text{Corrector} \quad \boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \epsilon_i s_\theta(\boldsymbol{x}_i, i) + \sqrt{2\epsilon_i} z, \quad z \sim \mathcal{N}(0, I)$$

where $\boldsymbol{x}_i$ is the state at sampling step i, s_θ is the learned score model approximating $\nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})$ and γ_i, ϵ_i are the step sizes for the predictor and corrector respectively. We see that both take similar forms, except the corrector adds in a factor $\sqrt{2}$ more Gaussian noise during the update step.

In discrete state spaces, rather than sampling using gradient guided stochastic steps as in the continuous state space case, we sample by simulating CTMCs with defined rates. When we take a predictor step, we simulate using $\hat{R}_t^\theta$ and when we take a corrector step we simulate using $R_t^{c\,\theta} = \hat{R}_t^\theta + R_t$. If we simulate the CTMC exactly, we have seen in the previous section that this amounts to sampling next states from the categorical distribution defined by normalizing the row of the rate matrix corresponding to the current state. Therefore, corrector sampling can be seen as sampling from a slightly noisier categorical distribution defined through $R_t^{c\,\theta}$ as compared to the predictor categorical distribution defined through $\hat{R}_t^\theta$. This is analogous to the increased Gaussian noise applied during a corrector step in continuous state spaces.

Adding corrector steps brings the marginal of the samples closer to $q_t(\boldsymbol{x})$ and continued application of the corrector will further explore the domain of $q_t(\boldsymbol{x})$. In previous work on continuous state predictor-corrector methods, the number of corrector steps has been small (e.g. 1 or 2 corrector steps per predictor step) or indeed the corrector steps have been removed altogether. In this work we have found that using up to 10 corrector steps per predictor steps can be beneficial during certain regions of the reverse generative process. Additionally, in continuous state spaces, it has been observed that too many corrector steps can result in unwanted noise in the generated data [34].

We hypothesize that corrector steps are better utilized in discrete state spaces to explore the domain of $q_t(\boldsymbol{x})$ than in continuous state spaces. This is because, the corrector update is defined largely through the reverse rate itself, $\hat{R}_t^\theta$, just with the categorical probabilities being annealed slightly more towards uniform through the addition of the forward rate R_t. This may be a more effective update than simply adding extra Gaussian noise in the continuous state space case. Furthermore, the denoising model in continuous state spaces can be seen as outputting a point estimate of $\boldsymbol{x}_0$ of dimension D. However,

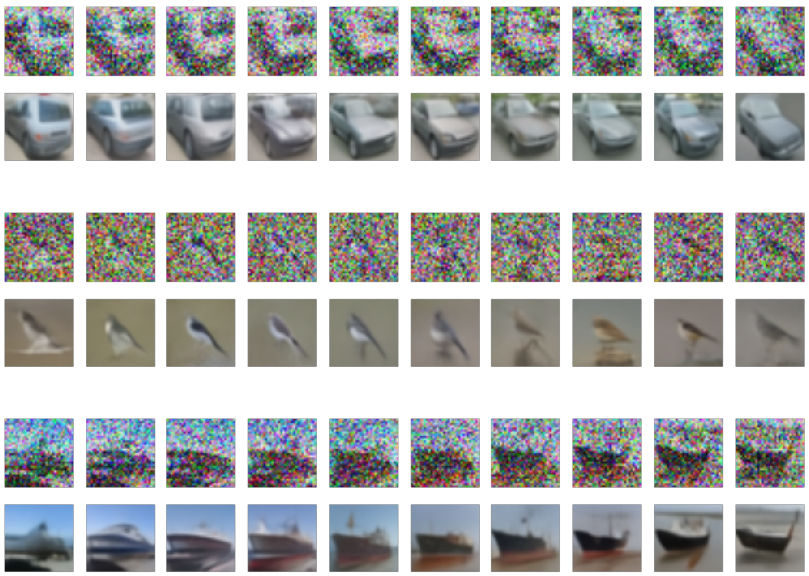

Figure 7: Progression of $\boldsymbol{x}_t$ for $t = 0.4$ by repeated application of corrector steps. In each pair of rows, the top row is $\boldsymbol{x}_t$ whilst the bottom row is the $\boldsymbol{x}_0$ prediction made by $p^\theta_{0|t}(\boldsymbol{x}_0|\boldsymbol{x}_t)$ (argmax of the categorical probabilities in each dimension). Each column represents an additional 100 corrector steps.

in discrete state spaces, the denoising model outputs a categorical distribution over every dimension (output dimension $D \times S$) allowing it to express some uncertainty information in the $\boldsymbol{x}_0$ prediction, albeit with conditional independence between the dimensions. Adding corrector steps in discrete state spaces would then allow information to mix between dimensions for the current time step, exploring modes of $q_t(\boldsymbol{x})$.

We explore this idea on the image modelling task in Figure 7. We run the reverse generative process until time $t = 0.4$ at which point we hold the time constant and apply 1000 corrector steps. We see that the resulting progression of $\boldsymbol{x}_t$ states explores potential local modes of $q_t(\boldsymbol{x})$ in the local region of image space.

G Implicit Dimensional Assumptions Made in Discrete Time

In discrete time, the parametric reverse kernel, $p^\theta_{k|k+1}$, is commonly defined through a denoising model $p^\theta_{0|k+1}$. Here, we examine this definition in the multi-dimensional case where the forward process factorizes, as in Appendix C.3 and previous discrete time work [8]. We begin by writing the true full dimensional reverse kernel, $q_{k|k+1}$, in terms of the true denoising distribution, $q_{0|k+1}$.

$$
\begin{aligned}
q_{k|k+1}(\boldsymbol{x}_k^{1:D}|\boldsymbol{x}_{k+1}^{1:D}) &= \prod_{d=1}^{D} q_{k|k+1}(x_k^d|\boldsymbol{x}_k^{1:d-1}, \boldsymbol{x}_{k+1}^{1:D}) \\
&= \prod_{d=1}^{D} \sum_{x_0^d} q_{k,0|k+1}(x_k^d, x_0^d|\boldsymbol{x}_k^{1:d-1}, \boldsymbol{x}_{k+1}^{1:D}) \\
&= \prod_{d=1}^{D} \sum_{x_0^d} q_{0|k+1}(x_0^d|\boldsymbol{x}_k^{1:d-1}, \boldsymbol{x}_{k+1}^{1:D}) q_{k|0,k+1}(x_k^d|x_0^d, x_{k+1}^d)
\end{aligned}
$$

where on the final line we have used the fact that the forward process is independent across dimensions. To create our approximate reverse kernel, $p^\theta_{k|k+1}$, we approximate $q_{0|k+1}(x_0^d|\boldsymbol{x}_k^{1:d-1}, \boldsymbol{x}_{k+1}^{1:D})$ with

$p^\theta_{0|k+1}(x_0^d|\boldsymbol{x}_{k+1}^{1:D})$,

$$p^\theta_{k|k+1}(\boldsymbol{x}_k^{1:D}|\boldsymbol{x}_{k+1}^{1:D}) = \prod_{d=1}^{D} \sum_{x_0^d} p^\theta_{0|k+1}(x_0^d|\boldsymbol{x}_{k+1}^{1:D}) q_{k|0,k+1}(x_k^d|x_0^d, x_{k+1}^d)$$

We throw away the extra $\boldsymbol{x}_k^{1:d-1}$ conditioning because we use a non-autoregressive model that takes in $\boldsymbol{x}_{k+1}^{1:D}$ and in a single forward pass gives conditionally independent probabilities over x_0^d, $d = 1, \ldots, D$. For finite K, this approximation can never match the true kernel because we are not conditioning on all relevant information. Of course, as K gets larger, this approximation becomes more accurate. Since we operate in the continuous regime, we do not have to make this approximation because the conditionally independent denoising model, $q_{0|t}(x_0^d|\boldsymbol{x}^{1:D})$, appears directly in our reverse rate, $\hat{R}_t^{1:D}$, when we factorize the forward process (see Proposition 3).

H Experimental Details

In this section, we provide additional details for the experiments we performed applying our method to practical problems. The code for our models is available at `https://github.com/andrew-cr/tauLDR`. Before describing the specifics for each experiment, we first explain the implementation details common to all.

When we evaluate the objective $\mathcal{L}_{\text{CT}}$ on each minibatch of training datapoints, we must sample a time for each from $t \sim \mathcal{U}(0, T)$ which represents the point in the forward process which we will noise to. Training instabilities can be found if t is sampled very close to 0 because the reverse rate, $\hat{R}_t$, becomes ill-conditioned in this region. This phenomenon is also observed in continuous state space models because the score, $\nabla_x \log q_t(x)$, becomes ill-conditioned close to $t = 0$. The reverse rate and score become ill conditioned close to the start of the forward process because the marginal probability, $q_t(x)$, will be highly peaked around the data manifold and $\log q_t(x)$ will explode in regions that are not close to the data. To avoid these issues, a common trick is to set a minimum time such that $t \sim \mathcal{U}(\epsilon, T)$. ϵ is set such that the level of noising at $t = \epsilon$ is very small and reverse sampling to this point will produce samples very close to p_{data}. In our experiments, we set $\epsilon = 0.01T$.

During reverse sampling, we use tau-leaping to simulate the reverse process from $t = T$ until $t = \epsilon$ because the reverse rate is not trained for $t < \epsilon$. This produces a sample close to p_{data}. We found improved performance in metrics such as FID if we then complete a final step to remove the small amount of noise that may still be present in the sample. Specifically, we pass the sample through the denoising model $p^\theta_{0|t}(x_0|x_t)$ with $t = \epsilon$ to obtain an output of shape $D \times S$ where D is the dimensionality of the problem. This is a probability distribution over the states for each of the dimensions. We set the value of each dimension to the state with the highest probability. This then produces a sample which has all of the noise removed.

The specific value of T within our model is arbitrary because the forward process can be scaled in the time axis to provide the same noising process for any T. Therefore, we simply set $T = 1$.

H.1 Demonstrative Example

Our 2d dataset is created by sampling 1M 2d points from a 32×32 state space with probability proportional to the pixel values of a 32×32 grayscale image of a τ character.

For our forward process, we use a Gaussian rate (see Appendix E) with stationary distribution standard deviation $\sigma_0 = 8$ and rate length scale $\sigma_r = 1$. We use a rate schedule of $\beta(t) = 5 \times 5^t \log(5)$.

To represent $p^\theta_{0|t}$ we use a residual MLP. The architecture consists of an input linear layer to lift the input dimension of 2 to the internal network dimension of 16. Then, there are 2 residual blocks each consisting of: a single hidden layer MLP of hidden dimension 32, a residual connection to the

input of the MLP, a layer norm, and finally a FiLM layer [35] modulated by the time embedding. At the output, there is a single linear layer with output size of $2 \times 32 = 64$ representing state probabilities in each of the 2 dimensions. The time is embedded using the Transformer sinusoidal position embedding [36] creating an embedding of size 32. Then, the embedding is further processed by a single hidden layer MLP with hidden layer size 32 and output size 128. To create the FiLM parameters in each residual block, the time embedding is passed through a linear layer with output of size 32 to provide a multiplicative and additive modulation to the state dimension of 16. We minimize the $\mathcal{L}_{\text{CT}}$ objective using Adam with a learning rate of 0.0001 and batch size of 32 for 1M steps.

For the exact simulation we use the next reaction method with modifications for time dependent transition rates [33]. This method steps through each transition in the exact simulation path individually by calculating the time to the next occurrence of each transition type and applying the transition that occurs soonest. Exact algorithmic details can be found in [33]. To calculate the time to the next occurrence for a transition, we need to integrate the reverse rate matrix (eq (13) in [33]). We do this with euler integration with a step size of 0.001.

H.2 Image Modeling

We train on the CIFAR10 training dataset that contains 50000 images of dimension $3 \times 32 \times 32$. We evaluate the test ELBO on the CIFAR10 test dataset which consists of 10000 images. For the forward noising process, we use the the Gaussian rate (see Appendix E) with stationary distribution standard deviation of $\sigma_0 = 512$ and rate length scale $\sigma_r = 6$. This effectively defines a uniform stationary distribution since the state space is of size 256. We found this performs better than a more concentrated Gaussian. Our β schedule is $\beta(t) = 3 \times 100^t \log 100$. This was selected in accordance with σ_r such that the overall shape of progression of the $q_{t|0}$ variances approximately matches that of the schedule proposed in [3].

Our $p^\theta_{0|t}$ model is parameterized with the standard U-net [37] architecture introduced in [3]. The network follows the PixelCNN++ backbone [38] with group normalization layers. There are four feature map resolutions (32×32 to 4×4) in the downsampling/upsampling stacks. At each resolution there are two convolutional residual blocks. There is a self-attention block between the residual blocks at the 16×16 resolution level [39]. The time is input into the network by first embedding with the Transformer sinusoidal position embedding [36]. This time embedding is passed into each residual block by passing it through a SiLU activation [40] and then a linear layer before adding it onto the hidden state within the residual block between the two convolution operations.

The original architecture of [3] has an output of dimension $3 \times 32 \times 32$ as it makes a point prediction of x_0 given x_t. In order for the model to output probabilities over x_0 (i.e. an output dimension of $3 \times 32 \times 32 \times 256$) we make the adjustments suggested in [8]. Specifically, we use their truncated logistic distribution parameterization where the model outputs the mean and log scale of a logistic distribution i.e. an output dimension of $3 \times 32 \times 32 \times 2$. The probability for a state is then the integral of this continuous distribution between this state and the next when mapped onto the real line. To impart a residual inductive bias on the output, the mean of the logistic distribution is taken to be $\tanh(x_t + \mu')$ where x_t is the normalized input into the model and μ' is mean outputted from the network. The normalization operation takes the input in the range $0, \dots, 255$ and maps it to $[-1, 1]$. In total, our network has approximately 35.7 million parameters.

We optimize with the auxiliary objective described in Appendix D with $\lambda = 0.001$. Within the auxiliary objective, we use the one-forward pass version of the continuous time ELBO, $\mathcal{L}_{\text{eCT}}$. We optimize with Adam for 2M steps with a learning rate of 0.0002 and batch size of 128. We use the standard set of training tricks to improve optimization [3, 4]. Throughout training we maintain an exponential moving average of the parameters with decay factor 0.9999. These average parameters are used during testing. At the start of optimization we use a linear learning rate warm-up for the first 5000 steps. We clip the gradient norm at a norm value of 1.0. We set the dropout rate for the network at 0.1. The skip connections for each residual block are rescaled by a factor of $\frac{1}{\sqrt{2}}$. The input images have random horizontal flips applied to them during training.

For sampling in Table 1 we set $\tau = 0.001$ for τLDR-0 and set $\tau = 0.002$ for τLDR-10. The 10 corrector steps per predictor steps for τLDR-10 are introduced after $t < 0.1T$. We found that introducing the corrector steps near the end of the reverse sampling process had the best improvement in sample quality for the smallest increase in computational cost. When performing tau-leaping with the corrector rate, R_t^c, we have control over what τ we use since we are sampling a different CTMC (with q_t as its stationary distribution) to the original reverse CTMC. We found that setting the corrector rate τ to be 1.5 times the original τ for the reverse CTMC achieves the best performance in this example.

We train using 4 V100 GPUs on an academic research cluster. To calculate Inception and FID values, we use pytorch-fid [41] and a further development [1]. We verified this library produced comparable values to previous work by calculating the Inception and FID scores for the published images from the DDPM [3] method.

We show a large array of unconditional samples from the τLDR-10 model in Figure 8. We now also present statistics from the reverse sampling process with standard tau-leaping with $\tau = 0.001$. Figure 9 shows the proportion of dimensions that transition during a single step of tau-leaping. We see that during the initial stages, every dimensions changes during every tau-leaping step, but nearer the end of the process, more dimensions will have settled in their final positions and the proportion is less. In Figure 10 we show the proportion of dimensions that are clipped due to proposing an out of bounds jump. Overall, the proportion is small. It is largest at the start of the process when we have initially sampled from the approximately uniform p_{ref} and there will be dimensions close to the boundary. As pixel values settle to their final values, the proportion reduces. Figure 11 shows the progression of a selection of dimensions during the reverse sampling process. A similar picture emerges where dimensions eventually settle in a region of the state space. We also note that larger jumps are made in a single tau-leaping step nearer the start of the reverse process and smaller jumps are made nearer the end.

H.3 Monophonic Music

We generate our training dataset from the Lakh pianoroll dataset [29, 30] (license CC By 4.0). This dataset consists of 174,154 multitrack pianorolls. We go through all songs and all tracks within each song and select sequences that match the following criteria: they are monophonic (only one note played at a time), there is not a period longer than one bar in which no note is played, there is more than one type of note played in the sequence and finally there is no one note played for more than 50 time steps out of the total 256 time steps. This removes the uninteresting and trivial sequences present within the dataset. We then remove any duplicates in the result. This leaves us with 6000 training examples and 950 testing examples. Each song consists of 256 time steps (16 per bar) and each time step takes one from 129 values i.e. we have $D = 256$ and $S = 129$. This state value represents either a note from 128 options or a rest. We scramble the ordering of this state space when mapping to the integers from 0 to 128. When we input into the denoising network, we input as one-hot 129 dimensional vectors.

For the forward noising process, we use a uniform rate matrix, $R_b = \mathbb{1}\mathbb{1}^T - S\text{Id}$ and set $\beta(t) = 0.03$. We found a constant in time $\beta(t)$ was sufficient for this dataset. In our comparison, we used a birth/death rate matrix defined as

$$\begin{bmatrix} -\lambda & \lambda & 0 & 0 & \dots & 0 & 0 \\ \lambda & -2\lambda & \lambda & 0 & \dots & 0 & 0 \\ 0 & \lambda & -2\lambda & \lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \lambda & -\lambda \end{bmatrix}$$

this is the rate matrix for a birth/death process. We set $\lambda = 1$ and $\beta(t) = \frac{1}{2} \times 10000^t \log 10000$. These hyperparameters were selected such that the forward process has a steady rate of noising whilst still having q_T very close to p_{ref}. We chose to compare these types of rate matrix because the birth/death rate is inappropriate for this categorical data as adjacent states have no meaning

[1] `https:/github.com/w86763777/pytorch-gan-metrics`

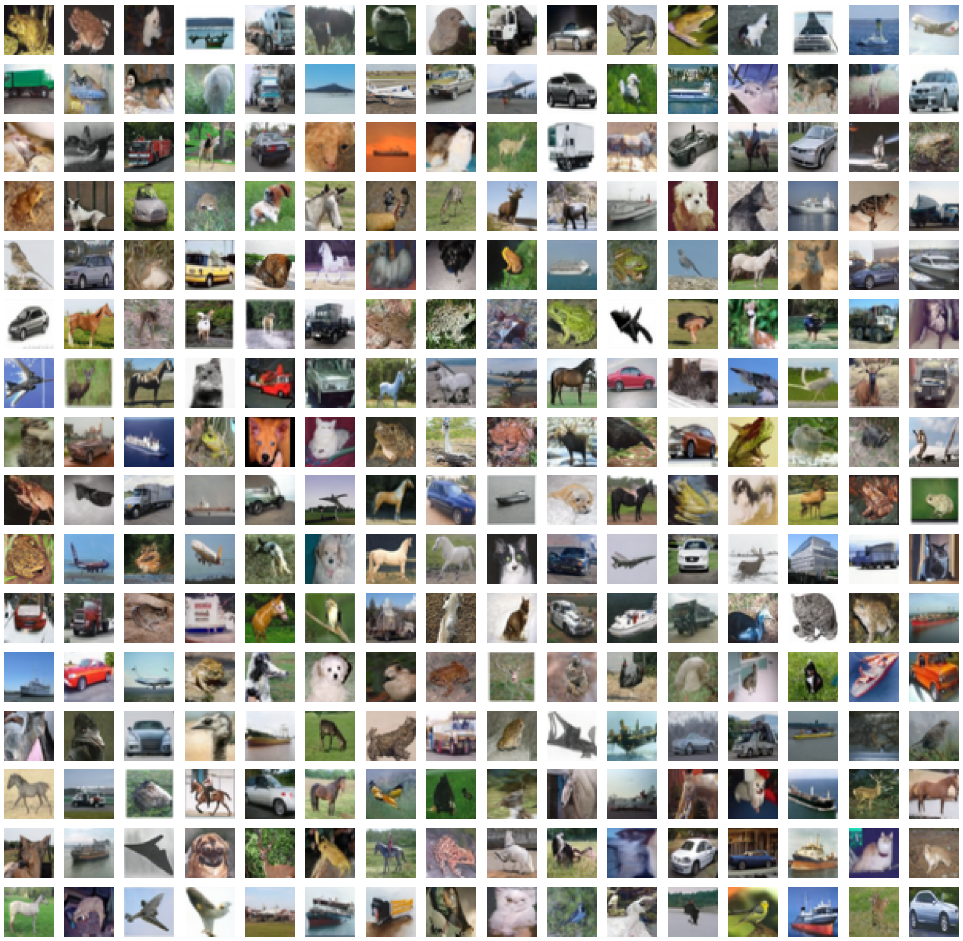

Figure 8: Unconditional CIFAR10 samples from our τLDR-10 model.

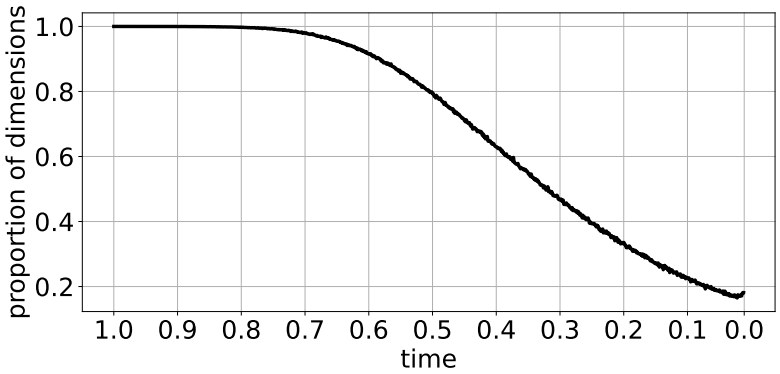

Figure 9: Proportion of dimensions that transition during a single step of tau-leaping during the reverse sampling process.

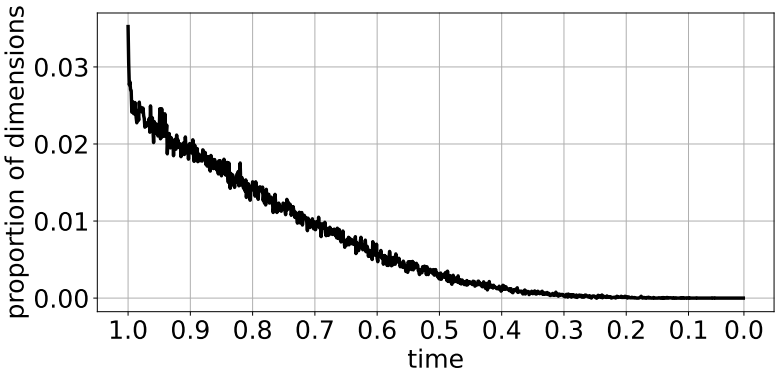

Figure 10: Proportion of dimensions that are clipped during a tau-leaping step due to proposing an out of bounds jump during the reverse sampling process.

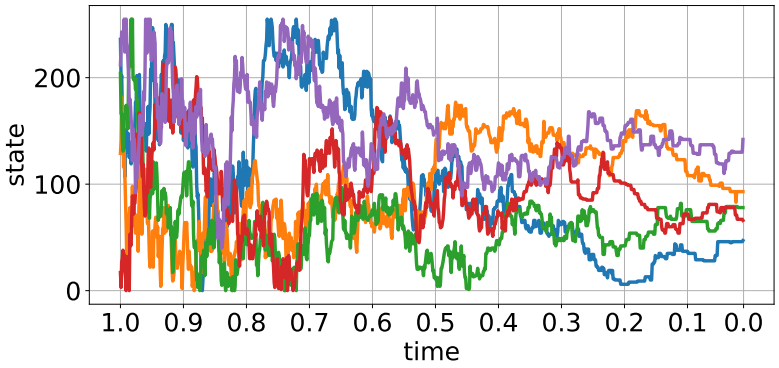

Figure 11: The progression of a selection of dimensions during the reverse sampling process.

since the mapping to the integers was arbitrary. The uniform rate is suitable for this categorical data because, during a time interval, it has a uniform probability to transition to any other state. The D3PM baseline was implemented also with a time homogeneous uniform forward kernel set such that the rate of noising is matched in the discrete and continuous time cases.

We define our conditional denoising network, $p^{\theta}_{0|t}(x_0|x, y)$ using a transformer architecture inspired by [31]. It takes an input of shape (B, D, S) where B is the batch size, D is the dimensionality (256) and S is the state size (129). This final dimension contains the one-hot vectors. The conditioning on the initial bars is achieved by concatenating the conditioning information y with the noisy input x. At the start of the network, there is an input embedding linear layer with output of size 128 which is our model dimension for the transformer. Then a transformer positional embedding is added to the hidden state. Next a stack of 6 transformer encoder layers are applied which consist of a self attention block and a one hidden layer MLP. The self attention block uses 8 heads and the MLP has a hidden layer size of 2048. At the output of each internal block, we apply dropout with rate 0.1. Finally, there is a stack of 2 residual MLP layers. Each consists of a one hidden layer MLP with a hidden dimension of 2048. There is a residual connection between the input and output of the MLP. A layer norm is applied to the output of the block. To create the output of the network, there is an output linear layer with an output shape of (B, D, S) where now the S dimension has logit probabilities. To instill a residual bias into the network, we add the one-hot input to the output logits. All activations are ReLU. The time is input into the network through FiLM layers [35]. First, the time is embedded using the sinusoidal transformer position embedding as in the U-net architecture used for image modeling to create an embedding size of 128. This is then passed into a single hidden layer MLP with hidden size

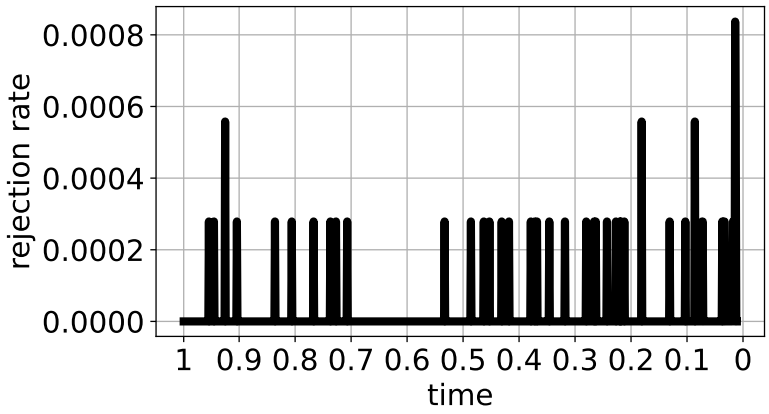

Figure 12: Proportion of jumps rejected during reverse sampling. The rejection rate is calculated as the proportion of dimensions in a tau leaping step that have their jump rejected. The results are averaged over a batch of 16.

2048 and output size 512. Within each encoder and residual MLP block, there is a FiLM linear layer which takes in the 512 time embedding and outputs two FiLM parameters each of size 128. These are the scale and offset applied to the hidden state. In the encoder blocks, this FiLM transform is applied after the self attention block and again after the fully connected block. In the residual MLP blocks, it is applied after the layer norm operation. Our network has approximately 7 million parameters in total.

We optimize using Adam for 1M steps with a batch size of 64 and learning rate of 0.0002. We use the conditional $\bar{\mathcal{L}}_{\text{CT}}$ objective with additional direct $p^{\theta}_{0|t}$ supervision as described in Appendix D with weight $\lambda = 0.001$. We also make the same one forward pass approximation as explained in Appendix C.4. We use the standard set of training tricks to improve optimization [3, 4]. Throughout training we maintain an exponential moving average of the parameters with decay factor 0.9999. These average parameters are used during testing. At the start of optimization we use a linear learning rate warm-up for the first 5000 steps. We clip the gradient norm at a norm value of 1.0. We train on a single V100 GPU on an academic cluster.

For sampling with τLDR-0 we use $\tau = 0.001$ and for sampling with τLDR-2 we include 2 corrector steps per predictor step after $t < 0.9T$. The corrector rate is simulated with $\tau = 0.0001$ which we found to perform best. We reject any dimension in which 2 or more jumps are proposed as this is categorical data. We plot the rejection rate in Figure 12. Most of the time, the rejection rate is zero and there are few steps for which it increases slightly. We show a large batch of samples from the first 10 songs in the test dataset in Figure 13. We see that there is variation between the sampled completions and they consistently follow the style of the conditioning first two bars of the song. Audio samples from the model are available at `https://github.com/andrew-cr/tauLDR`. Finally, we examine the progression of a random selection of dimensions during reverse sampling for the uniform and birth/death rate matrix cases. Figure 14 shows the progression for the uniform case, we see that large jumps through the state space are made throughout the reverse process. Figure 15 shows the progression for the birth/death case. At the start of reverse sampling, no dimensions move as the rejection rate is high in this case because the rate matrix is not suitable for categorical data. Nearer the end, small jumps are made between adjacent states but since large jumps between any category do not occur for this rate matrix, the performance will overall be worse.

I Ethical Considerations

Our work increases our theoretical understanding of denoising generative models and also improves generation capabilities within some discrete datasets. Deep generative models are generic methods

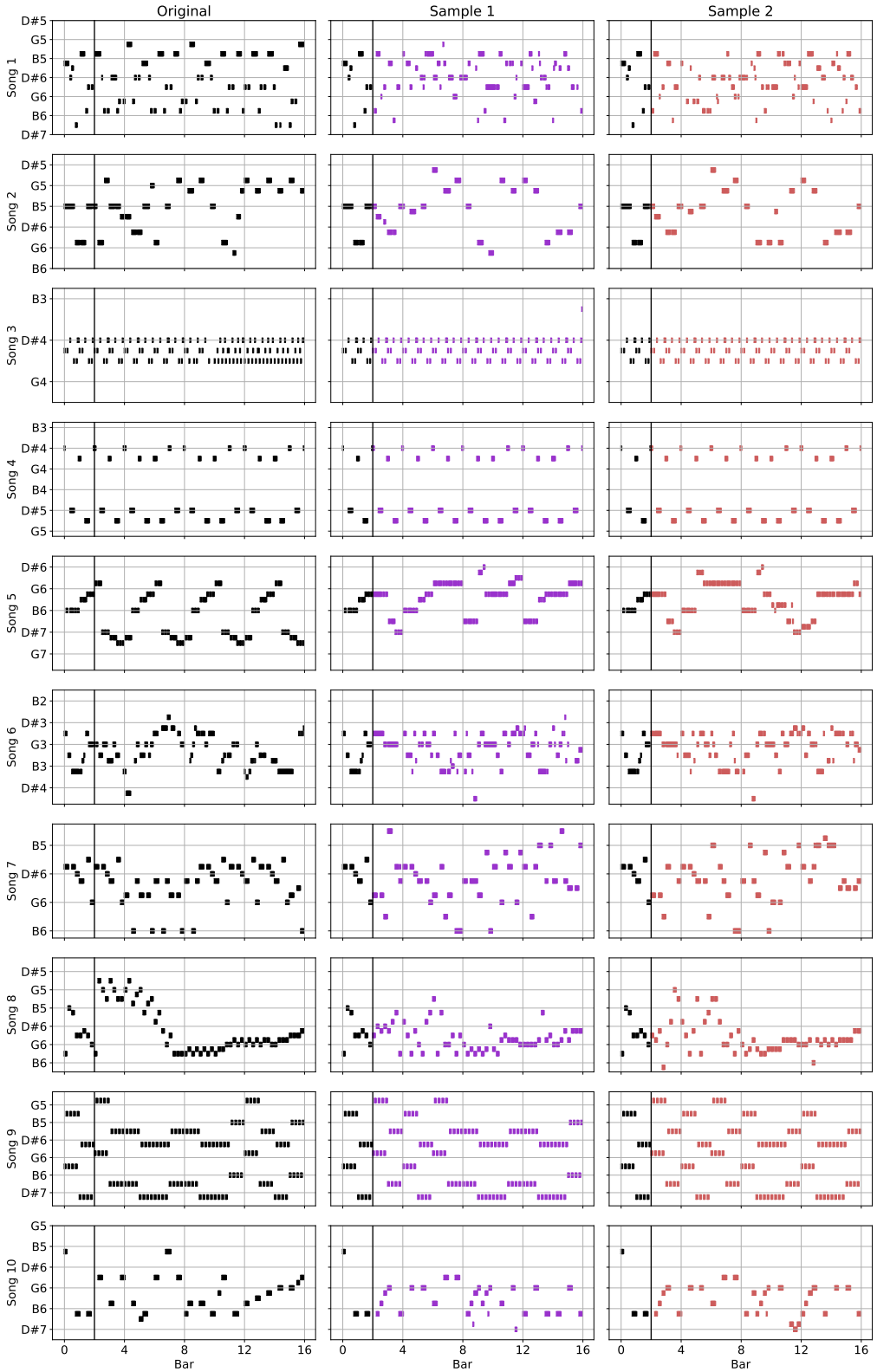

Figure 13: Two conditional samples from the τLDR-0 model for each of the first 10 songs in the test dataset.

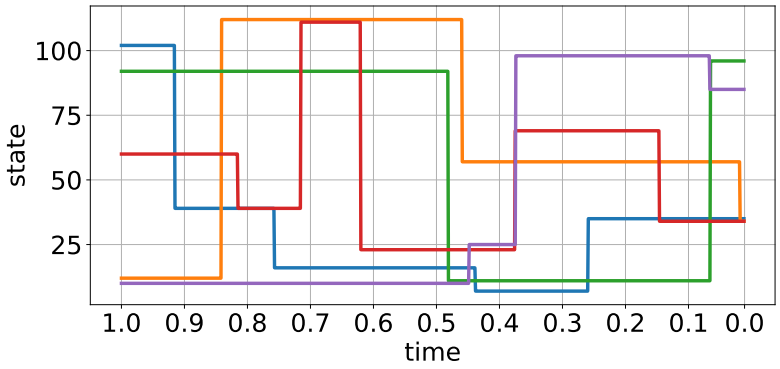

Figure 14: The progression of a selection of dimensions during the reverse sampling process for the uniform rate matrix.

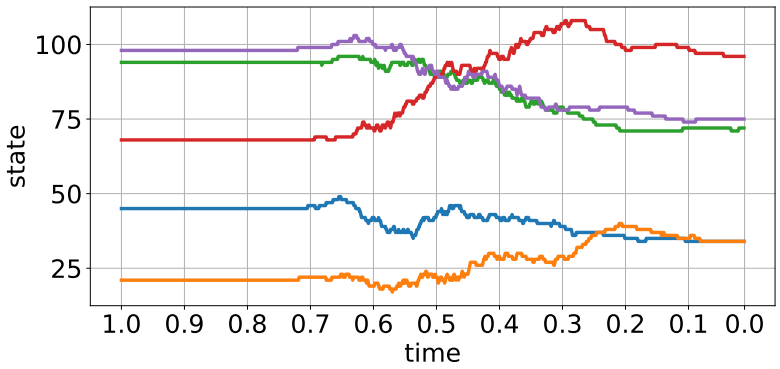

Figure 15: The progression of a selection of dimensions during the reverse sampling process for the birth/death rate matrix.

for learning from unstructured data and can have negative social impacts when misused. For example, they can be used to spread misinformation by reducing the resources required to create realistic fake content. Furthermore, generative models will produce samples that accurately reflect the statistics of their training dataset. Therefore, if samples from these models are interpreted as an objective truth without fully considering the biases present in the original data, then they can perpetuate discrimination against minority groups.

In this work, we train on datasets that contain less sensitive data such as pictures of objects and music samples. The methods we presented, however, could be used to model images of people or text from the internet which will contain biases and potentially harmful content that the model will then learn from and reproduce. Great care must be taken when training these models on real world datasets and when deploying them so as to mitigate and prevent the harms that they can cause.

Statement of Authorship

Title of paper:	A Continuous Time Framework for Discrete Denoising Models
Publication status:	Published
Publication details:	Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, Arnaud Doucet. *Advances in Neural Information Processing Systems*, 2022.

Student Confirmation

Student name:	Joe Benton
Contribution to the paper:	The idea to use tau-leaping for simulating the forward and reverse processes was provided by Andrew Campbell and Arnaud Doucet. I contributed to the theoretical framework for adapting diffusion models to discrete spaces, including providing the derivation of the ELBO, motivation for the choice of forward process, the predictor-corrector scheme and the error bound (Sections 3.2, 4.1, 4.4 and 4.5). Andrew Campbell implemented the experiments (Section 5) and lead the writing process.

Signature: J Benton Date: 29/10/2023

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: George Deligiannidis, Associate Professor of Statistics

Supervisor comments: None

Signature: Date: 31/10/2023

4

Error Bounds for Flow Matching Methods

Error Bounds for Flow Matching Methods

Joe Benton *benton@stats.ox.ac.uk*
Department of Statistics
University of Oxford

George Deligiannidis *deligian@stats.ox.ac.uk*
Department of Statistics
University of Oxford

Arnaud Doucet *doucet@stats.ox.ac.uk*
Department of Statistics
University of Oxford

Reviewed on OpenReview: *https://openreview.net/forum?id=uqQPyWFDhY*

Abstract

Score-based generative models are a popular class of generative modelling techniques relying on stochastic differential equations (SDEs). From their inception, it was realized that it was also possible to perform generation using ordinary differential equations (ODEs) rather than SDEs. This led to the introduction of the probability flow ODE approach and denoising diffusion implicit models. Flow matching methods have recently further extended these ODE-based approaches and approximate a flow between two arbitrary probability distributions. Previous work derived bounds on the approximation error of diffusion models under the stochastic sampling regime, given assumptions on the L^2 loss. We present error bounds for the flow matching procedure using fully deterministic sampling, assuming an L^2 bound on the approximation error and a certain regularity condition on the data distributions.

1 Introduction

Much recent progress in generative modelling has focused on learning a map from an easy-to-sample reference distribution π_0 to a target distribution π_1. Recent works have demonstrated how to learn such a map by defining a stochastic process transforming π_0 into π_1 and learning to approximate its marginal vector flow, a procedure known as flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023; Albergo et al., 2023; Heitz et al., 2023). Score-based generative models (Song et al., 2021b; Song & Ermon, 2019; Ho et al., 2020) can be viewed as a particular instance of flow matching where the interpolating paths are defined via Gaussian transition densities. However, the more general flow matching setting allows one to consider a broader class of interpolating paths, and leads to deterministic sampling schemes which are typically faster and require fewer steps (Song et al., 2021a; Zhang & Chen, 2023).

Given the striking empirical successes of these models at generating high-quality audio and visual data (Dhariwal & Nichol, 2021; Popov et al., 2021; Ramesh et al., 2022; Saharia et al., 2022), there has been significant interest in understanding their theoretical properties. Several works have sought convergence guarantees for score-based generative models with stochastic sampling (Block et al., 2022; De Bortoli, 2022; Lee et al., 2022), with recent work demonstrating that the approximation error of these methods decays polynomially in the relevant problem parameters under mild assumptions (Chen et al., 2023c; Lee et al., 2023; Chen et al., 2023a). Other works have explored bounds in the more general flow matching setting (Albergo & Vanden-Eijnden, 2023; Albergo et al., 2023). However these works either still require some stochasticity in the sampling procedure or do not hold for data distributions without full support.

In this work we present the first bounds on the error of the flow matching procedure that apply with fully deterministic sampling for data distributions without full support. Our results come in two parts: first, we control the error of the flow matching approximation under the 2-Wasserstein metric in terms of the L^2 training error and the Lipschitz constant of the approximate velocity field; second, we show that, under a smoothness assumption explained in Section 3.2, the true velocity field is Lipschitz and we bound the associated Lipschitz constant. Combining the two results, we obtain a bound on the approximation error of flow matching which depends polynomially on the L^2 training error.

1.1 Related work

Flow methods: Probability flow ODEs were originally introduced by Song et al. (2021b) in the context of score-based generative modelling. They can be viewed as an instance of the normalizing flow framework (Rezende & Mohamed, 2015; Chen et al., 2018), but where the additional Gaussian diffusion structure allows for much faster training and sampling schemes, for example using denoising score matching (Vincent, 2011), compared to previous methods (Grathwohl et al., 2019; Rozen et al., 2021; Ben-Hamu et al., 2022).

Diffusion models are typically expensive to sample and several works have introduced simplified sampling or training procedures. Song et al. (2021a) propose Denoising Diffusion Implicit Models (DDIM), a method using non-Markovian noising processes which allows for faster deterministic sampling. Alternatively, (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023) propose flow matching as a technique for simplifying the training procedure and allowing for deterministic sampling, while also incorporating a much wider class of possible noising processes. Albergo et al. (2023) provide an in depth study of flow matching methods, including a discussion of the relative benefits of different interpolating paths and stochastic versus deterministic sampling methods.

Error bounds: Bounds on the approximation error of diffusion models with stochastic sampling procedures have been extensively studied. Initial results typically relied either on restrictive assumptions on the data distribution (Lee et al., 2022; Yang & Wibisono, 2022) or produced non-quantitative or exponential bounds (Pidstrigach, 2022; Liu et al., 2022; De Bortoli et al., 2021; De Bortoli, 2022; Block et al., 2022). Recently, several works have derived polynomial convergence rates for diffusion models with stochastic sampling (Chen et al., 2023c; Lee et al., 2023; Chen et al., 2023a; Li et al., 2023; Benton et al., 2023; Conforti et al., 2023).

By comparison, the deterministic sampling regime is less well explored. On the one hand, Albergo & Vanden-Eijnden (2023) give a bound on the 2-Wasserstein distance between the endpoints of two flow ODEs which depends on the Lipschitz constants of the flows. However, their Lipschitz constant is uniform in time, whereas for most practical data distributions we expect the Lipschitz constant to explode as one approaches the data distribution; for example, this will happen for data supported on a submanifold. Additionally, Chen et al. (2023d) derive a polynomial bound on the error of the discretised exact reverse probability flow ODE, though their work does not treat learned approximations to the flow. Li et al. (2023) also provide bounds for fully deterministic probability flow sampling, but also require control of the difference between the derivatives of the true and approximate scores.

On the other hand, most other works studying error bounds for flow methods require at least some stochasticity in the sampling scheme. Albergo et al. (2023) provide bounds on the Kullback–Leibler (KL) error of flow matching, but introduce a small amount of random noise into their sampling procedure in order to smooth the reverse process. Chen et al. (2023b) derive polynomial error bounds for probability flow ODE with a learned approximation to the score function. However, they must interleave predictor steps of the reverse ODE with corrector steps to smooth the reverse process, meaning that their sampling procedure is also non-deterministic.

In contrast, we provide bounds on the approximation error of a fully deterministic flow matching sampling scheme, and show how those bounds behave for typical data distributions (for example, those supported on a manifold).

2 Background on flow matching methods

We give a brief overview of flow matching, as introduced by Lipman et al. (2023); Liu et al. (2023); Albergo & Vanden-Eijnden (2023); Heitz et al. (2023). Given two probability distributions π_0, π_1 on $\mathbb{R}^d$, flow matching is a method for learning a deterministic coupling between π_0 and π_1. It works by finding a time-dependent vector field $v : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ such that if $Z^{\mathbf{x}} = (Z_t^{\mathbf{x}})_{t\in[0,1]}$ is a solution to the ODE

$$\frac{\mathrm{d}Z_t^{\mathbf{x}}}{\mathrm{d}t} = v(Z_t^{\mathbf{x}}, t), \quad Z_0^{\mathbf{x}} = \mathbf{x} \tag{1}$$

for each $\mathbf{x} \in \mathbb{R}^d$ and if we define $Z = (Z_t)_{t\in[0,1]}$ by taking $\mathbf{x} \sim \pi_0$ and setting $Z_t = Z_t^{\mathbf{x}}$ for all $t \in [0,1]$, then $Z_1 \sim \pi_1$. When Z solves (1) for a given function v, we say that Z is a flow with velocity field v. If we have such a velocity field, then (Z_0, Z_1) is a coupling of (π_0, π_1). If we can sample efficiently from π_0 then we can generate approximate samples from the coupling by sampling $Z_0 \sim \pi_0$ and numerically integrating (1). This setup can be seen as an instance of the continuous normalizing flow framework (Chen et al., 2018).

In order to find such a vector field v, flow matching starts by specifying a path $I(\mathbf{x}_0, \mathbf{x}_1, t)$ between every two points $\mathbf{x}_0$ and $\mathbf{x}_1$ in $\mathbb{R}^d$. We do this via an interpolant function $I : \mathbb{R}^d \times \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$, which satisfies $I(\mathbf{x}_0, \mathbf{x}_1, 0) = \mathbf{x}_0$ and $I(\mathbf{x}_0, \mathbf{x}_1, 1) = \mathbf{x}_1$. In this work, we will restrict ourselves to the case of spatially linear interpolants, where $I(\mathbf{x}_0, \mathbf{x}_1, t) = \alpha_t \mathbf{x}_0 + \beta_t \mathbf{x}_1$ for functions $\alpha, \beta : [0,1] \to \mathbb{R}$ such that $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = 0$, but more general choices of interpolating paths are possible.

We then define the stochastic interpolant between π_0 and π_1 to be the stochastic process $X = (X_t)_{t\in[0,1]}$ formed by sampling $X_0 \sim \pi_0$, $X_1 \sim \pi_1$ and $Z \sim \mathcal{N}(0, I_d)$ independently and setting $X_t = I(X_0, X_1, t) + \gamma_t Z$ for each $t \in (0,1)$ where $\gamma : [0,1] \to [0,\infty)$ is a function such that $\gamma_0 = \gamma_1 = 0$ and which determines the amount of Gaussian smoothing applied at time t. The motivation for including $\gamma_t Z$ is to smooth the interpolant marginals, as was originally explained by Albergo et al. (2023). Liu et al. (2023) and Albergo & Vanden-Eijnden (2023) omit γ_t, leading to deterministic paths between a given X_0 and X_1, while Albergo et al. (2023) and Lipman et al. (2023) work in the more general case.

The process X is constructed so that its marginals deform smoothly from π_0 to π_1. However, X is not a suitable choice of flow between π_0 and π_1 since it is not causal – it requires knowledge of X_1 to construct X_t. So, we seek a causal flow with the same marginals. The key insight is to define the expected velocity of X by

$$v^X(\mathbf{x}, t) = \mathbb{E}\left[\dot{X}_t \mid X_t = \mathbf{x}\right], \quad \text{for all } \mathbf{x} \in \mathrm{supp}(X_t), \tag{2}$$

where $\dot{X}_t$ is the time derivative of X_t, and setting $v^X(\mathbf{x}, t) = 0$ for $\mathbf{x} \notin \mathrm{supp}(X_t)$ for each $t \in [0,1]$. Then, the following result shows that $v^X(\mathbf{x}, t)$ generates a deterministic flow Z between π_0 and π_1 with the same marginals as X (Liu et al., 2023).

Proposition 1. *Suppose that X is path-wise continuously differentiable, the expected velocity field $v^X(\mathbf{x}, t)$ exists and is locally bounded, and there exists a unique solution $Z^{\mathbf{x}}$ to (1) with velocity field v^X for each $\mathbf{x} \in \mathbb{R}^d$. If Z is the flow with velocity field $v^X(\mathbf{x}, t)$ starting in π_0, then $\mathrm{Law}(X_t) = \mathrm{Law}(Z_t)$ for all $t \in [0,1]$.*

Sufficient conditions for X to be path-wise continuously differentiable and for $v^X(\mathbf{x}, t)$ to exist and be locally bounded are that α, β, and γ are continuously differentiable and that π_0, π_1 have bounded support. We will assume that these conditions hold from now on. We also assume that all our data distributions are absolutely continuous with respect to the Lebesgue measure.

We can learn an approximation to v^X by minimising the objective function

$$\mathcal{L}(v) = \int_0^1 \mathbb{E}\left[w_t \left\|v(X_t, t) - v^X(X_t, t)\right\|^2\right] \mathrm{d}t,$$

over all $v \in \mathcal{V}$ where $\mathcal{V}$ is some class of functions, for some weighting function $w_t : [0,1] \to (0,\infty)$. (Typically, we will take $w_t = 1$ for all t.) As written, this objective is intractable, but we can show that

$$\mathcal{L}(v) = \int_0^1 \mathbb{E}\left[w_t \left\|v(X_t, t) - \dot{X}_t\right\|^2\right] \mathrm{d}t + \text{constant}, \tag{3}$$

where the constant is independent of v (Lipman et al., 2023). This last integral can be empirically estimated in an unbiased fashion given access to samples $X_0 \sim \pi_0$, $X_1 \sim \pi_1$ and the functions α_t, β_t, γ_t and w_t. In practice, we often take $\mathcal{V}$ to be a class of functions parameterised by a neural network $v_\theta(\mathbf{x}, t)$ and minimise $\mathcal{L}(v_\theta)$ over the parameters θ using (3) and stochastic gradient descent. Our hope is that if our approximation v_θ is sufficiently close to v^X, and Y is the flow with velocity field v_θ, then if we take $Y_0 \sim \pi_0$ the distribution of Y_1 is approximately π_1.

Most frequently, flow matching is used as a generative modelling procedure. In order to model a distribution π from which we have access to samples, we set $\pi_1 = \pi$ and take π_0 to be some reference distribution from which it is easy to sample, such as a standard Gaussian. Then, we use the flow matching procedure to learn an approximate flow $v_\theta(\mathbf{x}, t)$ between π_0 and π_1. We generate approximate samples from π_1 by sampling $Z_0 \sim \pi_0$ and integrating the flow equation (1) to find Z_1, which should be approximately distributed according to π_1.

3 Main results

We now present the main results of the paper. First, we show under three assumptions listed below that we can control the approximation error of the flow matching procedure under the 2-Wasserstein distance in terms of the quality of our approximation v_θ. We obtain bounds that depend crucially on the spatial Lipschitz constant of the approximate flow. Second, we show how to control this Lipschitz constant for the true flow v^X under a smoothness assumption on the data distributions, which is explained in Section 3.2. Combined with the first result, this will imply that for sufficiently regular π_0, π_1 there is a choice of $\mathcal{V}$ which contains the true flow such that, if we optimise $L(v)$ over all $v \in \mathcal{V}$, then we can bound the error of the flow matching procedure. Additionally, our bound will be polynomial in the L^2 approximation error. The results of this section will be proved under the following three assumptions.

Assumption 1 (Bound on L^2 approximation error)**.** *The true and approximate drifts $v^X(\mathbf{x}, t)$ and $v_\theta(\mathbf{x}, t)$ satisfy $\int_0^1 \mathbb{E}\left[\|v_\theta(X_t, t) - v^X(X_t, t)\|^2\right] \mathrm{d}t \leq \varepsilon^2$.*

Assumption 2 (Existence and uniqueness of smooth flows)**.** *For each $\mathbf{x} \in \mathbb{R}^d$ and $s \in [0, 1]$ there exist unique flows $(Y^{\mathbf{x}}_{s,t})_{t\in[s,1]}$ and $(Z^{\mathbf{x}}_{s,t})_{t\in[s,1]}$ starting in $Y^{\mathbf{x}}_{s,s} = \mathbf{x}$ and $Z^{\mathbf{x}}_{s,s} = \mathbf{x}$ with velocity fields $v_\theta(\mathbf{x}, t)$ and $v^X(\mathbf{x}, t)$ respectively. Moreover, $Y^{\mathbf{x}}_{s,t}$ and $Z^{\mathbf{x}}_{s,t}$ are continuously differentiable in $\mathbf{x}$, s and t.*

Assumption 3 (Regularity of approximate velocity field)**.** *The approximate flow $v_\theta(\mathbf{x}, t)$ is differentiable in both inputs. Also, for each $t \in (0, 1)$ there is a constant L_t such that $v_\theta(\mathbf{x}, t)$ is L_t-Lipschitz in $\mathbf{x}$.*

Assumption 1 is the natural assumption on the training error given we are learning with the L^2 training loss in (3). Assumption 2 is required since to perform flow matching we need to be able to solve the ODE (1). Without a smoothness assumption on $v_\theta(\mathbf{x}, t)$, it would be possible for the marginals Y_t of the solution to (1) initialised in $Y_0 \sim \pi_0$ to quickly concentrate on subsets of $\mathbb{R}^d$ of arbitrarily small measure under the distribution of X_t. Then, there would be choices of $v_\theta(\mathbf{x}, t)$ which were very different from $v^X(\mathbf{x}, t)$ on these sets of small measure while being equal to $v^X(\mathbf{x}, t)$ everywhere else. For these $v_\theta(\mathbf{x}, t)$ the L^2 approximation error can be kept arbitrarily small while the error of the flow matching procedure is made large. We therefore require some smoothness assumption on $v_\theta(\mathbf{x}, t)$ – we choose to make Assumption 3 since we can show that it holds for the true velocity field, as we do in Section 3.2.

3.1 Controlling the Wasserstein difference between flows

Our first main result is the following, which bounds the error of the flow matching procedure in 2-Wasserstein distance in terms of the L^2 approximation error and the Lipschitz constant of the approximate flow.

Theorem 1. *Suppose that π_0, π_1 are probability distributions on $\mathbb{R}^d$, that Y is the flow starting in π_0 with velocity field v_θ, and $\hat{\pi}_1$ is the law of Y_1. Then, under Assumptions 1-3, we have*

$$W_2(\hat{\pi}_1, \pi_1) \leq \varepsilon \exp\Big\{ \int_0^1 L_t \, \mathrm{d}t \Big\}.$$

We see that the error depends linearly on the L^2 approximation error ε and exponentially on the integral of the Lipschitz constant of the approximate flow L_t. Ostensibly, the exponential dependence on $\int_0^1 L_t \, \mathrm{d}t$

is undesirable, since v_θ may have a large spatial Lipschitz constant. However, we will show in Section 3.2 that the true velocity field is L_t^*-Lipschitz for a choice of L_t^* such that $\int_0^1 L_t^* \, \mathrm{d}t$ depends logarithmically on the amount of Gaussian smoothing. Thus, we may optimise (3) over a class of functions $\mathcal{V}$ which are all L_t^*-Lipschitz, knowing that this class contains the true velocity field, and that if v_θ lies in this class we get a bound on the approximation error which is polynomial in the L^2 approximation error, providing we make a suitable choice of α_t, β_t and γ_t.

We remark that our Theorem 1 is similar in content to Proposition 3 in Albergo & Vanden-Eijnden (2023). The crucial difference is that we work with a time-varying Lipschitz constant, which is required in practice since in many cases L_t explodes as $t \to 0$ or 1, as happens for example if π_0 or π_1 do not have full support.

A key ingredient in the proof of Theorem 1 will be the Alekseev–Gröbner formula, which provides a way to control the difference between the solutions to two different ODEs in terms of the difference between their drifts (Gröbner, 1960; Alekseev, 1961).

Proposition 2 (Alekseev–Gröbner)**.** *Let $\mu(\mathbf{x}, t) : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ be a function which is continuous in t and continuously differentiable in $\mathbf{x}$. Let $X : \mathbb{R}^d \times [0,T]^2 \to \mathbb{R}^d$ be a continuous solution of the integral equation*

$$X_{s,t}^{\mathbf{x}} = \mathbf{x} + \int_s^t \mu(r, X_{s,r}^{\mathbf{x}}) \, \mathrm{d}r,$$

and let $Y : [0,T] \to \mathbb{R}^d$ be another continuously differentiable process. Then

$$X_{0,T}^{Y_0} - Y_T = \int_0^T \left(\nabla_{\mathbf{x}} X_{r,T}^{Y_r} \right) \left(\mu(r, Y_r) - \frac{\mathrm{d}Y_r}{\mathrm{d}t} \right) \mathrm{d}r.$$

We now give the proof of Theorem 1.

Proof of Theorem 1. Define $Y_{s,t}^{\mathbf{x}}$ and $Z_{s,t}^{\mathbf{x}}$ as in Assumption 2. As $Y_{s,t}^{\mathbf{x}} \in C(\mathbb{R}^d \times [0,1]^2)$, $Z_{0,t}^{\mathbf{x}} \in C^1(\mathbb{R}^d \times [0,1])$ and $v_\theta(\mathbf{x}, t) \in C^{0,1}(\mathbb{R}^d \times [0,1])$ for any $\mathbf{x} \in \mathbb{R}^d$, the Alekseev–Gröbner formula implies that

$$Y_{0,t}^{\mathbf{x}} - Z_{0,t}^{\mathbf{x}} = \int_0^t \left(\nabla_{\mathbf{x}} Y_{s,t}^{Z_s} \right) \left(v_\theta(Z_s, s) - v^X(Z_s, s) \right) \mathrm{d}s. \tag{4}$$

We can write $Y_{s,t}^{\mathbf{x}} = \mathbf{x} + \int_s^t v_\theta(Y_{s,r}^{\mathbf{x}}, r) \, \mathrm{d}r$, and so

$$\nabla_{\mathbf{x}} Y_{s,t}^{\mathbf{x}} = I + \int_s^t \nabla_{\mathbf{x}} v_\theta(Y_{s,r}^{\mathbf{x}}, r) \nabla_{\mathbf{x}} Y_{s,r}^{\mathbf{x}} \, \mathrm{d}r.$$

It follows that

$$\frac{\partial}{\partial t} \left\| \nabla_{\mathbf{x}} Y_{s,t}^{\mathbf{x}} \right\|_{\mathrm{op}} \leq \left\| \frac{\partial}{\partial t} \nabla_{\mathbf{x}} Y_{s,t}^{\mathbf{x}} \right\|_{\mathrm{op}} = \left\| \nabla_{\mathbf{x}} v_\theta(Y_{s,t}^{\mathbf{x}}, t) \nabla_{\mathbf{x}} Y_{s,t}^{\mathbf{x}} \right\|_{\mathrm{op}} \leq L_t \left\| \nabla_{\mathbf{x}} Y_{s,t}^{\mathbf{x}} \right\|_{\mathrm{op}},$$

where $\| \cdot \|_{\mathrm{op}}$ denotes the operator norm with respect to the ℓ^2 metric on $\mathbb{R}^d$. Therefore, by Grönwall's lemma we have $\left\| \nabla_{\mathbf{x}} Y_{s,t}^{\mathbf{x}} \right\|_{\mathrm{op}} \leq \exp\left\{ \int_s^t L_r \, \mathrm{d}r \right\} \leq K$, where $K := \exp\left\{ \int_0^1 L_t \, \mathrm{d}t \right\}$. Applied to (4) at $t = 1$, we get

$$\| Y_1^{\mathbf{x}} - Z_1^{\mathbf{x}} \|_2 \leq \int_0^1 \left\| \nabla_{\mathbf{x}} Y_{s,1}^{Z_s} \right\|_{\mathrm{op}} \left\| v^X(Z_s, s) - v_\theta(Z_s, s) \right\|_2 \mathrm{d}s \leq K \int_0^1 \left\| v^X(Z_s, s) - v_\theta(Z_s, s) \right\|_2 \mathrm{d}s.$$

Letting $\mathbf{x} \sim \pi_0$ and taking expectations, we deduce

$$\begin{aligned} \mathbb{E}\left[\| Y_1 - Z_1 \|_2^2 \right] &\leq K^2 \mathbb{E}\left[\left(\int_0^1 \left\| v^X(Z_t, t) - v_\theta(Z_t, t) \right\|_2 \mathrm{d}t \right)^2 \right] \\ &\leq K^2 \int_0^1 \mathbb{E}\left[\left\| v_\theta(X_t, t) - v^X(X_t, t) \right\|^2 \right] \mathrm{d}t. \end{aligned}$$

Since $W_2(\hat{\pi}_1, \pi_1) = W_2(\mathrm{Law}(Y_1), \mathrm{Law}(Z_1)) \leq \mathbb{E}\left[\| Y_1 - Z_1 \|_2^2 \right]^{1/2}$, the result follows from our assumption on the L^2 error and the definition of K. □

The application of the Alekseev–Gröbner formula in (4) is not symmetric in Y and Z. Applying the formula with the roles of Y and Z reversed would mean we require a bound on $\left\|\nabla_{\mathbf{x}} Y^{\mathbf{x}}_{s,t}\right\|_{\mathrm{op}}$, which we could bound in terms of the Lipschitz constant of the true velocity field, leading to a more natural condition. However, we would then also be required to control the velocity approximation error $\|v_\theta(Y_t, t) - v^X(Y_t, t)\|_2$ along the paths of Y, which we have much less control over due to the nature of the training objective.

3.2 Smoothness of the velocity fields

As remarked in Section 3.1, the bound in Theorem 1 is exponentially dependent on the Lipschitz constant of v_θ. In this section, we control the corresponding Lipschitz constant of v^X. We prefer this to controlling the Lipschitz constant of v_θ directly since this is determined by our choice of $\mathcal{V}$, the class of functions over which we optimise (3).

In this section, we will relax the constraints from Section 2 that $\gamma_0 = \gamma_1 = 0$ and $\alpha_0 = \beta_1 = 1$. We do this because allowing $\gamma_t > 0$ for all $t \in [0, 1]$ means that we have some Gaussian smoothing at every t, which will help ensure the resulting velocity fields are sufficiently regular. This will mean that instead of learning a flow between π_0 and π_1 we will actually be learning a flow between $\tilde{\pi}_0$ and $\tilde{\pi}_1$, the distributions of $\alpha_0 X_0 + \gamma_0 Z$ and $\beta_1 X_1 + \gamma_1 Z$ respectively. However, if we centre X_0 and X_1 so that $\mathbb{E}[X_0] = \mathbb{E}[X_1] = 0$ and let R be such that $\|X_0\|_2, \|X_1\|_2 \leq R$, then $W_2(\tilde{\pi}_0, \pi_0)^2 \leq (1 - \alpha_0)^2 R^2 + d\gamma_0^2$ and similarly for $\tilde{\pi}_1, \pi_1$, so it follows that by taking γ_0, γ_1 sufficiently close to 0 and α_0, β_1 sufficiently close to 1 we can make $\tilde{\pi}_0, \tilde{\pi}_1$ very close in 2-Wasserstein distance to π_0, π_1 respectively. Note that if π_0 is easy to sample from then $\tilde{\pi}_0$ is also easy to sample from (we may simulate samples from $\tilde{\pi}_0$ by drawing $X_0 \sim \pi_0$, $Z \sim \mathcal{N}(0, I_d)$ independently, setting $\tilde{X}_0 = \alpha_0 X_0 + \gamma_0 Z$ and noting that $\tilde{X}_0 \sim \tilde{\pi}_0$), so we can run the flow matching procedure starting from $\tilde{\pi}_0$.

For arbitrary choices of π_0 and π_1 the expected velocity field $v^X(\mathbf{x}, t)$ can be very badly behaved, and so we will require an additional regularity assumption on the process X. This notion of regularity is somewhat non-standard, but is the natural one emerging from our proofs and bears some similarity to quantities controlled in stochastic localization (see discussion below).

Definition 1. *For a real number $\lambda \geq 1$, we say an $\mathbb{R}^d$-valued random variable W is λ-**regular** if, whenever we take $\tau \in (0, \infty)$ and $\xi \sim \mathcal{N}(0, \tau^2 I_d)$ independently of W and set $W' = W + \xi$, for all $\mathbf{x} \in \mathbb{R}^d$ we have*

$$\left\|\mathrm{Cov}_{\xi | W' = \mathbf{x}}(\xi)\right\|_{\mathrm{op}} \leq \lambda \tau^2.$$

We say that a distribution on $\mathbb{R}^d$ is λ-regular if the associated random variable is λ-regular.

Assumption 4 (Regularity of data distributions)**.** *For some $\lambda \geq 1$, $\alpha_t X_0 + \beta_t X_1$ is λ-regular for all $t \in [0, 1]$.*

To understand what it means for a random variable W to be λ-regular, note that before conditioning on W', we have $\|\mathrm{Cov}(\xi)\|_{\mathrm{op}} = \tau^2$. We can think of conditioning on $W' = \mathbf{x}$ as re-weighting the distribution of ξ proportionally to $p_W(\mathbf{x} - \xi)$, where $p_W(\cdot)$ is the density function of W. So, W is λ-regular if this re-weighting causes the covariance of ξ to increase by at most a factor of λ for any choice of τ.

Informally, if τ is much smaller than the scale over which $\log p_W(\cdot)$ varies then this re-weighting should have negligible effect, while for large τ we can write $\|\mathrm{Cov}_{\xi | W' = \mathbf{x}}(\xi)\|_{\mathrm{op}} = \|\mathrm{Cov}_{W | W' = \mathbf{x}}(W)\|_{\mathrm{op}}$ and we expect this to be less than τ^2 once τ is much greater than the typical magnitude of W. Thus $\lambda \approx 1$ should suffice for sufficiently small or large τ and the condition that W is λ-regular controls how much conditioning on W' can change the behavior of ξ as we transition between these two extremes.

If W is log-concave or alternatively Gaussian on a linear subspace of $\mathbb{R}^d$, then we show in Appendix A.1 that W is λ-regular with $\lambda = 1$. Additionally, we also show that a mixture of Gaussians $\pi = \sum_{i=1}^K \mu_i \mathcal{N}(\mathbf{x}_i, \sigma^2 I_d)$, where the weights μ_i satisfy $\sum_{i=1}^K \mu_i = 1$, $\sigma > 0$, and $\|\mathbf{x}_i\| \leq R$ for $i = 1, \ldots, K$, is λ-regular with $\lambda = 1 + (R^2/\sigma^2)$. This shows that λ-regularity can hold even for highly multimodal distributions. More generally, we show in Appendix A.2 that we always have $\|\mathrm{Cov}_{\xi | W' = \mathbf{x}}(\xi)_{\mathrm{op}}\| \leq O(d\tau^2)$ with high probability. Then, we can interpret Definition 1 as insisting that this inequality always holds, where the dependence on d is incorporated into the parameter λ. (We see that in the worst case λ may scale linearly with d, but we expect in practice that λ will be approximately constant in many cases of interest.)

Controlling covariances of random variables given noisy observations of those variables is also a problem which arises in stochastic localization (Eldan, 2013; 2020), and similar bounds to the ones we use here have been established in this context. For example, Alaoui & Montanari (2022) show that $\mathbb{E}\left[\mathrm{Cov}_{\xi|W'=\mathbf{x}}(W')\right] \leq \tau^2 I_d$ in our notation. However, the bounds from stochastic localization only hold in expectation over $\mathbf{x}$ distributed according to the law of W', whereas we require bounds that hold pointwise, or at least for $\mathbf{x}$ distributed according to the law of $Y^{\mathbf{x}}_{s,t}$, which is much harder to control.

We now aim to bound the Lipschitz constant of the true velocity field under Assumption 4. The first step is to show that v^X is differentiable and to get an explicit formula for its derivative. In the following sections, we abbreviate the expectation and covariance conditioned on $X_t = \mathbf{x}$ to $\mathbb{E}_{\mathbf{x}}[\,\cdot\,]$ and $\mathrm{Cov}_{\mathbf{x}}(\cdot)$ respectively.

Lemma 1. *If X is the stochastic interpolant between π_0 and π_1, then $v^X(\mathbf{x}, t)$ is differentiable with respect to $\mathbf{x}$ and*

$$\nabla_{\mathbf{x}} v^X(\mathbf{x}, t) = \frac{\dot\gamma_t}{\gamma_t} I_d - \frac{1}{\gamma_t}\mathrm{Cov}_{\mathbf{x}}(\dot X_t, Z).$$

The proof of Lemma 1 is given in Appendix B.1. Using Lemma 1, we can derive the following theorem, which provides a bound on the time integral of the Lipschitz constant of v_θ.

Theorem 2. *If X is the stochastic interpolant between two distributions π_0 and π_1 on $\mathbb{R}^d$ with bounded support, then under Assumption 4 for each $t \in (0,1)$ there is a constant L_t^* such that $v^X(\mathbf{x}, t)$ is L_t^*-Lipschitz in $\mathbf{x}$ and*

$$\int_0^1 L_t^* \,\mathrm{d}t \leq \lambda\left(\int_0^1 \frac{|\dot\gamma_t|}{\gamma_t}\,\mathrm{d}t\right) + \lambda^{1/2} R\left(\int_0^1 \frac{|\dot\alpha_t|}{\gamma_t}\,\mathrm{d}t + \int_0^1 \frac{|\dot\beta_t|}{\gamma_t}\,\mathrm{d}t\right),$$

where $\operatorname{supp}\pi_0,\ \operatorname{supp}\pi_1 \subseteq \bar B(0, R)$, the closed ball of radius R around the origin.

Proof. Using Lemma 1 plus the fact that $\dot X_t = \dot\alpha_t X_0 + \dot\beta_t X_1 + \dot\gamma_t Z$, we can write

$$\nabla_{\mathbf{x}} v^X(\mathbf{x}, t) = \frac{\dot\gamma_t}{\gamma_t}\Big(I_d - \mathrm{Cov}_{\mathbf{x}}(Z, Z)\Big) - \frac{\dot\alpha_t}{\gamma_t}\mathrm{Cov}_{\mathbf{x}}(X_0, Z) - \frac{\dot\beta_t}{\gamma_t}\mathrm{Cov}_{\mathbf{x}}(X_1, Z).$$

Since we can express $X_t = (\alpha_t X_0 + \beta_t X_1) + \gamma_t Z$ where $\alpha_t X_0 + \beta_t X_1$ is λ-regular by Assumption 4 and $\gamma_t Z \sim \mathcal{N}(0, \gamma_t^2 I_d)$, we have $\|\mathrm{Cov}_{\mathbf{x}}(\gamma_t Z)\|_{\mathrm{op}} \leq \lambda\gamma_t^2$. It follows that

$$\|I_d - \mathrm{Cov}_{\mathbf{x}}(Z, Z)\|_{\mathrm{op}} \leq \max\Big(1, \|\mathrm{Cov}_{\mathbf{x}}(Z, Z)\|_{\mathrm{op}}\Big) \leq \lambda.$$

Also, using Cauchy–Schwarz we have

$$\|\mathrm{Cov}_{\mathbf{x}}(X_0, Z)\|_{\mathrm{op}} \leq \|\mathrm{Cov}_{\mathbf{x}}(X_0)\|_{\mathrm{op}}^{1/2}\,\|\mathrm{Cov}_{\mathbf{x}}(Z)\|_{\mathrm{op}}^{1/2} \leq \lambda^{1/2} R,$$

and a similar result holds for X_1 in place of X_0.

Putting this together, we get

$$\begin{aligned}\big\|\nabla_{\mathbf{x}} v^X(\mathbf{x}, t)\big\|_{\mathrm{op}} &\leq \frac{|\dot\gamma_t|}{\gamma_t}\,\|I_d - \mathrm{Cov}_{\mathbf{x}}(Z, Z)\|_{\mathrm{op}} + \frac{|\dot\alpha_t|}{\gamma_t}\,\|\mathrm{Cov}_{\mathbf{x}}(X_0, Z)\|_{\mathrm{op}} + \frac{|\dot\beta_t|}{\gamma_t}\,\|\mathrm{Cov}_{\mathbf{x}}(X_1, Z)\|_{\mathrm{op}} \\ &\leq \lambda\frac{|\dot\gamma_t|}{\gamma_t} + \lambda^{1/2} R\left(\frac{|\dot\alpha_t|}{\gamma_t} + \frac{|\dot\beta_t|}{\gamma_t}\right). \end{aligned} \tag{5}$$

Finally, since $v^X(\mathbf{x}, t)$ is differentiable with uniformly bounded derivative, it follows that $v^X(\mathbf{x}, t)$ is L_t^*-Lipschitz with $L_t^* = \sup_{\mathbf{x}\in\mathbb{R}^d}\big\|\nabla_{\mathbf{x}} v^X(\mathbf{x}, t)\big\|_{\mathrm{op}}$. Integrating (5) from $t = 0$ to $t = 1$, the result follows. □

We now provide some intuition for the bound in Theorem 2. The key term on the RHS is the first one, which we typically expect to be on the order of $\log(\gamma_{\max}/\gamma_{\min})$ where $\gamma_{\max}$ and $\gamma_{\min}$ are the maximum and minimum values taken by γ_t on the interval $[0, 1]$ respectively. This is suggested by the following lemma.

Lemma 2. *Suppose that $\gamma_0 = \gamma_1 = \gamma_{\min}$ and γ_t increases smoothly from $\gamma_{\min}$ at $t = 0$ to $\gamma_{\max}$ before decreasing back to $\gamma_{\min}$ at $t = 1$. Then $\int_0^1 (|\dot\gamma_t|/\gamma_t)\,\mathrm{d}t = 2\log(\gamma_{\max}/\gamma_{\min})$.*

Proof. Note that we can write $|\dot{\gamma}_t|/\gamma_t = |\mathrm{d}(\log \gamma_t)/\mathrm{d}t|$, so $\int_0^1 (|\dot{\gamma}_t|/\gamma_t)\,\mathrm{d}t$ is simply the total variation of $\log \gamma_t$ over the interval $[0,1]$. By the assumptions on γ_t, we see this total variation is equal to $2\log(\gamma_{\max}/\gamma_{\min})$. □

The first term on the RHS in Theorem 2 also explains the need to relax the boundary conditions, since $\gamma_0 = 0$ or $\gamma_1 = 0$ would cause $\int_0^1 (|\dot{\gamma}_t|/\gamma_t)\,\mathrm{d}t$ to diverge.

We can ensure the second terms in Theorem 2 are relatively small through a suitable choice of α_t, β_t and γ_t. Since α_t and β_t are continuously differentiable, $|\dot{\alpha}_t|$ and $|\dot{\beta}_t|$ are bounded, so to control these terms we should pick γ_t such that $R\int_0^1 (1/\gamma_t)\,\mathrm{d}t$ is small, while respecting the fact that we need $\gamma_t \ll 1$ at the boundaries. One sensible choice among many would be $\gamma_t = 2R\sqrt{(\delta + t)(1 + \delta - t)}$ for some $\delta \ll 1$, where $\int_0^1 (|\dot{\gamma}_t|/\gamma_t)\,\mathrm{d}t \approx 2\log(1/\sqrt{\delta})$ and $\int_0^1 (1/\gamma_t)\,\mathrm{d}t \approx 2\pi$. In this case, the bound of Theorem 2 implies $\int_0^1 L_t^*\,\mathrm{d}t \le \lambda \log(1/\delta) + 2\pi\lambda^{1/2} \sup_{t\in[0,1]}(|\dot{\alpha}_t| + |\dot{\beta}_t|)$, so if $\delta \ll 1$ the dominant term in our bound is $\lambda \log(1/\delta)$.

3.3 Bound on the Wasserstein error of flow matching

Now, we demonstrate how to combine the results of the previous two sections to get a bound on the error of the flow matching procedure that applies in settings of practical interest. In Section 3.2, we saw that we typically have $\int_0^1 L_t^*\,\mathrm{d}t \sim \lambda \log(\gamma_{\max}/\gamma_{\min})$. The combination of this logarithm and the exponential in Theorem 1 should give us bounds on the 2-Wasserstein error which are polynomial in ε and $\gamma_{\max}, \gamma_{\min}$.

The main idea is that in order to ensure that we may apply Theorem 1, we should optimise $\mathcal{L}(v)$ over a class of functions $\mathcal{V}$ which all satisfy Assumption 3. (Technically, we only need Assumption 3 to hold at the minimum of $\mathcal{L}(v)$, but since it is not possible to know what this minimum will be ex ante, it is easier in practice to enforce that Assumption 3 holds for all $v \in \mathcal{V}$.) Given distributions π_0, π_1 satisfying Assumption 4 as well as specific choices of α_t, β_t and γ_t we define $K_t = \lambda(|\dot{\gamma}_t|/\gamma_t) + \lambda^{1/2} R\{(|\dot{\alpha}_t|/\alpha_t) + (|\dot{\beta}_t|/\beta_t)\}$ and let $\mathcal{V}$ be the set of functions $v : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ which are K_t-Lipschitz in $\mathbf{x}$ for all $t \in [0,1]$.

Theorem 3. *Suppose that Assumptions 1-4 hold and that γ_t is concave on $[0,1]$. Then $v^X \in \mathcal{V}$ and for any $v_\theta \in \mathcal{V}$, if Y is a flow starting in $\tilde{\pi}_0$ with velocity field v_θ and $\hat{\pi}_1$ is the law of Y_1,*

$$W_2(\hat{\pi}_1, \tilde{\pi}_1) \le C^{\lambda^{1/2}} \varepsilon \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right)^{2\lambda},$$

where $\gamma_{\min} = \inf_{t\in[0,1]} \gamma_t$, $\gamma_{\max} = \sup_{t\in[0,1]} \gamma_t$ and $C = \exp\left\{R\left\{\int_0^1 (|\dot{\alpha}_t|/\gamma_t)\,\mathrm{d}t + \int_0^1 (|\dot{\beta}_t|/\gamma_t)\,\mathrm{d}t\right\}\right\}$.

Proof. First, note that v^X is K_t-Lipschitz in $\mathbf{x}$ for all $t \in [0,1]$ by the proof of Theorem 2, so $v^X \in \mathcal{V}$. Then, since γ_t is concave, $\int_0^1 (|\dot{\gamma}_t|/\gamma_t)\,\mathrm{d}t \le 2\log(\gamma_{\max}/\gamma_{\min})$ by Lemma 2. Thus, $\int_0^1 K_t\,\mathrm{d}t \le 2\lambda \log(\gamma_{\max}/\gamma_{\min}) + \lambda^{1/2} \log C$. Hence for any $v_\theta \in \mathcal{V}$ we may apply Theorem 1 to get

$$W_2(\hat{\pi}_1, \tilde{\pi}_1) \le \varepsilon \exp\left\{\int_0^1 K_t\,\mathrm{d}t\right\} \le C^{\lambda^{1/2}} \varepsilon \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right)^{2\lambda},$$

where $\tilde{\pi}_0, \tilde{\pi}_1$ are replacing π_0, π_1 since we have relaxed the boundary conditions. □

Theorem 3 shows that if we optimise $\mathcal{L}(v)$ over the class $\mathcal{V}$, we can bound the resulting distance between the flow matching paths. We typically choose γ_t to scale with R, so C should be thought of as a constant that is $\Theta(1)$ and depends only on the smoothness of the interpolating paths. For a given distribution, λ is fixed and our bound becomes polynomial in ε and $\gamma_{\max}/\gamma_{\min}$.

Finally, our choice of $\gamma_{\min}$ controls how close the distributions $\tilde{\pi}_0$, $\tilde{\pi}_1$ are to π_0, π_1 respectively. We can combine this with our bound on $W_2(\hat{\pi}_1, \tilde{\pi}_1)$ to get the following result.

Theorem 4. *Suppose that Assumptions 1-4 hold, that γ_t is concave on $[0,1]$, and that $\alpha_0 = \beta_1 = 1$ and $\gamma_0 = \gamma_1 = \gamma_{\min}$. Then, for any $v_\theta \in \mathcal{V}$, if Y is a flow starting in $\tilde{\pi}_0$ with velocity field v_θ and $\hat{\pi}_1$ is the law of Y_1,*

$$W_2(\hat{\pi}_1, \pi_1) \le C^{\lambda^{1/2}} \varepsilon \left(\frac{\gamma_{\max}}{\gamma_{\min}}\right)^{2\lambda} + \sqrt{d}\gamma_{\min}$$

where $\gamma_{\min}$, $\gamma_{\max}$ and C are as before.

Proof. We have $W_2(\hat{\pi}_1, \pi_1) \leq W_2(\hat{\pi}_1, \tilde{\pi}_1) + W_2(\tilde{\pi}_1, \pi_1)$. The first term is controlled by Theorem 3, while for the second term we have $W_2(\tilde{\pi}_1, \pi_1)^2 \leq (1 - \beta_1)R^2 + d\gamma_0^2$, as noted previously. Using $\beta_1 = 1$ and combining these two results completes the proof. □

Optimizing the expression in Theorem 4 over $\gamma_{\min}$, we see that for a given L^2 error tolerance ε, we should take $\gamma_{\min} \sim d^{-1/(4\lambda+2)}\varepsilon^{1/(2\lambda+1)}$. We deduce the following corollary.

Corollary 1. *In the setting of Theorem 4, if we take $\gamma_{\min} \sim d^{-1/(4\lambda+2)}\varepsilon^{1/(2\lambda+1)}$ then the total Wasserstein error of the flow matching procedure is of order $W_2(\hat{\pi}_1, \pi_1) \lesssim d^{2\lambda/(4\lambda+2)}\varepsilon^{1/(2\lambda+1)}$.*

4 Application to probability flow ODEs

For generative modelling applications, π_1 is the target distribution and π_0 is typically chosen to be a simple reference distribution. A common practical choice for π_0 is a standard Gaussian distribution, and in this setting the flow matching framework reduces to the probability flow ODE (PF-ODE) framework for diffusion models (Song et al., 2021b). (Technically, this corresponds to running the PF-ODE framework for infinite time. Finite time versions of the PF-ODE framework can be recovered by taking β_0 to be positive but small.) Previously, this correspondence has been presented by taking $\gamma_t = 0$ for all $t \in [0, 1]$ and $\pi_0 = \mathcal{N}(0, I_d)$ in our notation from Section 2 (Liu et al., 2023). Instead, we choose to set $\alpha_t = 0$ for all $t \in [0, 1]$ and have $\gamma_t > 0$, which recovers exactly the same framework, just with Z playing the role of the reference random variable rather than X_0. We do this because it allows us to apply the results of Section 3 directly.

Because we have this alternative representation, we can strengthen our results in the PF-ODE setting. First, note that Assumption 4 simplifies, so that we only need to assume that π_1 is λ-regular. We thus replace Assumption 4 with Assumption 4' for the rest of this section.

Assumption 4' (Regularity of data distribution, Gaussian case)**.** *For some $\lambda \geq 1$, the distribution π_1 is λ-regular.*

We also get the following alternative form of Lemma 1 in this setting, which is proved in Appendix B.2.

Lemma 3. *If X is the stochastic interpolant in the PF-ODE setting above, then $v^X(\mathbf{x}, t)$ is differentiable with respect to $\mathbf{x}$ and*

$$\nabla_{\mathbf{x}} v^X(\mathbf{x}, t) = \frac{\dot{\gamma}_t}{\gamma_t} I_d - \left(\frac{\dot{\gamma}_t}{\gamma_t} - \frac{\dot{\beta}_t}{\beta_t} \right) \mathrm{Cov}_{\mathbf{x}}(Z).$$

Using Lemma 3, we can follow a similar argument to the proof of Theorem 2 to get additional control on the terms L_t^*.

Theorem 5. *If X is the stochastic interpolant in the PF-ODE setting above, then under Assumption 4' for each $t \in (0, 1)$ there is a constant L_t^* such that $v^X(\mathbf{x}, t)$ is L_t^*-Lipschitz in $\mathbf{x}$ and*

$$\int_0^1 L_t^* \, \mathrm{d}t \leq \lambda \left(\int_0^1 \frac{|\dot{\gamma}_t|}{\gamma_t} \, \mathrm{d}t \right) + \int_0^1 \min \left\{ \lambda \frac{|\dot{\beta}_t|}{\beta_t}, \, \lambda^{1/2} R \, \frac{|\dot{\beta}_t|}{\gamma_t} \right\} \, \mathrm{d}t.$$

Proof. From Theorem 2 we know that $v^X(\mathbf{x}, t)$ is differentiable and Lipschitz with some constant L_t^*. From (5) in the proof of Theorem 2, for any $\mathbf{x} \in \mathbb{R}^d$ we have

$$\left\| \nabla_{\mathbf{x}} v^X(\mathbf{x}, t) \right\|_{\mathrm{op}} \leq \lambda \frac{|\dot{\gamma}_t|}{\gamma_t} + \lambda^{1/2} R \, \frac{|\dot{\beta}_t|}{\gamma_t}, \tag{6}$$

since $\dot{\alpha}_t = 0$ in this setting. Also, using Lemma 3,

$$\left\| \nabla_{\mathbf{x}} v^X(\mathbf{x}, t) \right\|_{\mathrm{op}} \leq \frac{|\dot{\gamma}_t|}{\gamma_t} \left\| I_d - \mathrm{Cov}_{\mathbf{x}}(Z) \right\|_{\mathrm{op}} + \frac{|\dot{\beta}_t|}{\beta_t} \left\| \mathrm{Cov}_{\mathbf{x}}(Z) \right\|_{\mathrm{op}} \leq \lambda \frac{|\dot{\gamma}_t|}{\gamma_t} + \lambda \frac{|\dot{\beta}_t|}{\beta_t}. \tag{7}$$

As $L_t^* = \sup_{\mathbf{x} \in \mathbb{R}^d} \left\| \nabla_{\mathbf{x}} v^X(\mathbf{x}, t) \right\|_{\mathrm{op}}$, combining (6), (7) and integrating from $t = 0$ to $t = 1$ gives the result. □

To interpret Theorem 5, recall that the boundary conditions we are operating under are $\gamma_0 = \beta_1 = 1$, $\beta_0 = 0$ and $\gamma_1 \ll 1$, so that $\tilde{\pi}_0 = \mathcal{N}(0, I_d)$ and $\tilde{\pi}_1$ is π_1 plus a small amount of Gaussian noise at scale γ_1. The following corollary gives the implications of Theorem 5 in the standard variance-preserving (VP) and variance-exploding (VE) PF-ODE settings of Song et al. (2021b).

Corollary 2. *Suppose that we are in the setting of Theorem 5, so that $v^X(\mathbf{x}, t)$ is L_t^*-Lipschitz in $\mathbf{x}$. Then,*

(i) if $\gamma_t = R\cos((\frac{\pi}{2} - \delta)t)$ and $\beta_t = \sin((\frac{\pi}{2} - \delta)t)$ for $\delta \ll 1$, corresponding to the VP ODE framework of Song et al. (2021b), then $\int_0^1 L_t^* \, \mathrm{d}t \leq \lambda(1 + \log(1/\gamma_1))$;

(ii) if $\beta_t = 1$ for all $t \in [0, 1]$ and γ_t is decreasing, corresponding to the VE ODE framework of Song et al. (2021b), then $\int_0^1 L_t^* \, \mathrm{d}t \leq \lambda \log(1/\gamma_1)$.

Proof. In both cases, we apply Theorem 5 to bound $\int_0^1 L_t^*$. As γ_t is decreasing, we have $\int_0^1 |\dot{\gamma}_t|/\gamma_t \, \mathrm{d}t = \log(\gamma_0/\gamma_1)$, similarly to in the proof of Lemma 2. For (i), the second term on the RHS of Theorem 5 can be bounded above by λ, while for (ii) it vanishes entirely. □

In each case, we can plug the resulting bound into Theorem 1 to get a version of Theorem 3 for the PF-ODE setting with the given noising schedule. We define $K_t = \lambda(|\dot{\gamma}_t|/\gamma_t) + \min\{\lambda(|\dot{\beta}_t|/\beta_t), \lambda^{1/2} R(|\dot{\beta}_t|/\gamma_t)\}$ and let $\mathcal{V}$ be the set of functions $v : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ which are K_t-Lipschitz in $\mathbf{x}$ for all $t \in [0, 1]$ as before.

Theorem 6. *Suppose that Assumptions 1-3 and 4' hold and we are in either (i) the VP ODE or (ii) the VE ODE setting above. Then $v^X \in \mathcal{V}$ and for any $v_\theta \in \mathcal{V}$, if Y is a flow starting in $\tilde{\pi}_0$ with velocity field v_θ and $\hat{\pi}_1$ is the law of Y_1, then*

(i) in the VP ODE setting, we have $W_2(\hat{\pi}_1, \tilde{\pi}_1) \leq \varepsilon(e/\gamma_1)^\lambda$;

(ii) in the VE ODE setting, we have $W_2(\hat{\pi}_1, \tilde{\pi}_1) \leq \varepsilon(1/\gamma_1)^\lambda$.

Proof. That $v^X \in \mathcal{V}$ follows analogously to the proof for Theorem 3. The results then follow from combining Theorem 1 with the bounds in Corollary 2, as in the proof of Theorem 3. □

5 Conclusion

We have provided the first bounds on the error of the general flow matching procedure that apply in the case of completely deterministic sampling. Under the smoothness criterion of Assumption 4 on the data distributions, we have derived bounds which for a given sufficiently smooth choice of α_t, β_t, γ_t and fixed data distribution are polynomial in the level of Gaussian smoothing $\gamma_{\max}/\gamma_{\min}$ and the L^2 error ε of our velocity approximation.

However, our bounds still depend exponentially on the parameter λ from Assumption 4. It is therefore a key question how λ behaves for typical distributions or as we scale up the dimension. In particular, the informal argument we provide in Section A.2 suggests that λ may scale linearly with d. However, we expect that in many practical cases λ should be $\Theta(1)$ even as d scales. Furthermore, even if this is not the case, in the proof of Theorem 1 we only require control of the norm of $\nabla_{\mathbf{x}} v_\theta(\mathbf{x}, t)$ when applied to $\nabla_{\mathbf{x}} Y_{s,t}^{\mathbf{x}}$. In cases such as these, the operator norm bound is typically loose unless $\nabla_{\mathbf{x}} Y_{s,t}^{\mathbf{x}}$ is highly correlated with the largest eigenvectors of $\nabla_{\mathbf{x}} v_\theta(\mathbf{x}, t)$. We see no a priori reason why these two should be highly correlated in practice, and so we expect in most practical applications to get behaviour which is much better than linear in d. Quantifying this behaviour remains a question of interest.

Finally, the fact that we get weaker results in the fully deterministic setting than Albergo et al. (2023) do when adding a small amount of Gaussian noise in the reverse sampling procedure suggests that some level of Gaussian smoothing is helpful for sampling and leads to the suppression of the exploding divergences of paths.

Acknowledgements

We thank Michael Albergo, Nicholas Boffi and Eric Vanden-Eijnden for their comments on an early version of this paper. Joe Benton was supported by the EPSRC through the StatML CDT (EP/S023151/1). Arnaud Doucet acknowledges support from EPSRC grants EP/R034710/1 and EP/R018561/1.

References

Ahmed El Alaoui and Andrea Montanari. An Information-Theoretic View of Stochastic Localization. *IEEE Transactions on Information Theory*, 68(11):7423–7426, 2022.

Michael S Albergo and Eric Vanden-Eijnden. Building Normalizing Flows with Stochastic Interpolants. In *International Conference on Learning Representations*, 2023.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Vladimir M Alekseev. An estimate for the perturbations of the solutions of ordinary differential equations. *Vestn. Mosk. Univ. Ser. I. Math. Mekh*, 2:28–36, 1961.

Heli Ben-Hamu, Samuel Cohen, Joey Bose, Brandon Amos, Aditya Grover, Maximilian Nickel, Ricky T Q Chen, and Yaron Lipman. Matching Normalizing Flows and Probability Paths on Manifolds. In *International Conference on Machine Learning*, 2022.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-Linear Convergence Bounds for Diffusion Models via Stochastic Localization. *arXiv preprint arXiv:2308.03686*, 2023.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative Modeling with Denoising Auto-Encoders and Langevin Sampling. *arXiv preprint arXiv:2002.00107*, 2022.

Herm Jan Brascamp and Elliott H Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions. In *International Conference on Machine Learning*, 2023a.

Ricky T Q Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, 2018.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. *arXiv preprint arXiv:2305.11798*, 2023b.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023c.

Sitan Chen, Giannis Daras, and Alexandros G Dimakis. Restoration-Degradation Beyond Linear Diffusions: A Non-Asymptotic Analysis For DDIM-Type Samplers. In *International Conference on Machine Learning*, 2023d.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite Fisher information is all you need. *arXiv preprint arXiv:2308.12240*, 2023.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems*, 2021.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Ronen Eldan. Thin Shell Implies Spectral Gap Up to Polylog via a Stochastic Localization Scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3-4):737–755, 2020.

Will Grathwohl, Ricky T Q Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form Continuous Dynamics for Scalable Reverse Generative Models. In *International Conference on Learning Representations*, 2019.

Wolfgang Gröbner. *Die Lie-Reihen und ihre Anwendungen.* Berlin: VEB Deutscher Verlag der Wissenschaften, 1960.

Eric Heitz, Laurent Belcour, and Thomas Chambon. Iterative α-(de)Blending: a Minimalist Deterministic Diffusion Model. In *ACM SIGGRAPH Conference Prooceedings*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 2020.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, 2023.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models. *arXiv preprint arXiv:2306.09251*, 2023.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us Build Bridges: Understanding and Extending Diffusion Generative Models. *arXiv preprint arXiv:2208.14699*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *International Conference on Learning Representations*, 2023.

Jakiw Pidstrigach. Score-Based Generative Models Detect Manifolds. In *Advances in Neural Information Processing Systems*, 2022.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In *International Conference on Machine Learning*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022.

Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning*, 2015.

Noam Rozen, Aditya Grover, Maximilian Nickel, and Yaron Lipman. Moser Flow: Divergence-based Generative Modeling on Manifolds. In *Advances in Neural Information Processing Systems*, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2021a.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021b.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Kaylee Yingxi Yang and Andre Wibisono. Convergence in KL and Rényi Divergence of the Unadjusted Langevin Algorithm Using Estimated Score. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Qinsheng Zhang and Yongxin Chen. Fast Sampling of Diffusion Models with Exponential Integrator. In *International Conference on Learning Representations*, 2023.

A Exploration of Definition 1

A.1 Special cases of λ-regularity

First, we show that all log-concave random variables are λ-regular for $\lambda = 1$. The key ingredient in the proof is the following result of Brascamp & Lieb (1976).

Proposition 3 (Brascamp-Lieb 1976)**.** *Suppose that W is an $\mathbb{R}^d$-valued random variable with density function $p_W(\mathbf{x}) = e^{-\varphi(\mathbf{x})}$, where φ is strictly convex on $\mathbb{R}^d$ and twice differentiable. Assume that $D^2\varphi \geq \mu I_d$ with $\mu > 0$ and $g \in C^1(\mathbb{R}^d)$. Then,*

$$\mathrm{Var}_W(g(W)) \leq \mathbb{E}\left[\langle (D^2\varphi)^{-1}\nabla g(W), \nabla g(W)\rangle\right] \leq \frac{1}{\mu}\mathbb{E}\left[\|\nabla g(W)\|^2\right].$$

Lemma 4. *Suppose that W is an $\mathbb{R}^d$-valued log-concave random variable. Then W is λ-regular for $\lambda = 1$.*

Proof. Fix $\tau > 0$ and denote the density of W by $p_W(\mathbf{x}) = e^{-\varphi(\mathbf{x})}$ for some convex function φ. Take $\xi \sim \mathcal{N}(0, \tau^2 I_d)$ and set $W' = W + \xi$. Then, the density of ξ conditional on W' is given by

$$p_{\xi|W'}(\xi|\mathbf{x}') \propto e^{-\|\xi\|^2/(2\tau^2)} p_W(\mathbf{x}' - \xi) = \exp\left\{-\tfrac{\|\xi\|^2}{2\tau^2} - \varphi(\mathbf{x}' - \xi)\right\}.$$

Let $\mathbf{u} \in \mathbb{R}^d$ be an arbitrary unit vector, and consider $g(W) = \mathbf{u}^T W$. Since $\varphi'(\xi) = \frac{\|\xi\|^2}{2\tau^2} + \varphi(\mathbf{x}' - \xi)$ is strictly convex in ξ, and

$$D^2\varphi' = \tau^{-2} I_d + D^2\varphi(\mathbf{x}' - \xi) \geq \tau^{-2} I_d,$$

we can apply Theorem 3 to the random variable ξ conditional on W' with $\mu = \tau^{-2}$ to get

$$\mathrm{Var}_{\xi|W'}(\mathbf{u}^T\xi) \leq \tau^2 \mathbb{E}\left[\|\mathbf{u}\|_2^2\right] = \tau^2.$$

Then,

$$\left\|\mathrm{Cov}_{\xi|W'}(\xi)\right\|_{\mathrm{op}} \leq \sup_{\|\mathbf{u}\|_2=1} \mathrm{Var}_{\xi|W'}(\mathbf{u}^T\xi) \leq \tau^2.$$

□

Second, we show all random variables which are Gaussian on a linear subspace of $\mathbb{R}^d$ are λ-regular for $\lambda = 1$.

Lemma 5. *Suppose that W is an $\mathbb{R}^d$-valued random variable supported on some linear subspace $\mathcal{S} \subseteq \mathbb{R}^d$ and that W restricted to $\mathcal{S}$ is Gaussian with positive definite covariance matrix. Then W is λ-regular for $\lambda = 1$.*

Proof. We decompose ξ into two orthogonal components $\xi_\perp$ and $\xi_\parallel$, so that $\xi = \xi_\perp + \xi_\parallel$, and $\xi_\perp$ is perpendicular to $\mathcal{S}$ while $\xi_\parallel$ is parallel to $\mathcal{S}$. Then, we can write $W' = (W + \xi_\parallel) + \xi_\perp$. We denote $W'_\parallel = W + \xi_\parallel$ and note that $W'_\parallel \in \mathcal{S}$. From observing $W' = \mathbf{x}$ we can deduce the values of both $W'_\parallel$ and $\xi_\perp$. Therefore, $\mathrm{Cov}_{\xi|W'=\mathbf{x}}(\xi) = \mathrm{Cov}_{\xi|W'=\mathbf{x}}(\xi_\parallel) = \mathrm{Cov}_{\xi_\parallel|W'_\parallel=\mathbf{x}_\parallel}(\xi_\parallel)$, where $\mathbf{x}_\parallel$ denotes the projection of $\mathbf{x}$ onto $\mathcal{S}$.

By restricting our attention to the subspace $\mathcal{S}$, we may therefore reduce to the case where W is Gaussian with full support. The result then follows from Lemma 4, since Gaussians with full support are log-concave. □

Third, we show that all random variables which are locally at least as smooth as a Gaussian of covariance $\sigma^2 I_d$ and are bounded up to Gaussian tails of covariance $\sigma^2 I_d$ are λ-regular.

Lemma 6. *Suppose that W is an $\mathbb{R}^d$-valued random variable which can be decomposed as $W = U + \eta$, where U and η are independent random variables such that $\|U\| \leq R$ for some $R > 0$ and $\eta \sim \mathcal{N}(0, \sigma^2 I_d)$ for some $\sigma > 0$. Then W is λ-regular for $\lambda = 1 + (R^2/\sigma^2)$.*

Proof. Fix $\tau > 0$, so that $W' = W + \xi = U + \eta + \xi$ where $\eta \sim \mathcal{N}(0, \sigma^2 I_d)$ and $\xi \sim \mathcal{N}(0, \tau^2 I_d)$ and U, η, ξ are all independent. By the law of total variance, we have

$$\mathrm{Cov}_{\xi|W'}(\xi) = \mathbb{E}\left[\mathrm{Cov}_{\xi|W',U}(\xi) \mid W'\right] + \mathrm{Cov}(\mathbb{E}\left[\xi \mid W', U\right] \mid W').$$

The distribution of $\xi \mid W', U$ is the same as the distribution of $\xi \mid (\eta + \xi)$, so it suffices to understand the latter. To this end, we define $\rho = \eta + \xi$ and $\omega = \sigma^2 \xi - \tau^2 \eta$. It is straightforward to check that ρ and ω are independent centered Gaussians with covariances $(\sigma^2 + \tau^2) I_d$ and $\sigma^2 \tau^2 (\sigma^2 + \tau^2) I_d$ respectively. Then, we can write $\xi = \frac{1}{\sigma^2 + \tau^2}(\tau^2 \rho + \omega)$, from which it follows that

$$\mathrm{Cov}_{\xi \mid W', U}(\xi) = \left(\tfrac{1}{\sigma^2+\tau^2}\right)^2 \mathrm{Cov}(\tau^2 \rho + \omega \mid \rho) = \left(\tfrac{\sigma^2 \tau^2}{\sigma^2+\tau^2}\right) I_d.$$

Therefore, $\mathbb{E}\left[\mathrm{Cov}_{\xi \mid W', U}(\xi) \mid W'\right] = \left(\frac{\sigma^2 \tau^2}{\sigma^2+\tau^2}\right) I_d$. In addition, we have

$$\mathbb{E}\left[\xi \mid W', U\right] = \left(\tfrac{1}{\sigma^2+\tau^2}\right) \mathbb{E}\left[\tau^2 \rho + \omega \mid \rho\right] = \left(\tfrac{\tau^2}{\sigma^2+\tau^2}\right) \rho,$$

so $\mathrm{Cov}(\mathbb{E}\left[\xi \mid W', U\right] \mid W') = \left(\frac{\tau^2}{\sigma^2+\tau^2}\right)^2 \mathrm{Cov}_{\eta, \xi \mid W'}(\eta + \xi)$. Finally, we have

$$\left\|\mathrm{Cov}_{\eta, \xi \mid W'}(\eta + \xi)\right\|_{\mathrm{op}} = \left\|\mathrm{Cov}_{U \mid W'}(W' - U)\right\|_{\mathrm{op}} = \left\|\mathrm{Cov}_{U \mid W'}(U)\right\|_{\mathrm{op}} \le R^2.$$

Putting this all together, we see that

$$\begin{aligned}
\left\|\mathrm{Cov}_{\xi \mid W'}(\xi)\right\|_{\mathrm{op}} &\le \left\|\mathbb{E}\left[\mathrm{Cov}_{\xi \mid W', U}(\xi) \mid W'\right]\right\|_{\mathrm{op}} + \left\|\mathrm{Cov}(\mathbb{E}\left[\xi \mid W', U\right] \mid W')\right\|_{\mathrm{op}} \\
&\le \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} + R^2 \left(\frac{\tau^2}{\sigma^2 + \tau^2}\right)^2 \\
&\le \lambda \tau^2
\end{aligned}$$

for $\lambda = 1 + (R^2/\sigma^2)$, as required. □

We note in particular that Lemma 6 can be applied in the case where our target distribution is a bounded mixture of Gaussian components with covariance $\sigma^2 I_d$ for some $\sigma > 0$. We obtain the following corollary.

Corollary 3. *Suppose that $\pi = \sum_{i=1}^K \mu_i \mathcal{N}(\mathbf{x}_i, \sigma^2 I_d)$ is a mixture of Gaussian components of covariance $\sigma^2 I_d$ for some $\sigma > 0$, where the weights μ_i satisfy $\sum_{i=1}^K \mu_i = 1$ and we have $\|\mathbf{x}_i\| \le R$ for all $i = 1, \ldots, K$. Then π is λ-regular for $\lambda = 1 + (R^2/\sigma^2)$.*

Proof. This follows immediately from the fact that π can be represented in the form required to apply Lemma 6, by taking U to have distribution $\sum_{i=1}^K \mu_i \delta_{\mathbf{x}_i}$. □

A.2 High-probability bounds

Lemma 7. *Suppose that W is an $\mathbb{R}^d$-valued random variable. For any $\tau > 0$, if we take $\xi \sim \mathcal{N}(0, \tau^2 I_d)$ and set $W' = W + \xi$, then we have*

$$\left\|\mathrm{Cov}_{\xi \mid W'}(\xi)\right\|_{\mathrm{op}} \le 2dc^2\tau^2$$

with probability at least $1 - 6de^{-c^2/2}$ for any $c \ge 1$.

Proof. First, we have

$$\left\|\mathrm{Cov}_{\xi \mid W'}(\xi)\right\|_{\mathrm{op}} \le \sup_{\|\mathbf{u}\|_2 = 1} \mathrm{Var}_{\xi \mid W'}(\mathbf{u}^T \xi) \le \sup_{\|\mathbf{u}\|_2 = 1} \mathbb{E}_{\xi \mid W'}\left[(\mathbf{u}^T \xi)^2\right] \le \mathbb{E}\left[\|\xi\|_2^2 \mid W'\right].$$

Next, we bound the quantity $\mathbb{E}\left[\|\xi\|_2^2 \mid W'\right]$ using Markov's inequality. If $\|\xi\|_2^2 > dc^2\tau^2$, then we must have $\xi_i^2 > c^2\tau^2$ for some i between 1 and d. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|\xi\|_2^2 \mathbb{1}_{\|\xi\|_2^2 > dc^2\tau^2}\right] &\le d\, \mathbb{E}\left[\|\xi\|_2^2 \mathbb{1}_{\xi_1^2 > c^2\tau^2}\right] \\
&= d\, \mathbb{E}\left[\xi_1^2 \mathbb{1}_{\xi_1^2 > c^2\tau^2}\right] + d \sum_{i=2}^d \mathbb{E}\left[\xi_i^2 \mathbb{1}_{\xi_1^2 > c^2\tau^2}\right] \\
&= d\, \mathbb{E}\left[\xi_1^2 \mathbb{1}_{\xi_1^2 > c^2\tau^2}\right] + d(d-1)\tau^2 \mathbb{P}\left(\xi_1^2 > c^2\tau^2\right).
\end{aligned}$$

A standard Chernoff bound gives $\mathbb{P}\left(\xi_1^2 > c^2\tau^2\right) \leq 2e^{-c^2/2}$, and

$$\begin{aligned}
\mathbb{E}\left[\xi_1^2 \mathbb{1}_{\xi_1^2 > c^2\tau^2}\right] &= 2\int_{c\tau}^{\infty} \frac{z^2}{\sqrt{2\pi\tau^2}} e^{-z^2/(2\tau^2)}\, \mathrm{d}z \\
&= 2\tau^2 \int_{c}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-z^2/2}\, \mathrm{d}z \\
&= 2\tau^2 \left[-\frac{z}{\sqrt{2\pi}} e^{-z^2/2}\right]_c^{\infty} + 2\tau^2 \int_c^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, \mathrm{d}z \\
&= \frac{2\tau^2 c}{\sqrt{2\pi}} e^{-c^2/2} + 2\tau^2 \mathbb{P}\left(\xi_1 > c\tau\right) \\
&\leq 2\tau^2 e^{-c^2/2}\left(\frac{c}{\sqrt{2\pi}} + 1\right) \\
&\leq 4c\tau^2 e^{-c^2/2}.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}\left[\|\xi\|_2^2 \mathbb{1}_{\|\xi\|_2^2 > dc^2\tau^2}\right] &\leq 4dc\tau^2 e^{-c^2/2} + 2d^2\tau^2 e^{-c^2/2} \\
&\leq 6d^2 c\tau^2 e^{-c^2/2}.
\end{aligned}$$

It follows by Markov's inequality that

$$\begin{aligned}
\mathbb{P}\left(\mathbb{E}\left[\|\xi\|_2^2 \mathbb{1}_{\|\xi\|_2^2 > dc^2\tau^2} \mid W'\right] \geq dc^2\tau^2\right) &\leq \frac{\mathbb{E}\left[\|\xi\|_2^2 \mathbb{1}_{\|\xi\|_2^2 > dc^2\tau^2}\right]}{dc^2\tau^2} \\
&\leq 6de^{-c^2/2}.
\end{aligned}$$

Finally, we can write

$$\mathbb{E}\left[\|\xi\|_2^2 \mid W'\right] \leq \mathbb{E}\left[\|\xi\|_2^2 \mathbb{1}_{\|\xi\|_2^2 > dc^2\tau^2} \mid W'\right] + dc^2\tau^2,$$

and so $\left\|\mathrm{Cov}_{\xi|W'}(\xi)\right\|_{\mathrm{op}} \leq \mathbb{E}\left[\|\xi\|_2^2 \mid W'\right] \leq 2dc^2\tau^2$ with probability at least $1 - 6de^{-c^2/2}$, as required. □

B Derivation of formulae for gradients of velocity fields

B.1 Proof of Lemma 1

In order to prove Lemma 1, we use the following intermediate result.

Lemma 8. *If X is the stochastic interpolant between π_0 and π_1, then*

$$\nabla_{\mathbf{x}} \mathbb{E}\big[X_0 \mid X_t = \mathbf{x}\big] = -\frac{1}{\gamma_t}\mathrm{Cov}_{\mathbf{x}}(X_0, Z).$$

Moreover, a similar expression holds for X_1 in place of X_0.

Proof. First, note that $X_t|X_0, X_1 \sim \mathcal{N}(\alpha_t X_0 + \beta_t X_1, \gamma_t^2 I_d)$, so

$$\nabla_{\mathbf{x}} \log(p_{X_t|X_0,X_1}(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1)) = -\frac{1}{\gamma_t^2}(\mathbf{x} - \alpha_t\mathbf{x}_0 - \beta_t\mathbf{x}_1)^T.$$

Therefore,

$$\begin{aligned}
\nabla_{\mathbf{x}} \log(p_{X_t}(\mathbf{x})) &= \frac{1}{p_{X_t}(\mathbf{x})} \cdot \nabla_{\mathbf{x}} \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} p_{X_t|X_0,X_1}(\mathbf{x}|\mathbf{x}_0,\mathbf{x}_1) p_{X_0,X_1}(\mathbf{x}_0,\mathbf{x}_1)\, \mathrm{d}\mathbf{x}_0\mathrm{d}\mathbf{x}_1\right) \\
&= \frac{1}{p_{X_t}(\mathbf{x})}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} p_{X_t|X_0,X_1}(\mathbf{x}|\mathbf{x}_0,\mathbf{x}_1) p_{X_0,X_1}(\mathbf{x}_0,\mathbf{x}_1) \nabla_{\mathbf{x}} \log(p_{X_t|X_0,X_1}(\mathbf{x}|\mathbf{x}_0,\mathbf{x}_1))\, \mathrm{d}\mathbf{x}_0\mathrm{d}\mathbf{x}_1 \\
&= -\frac{1}{\gamma_t^2}\mathbb{E}\big[(X_t - \alpha_t X_0 - \beta_t X_1)^T \mid X_t = \mathbf{x}\big]
\end{aligned}$$

We can then calculate

$$\begin{aligned}
\nabla_{\mathbf{x}} \mathbb{E}\big[X_0 \mid X_t = \mathbf{x}\big] &= \nabla_{\mathbf{x}} \left(\frac{1}{p_{X_t}(\mathbf{x})} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbf{x}_0 \, p_{X_t|X_0,X_1}(\mathbf{x}|\mathbf{x}_0,\mathbf{x}_1) p_{X_0,X_1}(\mathbf{x}_0,\mathbf{x}_1) \, \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_1 \right) \\
&= \frac{1}{p_{X_t}(\mathbf{x})} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbf{x}_0 \, p_{X_t|X_0,X_1}(\mathbf{x}|\mathbf{x}_0,\mathbf{x}_1) p_{X_0,X_1}(\mathbf{x}_0,\mathbf{x}_1) \nabla_{\mathbf{x}} \left(\log p_{X_t|X_0,X_1}(\mathbf{x}|\mathbf{x}_0,\mathbf{x}_1) \right) \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_1 \\
&\quad - \big(\nabla_{\mathbf{x}} \log p_{X_t}(\mathbf{x})\big) \left(\frac{1}{p_{X_t}(\mathbf{x})} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbf{x}_0 \, p_{X_t|X_0,X_1}(\mathbf{x}|\mathbf{x}_0,\mathbf{x}_1) p_{X_0,X_1}(\mathbf{x}_0,\mathbf{x}_1) \, \mathrm{d}\mathbf{x}_0 \mathrm{d}\mathbf{x}_1 \right) \\
&= -\frac{1}{\gamma_t^2} \mathbb{E}_{\mathbf{x}}\big[X_0 (X_t - \alpha_t X_0 - \beta_t X_1)^T\big] + \frac{1}{\gamma_t^2} \mathbb{E}_{\mathbf{x}}\big[(X_t - \alpha_t X_0 - \beta_t X_1)^T\big] \mathbb{E}_{\mathbf{x}}\big[X_0\big] \\
&= -\frac{1}{\gamma_t} \mathrm{Cov}_{\mathbf{x}}(X_0, Z).
\end{aligned}$$

□

Lemma 1. *If X is the stochastic interpolant between π_0 and π_1, then $v^X(\mathbf{x}, t)$ is differentiable with respect to $\mathbf{x}$ and*

$$\nabla_{\mathbf{x}} v^X(\mathbf{x}, t) = \frac{\dot{\gamma}_t}{\gamma_t} I_d - \frac{1}{\gamma_t} \mathrm{Cov}_{\mathbf{x}}(\dot{X}_t, Z).$$

Proof. Since $X_t = \alpha_t X_0 + \beta_t X_1 + \gamma_t Z$ and $\dot{X}_t = \dot{\alpha}_t X_0 + \dot{\beta}_t X_1 + \dot{\gamma}_t Z$, we can write

$$\mathbb{E}\left[\dot{X}_t \mid X_t = \mathbf{x}, X_0 = \mathbf{x}_0, X_1 = \mathbf{x}_1\right] = \dot{\alpha}_t \mathbf{x}_0 + \dot{\beta}_t \mathbf{x}_1 + \frac{\dot{\gamma}_t}{\gamma_t} \left(\mathbf{x} - \alpha_t \mathbf{x}_0 - \beta_t \mathbf{x}_1\right)$$

and therefore

$$\mathbb{E}\left[\dot{X}_t \mid X_t = \mathbf{x}\right] = \frac{\dot{\gamma}_t}{\gamma_t} \mathbf{x} + \frac{(\dot{\alpha}_t \gamma_t - \dot{\gamma}_t \alpha_t)}{\gamma_t} \cdot \mathbb{E}\big[X_0 \mid X_t = \mathbf{x}\big] + \frac{(\dot{\beta}_t \gamma_t - \dot{\gamma}_t \beta_t)}{\gamma_t} \cdot \mathbb{E}\big[X_1 \mid X_t = \mathbf{x}\big].$$

Taking gradients with respect to $\mathbf{x}$ and applying Lemma 8,

$$\begin{aligned}
\nabla_{\mathbf{x}} v^X(\mathbf{x}, t) &= \frac{\dot{\gamma}_t}{\gamma_t} I_d - \frac{(\dot{\alpha}_t \gamma_t - \dot{\gamma}_t \alpha_t)}{\gamma_t^2} \cdot \mathrm{Cov}_{\mathbf{x}}(X_0, Z) - \frac{(\dot{\beta}_t \gamma_t - \dot{\gamma}_t \beta_t)}{\gamma_t^2} \cdot \mathrm{Cov}_{\mathbf{x}}(X_1, Z) \\
&= \frac{\dot{\gamma}_t}{\gamma_t} I_d + \frac{\dot{\gamma}_t}{\gamma_t^2} \mathrm{Cov}_{\mathbf{x}}(X_t - \gamma_t Z, Z) - \frac{1}{\gamma_t} \mathrm{Cov}_{\mathbf{x}}(\dot{\alpha}_t X_0 + \dot{\beta}_t X_1, Z) \\
&= \frac{\dot{\gamma}_t}{\gamma_t} I_d - \frac{1}{\gamma_t} \mathrm{Cov}_{\mathbf{x}}(\dot{X}_t, Z).
\end{aligned}$$

□

B.2 Proof of Lemma 3

Lemma 3. *If X is the stochastic interpolant in the PF-ODE setting above, then $v^X(\mathbf{x}, t)$ is differentiable with respect to $\mathbf{x}$ and*

$$\nabla_{\mathbf{x}} v^X(\mathbf{x}, t) = \frac{\dot{\gamma}_t}{\gamma_t} I_d - \left(\frac{\dot{\gamma}_t}{\gamma_t} - \frac{\dot{\beta}_t}{\beta_t} \right) \mathrm{Cov}_{\mathbf{x}}(Z).$$

Proof. From Lemma 1, we have

$$\nabla_{\mathbf{x}} v^X(\mathbf{x}, t) = \frac{\dot{\gamma}_t}{\gamma_t} I_d - \frac{1}{\gamma_t} \mathrm{Cov}_{\mathbf{x}}(\dot{X}_t, Z).$$

Then, since $\alpha_t = 0$ we have $X_t = \beta_t X_1 + \gamma_t Z$ and $\dot{X}_t = \dot{\beta}_t X_1 + \dot{\gamma}_t Z$, so we can write

$$\begin{aligned}
\nabla_{\mathbf{x}} v^X(\mathbf{x}, t) &= \frac{\dot{\gamma}_t}{\gamma_t} I_d - \frac{1}{\gamma_t} \mathrm{Cov}_{\mathbf{x}} \left(\frac{\dot{\beta}_t}{\beta_t} \left(X_t - \gamma_t Z\right) + \dot{\gamma}_t Z, Z \right) \\
&= \frac{\dot{\gamma}_t}{\gamma_t} I_d - \left(\frac{\dot{\gamma}_t}{\gamma_t} - \frac{\dot{\beta}_t}{\beta_t} \right) \mathrm{Cov}_{\mathbf{x}}(Z),
\end{aligned}$$

noting that we may discard the X_t term since we are conditioning on $X_t = \mathbf{x}$. □

Statement of Authorship

Title of paper:	Error Bounds for Flow Matching Methods
Publication status:	Published
Publication details:	Joe Benton, George Deligiannidis, Arnaud Doucet. *Transactions on Machine Learning Research*, February 2024.

Student Confirmation

Student name:	Joe Benton
Contribution to the paper:	The idea for the work was developed in discussion with Arnaud Doucet and George Deligiannidis. I provided all of the theoretical results and wrote the paper, with some direction from Arnaud Doucet and George Deligiannidis.

Signature:  Date: 29/10/2023

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: George Deligiannidis, Associate Professor of Statistics

Supervisor comments: None

Signature:  Date: 31/10/2023

5

Nearly d-Linear Convergence Bounds for Diffusion Models via Stochastic Localization

NEARLY d-LINEAR CONVERGENCE BOUNDS FOR DIFFUSION MODELS VIA STOCHASTIC LOCALIZATION

Joe Benton[*], **Valentin De Bortoli**[†], **Arnaud Doucet**[*], **George Deligiannidis**[*]

ABSTRACT

Denoising diffusions are a powerful method to generate approximate samples from high-dimensional data distributions. Recent results provide polynomial bounds on their convergence rate, assuming L^2-accurate scores. Until now, the tightest bounds were either superlinear in the data dimension or required strong smoothness assumptions. We provide the first convergence bounds which are linear in the data dimension (up to logarithmic factors) assuming only finite second moments of the data distribution. We show that diffusion models require at most $\tilde{O}(\frac{d \log^2(1/\delta)}{\varepsilon^2})$ steps to approximate an arbitrary distribution on $\mathbb{R}^d$ corrupted with Gaussian noise of variance δ to within ε^2 in KL divergence. Our proof extends the Girsanov-based methods of previous works. We introduce a refined treatment of the error from discretizing the reverse SDE inspired by stochastic localization.

1 INTRODUCTION

Denoising diffusion models are a recent advance in generative modeling which have produced state-of-the-art results in many domains (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021b), including image and text generation (Dhariwal & Nichol, 2021; Austin et al., 2021; Ramesh et al., 2022; Saharia et al., 2022), text-to-speech synthesis (Popov et al., 2021), and molecular structure modeling (Xu et al., 2022; Trippe et al., 2023; Watson et al., 2023). Denoising diffusion models take data samples, corrupt them through the iterated application of noise, and learn to reverse this noising procedure. For data on $\mathbb{R}^d$, we typically use a stochastic differential equation (SDE) as the noising process. Then, learning the reverse process is equivalent to learning the score of the noised distributions (Vincent, 2011; Song & Ermon, 2019; Song et al., 2021b).

Recently, significant progress has been made on improving our theoretical understanding of diffusion models, including several works which have established polynomial convergence bounds for such models (Chen et al., 2023a;d; Lee et al., 2023; Li et al., 2023). The current state-of-the-art bound without Lipschitz assumptions on the score of the data distribution is provided by Chen et al. (2023a), who show that diffusion models require at most $\tilde{O}\big(\frac{d^2 \log^2(1/\delta)}{\varepsilon^2}\big)$ steps to approximate a given distribution to within ε^2 in KL divergence, assuming only a finite second moment of the target distribution and early stopping at time δ. However, Chen et al. (2023a;d) suggest that the iteration complexity ought to scale linearly in the data dimension (see e.g. Chen et al. (2023d, Theorem 6)).

In this work, we close this gap. We derive the first convergence bounds for diffusion models which are linear in the data dimension (up to logarithmic factors) without smoothness assumptions. We deduce that an early stopped diffusion model has iteration complexity $\tilde{O}(\frac{d \log^2(1/\delta)}{\varepsilon^2})$. Our proof builds on Chen et al. (2023a;d), who use Girsanov's theorem to measure the distance between the true and approximate reverse paths. However, we provide a refined treatment of the time discretization error, using techniques from stochastic calculus to derive a differential inequality for the expected difference between the drift terms at different times. By bounding the terms of this differential inequality, we achieve tighter bounds on the difference between the true and approximate path measures.

A key ingredient in our proof will be Lemma 1, which allows us to perform several calculations explicitly. This result is inspired by stochastic localization (Eldan, 2013; 2020), developed to study the

[*]Department of Statistics, University of Oxford, {benton,doucet,deligian}@stats.ox.ac.uk
[†]CNRS, ENS Ulm, Paris, valentin.debortoli@gmail.com

KLS conjecture and applied to sampling. Stochastic localization sampling is equivalent to diffusion models (Montanari, 2023), letting us transfer insights from stochastic localization to our proof.

1.1 DIFFUSION MODELS

A diffusion model starts with a stochastic process $(X_t)_{t\in[0,T]}$ constructed by initializing X_0 in the data distribution p_{data} and then evolving according to the Ornstein–Uhlenbeck (OU) SDE

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \text{for } 0 \le t \le T, \tag{1}$$

where $(B_t)_{t\in[0,T]}$ is a Brownian motion on $\mathbb{R}^d$ (Song et al., 2021b). We then learn the dynamics of the reverse process $(Y_t)_{t\in[0,T]}$ defined by $Y_t = X_{T-t}$. If we let $q_t(\mathbf{x}_t)$ denote the marginals of the forward process, then under mild regularity conditions on p_{data}, the reverse process satisfies the SDE

$$\mathrm{d}Y_t = \{Y_t + 2\nabla \log q_{T-t}(Y_t)\}\mathrm{d}t + \sqrt{2}\mathrm{d}B'_t, \qquad Y_0 \sim q_T, \tag{2}$$

where $(B'_t)_{t\in[0,T]}$ is another Brownian motion (Anderson, 1982; Cattiaux et al., 2022). We can thus generate samples $\xi \sim p_{\text{data}}$ by sampling $Y_0 \sim q_T$, running the reverse SDE, and setting $\xi = Y_T$.

The OU process is a convenient choice of forward process as its transition densities are analytically tractable, with $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0 e^{-t}, \sigma_t^2 I_d)$ where $\sigma_t^2 := 1 - e^{-2t}$. In what follows, we will denote the posterior mean $\mathbf{m}_t(\mathbf{x}_t) := \mathbb{E}_{q_{0|t}(\cdot|\mathbf{x}_t)}[X_0]$ and variance $\mathbf{\Sigma}_t(\mathbf{x}_t) := \mathrm{Cov}_{q_{0|t}(\cdot|\mathbf{x}_t)}(X_0)$; $\mathbf{m}_t$ is related to the score function via $\nabla \log q_t(\mathbf{x}_t) = -\sigma_t^{-2}\mathbf{x}_t + e^{-t}\sigma_t^{-2}\mathbf{m}_t$ (see Lemma 5 below).

For convenience, we will assume throughout that our data distribution p_{data} has identity covariance matrix (as is standard in applications), though our analysis holds similarly without this assumption. We also focus on an OU noising process rather than a general noising SDE (as in e.g. Chen et al. (2023e)), since the OU process is most common in practice. However, our results can be straightforwardly extended to any linear SDE (including in particular the VE SDE (Song et al., 2021b)).

To simulate (2), we must make three approximations. First, as we do not have access to $\nabla \log q_t(\mathbf{x}_t)$ directly, we learn an approximation $s_\theta(\mathbf{x}_t, t)$ to $\nabla \log q_t(\mathbf{x}_t)$ for $t \in [0, T]$ by minimising

$$\mathcal{L}(s_\theta) = \int_0^T \mathbb{E}_{q_t}\left[\|s_\theta(X_t, t) - \nabla \log q_t(X_t)\|^2\right] \mathrm{d}t. \tag{3}$$

Although $\mathcal{L}(s_\theta)$ cannot be estimated directly, there are techniques such as denoising or implicit score matching which provide equivalent tractable objectives (Hyvärinen, 2005; Vincent, 2011). In practice, we empirically estimate these objectives using samples drawn from the forward process, which can be analytically simulated. We typically parameterize $s_\theta(\mathbf{x}_t, t)$ via a neural network for a vector of parameters $\theta \in \mathbb{R}^D$ and minimise the objective function using stochastic gradient descent.

Second, since we do not have access to the initial distribution q_T of the reverse process, we instead initialize the reverse process in the standard Gaussian, which we denote π_d. This is a reasonable approximation since the OU process converges exponentially quickly to π_d (Bakry et al., 2014).

Third, since (2) is a continuous-time process we must discretize time in order to simulate it. We pick time steps $0 = t_0 < t_1 < \cdots < t_N \le T$, sample $\hat{Y}_0 \sim \pi_d$ and define $(\hat{Y}_t)_{t\in[0,T]}$ via the SDE

$$\mathrm{d}\hat{Y}_t = \{\hat{Y}_t + 2s_\theta(\hat{Y}_{t_k}, T - t_k)\}\mathrm{d}t + \mathrm{d}\hat{B}_t \tag{4}$$

for each interval $[t_k, t_{k+1}]$ and $k = 0, \ldots, N-1$, for a Brownian motion $(\hat{B}_t)_{t\in[0,T]}$. We use the notation $\gamma_k := t_{k+1} - t_k$ for the step size and denote the marginals of the approximate reverse process by $p_t(\mathbf{x})$. The sampling scheme defined via (4) is known as the exponential integrator (Zhang & Chen, 2023; De Bortoli, 2022; Chen et al., 2023a). An alternative is the Euler–Maruyama (EM) scheme, which replaces the time-dependent drift term in (4) with its value at the start of the associated interval (Song et al., 2021b; Chen et al., 2023a). Both methods give similar asymptotic convergence rates, as indicated by Chen et al. (2023a), so we focus on the exponential integrator.

Finally, instead of running (4) all the way back to the start time, we perform early stopping and set $t_N = T - \delta$ for some small δ. We do this because for non-smooth data distributions $\nabla \log q_t$ can blow up as $t \to 0$. This means that our model will approximate q_δ rather than $q_0 = p_{\text{data}}$, which is acceptable since for small δ the distance (e.g. in Wasserstein-p metric) between q_δ and p_{data} is small. Early stopping is frequently used in practical applications of diffusion models (Song et al., 2021b).

1.2 STOCHASTIC LOCALIZATION

Stochastic localization sampling schemes were developed by El Alaoui et al. (2022); Montanari (2023). To sample from a distribution p_{data}, we construct a measure-valued stochastic process $(\mu_s)_{s\geq 0}$ such that μ_s "localizes" as $s \to \infty$, meaning that there a.s. exists some ξ such that $\mu_s \to \delta_\xi$ as $s \to \infty$, and $\xi \sim p_{\text{data}}$. We construct this process by sampling $\xi \sim p_{\text{data}}$ and defining a sequence $(U_s)_{s\geq 0}$ of noisy observations of ξ via

$$U_s = s\xi + W_s, \tag{5}$$

where $(W_s)_{s\geq 0}$ is a Brownian motion on $\mathbb{R}^d$ (Montanari, 2023). We then define $\mu_s = \text{Law}(\xi \mid U_s)$, so that μ_s is a random measure depending on U_s. Since $U_s/s \to \xi$ almost surely as $s \to \infty$, we see $(\mu_s)_{s\geq 0}$ does indeed localize and $\lim_{s\to\infty} U_s/s$ is distributed according to p_{data}.

However, constructing $(U_s)_{s\geq 0}$ via (5) requires sampling $\xi \sim p_{\text{data}}$, which we cannot do. Fortunately, we avoid this using the observation that if p_{data} has finite second moments, then $(U_s)_{s\geq 0}$ is equivalent in law to the unique solution to the SDE

$$\mathrm{d}U_s = \mathbf{a}_s(U_s)\mathrm{d}s + \mathrm{d}W'_s, \tag{6}$$

where $(W'_s)_{s\geq 0}$ is a Brownian motion and $\mathbf{a}_s(U_s) = \mathbb{E}_{\mu_s}[\xi] = \mathbb{E}[\xi \mid U_s]$ (Liptser & Shiryaev, 1977; Montanari, 2023). This allows us to construct the stochastic localization process without direct access to p_{data}, so long as we have access to the function $\mathbf{a}_s(U_s)$. We also define $\mathbf{A}_s(U_s) = \text{Cov}(\mu_s)$.

Diffusion models and stochastic localization are equivalent under a time change (Montanari, 2023). If we define $(X_t)_{t\geq 0}$, $(U_s)_{s\geq 0}$ according to (1), (5) respectively and let $t(s) := \frac{1}{2}\log(1+s^{-1})$, then $(U_s)_{s\geq 0}$ and $(se^{t(s)}X_{t(s)})_{s\geq 0}$ have the same law. In addition, $\mathbf{a}_s(U_s)$ and $\mathbf{m}_t(X_t)$ have the same law and $\mathbf{A}_s(U_s)$ and $\mathbf{\Sigma}_t(X_t)$ have the same law when $t = t(s)$. We will sometimes suppress the dependence of $\mathbf{a}_s, \mathbf{A}_s$ and $\mathbf{m}_t, \mathbf{\Sigma}_t$ on U_s and X_t respectively when the meaning is clear. For more details and an explicit derivation of the equivalence, see Appendix A.

We now recall two lemmas from the stochastic localization literature. The first can be found in Eldan (2013); Alaoui & Montanari (2022). The second follows from the first plus the argument in Eldan (2020). We provide proofs for both results in Appendix B for the reader's convenience.

Proposition 1 (Alaoui & Montanari (2022), Theorem 2)**.** *If we define* $L_s(\mathbf{x}) = \frac{\mathrm{d}\mu_s}{\mathrm{d}p_{\text{data}}}(\mathbf{x})$*, then* $\mathrm{d}L_s(\mathbf{x}) = L_s(\mathbf{x})(\mathbf{x} - \mathbf{a}_s)\cdot \mathrm{d}W'_s$ *for all* $s \geq 0$*.*

Proposition 2 (Eldan (2020), Equation 11)**.** *For all* $s \geq 0$*,* $\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}[\mathbf{A}_s] = -\mathbb{E}\left[\mathbf{A}_s^2\right]$*.*

Translating Proposition 2 into the language of diffusion models, using that $\mathbf{A}_s$ and $\mathbf{\Sigma}_t$ are equal in law when $t = \frac{1}{2}\log(1+s^{-1})$, we get the following directly from Proposition 2 and the chain rule.

Lemma 1. *For all* $t > 0$*,* $\frac{\sigma_t^3}{2\dot{\sigma}_t}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[\mathbf{\Sigma}_t] = \mathbb{E}\left[\mathbf{\Sigma}_t^2\right]$*.*

Lemma 1 is the key insight that allows us to control the discretization error more precisely than previous works. In Lemma 5, we will see that $\mathbf{\Sigma}_t$ is related to the Jacobian of the score. Control of $\mathbf{\Sigma}_t$ via Lemma 1 will thus allow us to control time-discretization terms $E_{s,t}$ (defined in Section 3).

1.3 RELATED WORK

Convergence of diffusion models Initial results on the convergence of diffusion models required restrictive assumptions on the data distribution such as a log-Sobolev inequality (Lee et al., 2022; Yang & Wibisono, 2022), or produced bounds that were non-quantitative (Pidstrigach, 2022; Liu et al., 2022) or exponential in the problem parameters (De Bortoli et al., 2021; De Bortoli, 2022; Block et al., 2022). Recently however, several works have proven polynomial convergence rates for diffusion models (Chen et al., 2023a;d; Lee et al., 2023; Li et al., 2023).

First, Chen et al. (2023d) gave polynomial TV error bounds, assuming a Lipschitz score for all $t \geq 0$. They used Girsanov's theorem to measure the KL divergence between the true and approximate reverse processes (similar to Song et al. (2021a)), with an approximation argument since the standard assumptions for Girsanov's theorem do not hold in their setting. Second, Chen et al. (2023a) developed this method, introducing a tighter bound on the drift terms and an exponentially

Regularity condition	Metric	Complexity	Result
$\forall t$, $\nabla \log q_t$ L-Lipschitz	$\mathrm{TV}(q_0, p_T)^2$	$\tilde{O}\big(\frac{dL^2}{\varepsilon^2}\big)$	(Chen et al., 2023d, Thm 2)
$\forall t$, $\nabla \log q_t$ L-Lipschitz	$\mathrm{KL}(q_0 \| p_T)$	$\tilde{O}\big(\frac{dL^2}{\varepsilon^2}\big)$	(Chen et al., 2023a, Thm 1)
None	$\mathrm{KL}(q_\delta \| p_{t_N})$	$\tilde{O}\big(\frac{d^2 \log^2(1/\delta)}{\varepsilon^2}\big)$	(Chen et al., 2023a, Thm 2)
None	$\mathrm{KL}(q_\delta \| p_{t_N})$	$\tilde{O}\big(\frac{d \log^2(1/\delta)}{\varepsilon^2}\big)$	This work: Corollary 1

Table 1: Summary of previous bounds and our results. Bounds expressed in terms of the number of steps required to guarantee an error of at most ε^2 in the stated metric, assuming perfect score estimation. We assume p_{data} has finite second moments and is normalized so that $\mathrm{Cov}(p_{\text{data}}) = I_d$.

decaying sequence of time steps. They provide two bounds on the KL error. The first (Theorem 1) is linear in the data dimension d but requires Lipschitz scores for $t \geq 0$ and depends quadratically on the Lipschitz constant. This is unideal since the Lipschitz assumption excludes many distributions of interest, such as those supported on a submanifold. In addition, the Lipschitz constant can hide additional dimension dependence in some cases (such as when the data are approximately supported on a submanifold). The second (Theorem 2) uses early stopping and applies to any data distribution with finite second moments but is quadratic in d. The latter gives the current best bound on the iteration complexity of diffusion models without smoothness assumptions of $\tilde{O}\big(\frac{d^2 \log^2(1/\delta)}{\varepsilon^2}\big)$[1].

In parallel, Lee et al. (2023) derived weaker polynomial convergence bounds using a χ^2-based analysis and a method to convert L^∞-accurate score estimates into L^2-accurate estimates. Most recently, Li et al. (2023) demonstrated bounds for several deterministic and non-deterministic sampling methods using elementary techniques, directly in discrete time and assuming a perfectly accurate score estimate. We summarise the results of Chen et al. (2023a;d) and compare them to ours in Table 1.

In addition, several works have studied the convergence properties of deterministic or approximately deterministic sampling schemes based on diffusion models (Chen et al., 2023c;e; Albergo & Vanden-Eijnden, 2023; Albergo et al., 2023; Benton et al., 2023; Li et al., 2023). Other work has focused on the problem of score estimation. In particular, Oko et al. (2023) bound the error when approximating the score with a neural network and show that diffusion models are approximately minimax optimal for a certain class of target distributions, while Chen et al. (2023b) study the sample complexity and convergence properties of diffusion models when the data lies on a linear submanifold.

Stochastic localization Stochastic localization was developed by Ronen Eldan to study isoperimetric inequalities (Eldan, 2013) such as the Kannan–Lovász–Simonovits (KLS) conjecture (Kannan et al., 1995) and the thin shell conjecture (Anttila et al., 2003; Bobkov & Koldobsky, 2003). It was based on the original localization methodologies of Lovász, Simonovits and Kannan (Lovász & Simonovits, 1993; Kannan et al., 1995) used to study high-dimensional and isoperimetric inequalities. Subsequently, stochastic localization has been used to make significant progress towards the KLS conjecture (Lee & Vempala, 2017; Chen, 2021), as well as to prove measure decomposition results (Eldan, 2020; Alaoui & Montanari, 2022) and for studying mixing times of Markov Chains (Chen & Eldan, 2022). Recently, several works developed algorithmic sampling techniques based on stochastic localization (El Alaoui et al., 2022; Montanari & Wu, 2023), and Montanari (2023) has shown that these sampling approaches are equivalent in the Gaussian setting to diffusion models.

Concurrent work: After the first version of our work was made publicly available, an independent work appeared by Conforti et al. (2023) who derive similar bounds. In contrast to our work, they take a stochastic control perspective and work under a finite Fisher information smoothness condition. This condition can be removed with early stopping, giving bounds that are linear in data dimension up to logarithmic factors, similar to our own.

2 Main Results

The core assumption required for our main result is the following control of the score approximation.

[1] Note that their bound is stated incorrectly in their abstract; the correct bound can be found in their Table 1.

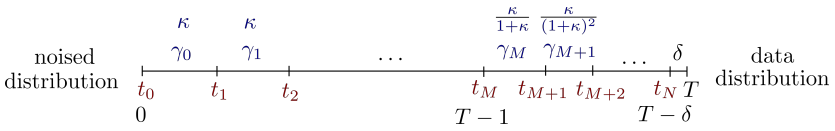

Figure 1: Illustration of a typical choice of step sizes satisfying $\gamma_k \le \kappa \min\{1, T - t_{k+1}\}$.

Assumption 1. *The score approximation function $s_\theta(\mathbf{x}, t)$ satisfies*

$$\sum_{k=0}^{N-1} \gamma_k \mathbb{E}_{q_{t_k}} \left[\|\nabla \log q_{T-t_k}(X_{t_k}) - s_\theta(X_{t_k}, T - t_k)\|^2 \right] \le \varepsilon_{\text{score}}^2. \tag{7}$$

We can view (7) as a time-discretized version of $\mathcal{L}(s_\theta)$, so Assumption 1 suggests we learn a score approximation with L^2 error at most $\varepsilon_{\text{score}}$, adjusting for the discretization of the reverse process. Though this assumption is standard in the literature (Chen et al., 2023a;b), there may be statistical barriers to efficient estimation of the score in high dimensions (Biroli & Mézard, 2023; Ghio et al., 2023) and Assumption 1 may not capture the full practical difficulty of learning the score.

Assumption 2. *The data distribution p_{data} has finite second moments, and $\text{Cov}(p_{\text{data}}) = I_d$.*

The first part of Assumption 2 is required for the convergence of the forward SDE. We include the second part for convenience when stating our results. Our analysis is not dependent on it, and we outline how to adapt our proofs for a general covariance matrix in Appendix C.

Theorem 1. *Suppose that Assumptions 1 and 2 hold, that $T \ge 1$, and that there is some $\kappa > 0$ such that for each $k = 0, \dots, N-1$ we have $\gamma_k \le \kappa \min\{1, T - t_{k+1}\}$. Then,*

$$\text{KL}(q_\delta || p_{t_N}) \lesssim \varepsilon_{\text{score}}^2 + \kappa^2 dN + \kappa dT + de^{-2T},$$

where $f_1 \lesssim f_2$ denotes that there is a universal constant C such that $f_1 \le C f_2$.

This is our main bound, and it consists of three parts. The first term $\varepsilon_{\text{score}}^2$ measures the error from using a learned rather than exact score. The second terms $\kappa^2 dN + \kappa dT$ are due to the discretization of the reverse SDE. The final term de^{-2T} controls the convergence of the forward SDE.

We interpret κ as controlling the maximum step size; γ_k is bounded by κ for $t \in [0, T-1]$, and for $t \in [T-1, T]$ the condition $\gamma_k \le \kappa(T - t_{k+1})$ forces γ_k to decay exponentially at rate $(1+\kappa)^{-1}$. We visualise this in Figure 1. As in Corollary 1 below, for any N there is a choice of time steps such that $\kappa = \tilde{O}(1/N)$, up to factors which are linear in T and logarithmic in $1/\delta$. Since T will scale logarithmically in d and $\varepsilon_{\text{score}}$, we think of the second terms in Theorem 1 as scaling like $\tilde{O}(d/N)$.

We expect the first term $\varepsilon_{\text{score}}^2$ to scale linearly in d in many cases, e.g. if the target distribution were the product of i.i.d. components. The convergence of the forward process is also linear in d. As such, Theorem 1 improves upon the previous state-of-the-art bounds (Chen et al., 2023a;d), which either require strong smoothness assumptions on p_{data} or have at least quadratic dependence on d.

We next show that given N we can choose a suitable sequence of time steps. This results in a bound on the iteration complexity of the diffusion model.

Corollary 1. *For $T \ge 1$, $\delta < 1$ and $N \ge \log(1/\delta)$, there exist $0 = t_0 < t_1 < \cdots < t_N = T - \delta$ such that for some $\kappa = \Theta\big(\frac{T + \log(1/\delta)}{N}\big)$ we have $\gamma_k \le \kappa \min\{1, T - t_{k+1}\}$ for each $k = 0, \dots, N-1$. Then, if we take $T = \frac{1}{2}\log\big(\frac{d}{\varepsilon_{\text{score}}^2}\big)$ and $N = \Theta\big(\frac{d(T + \log(1/\delta))^2}{\varepsilon_{\text{score}}^2}\big)$, we have $\text{KL}(q_\delta || p_{t_N}) = O(\varepsilon_{\text{score}}^2)$. Hence, the diffusion model requires at most $\tilde{O}\big(\frac{d \log^2(1/\delta)}{\varepsilon^2}\big)$ steps to approximate q_δ to within ε^2 in KL divergence, assuming a sufficiently accurate score estimator.*

Corollary 1 shows that by making a suitable choice of T, N and $t_0, \dots, t_N$ we achieve a bound on the iteration complexity which is linear in the data dimension (up to logarithmic factors) under minimal smoothness assumptions, resolving the question raised in Chen et al. (2023a). The proof of Corollary 1 is deferred to Appendix D, where we show that taking half the time steps to be linearly spaced between 0 and $T-1$ and half to be exponentially spaced between $T-1$ and $T-\delta$, as illustrated in Figure 1 is sufficient. This choice is similar to that made in Chen et al. (2023a), reflecting their observation that exponentially decaying time steps appear optimal, at least theoretically.

The tightest previous bounds on the iteration complexity were obtained by Chen et al. (2023a;d). Assuming only finite second moments of p_{data}, Chen et al. (2023a) showed that diffusion models with early stopping have an iteration complexity of $\tilde{O}\big(\frac{d^2 \log^2(1/\delta)}{\varepsilon^2}\big)$. Alternatively, Chen et al. (2023a;d) showed that if the score is L-Lipschitz for all $t \geq 0$ then the iteration complexity is $\tilde{O}\big(\frac{dL^2}{\varepsilon^2}\big)$. The Lipschitz assumption can be removed using early stopping at the cost of an additional factor of $1/\delta^4$, arising from the fact that for an arbitrary data distribution the Lipschitz constant of $\nabla \log q_t$ explodes at rate $1/t^2$ as $t \to 0$ (Chen et al., 2023d, Lemma 20). As the distance between q_0 and q_δ scales with $d^{1/2}$ (e.g. in Wasserstein-p metric), for a fixed Wasserstein-plus-KL (or Wasserstein-plus-TV) error we should scale δ proportionally to $d^{-1/2}$. As a result, for a fixed approximation error for p_{data}, the bounds of Chen et al. (2023d) and Chen et al. (2023a, Theorem 2) also scale superlinearly with d.

3 Proof of Theorem 1

We now provide the proof of Theorem 1, which comes in three steps. We view Step 1 as our main novel contribution, while Steps 2 and 3 closely follow Chen et al. (2023a;d).

Step 1: We bound the error from discretizing the reverse SDE (see Lemma 2 below). Previous works bound $E_{s,t} := \mathbb{E}\left[\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)\|^2\right]$ using a Lipschitz assumption. We use a new Itô calculus argument to get a differential inequality for $E_{s,t}$ (Lemmas 3 and 4), relate the coefficients of this inequality to $\mathbf{m}_t$ and $\mathbf{\Sigma}_t$ using known results (Lemma 5), and bound the differential inequality coefficients using our key Lemma 1 (resulting in Lemmas 6 and 7).

Step 2: We bound the KL distance between the path measures of the true and approximate reverse process (Lemma 8). We use an analogous Girsanov-based method to Chen et al. (2023a;d).

Step 3: We use the data processing inequality to bound $\mathrm{KL}(q_\delta || p_{t_N})$ in terms of the distance between reverse path measures and the distance between q_T and π_d. We bound the former using Step 2 and the latter using the convergence of the OU process (Proposition 4), as in Chen et al. (2023a;d).

Rather than work with two processes $(Y_t)_{t\in[0,T]}$ and $(\hat{Y}_t)_{t\in[0,T]}$ which are solutions to (2) and (4) under the same probability measure, it is more convenient to follow Chen et al. (2023a;d) and fix a single process $(Y_t)_{t\in[0,T]}$ and then define Q and P^{π_d} to be two different probability measures such that $(Y_t)_{t\in[0,T]}$ is a solution to (2) under Q and to (4) under P^{π_d}. In addition, we define a probability measure P^{q_T} under which $(Y_t)_{t\in[0,T]}$ is a solution to (4) but with the initial condition $Y_0 \sim q_T$.

3.1 Bounding the Discretization Error

Lemma 2 (Bound on discretization error)**.** *If $(Y_t)_{t\in[0,T]}$ is the solution to the SDE* (2)*, then we have*

$$\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-t_k}(Y_{t_k})\|^2\right] \mathrm{d}t \lesssim \kappa^2 dN + \kappa dT.$$

Proof. We start by controlling $E_{s,t}$ for $0 \leq s \leq t < T$, where $(Y_t)_{t\geq 0}$ follows the law Q of the exact reverse process (2). Since $\nabla \log q_{T-t}(\mathbf{x})$ is smooth, we may apply Itô's lemma to get the following result, proved in Appendix E.

Lemma 3. *If $(Y_t)_{t\in[0,T]}$ is the solution to the SDE* (2)*, then for all $t \in [0, T)$ we have*

$$\mathrm{d}(\nabla \log q_{T-t}(Y_t)) = -\nabla \log q_{T-t}(Y_t)\mathrm{d}t + \sqrt{2}\nabla^2 \log q_{T-t}(Y_t) \cdot \mathrm{d}B'_t.$$

From Lemma 3 and the product rule, we have $\mathrm{d}(e^t \nabla \log q_{T-t}(Y_t)) = \sqrt{2}e^t \nabla^2 \log q_{T-t}(Y_t) \cdot \mathrm{d}B'_t$. Since, by (13) below, $\int_s^t e^{2r}\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-r}(Y_r)\|_F^2\right] \mathrm{d}r < \infty$ for $0 \leq s \leq t < T$, where we denote by $\|A\|_F = \mathrm{Tr}(A^T A)^{1/2}$ the Frobenius norm of a matrix A, the RHS is a square-integrable martingale. So, we may apply Itô's isometry (Le Gall, 2016, Equation 5.8) and differentiate to get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_Q\left[\|e^t \nabla \log q_{T-t}(Y_t) - e^s \nabla \log q_{T-s}(Y_s)\|^2\right] = 2e^{2t}\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right]. \quad (8)$$

In addition, for $0 \le s \le t < T$ where s is considered fixed and t may vary, from Lemma 3 we have

$$\mathrm{d}(\nabla \log q_{T-s}(Y_s) \cdot \nabla \log q_{T-t}(Y_t)) = -\nabla \log q_{T-s}(Y_s) \cdot \nabla \log q_{T-t}(Y_t)\mathrm{d}t + \sqrt{2}\nabla \log q_{T-s}(Y_s) \cdot \nabla^2 \log q_{T-t}(Y_t) \cdot \mathrm{d}B'_t.$$

Since $\int_s^t \mathbb{E}_Q\left[\|\nabla^2 \log q_{T-r}(Y_r)\|_F^2\right] \mathrm{d}r < \infty$ for $0 \le s \le t < T$, the final term is a square-integrable martingale. Integrating and taking expectations, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_Q\left[\nabla \log q_{T-s}(Y_s) \cdot \nabla \log q_{T-t}(Y_t)\right] = -\mathbb{E}_Q\left[\nabla \log q_{T-s}(Y_s) \cdot \nabla \log q_{T-t}(Y_t)\right], \quad (9)$$

where again we may interchange integration and expectation using Fubini. Combining (8) and (9), we deduce the following differential inequality for $E_{s,t}$. (See Appendix E for full derivation.)

Lemma 4. *For all $0 \le s \le t < T$, we have*

$$\frac{\mathrm{d}E_{s,t}}{\mathrm{d}t} = 2\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right] - 2\mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)\|^2\right] + 2\mathbb{E}_Q\left[\{\nabla \log q_{T-s}(Y_s) - \nabla \log q_{T-t}(Y_t)\} \cdot \nabla \log q_{T-s}(Y_s)\right]. \quad (10)$$

To further bound the RHS of (10), note that by Young's inequality

$$\mathbb{E}_Q\left[\{\nabla \log q_{T-s}(Y_s) - \nabla \log q_{T-t}(Y_t)\} \cdot \nabla \log q_{T-s}(Y_s)\right] \le \frac{1}{2}\Big\{\mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-s}(Y_s)\|^2\right] + \mathbb{E}_Q\left[\|\nabla \log q_{T-s}(Y_s)\|^2\right]\Big\}.$$

Therefore,

$$\frac{\mathrm{d}E_{s,t}}{\mathrm{d}t} \le \mathbb{E}_Q\left[\|\nabla \log q_{T-s}(Y_s)\|^2\right] + 2\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right]. \quad (11)$$

We must thus bound $\mathbb{E}_Q\left[\|\nabla \log q_{T-s}(Y_s)\|^2\right]$ and $\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right]$. We make use of the following lemma, which is found in previous work on diffusion models (see e.g. De Bortoli (2022); Lee et al. (2023); Benton et al. (2023)) and which we prove for completeness in Appendix E.

Lemma 5. *For all $t > 0$, we have $\nabla \log q_t(\mathbf{x}_t) = -\sigma_t^{-2}\mathbf{x}_t + e^{-t}\sigma_t^{-2}\mathbf{m}_t$ and $\nabla^2 \log q_t(\mathbf{x}_t) = -\sigma_t^{-2}I + e^{-2t}\sigma_t^{-4}\mathbf{\Sigma}_t$.*

We may use Lemma 5 to rewrite $\|\nabla \log q_{T-s}(Y_s)\|^2$ and $\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2$ in terms of Y_t, $\mathbf{m}_t$ and $\mathbf{\Sigma}_t$. Expanding out the resulting expressions, the first can be bounded using elementary properties of the OU process, while the second can be bounded using properties of the OU process plus our key Lemma 1. We obtain the following inequalities (see Appendix E for derivations).

Lemma 6. *If $(Y_t)_{t\in[0,T]}$ is the solution to the reverse SDE* (2)*, then for all $t, s \in [0, T)$ we have*

$$\mathbb{E}_Q\left[\|\nabla \log q_{T-s}(Y_s)\|^2\right] \le d\sigma_{T-s}^{-2}, \quad (12)$$

and

$$\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right] \le d\sigma_{T-t}^{-4} - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}r}\big(\sigma_{T-r}^{-4}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-r})\right]\big)|_{r=t}. \quad (13)$$

Putting the bounds in (12) and (13) together, we get

$$\mathbb{E}_Q\left[\|\nabla \log q_{T-s}(Y_s)\|^2\right] + 2\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right] \le d\sigma_{T-s}^{-2} + 2d\sigma_{T-t}^{-4} - \frac{\mathrm{d}}{\mathrm{d}r}\big(\sigma_{T-r}^{-4}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-r})\right]\big)|_{r=t}.$$

Let us denote the two parts of this error term by

$$E_{s,t}^{(1)} := d\sigma_{T-s}^{-2} + 2d\sigma_{T-t}^{-4}, \qquad E_{s,t}^{(2)} := -\frac{\mathrm{d}}{\mathrm{d}r}\big(\sigma_{T-r}^{-4}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-r})\right]\big)|_{r=t}.$$

From (11), it follows that

$$E_{t_k,t} \le \int_{t_k}^t \mathbb{E}_Q\left[\|\nabla \log q_{T-t_k}(Y_{t_k})\|^2\right] + 2\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-s}(Y_s)\|_F^2\right]\mathrm{d}s \le \int_{t_k}^t E_{t_k,s}^{(1)} + E_{t_k,s}^{(2)}\mathrm{d}s. \quad (14)$$

We now bound the contributions to $E_{t_k,t}$ from $E^{(1)}_{t_k,s}$ and $E^{(2)}_{t_k,s}$. It will be convenient to divide our time period into intervals $[0, T-1]$ and $[T-1, T-\delta]$ and treat these separately. We therefore assume there is an index M with $t_M = T-1$. This assumption is purely for presentation clarity and our argument works similarly without it. Then, the following lemma bounds the various contributions to $E_{t_k,t}$. The proof of Lemma 7 is elementary but technical, so we defer it to Appendix E.

Lemma 7. *The error terms $E^{(1)}_{t_k,s}$ satisfy*

$$\sum_{k=0}^{M-1} \int_{t_k}^{t_{k+1}} \left(\int_{t_k}^{t} E^{(1)}_{t_k,s} \mathrm{d}s \right) \mathrm{d}t \lesssim \kappa d T, \qquad \sum_{k=M}^{N-1} \int_{t_k}^{t_{k+1}} \left(\int_{t_k}^{t} E^{(1)}_{t_k,s} \mathrm{d}s \right) \mathrm{d}t \lesssim \kappa^2 d N, \tag{15}$$

and the error terms $E^{(2)}_{t_k,s}$ satisfy

$$\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \left(\int_{t_k}^{t} E^{(2)}_{t_k,s} \mathrm{d}s \right) \mathrm{d}t \lesssim \kappa d + \kappa^2 d N. \tag{16}$$

Finally, by combining (14) with (15) and (16) we complete the proof of Lemma 2. □

3.2 Bounding the KL Distance Between Path Measures

Lemma 8 (Bound on distance between path measures)**.** *If Q and P^{q_T} are the true and approximate path measures respectively then* $\mathrm{KL}(Q||P^{q_T}) \lesssim \varepsilon^2_{\text{score}} + \kappa^2 d N + \kappa d T$.

Proof. We use the following result from Chen et al. (2023d) (proof recalled in Appendix F).

Proposition 3 (Section 5.2 of Chen et al. (2023d))**.** *Let Q and P^{q_T} be the path measures of the solutions to* (2) *and* (4) *respectively, both started in $Y_0 \sim q_T$ and run from $t = 0$ to $t = t_N$. Assume that*

$$\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q \left[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T - t_k)\|^2 \right] \mathrm{d}t < \infty.$$

Then, we have

$$\mathrm{KL}(Q||P^{q_T}) \leq \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q \left[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T - t_k)\|^2 \right] \mathrm{d}t.$$

To apply Proposition 3, we note that by Lemma 2 and Assumption 1 we have

$$\begin{aligned}
&\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q \left[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T - t_k)\|^2 \right] \mathrm{d}t \\
&\qquad \lesssim \sum_{k=0}^{N-1} \gamma_k \mathbb{E}_Q \left[\|\nabla \log q_{T-t_k}(Y_{t_k}) - s_\theta(Y_{t_k}, T - t_k)\|^2 \right] \\
&\qquad\qquad + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_Q \left[\|\nabla \log q_{T-t}(Y_t) - \nabla \log q_{T-t_k}(Y_{t_k})\|^2 \right] \mathrm{d}t \\
&\qquad \lesssim \varepsilon^2_{\text{score}} + \kappa^2 d N + \kappa d T < \infty.
\end{aligned}$$

Therefore, the conditions of Proposition 3 hold and Lemma 8 follows by applying Proposition 3. □

3.3 Completing the Proof

Finally, we show how to complete the proof of Theorem 1 from Lemma 8. As a corollary of Lemma 8, we see that Q is absolutely continuous with respect to P^{q_T}. Since P^{q_T} and P^{π_d} differ only by a change of starting distribution, we can write $\frac{\mathrm{d}P^{q_T}}{\mathrm{d}P^{\pi_d}}(\mathbf{y}) = \frac{\mathrm{d}q_T}{\mathrm{d}\pi_d}(\mathbf{y}_0)$ for a path $\mathbf{y} = (\mathbf{y}_t)_{t \in [0, t_N]}$,

and deduce that P^{q_T} and P^{π_d} are mutually absolutely continuous. It follows that Q is absolutely continuous with respect to P^{π_d} and $\frac{\mathrm{d}Q}{\mathrm{d}P^{\pi_d}}(\mathbf{y}) = \frac{\mathrm{d}Q}{\mathrm{d}P^{q_T}}(\mathbf{y})\frac{\mathrm{d}P^{q_T}}{\mathrm{d}P^{\pi_d}}(\mathbf{y}) = \frac{\mathrm{d}Q}{\mathrm{d}P^{q_T}}(\mathbf{y})\frac{\mathrm{d}q_T}{\mathrm{d}\pi_d}(\mathbf{y}_0)$. Therefore,

$$\mathrm{KL}(Q||P^{\pi_d}) = \mathbb{E}_Q\left[\log\left(\frac{\mathrm{d}Q}{\mathrm{d}P^{q_T}}(Y)\frac{\mathrm{d}q_T}{\mathrm{d}\pi_d}(Y_0)\right)\right] = \mathrm{KL}(Q||P^{q_T}) + \mathrm{KL}(q_T||\pi_d) \tag{17}$$

The first term is bounded by Lemma 8. The second is controlled by the convergence of the forward process in KL divergence and can be bounded using the following proposition (which is very similar to Lemma 9 in Chen et al. (2023a)). We defer the proof of Proposition 4 to Appendix G.

Proposition 4. *Under Assumption 2, we have* $\mathrm{KL}(q_T||\pi_d) \lesssim de^{-2T}$ *for* $T \geq 1$.

Since q_δ and p_{t_N} are the pushfowards of the path measures Q and P^{π_d} under $f : (\omega_t)_{t\in[0,t_N]} \mapsto \omega_{t_N}$, the data processing inequality implies that $\mathrm{KL}(q_\delta||p_{t_N}) \leq \mathrm{KL}(Q||P^{\pi_d})$. Finally, combining this with (17), Lemma 8, and Proposition 4 completes the proof of Theorem 1.

4 Discussion

Inspecting our bound in Theorem 1, the error from approximating q_T by π_d decays exponentially in T and so is typically negligible. If the L^2 error of our score approximation is $\varepsilon^2_{\text{score}}$, then we cannot hope to improve on the term of order $\varepsilon^2_{\text{score}}$ for the KL error due to using an approximate score. It remains to consider how tight the term corresponding to the discretization of the reverse process is.

Our proof shows that under our assumptions, with perfect score approximation and initializing the reverse SDE in q_T, the KL error induced by discretizing time is of order $\kappa^2 dN + \kappa dT$. As explained in Section 2, we can think of this as being $\tilde{O}(d/N)$, assuming a suitable choice of time steps and κ, or equivalently $\tilde{O}(d\eta)$ where η is the average step size. We note that the linear dependence on d (up to logarithmic factors) here is optimal (for justification, see Appendix H).

However, it is unclear whether the linear dependence on η is optimal here. On the one hand, the KL divergence between the true and approximate reverse path measures is $\Theta(d\eta)$ in the worst case (consider the case where p_{data} is a point mass, or see Theorem 7 in Chen et al. (2023d) in the critically damped Langevin setting). Thus, our Girsanov-based method cannot improve upon the rate of $\tilde{O}(d\eta)$ without significant modification. In addition, the best known convergence rates in KL divergence for Langevin Monte Carlo (LMC) under various functional inequalities are $\tilde{O}(d\eta)$ (Cheng & Bartlett, 2018; Vempala & Wibisono, 2019; Chewi et al., 2022; Yang & Wibisono, 2022).

On the other hand, the increasing noise schedule of diffusion models may allow for improved convergence rates compared to LMC. As evidence, Mou et al. (2022) show that the EM discretization of an SDE has error $O(\eta^2)$ in reverse KL divergence, under smoothness assumptions which should be satisfied under early stopping, making only mild additional assumptions on the data distribution such as bounded support. We could apply their results to get $O(\eta^2)$ convergence bounds for the diffusion model, but the implicit constant would depend on d and on the smoothness parameters, which in turn depend polynomially on d and $1/\delta$. This leads to bounds which are quadratic in η but superlinear in d. We expect such bounds to be tighter than our own when $\eta^{-1} \gg \mathrm{poly}(d)$.

We may also get better convergence bounds by working in weaker metrics. Under some smoothness assumptions on p_{data}, our KL error bounds imply a bound of $\tilde{O}(\sqrt{\eta})$ in Wasserstein-2 distance via a Talagrand inequality (Otto & Villani, 2000). However, there is evidence that this rate is suboptimal. Under smoothness assumptions on the drift, the EM discretization has error $\tilde{O}(\eta)$ in Wasserstein-ρ metric for $\rho \geq 1$ (Alfonsi et al., 2015). Under smoothness and convexity assumptions on the data distribution, LMC converges at rate $O(\eta)$ (Durmus & Moulines, 2019; Li et al., 2022). However, these results require additional smoothness assumptions and have superlinear dependence on d.

Ultimately, there appears to be a trade off between the dependence on the data dimension and the step size. Our key to obtaining bounds which are tight in d was to use bounds on the drift coefficient of the reverse SDE which hold in expectation. All methods of which we are aware that achieve a better dependence on η require stronger (e.g. L^∞) control on the drift. These necessitate worse dimension dependence and additional smoothness assumptions. We leave bridging the gap between these two strands of proofs to future work.

ACKNOWLEDGMENTS

We thank Tyler Farghly for pointing out a small gap in the original version of this work. Joe Benton was supported by the EPSRC through the StatML CDT (EP/S023151/1). Arnaud Doucet acknowledges support from EPSRC grants EP/R034710/1 and EP/R018561/1.

REFERENCES

Ahmed El Alaoui and Andrea Montanari. An Information-Theoretic View of Stochastic Localization. *IEEE Transactions on Information Theory*, 68(11):7423–7426, 2022.

Michael S Albergo and Eric Vanden-Eijnden. Building Normalizing Flows with Stochastic Interpolants. In *International Conference on Learning Representations*, 2023.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Aurélien Alfonsi, Benjamin Jourdain, and Arturo Kohatsu-Higa. Optimal transport bounds between the time-marginals of a multidimensional diffusion and its Euler scheme. *Electronic Journal of Probability*, 20:1–31, 2015.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

Brian D O Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Milla Anttila, Keith Ball, and Irini Perissinaki. The central limit problem for convex bodies. *Transactions of the American Mathematical Society*, 355(12):4723–4735, 2003.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. In *Advances in Neural Information Processing Systems*, 2021.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

Joe Benton, George Deligiannidis, and Arnaud Doucet. Error Bounds for Flow Matching Methods. *arXiv preprint arXiv:2305.16860*, 2023.

Giulio Biroli and Marc Mézard. Generative diffusion in very large dimensions. *arXiv preprint arXiv:2306.03518*, 2023.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative Modeling with Denoising Auto-Encoders and Langevin Sampling. *arXiv preprint arXiv:2002.00107*, 2022.

Sergey G Bobkov and Alexander Koldobsky. On the Central Limit Property of Convex Bodies. In *Geometric Aspects of Functional Analysis: Israel Seminar 2001-2002*, pp. 44–52. Springer, 2003.

Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 2022.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions. In *International Conference on Machine Learning*, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data. *arXiv preprint arXiv:2302.07194*, 2023b.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ODE is provably fast. *arXiv preprint arXiv:2305.11798*, 2023c.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023d.

Sitan Chen, Giannis Daras, and Alexandros G Dimakis. Restoration-Degradation Beyond Linear Diffusions: A Non-Asymptotic Analysis For DDIM-Type Samplers. In *International Conference on Machine Learning*, 2023e.

Yuansi Chen. An Almost Constant Lower Bound of the Isoperimetric Coefficient in the KLS Conjecture. *Geometric and Functional Analysis*, 31:34–61, 2021.

Yuansi Chen and Ronen Eldan. Localization Schemes: A Framework for Proving Mixing Bounds for Markov Chains. *IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 110–122, 2022.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83, pp. 186–211, may 2018.

Sinho Chewi, Murat A Erdogdu, Mufan Bill Li, Ruoqi Shen, and Matthew Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. In *Proceedings of Thirty Fifth Conference on Learning Theory*, 2022.

Giovanni Conforti, Alain Durmus, and Marta Gentiloni Silveri. Score diffusion models without early stopping: finite Fisher information is all you need. *arXiv preprint arXiv:2308.12240*, aug 2023.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems*, 2021.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 2021.

Alain Durmus and Éric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the Sherrington-Kirkpatrick Gibbs measure via algorithmic stochastic localization. *IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 323–334, 2022.

Ronen Eldan. Thin Shell Implies Spectral Gap Up to Polylog via a Stochastic Localization Scheme. *Geometric and Functional Analysis*, 23(2):532–569, 2013.

Ronen Eldan. Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probability Theory and Related Fields*, 176(3-4):737–755, 2020.

Davide Ghio, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Sampling with flows, diffusion and autoregressive neural networks: A spin-glass perspective. *arXiv preprint arXiv:2308.14085*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, 2020.

Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13:541–559, 1995.

Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*. Springer, 2016.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pp. 946–985, 2023.

Yin Tat Lee and Santosh S Vempala. Eldan's Stochastic Localization and the KLS Hyperplane Conjecture: An Improved Lower Bound for Expansion. *IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 998–1007, 2017.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models. *arXiv preprint arXiv:2306.09251*, 2023.

Ruilin Li, Hongyuan Zha, and Molei Tao. Sqrt(d) Dimension Dependence of Langevin Monte Carlo. In *International Conference on Learning Representations*, 2022.

Robert S Liptser and Albert N Shiryaev. *Statistics of Random Processes: General Theory*. Springer, 1977.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Let us Build Bridges: Understanding and Extending Diffusion Generative Models. *arXiv preprint arXiv:2208.14699*, 2022.

László Lovász and Miklós Simonovits. Random walks in a convex body and an improved volume algorithm. *Random Structures & Algorithms*, 4(4):359–412, 1993.

Andrea Montanari. Sampling, Diffusions, and Stochastic Localization. *arXiv preprint arXiv:2305.10690*, 2023.

Andrea Montanari and Yuchen Wu. Posterior Sampling from the Spiked Models via Diffusion Processes. *arXiv preprint arXiv:2304.11449*, 2023.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved Bounds for Discretization of Langevin Diffusions: Near-Optimal Rates without Convexity. *Bernoulli*, 28 (3):1577–1601, 2022.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion Models are Minimax Optimal Distribution Estimators. In *ICLR 2023 Workshop on Understanding Foundation Models*, 2023.

Felix Otto and Cédric Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Jakiw Pidstrigach. Score-Based Generative Models Detect Manifolds. In *Advances in Neural Information Processing Systems*, 2022.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In *International Conference on Machine Learning*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022.

Herbert Robbins. An Empirical Bayes Approach to Statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–164, 1956.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, 2022.

Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*, volume 32. Neural information processing systems foundation, 2019.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021b.

Brian L Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. In *International Conference on Learning Representations*, 2023.

Santosh S Vempala and Andre Wibisono. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices. In *Advances in Neural Information Processing Systems*, 2019.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A Geometric Diffusion Model for Molecular Conformation Generation. In *International Conference on Learning Representations*, 2022.

Kaylee Yingxi Yang and Andre Wibisono. Convergence in KL and Rényi Divergence of the Unadjusted Langevin Algorithm Using Estimated Score. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

Qinsheng Zhang and Yongxin Chen. Fast Sampling of Diffusion Models with Exponential Integrator. In *International Conference on Learning Representations*, 2023.

A EQUIVALENCE OF DIFFUSION MODELS AND STOCHASTIC LOCALIZATION

Suppose that $(X_t)_{t\geq 0}$ follows the OU SDE defined in (1). Using the integration by parts formula for continuous semimartingales (Le Gall, 2016),

$$\mathrm{d}(e^t X_t) = e^t X_t \mathrm{d}t + e^t\{-X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t\} = \sqrt{2}e^t \mathrm{d}B_t.$$

By the Dubins–Schwarz theorem (Le Gall, 2016, Theorem 5.13), there is a process $(\hat{W}_s)_{s\geq 0}$ such that

$$\hat{W}_{e^{2s}-1} = \int_0^s \sqrt{2}e^r \mathrm{d}B_r$$

and $(\hat{W}_s)_{s\geq 0}$ is a standard Brownian motion with respect to the filtration $(\mathcal{F}_{\tau(s)})_{s\geq 0}$ where $\tau(s) = \frac{1}{2}\log(1+s)$. Then, for all $s \in (0,\infty)$ we can write

$$e^{\tau(s)} X_{\tau(s)} = X_0 + \hat{W}_s.$$

If we set $U_0 = 0$ and

$$U_s := s e^{\tau(1/s)} X_{\tau(1/s)} = s X_0 + s\hat{W}_{1/s}$$

for $s \in (0,\infty)$, then we observe that $(U_s)_{s\geq 0}$ satisfies the definition of the stochastic localization process in (5), since the law of $(s\hat{W}_{1/s})_{s\geq 0}$ is the same as the law of $(W_s)_{s\geq 0}$. Thus the forward diffusion process and the stochastic localization process are equivalent under the chance of time variables $t(s) = \frac{1}{2}\log(1+s^{-1})$.

In addition, we see that conditioning on U_s is equivalent to conditioning on $X_{t(s)}$ and thus μ_s and $q_{0|t}(\,\cdot\,|X_t)$ define the same distributions when $t = t(s)$. It follows that $\mathbf{a}_s(U_s)$ and $\mathbf{m}_t(X_t)$ have the same law and $\mathbf{A}_s(U_s)$ and $\mathbf{\Sigma}_t(X_t)$ have the same law when $t = t(s)$.

B PROOFS OF STOCHASTIC LOCALIZATION RESULTS

Propositions 1 and 2 are well-known and we reproduce the proofs here for convenience. Proposition 1 can be found for example in Eldan (2013); Alaoui & Montanari (2022); El Alaoui et al. (2022) and our proof of Proposition 2 is based on the argument on pages 8–9 of Eldan (2020).

Proof of Proposition 1. From (5), we can deduce that

$$\mu_s(\mathrm{d}\mathbf{x}) = \frac{1}{Z_s}\exp\Big\{\mathbf{x}\cdot U_s - \frac{s}{2}\|\mathbf{x}\|^2\Big\} p_{\text{data}}(\mathrm{d}\mathbf{x}),$$

where $Z_s = \int_{\mathbb{R}^d} \exp\big\{\mathbf{x}\cdot U_s - \frac{s}{2}\|\mathbf{x}\|^2\big\}\, p_{\text{data}}(\mathrm{d}\mathbf{x})$ is the normalizing constant. Therefore,

$$\mathrm{d}\log L_s(\mathbf{x}) = \mathbf{x}\cdot \mathrm{d}U_s - \frac{1}{2}\|\mathbf{x}\|^2 \mathrm{d}s - \mathrm{d}\log Z_s. \tag{18}$$

Writing $h_s(\mathbf{x}) = \mathbf{x}\cdot U_s - \frac{s}{2}\|\mathbf{x}\|^2$ and using the definition of U_s in (5) plus $\mathrm{d}[W,W]_s = \mathrm{d}s$, we have $\mathrm{d}h_s(\mathbf{x}) = \mathbf{x}\cdot \mathrm{d}U_s - \frac{1}{2}\|\mathbf{x}\|^2\mathrm{d}s$ and $\mathrm{d}[h(\mathbf{x}), h(\mathbf{x})]_s = \|\mathbf{x}\|^2\mathrm{d}s$. Since $h_s(\mathbf{x})$ is a continuous semi-martingale and $\exp$ is smooth, we may apply Itô's lemma to $\exp\{h_s(\mathbf{x})\}$ and integrate with respect to $p_{\text{data}}(\mathbf{x})$ to get

$$\begin{aligned}
\mathrm{d}Z_s &= \int_{\mathbb{R}^d} \Big(\mathrm{d}h_s(\mathbf{x}) + \frac{1}{2}\mathrm{d}[h(\mathbf{x}), h(\mathbf{x})]_s\Big) e^{h_s(\mathbf{x})} p_{\text{data}}(\mathrm{d}\mathbf{x}) \\
&= \int_{\mathbb{R}^d} \mathbf{x}\cdot \mathrm{d}U_s e^{h_s(\mathbf{x})} p_{\text{data}}(\mathrm{d}\mathbf{x}) \\
&= Z_s(\mathbf{a}_s \cdot \mathrm{d}U_s).
\end{aligned}$$

Then, via another application of Itô's lemma, since Z_s is continuous and $\log$ is smooth,

$$\mathrm{d}\log Z_s = \frac{\mathrm{d}Z_s}{Z_s} - \frac{1}{2}\frac{d[Z]_s}{Z_s^2} = \mathbf{a}_s\cdot \mathrm{d}U_s - \frac{1}{2}\|\mathbf{a}_s\|^2 \mathrm{d}s.$$

Substituting this into (18), we see that

$$\begin{aligned}\mathrm{d}\log L_s(\mathbf{x}) &= (\mathbf{x}-\mathbf{a}_s)\cdot(\mathrm{d}U_s - \mathbf{a}_s\mathrm{d}s) - \frac{1}{2}\|\mathbf{x}-\mathbf{a}_s\|^2\mathrm{d}s \\ &= (\mathbf{x}-\mathbf{a}_s)\cdot\mathrm{d}W'_s - \frac{1}{2}\|\mathbf{x}-\mathbf{a}_s\|^2\mathrm{d}s\end{aligned}$$

where we recall the definition of $(W'_s)_{s\geq 0}$ from (6). The result then follows via a final application of Itô's lemma. □

Proof of Proposition 2. First, using Proposition 1 we have

$$\begin{aligned}\mathrm{d}\mathbf{a}_s &= \mathrm{d}\left(\int_{\mathbb{R}^d}\mathbf{x}L_s(\mathbf{x})p_{\text{data}}(\mathrm{d}\mathbf{x})\right) \\ &= \int_{\mathbb{R}^d}\mathbf{x}\mathrm{d}L_s(\mathbf{x})p_{\text{data}}(\mathrm{d}\mathbf{x}) \\ &= \int_{\mathbb{R}^d}\mathbf{x}\otimes(\mathbf{x}-\mathbf{a}_s)L_s(\mathbf{x})\cdot\mathrm{d}W'_s\, p_{\text{data}}(\mathrm{d}\mathbf{x}).\end{aligned}$$

This implies that

$$\mathrm{d}\mathbf{a}_s = \mathbb{E}_{\mu_s}\left[\xi\otimes(\xi-\mathbf{a}_s)\right]\mathrm{d}W'_s = \mathbf{A}_s\cdot\mathrm{d}W'_s.$$

It then follows from Itô's isometry that

$$\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[\mathbf{a}_s^{\otimes 2}\right] = \mathbb{E}\left[\mathbf{A}_s^2\right].$$

The result then follows from the fact that $\mathbb{E}\left[\mathbf{A}_s\right] = \mathbb{E}_{\mu_s}\left[\xi^{\otimes 2}\right] - \mathbb{E}\left[\mathbf{a}_s^{\otimes 2}\right]$. □

C Adaptations Required to Handle a General Covariance of p_{data}

We briefly outline the changes required in our proofs to handle a data distribution with a general covariance matrix. The main changes required are in the proofs of Lemmas 6 and 7 and Proposition 4. In this section, we replace Assumption 2 with the following.

Assumption 3. *The data distribution p_{data} has finite second moments, with $M_2 := \mathbb{E}_{p_{\text{data}}}\left[\|X_0\|^2\right]$.*

First, we note that the proofs of Lemmas 3, 4 and 5 go through unchanged, and so these results also hold in the more general setting. For Lemma 6, the proof of (12) can be adapted, replacing each time we use $\mathbb{E}\left[\|X_0\|^2\right] = d$ with $\mathbb{E}\left[\|X_0\|^2\right] = M_2$ and noting that since X_0 and $X_t - e^{-t}X_0$ are independent, we have $\mathbb{E}\left[\|X_t\|^2\right] = \mathbb{E}\left[\|(X_t - e^{-t}X_0) + e^{-t}X_0\|^2\right] = d\sigma_t^2 + e^{-2t}M_2$. We find that all instances of M_2 cancel in the final result and (12) holds unchanged. In addition, the proof of (13) needs no alteration.

The only change in the proof of Lemma 7 comes towards the end, when we assert that $\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] = \mathbb{E}\left[\mathbb{E}\left[\|X_0\|^2|X_t\right] - \|\mathbb{E}\left[X_0|X_t\right]\|^2\right] \leq \mathbb{E}\left[\|X_0\|^2\right] = d$ for $t\in[1,T]$. Under Assumption 3, this becomes $\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] \leq \mathbb{E}\left[\|X_0\|^2\right] = M_2$. Propagating this through the rest of the proof, we find that (16) should be replaced by

$$\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\left(\int_{t_k}^{t}E^{(2)}_{t_k,s}\mathrm{d}s\right)\mathrm{d}t \lesssim \kappa M_2 + \kappa^2 dN. \tag{19}$$

The final change that must be made is to Proposition 4, which must be replaced by the following more general version that can be proved along identical lines.

Proposition 5. *Under Assumption 3, we have* $\mathrm{KL}(q_T||\pi_d) \lesssim (d+M_2)e^{-2T}$ *for* $T\geq 1$.

Putting all of these changes together, we arrive at the following more general version of Theorem 1.

Theorem 2. *Suppose that Assumptions 1 and 3 hold, that $T\geq 1$, and that there is some $\kappa > 0$ such that for each $k = 0,\dots,N-1$ we have $\gamma_k \leq \kappa\min\{1, T-t_{k+1}\}$. Then,*

$$\mathrm{KL}(q_\delta||p_{t_N}) \lesssim \varepsilon^2_{\text{score}} + \kappa^2 dN + \kappa dT + \kappa M_2 + (d+M_2)e^{-2T}.$$

Theorem 2 holds even in the case where the covariance is not strictly positive definite, and in the case where it is unknown to our diffusion model algorithm. Since we expect M_2 to scale linearly in d in most cases, we consider this result to be in essentially the same spirit as Theorem 1.

D PROOF OF COROLLARY 1

Proof of Corollary 1. For convenience, we assume that N is even. We take $t_0 = 0$, $t_{N/2} = T - 1$, and $t_N = T - \delta$, and pick $t_1, \dots, t_{N-1}$ such that $t_0, \dots, t_{N/2}$ are linearly spaced on $[0, T-1]$ and $T - t_{N/2}, \dots, T - t_N$ are an exponentially decaying sequence from 1 to δ (illustrated in Figure 1).

Then, $\gamma_k \le \kappa \min\{1, T - t_{k+1}\}$ for each $k = 0, \dots, N-1$ if and only if $\kappa \ge (T-1)/(N/2)$ and $\kappa \ge (1/\delta)^{1/N} - 1$. The former is satisfied if we take $\kappa = \Omega\big(\frac{T}{N}\big)$ and the second is satisfied provided that $\kappa = \Omega\big(\frac{\log(1/\delta)}{N}\big)$, since $N \ge \log(1/\delta)$ and $e^x \le 1 + (e-1)x$ for $x \le 1$. Therefore, for some $\kappa = \Theta\big(\frac{T + \log(1/\delta)}{N}\big)$ we have $\gamma_k \le \kappa \min\{1, T - t_{k+1}\}$ for each $k = 0, \dots, N-1$, proving the first part of Corollary 1.

For the second part, suppose that we set $T = \frac{1}{2}\log\big(\frac{d}{\varepsilon_{\text{score}}^2}\big)$ and $N = \Theta\big(\frac{d(T + \log(1/\delta))^2}{\varepsilon_{\text{score}}^2}\big)$. Then, we have $\kappa^2 dN = O(\varepsilon_{\text{score}}^2)$, $\kappa dT = O(\varepsilon_{\text{score}}^2)$, and $de^{-2T} = O(\varepsilon_{\text{score}}^2)$. We may therefore apply Theorem 1 to get that $\mathrm{KL}(q_\delta || p_{t_N}) = O(\varepsilon_{\text{score}}^2)$. The bound on the iteration complexity then follows since T depends only logarithmically on d and $\varepsilon_{\text{score}}$. □

E OMITTED PROOFS FROM SECTION 3

Here, we provide the proofs of Lemmas 3, 4, 5, 6 and 7 which were omitted from Section 3.

Proof of Lemma 3. Recall that the reverse process $(Y_t)_{t \in [0,T]}$ satisfies

$$\mathrm{d}Y_t = \{Y_t + 2\nabla \log q_{T-t}(Y_t)\}\mathrm{d}t + \sqrt{2}\mathrm{d}B'_t.$$

Since $\nabla \log q_{T-t}(\mathbf{x})$ is smooth for $t \in [0, T)$, by Itô's lemma we can write

$$\begin{aligned}\mathrm{d}(\nabla \log q_{T-t}(Y_t)) = \big\{\nabla^2 \log q_{T-t}(X_t) \cdot \{Y_t + 2\nabla \log q_{T-t}(Y_t)\} + \Delta(\nabla \log q_{T-t})(Y_t)\big\}\,\mathrm{d}t \\ + \frac{\mathrm{d}(\nabla \log q_{T-t})(Y_t)}{\mathrm{d}t}\mathrm{d}t + \sqrt{2}\nabla^2 \log q_{T-t}(Y_t) \cdot \mathrm{d}B'_t. \quad (20)\end{aligned}$$

The Fokker–Planck equation for the forward process is

$$\mathrm{d}q_t(\mathbf{x}) = \{-\nabla \cdot (-\mathbf{x} q_t(\mathbf{x})) + \Delta q_t(\mathbf{x})\}\mathrm{d}t,$$

from which we can deduce

$$\mathrm{d}(\log q_t)(\mathbf{x}) = \{d + \mathbf{x} \cdot \nabla \log q_t(\mathbf{x}) + \Delta \log q_t(\mathbf{x}) + \|\nabla \log q_t(\mathbf{x})\|^2\}\mathrm{d}t.$$

It follows that

$$\begin{aligned}\frac{\mathrm{d}(\nabla \log q_{T-t})}{\mathrm{d}t}(\mathbf{x}) = -\{\nabla \log q_{T-t}(\mathbf{x}) + \nabla^2 \log q_{T-t}(\mathbf{x}) \cdot \mathbf{x} + \nabla(\Delta \log q_{T-t}(\mathbf{x})) \\ + 2\nabla^2 \log q_{T-t}(\mathbf{x}) \cdot \nabla \log q_{T-t}(\mathbf{x})\},\end{aligned}$$

by interchanging the order of the derivative operators. Substituting this into (20) and simplifying, we obtain the desired result. □

Proof of Lemma 4. First, expanding (8) we get

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t)\|^2\right] + 2\mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t)\|^2\right] \\ &\qquad - 2e^{-(t-s)}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_Q\left[\nabla \log q_{T-s}(Y_s) \cdot \nabla \log q_{T-t}(Y_t)\right] \\ &\qquad\qquad - 2e^{-(t-s)}\mathbb{E}_Q\left[\nabla \log q_{T-s}(Y_s) \cdot \nabla \log q_{T-t}(Y_t)\right] \\ &= 2\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right].\end{aligned}$$

Combined with (9), this shows that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t)\|^2\right] + 2\mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t)\|^2\right] = 2\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right].$$

Then,

$$\begin{aligned}\frac{\mathrm{d}E_{s,t}}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t)\|^2\right] - 2\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}_Q\left[\nabla \log q_{T-s}(Y_s)\cdot \nabla \log q_{T-t}(Y_t)\right] \\ &= 2\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right] - 2\mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t)\|^2\right] \\ &\quad + 2\mathbb{E}_Q\left[\nabla \log q_{T-s}(Y_s)\cdot \nabla \log q_{T-t}(Y_t)\right]\end{aligned}$$

which rearranges to give (10). □

Proof of Lemma 5. Part (i) is a classical result, sometimes known as Tweedie's formula (Robbins, 1956). Part (ii) has been established in previous works (see e.g. De Bortoli (2022, Lemma C.2) or Lee et al. (2023, Lemma 4.13)). We provide proofs of both results for reference.

For (i), we have

$$\nabla \log q_t(\mathbf{x}_t) = \frac{1}{q_t(\mathbf{x}_t)}\int_{\mathbb{R}^d} \nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)\mathrm{d}\mathbf{x}_0.$$

Since $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0 e^{-t}, \sigma_t^2 I)$, it follows that $\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = -\sigma_t^{-2}(\mathbf{x}_t - \mathbf{x}_0 e^{-t})$. Therefore,

$$\begin{aligned}\nabla \log q_t(\mathbf{x}_t) &= \mathbb{E}_{q_{0|t}(\cdot|\mathbf{x}_t)}\left[-\sigma_t^{-2}(\mathbf{x}_t - X_0 e^{-t})\right] \\ &= -\sigma_t^{-2}\mathbf{x}_t + e^{-t}\sigma_t^{-2}\mathbf{m}_t.\end{aligned}$$

For (ii), we can write

$$\begin{aligned}&\nabla^2 \log q_t(\mathbf{x}_t) \\ &= \frac{1}{q_t(\mathbf{x}_t)}\int_{\mathbb{R}^d} \nabla^2 \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)\mathrm{d}\mathbf{x}_0 \\ &\quad + \frac{1}{q_t(\mathbf{x}_t)}\int_{\mathbb{R}^d} (\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0)(\nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0))^T q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)\mathrm{d}\mathbf{x}_0 \\ &\quad - \frac{1}{q_t(\mathbf{x}_t^2)}\left(\int_{\mathbb{R}^d} \nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)\mathrm{d}\mathbf{x}_0\right)\left(\int_{\mathbb{R}^d} \nabla \log q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) q_{0,t}(\mathbf{x}_0,\mathbf{x}_t)\mathrm{d}\mathbf{x}_0\right)^T \\ &= -\frac{1}{\sigma_t^2} I + \mathbb{E}_{q_{0|t}(\cdot|\mathbf{x}_t)}\left[\sigma_t^{-4}(\mathbf{x}_t - X_0 e^{-t})(\mathbf{x}_t - X_0 e^{-t})^T\right] \\ &\quad - \mathbb{E}_{q_{0|t}(\cdot|\mathbf{x}_t)}\left[-\sigma_t^2(\mathbf{x}_t - X_0 e^{-t})\right]\mathbb{E}_{q_{0|t}(\cdot|\mathbf{x}_t)}\left[-\sigma_t^2(\mathbf{x}_t - X_0 e^{-t})\right]^T \\ &= -\sigma_t^{-2} I + \sigma_t^{-4}\mathrm{Cov}_{q_{0|t}(\cdot|\mathbf{x}_t)}(\mathbf{x}_t - X_0 e^{-t}) \\ &= -\sigma_t^{-2} I + e^{-2t}\sigma_t^{-4}\boldsymbol{\Sigma}_t.\end{aligned}$$

□

Proof of Lemma 6. First, using the first part of Lemma 5 and expanding, we see that

$$\mathbb{E}_{q_t}\left[\|\nabla \log q_t(X_t)\|^2\right] = \sigma_t^{-4}\mathbb{E}\left[\|X_t\|^2\right] - 2e^{-t}\sigma_t^{-4}\mathbb{E}\left[X_t\cdot \mathbf{m}_t\right] + e^{-2t}\sigma_t^{-4}\mathbb{E}\left[\|\mathbf{m}_t\|^2\right],$$

where all expectations are with respect to $X_t \sim q_t$. Then,

$$\mathbb{E}\left[X_t\cdot \mathbf{m}_t\right] = \mathbb{E}\left[X_t\cdot \mathbb{E}\left[X_0|X_t\right]\right] = \mathbb{E}\left[X_t\cdot X_0\right] = \mathbb{E}\left[X_0\cdot \mathbb{E}\left[X_t|X_0\right]\right] = e^{-t}\mathbb{E}\left[\|X_0\|^2\right] = de^{-t}.$$

Also, $\mathrm{Tr}(\boldsymbol{\Sigma}_t) = \mathbb{E}\left[\|X_0\|^2|\mathbf{x}_t\right] - \|\mathbf{m}_t\|^2$, so $\mathbb{E}\left[\|\mathbf{m}_t\|^2\right] = d - \mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Sigma}_t)\right]$. We conclude that

$$\mathbb{E}_{q_t}\left[\|\nabla \log q_t(X_t)\|^2\right] = d\sigma_t^{-2} - \dot{\sigma}_t\sigma_t^{-3}\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Sigma}_t)\right],$$

where we have used $\dot{\sigma}_t\sigma_t = e^{-2t}$. This implies that

$$\mathbb{E}_Q\left[\|\nabla \log q_{T-s}(Y_s)\|^2\right] \leq d\sigma_{T-s}^{-2}.$$

Second, using the second part of Lemma 5 along with the definition $\|A\|_F^2 = \mathrm{Tr}(A^T A)$ and expanding, we see that

$$\mathbb{E}_{q_t}\left[\|\nabla^2 \log q_t(X_t)\|_F^2\right] = d\sigma_t^{-4} - 2\dot{\sigma}_t\sigma_t^{-5}\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Sigma}_t)\right] + \dot{\sigma}_t^2\sigma_t^{-6}\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{\Sigma}_t^2)\right].$$

Taking traces in Lemma 1, we get

$$\frac{\sigma_t^4}{2e^{-2t}}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] = \frac{\sigma_t^4}{2\dot\sigma_t\sigma_t}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] = \frac{\sigma_t^3}{2\dot\sigma_t}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] = \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t^2)\right],$$

and since $\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t^2)\right] \geq 0$, it follows that $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] \geq 0$. Therefore,

$$\begin{aligned}\mathbb{E}_{q_t}\left[\|\nabla^2 \log q_t(X_t)\|_F^2\right] &= d\sigma_t^{-4} - 2\dot\sigma_t\sigma_t^{-5}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] + \frac{1}{2}\dot\sigma_t\sigma_t^{-3}\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] \\ &\leq d\sigma_t^{-4} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big(\sigma_t^{-4}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right]\Big),\end{aligned}$$

where we have used that $\sigma_t\dot\sigma_t \leq 1$. This implies that

$$\begin{aligned}\mathbb{E}_Q\left[\|\nabla^2 \log q_{T-t}(Y_t)\|_F^2\right] &\leq d\sigma_{T-t}^{-4} + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}r}\big(\sigma_r^{-4}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_r)\right]\big)|_{r=T-t} \\ &\leq d\sigma_{T-t}^{-4} - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}r}\big(\sigma_{T-r}^{-4}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-r})\right]\big)|_{r=t}.\end{aligned}$$

□

Proof of Lemma 7. First we control the error terms $E^{(1)}_{t_k,s}$. If $s,t \in [0, T-1]$ then $\sigma^2_{T-s}, \sigma^2_{T-t} \geq 1/2$ and so $E^{(1)}_{s,t} \leq 10d$. We therefore have

$$\begin{aligned}\sum_{k=0}^{M-1}\int_{t_k}^{t_{k+1}}\left(\int_{t_k}^{t} E^{(1)}_{t_k,s}\mathrm{d}s\right)\mathrm{d}t &\leq 5d\sum_{k=0}^{M-1}\gamma_k^2 \\ &\lesssim \kappa d T,\end{aligned}$$

since we have assumed that $\gamma_k \leq \kappa$. This proves the first part of (15).

If $s,t \in [T-1, T-\delta]$ then $(T-s)/2 \leq \sigma^2_{T-s} \leq 2(T-s)$ and similarly for t. Therefore,

$$\begin{aligned}\sum_{k=M}^{N-1}\int_{t_k}^{t_{k+1}}\left(\int_{t_k}^{t} E^{(1)}_{t_k,s}\mathrm{d}s\right)\mathrm{d}t &\leq 12d\sum_{k=M}^{N-1}\int_{t_k}^{t_{k+1}}\left(\int_{t_k}^{t}(T-s)^{-2}\mathrm{d}s\right) \\ &\leq 12d\sum_{k=M}^{N-1}\frac{\gamma_k^2}{(T-t_{k+1})^2} \\ &\lesssim \kappa^2 d N\end{aligned}$$

since we have assumed that $\gamma_k \leq \kappa(T - t_{k+1})$. This proves the second part of (15).

Next, we control the error terms $E^{(2)}_{t_k,s}$. Note that σ^{-4}_{T-t} is increasing in t and $\mathbb{E}\left[\mathrm{Tr}(\Sigma_{T-t})\right]$ is decreasing in t by Lemma 1. Therefore,

$$\begin{aligned}\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\left(\int_{t_k}^{t} E^{(2)}_{t_k,s}\mathrm{d}s\right)\mathrm{d}t &\leq \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\left(\sigma^{-4}_{T-t_k}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right] - \sigma^{-4}_{T-t}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t})\right]\right)\mathrm{d}t \\ &\leq \sum_{k=0}^{N-1}\gamma_k\left(\sigma^{-4}_{T-t_k}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right] - \sigma^{-4}_{T-t_k}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_{k+1}})\right]\right).\end{aligned}$$

We split this sum into $k = 0, \ldots, M-1$ and $k = M, \ldots, N-1$. For $k = 0, \ldots, M-1$, we have $\gamma_k\sigma^{-4}_{T-t_k} \leq 4\gamma_k \leq 4\kappa$ and so

$$\begin{aligned}&\sum_{k=0}^{M-1}\gamma_k\left(\sigma^{-4}_{T-t_k}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right] - \sigma^{-4}_{T-t_k}\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_{k+1}})\right]\right) \\ &\leq 4\kappa\sum_{k=0}^{N-1}\left(\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right] - \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_{k+1}})\right]\right) \\ &\leq 4\kappa\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_T)\right].\end{aligned}$$

For $k = M, \dots, N-1$ we have $\gamma_k \sigma_{T-t_k}^{-4} \leq 4\gamma_k/(T-t_k)^2 \leq 4\kappa/(T-t_k)$ and so

$$\begin{aligned}
&\sum_{k=M}^{N-1} \gamma_k \left(\sigma_{T-t_k}^{-4} \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right] - \sigma_{T-t_k}^{-4} \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_{k+1}})\right] \right) \\
&\leq 4\kappa \sum_{k=M}^{N-1} \frac{1}{(T-t_k)} \left(\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right] - \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_{k+1}})\right] \right) \\
&\leq 4\kappa \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_1)\right] + 4\kappa \sum_{k=M+1}^{N-1} \frac{\gamma_{k-1}}{(T-t_k)(T-t_{k-1})} \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right] \\
&\leq 4\kappa \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_1)\right] + 4\kappa^2 \sum_{k=M+1}^{N-1} \frac{1}{(T-t_k)} \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right].
\end{aligned}$$

We then have $\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] = \mathbb{E}\left[\mathbb{E}\left[\|X_0\|^2 | X_t\right] - \|\mathbb{E}\left[X_0|X_t\right]\|^2\right] \leq \mathbb{E}\left[\|X_0\|^2\right] = d$ for $t \in [1, T]$ and

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_t)\right] &= e^{2t} \mathbb{E}\left[\mathrm{Tr}(\mathrm{Cov}(X_0 e^{-t} - X_t | X_t))\right] \\
&= e^{2t} \mathbb{E}\left[\mathbb{E}\left[\|X_0 e^{-t} - X_t\|^2 | X_t\right] - \|\mathbb{E}\left[X_0 e^{-t} - X_t | X_t\right]\|^2\right] \\
&\leq e^{2t} \mathbb{E}\left[\|X_0 e^{-t} - X_t\|^2\right] \\
&\leq d e^{2t} \sigma_t^2 \\
&\leq 16 dt
\end{aligned}$$

for $t \in (0, 1]$. Putting this together, we conclude that

$$\begin{aligned}
\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \left(\int_{t_k}^{t} E_{t_k,s}^{(2)} \mathrm{d}s \right) \mathrm{d}t &\leq 4\kappa \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_T)\right] + 4\kappa \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_1)\right] \\
&\quad + 4\kappa^2 \sum_{k=M+1}^{N-1} \frac{1}{(T-t_k)} \mathbb{E}\left[\mathrm{Tr}(\mathbf{\Sigma}_{T-t_k})\right] \\
&\lesssim \kappa d + \kappa^2 dN.
\end{aligned}$$

This proves completes the proof of (16). □

F Application of Girsanov's Theorem

We now recall the proof of Proposition 3. The following approximation argument is essentially identical to that of Chen et al. (2023d, Section 5.2) and we reproduce it here simply for clarity of presentation.

The main ingredient in the proof will be Girsanov's theorem, which we recall below. The version we state can be obtained from Theorem 4.13 combined with Theorem 5.22 and Pages 136–139 in Le Gall (2016).

Proposition 6 (Girsanov's Theorem)**.** *Suppose that $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, Q)$ is a filtered probability space and $(b_t)_{t\in[0,T]}$ is an adapted process on this space such that $\mathbb{E}_Q\left[\int_0^T \|b_s\|^2 \mathrm{d}s\right] < \infty$. Let $(B_t)_{t\geq 0}$ be a Q-Brownian motion and define $\mathcal{L}_t = \int_0^t b_s \mathrm{d}B_s$. Then, $\mathcal{L}$ is a square-integrable Q-martingale. Moreover, if we define*

$$\mathcal{E}(\mathcal{L})_t = \exp\left\{ \int_0^t b_s \mathrm{d}B_s - \frac{1}{2} \int_0^t \|b_s\|^2 \mathrm{d}s \right\}$$

for $t \in [0, T]$ and suppose that $\mathbb{E}_Q\left[\mathcal{E}(\mathcal{L})_T\right] = 1$ then $\mathcal{E}(\mathcal{L})$ is a Q-martingale and so we may define a new measure $P = \mathcal{E}(\mathcal{L})_T Q$. Then, the process

$$\beta_t = B_t - \int_0^t b_s \mathrm{d}s$$

is a P-Brownian motion.

Proof of Proposition 3. We will apply Girsanov's theorem on the interval $[0, t_N]$ in the case where Q is the path measure of the solution to (2), $(B'_t)_{t\geq 0}$ is our Q-Brownian motion, and

$$b_t = \sqrt{2}\left\{s_\theta(Y_{t_k}, T - t_k) - \nabla \log q_{T-t}(Y_t)\right\}$$

for $t \in [t_k, t_{k+1}]$ and each $k = 0, \ldots, N-1$. Note that $(b_t)_{t\in[0,t_N]}$ is an adapted process and we have

$$\mathbb{E}_Q\left[\int_0^{t_N} \|b_s\|^2 \mathrm{d}s\right] = 2\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T-t_k)\|^2\right]\mathrm{d}t < \infty$$

by the assumptions of Proposition 3. Therefore, if we define $\mathcal{L}_t = \int_0^t b_s \mathrm{d}B'_s$ as in Proposition 6 then $(\mathcal{E}(\mathcal{L})_t)_{t\in[0,T]}$ is a continuous local martingale (Le Gall, 2016, Proposition 5.11). So, we can find an increasing sequence of stopping times $(T_n)_{n\geq 1}$ such that $T_n \to t_N$ almost surely and $(\mathcal{E}(\mathcal{L})_{t\wedge T_n})_{t\in[0,t_N]}$ is a continuous martingale for each n.

Let us define $\mathcal{L}^n_t = \int_0^t b_s \mathbb{1}_{[0,T_n]}(s)\mathrm{d}B'_s$ for all $t \in [0, t_N]$ and $n \geq 1$. Then we have $\mathcal{E}(\mathcal{L})_{t\wedge T_n} = \mathcal{E}(\mathcal{L}^n)_t$, so $\mathcal{E}(\mathcal{L}^n)$ is a continuous martingale, and it follows that $\mathbb{E}_Q\left[\mathcal{E}(\mathcal{L}^n)_{t_N}\right] = 1$. Therefore, we may apply Girsanov's theorem (Proposition 6) to $\mathcal{L}^n$ on the interval $[0, t_N]$. We deduce that we may define a new probability measure $P^n := \mathcal{E}(\mathcal{L}^n)_{t_N} Q$ and a new process

$$\beta^n_t = B'_t - \int_0^t b_s \mathbb{1}_{[0,T_n]}(s)\mathrm{d}s$$

such that $(\beta^n_t)_{t\in[0,t_N]}$ is a P^n-Brownian motion.

Since (2) holds almost surely under Q, we see that

$$\mathrm{d}Y_t = \{Y_t + 2s_\theta(Y_{t_k}, T-t_k)\}\mathbb{1}_{[0,T_n]}(t)\mathrm{d}t + \{Y_t + 2\nabla \log q_{T-t}(Y_t)\}\mathbb{1}_{[T_n,t_N]}(t)\mathrm{d}t + \sqrt{2}\mathrm{d}\beta^n_t.$$

In addition,

$$\begin{aligned}
\mathrm{KL}(Q||P^n) &= \mathbb{E}_Q\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P^n}\right] = -\mathbb{E}_Q\left[\log \mathcal{E}(\mathcal{L}^n)_{t_N}\right] \\
&= \mathbb{E}_Q\left[-\mathcal{L}_{T_n} + \frac{1}{2}\int_0^{T_n} \|b_s\|^2 \mathrm{d}s\right] \\
&\leq \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T-t_k)\|^2\right]\mathrm{d}t, \qquad (21)
\end{aligned}$$

since $\mathcal{L}$ is a Q-martingale.

Now, we consider coupling P^n for each n and P^{q_T} by taking a fixed probability space and a single Brownian motion $(W_t)_{t\geq 0}$ on that space and defining the processes $(Y^n_t)_{t\in[0,t_N]}$ and $(Y_t)_{t\in[0,t_N]}$ via

$$\mathrm{d}Y^n_t = \{Y^n_t + 2s_\theta(Y^n_{t_k}, T-t_k)\}\mathbb{1}_{[0,T_n]}(t)\mathrm{d}t + \{Y^n_t + 2\nabla \log q_{T-t}(Y^n_t)\}\mathbb{1}_{[T_n,t_N]}(t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

and

$$\mathrm{d}Y_t = \{Y_t + 2s_\theta(Y_{t_k}, T-t_k)\}\mathrm{d}t + \sqrt{2}\mathrm{d}W_t,$$

and taking $X_0 \sim q_T$ and setting $X^n_0 = X_0$ for each n. Then, the law of Y^n is P^n for each n and the law of Y is P^{q_T}.

Fix $\varepsilon > 0$ and define $\pi_\varepsilon : \mathcal{C}([0,t_N];\mathbb{R}^d) \to \mathcal{C}([0,t_N];\mathbb{R}^d)$ by $\pi_\varepsilon(\omega)(t) = \omega(t \wedge (t_N - \varepsilon))$ for $t \in [0, t_N]$. Then, $\pi_\varepsilon(Y^n) \to \pi_\varepsilon(Y)$ uniformly over $[0, t_N]$ almost surely and hence $(\pi_\varepsilon)_\# P^n \to (\pi_\varepsilon)_\# P^{q_T}$ weakly (Chen et al., 2023d, Lemma 12). We then have that

$$\begin{aligned}
\mathrm{KL}((\pi_\varepsilon)_\# Q||(\pi_\varepsilon)_\# P^{q_T}) &\leq \liminf_{n\to\infty} \mathrm{KL}((\pi_\varepsilon)_\# Q||(\pi_\varepsilon)_\# P^n) \\
&\leq \liminf_{n\to\infty} \mathrm{KL}(Q||P^n) \\
&\leq \sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}} \mathbb{E}_Q\left[\|\nabla \log q_{T-t}(Y_t) - s_\theta(Y_{t_k}, T-t_k)\|^2\right]\mathrm{d}t, \qquad (22)
\end{aligned}$$

where in the first line we have used the lower semicontinuity of the KL divergence (Ambrosio et al., 2005, Lemma 9.4.3), in the second line we have used the data processing inequality, and in the third line we have used (21).

Finally, letting $\varepsilon \to 0$ we have that $\pi_\varepsilon(\omega) \to \omega$ uniformly on $[0, T]$ (Chen et al., 2023d, Lemma 13), and hence $\mathrm{KL}((\pi_\varepsilon)_\# Q || (\pi_\varepsilon)_\# P^{q_T}) \to \mathrm{KL}(Q||P^{q_T})$ (Ambrosio et al., 2005, Corollary 9.4.6). We therefore conclude by taking $\varepsilon \to 0$ in (22). □

G Convergence of Forward Process

We show that the forward OU process converges exponentially in KL divergence under Assumption 2. We note that the exponential convergence of the OU process under various metrics is well-established, and the particular result we prove here is very similar to Lemma 9 in Chen et al. (2023a).

Proof of Proposition 4. Since $q_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0 e^{-t}, \sigma_t^2 I_d)$, we have

$$\mathrm{KL}(q_{t|0}(\,\cdot\,|\mathbf{x}_0)||\pi_d) = \frac{1}{2}\left\{d \log \sigma_t^{-2} - d + d\sigma_t^2 + \|e^{-t}\mathbf{x}_0\|^2\right\}.$$

By the convexity of the KL divergence,

$$\begin{aligned}
\mathrm{KL}(q_T||\pi_d) &= \mathrm{KL}\Big(\int_{\mathbb{R}^d} q_{T|0}(\,\cdot\,|\mathbf{x}_0) p_{\text{data}}(\mathrm{d}\mathbf{x}_0) \Big\| \pi_d\Big) \\
&\le \int_{\mathbb{R}^d} \mathrm{KL}(q_{T|0}(\,\cdot\,|\mathbf{x}_0)||\pi_d) p_{\text{data}}(\mathrm{d}\mathbf{x}_0) \\
&= \frac{1}{2}\left\{d \log \sigma_T^{-2} - d + d\sigma_T^2 + e^{-2T}\mathbb{E}_{p_{\text{data}}}\left[\|X_0\|^2\right]\right\} \\
&= -d\log(1 - e^{-2T}) \\
&\lesssim de^{-2T}
\end{aligned}$$

for $T \ge 1$, where we have used that $\mathbb{E}_{p_{\text{data}}}\left[\|X_0\|^2\right] = d$ since $\mathrm{Cov}(p_{\text{data}}) = I_d$. □

H Linear Dependence on Data Dimension is Optimal

Suppose that we have a data distribution p_* on $\mathbb{R}^d$ such that a diffusion model approximating p_* using the exact scores, initialized from q_T rather than π_d, and using a given sequence of discretization times $t_0, \ldots, t_N$ has a KL error of ε_*^2. Then, if we consider approximating the product measure $p_*^{\otimes m}$ on $(\mathbb{R}^d)^m$, again using the exact score, initializing from $q_T^{\otimes m}$, and using the same sequence of discretization times, the reverse process will factorize across the m copies of $\mathbb{R}^d$. Consequently, the total KL error will be $m\varepsilon_*^2$ by tensorization of the KL divergence. Thus the KL error of the diffusion model scales linearly in the data dimension in this case, demonstrating that the linear dependence of our KL bounds on d is optimal.

Statement of Authorship

Title of paper:	Nearly d-Linear Convergence Bounds for Diffusion Models via Stochastic Localization
Publication status:	Published
Publication details:	Joe Benton, Valentin De Bortoli, Arnaud Doucet, George Deligiannidis. *International Conference on Learning Representations*, 2024.

Student Confirmation

Student name:	Joe Benton
Contribution to the paper:	The idea for the paper came out of discussions with Valentin De Bortoli and George Deligiannidis. All of the theoretical results were provided by me, with some guidance on the presentation from George Deligiannidis, Arnaud Doucet and Valentin De Bortoli.

Signature:  Date: 29/10/2023

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: George Deligiannidis, Associate Professor of Statistics

Supervisor comments: None

Signature:  Date: 31/10/2023

6

Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders

Journal of Machine Learning Research 24 (2023) 1-83 Submitted 10/22; Revised 6/23; Published 8/23

Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics

Kamélia Daudel KAMELIA.DAUDEL@STATS.OX.AC.UK
Joe Benton* BENTON@STATS.OX.AC.UK
Yuyang Shi* YSHI@STATS.OX.AC.UK
Arnaud Doucet DOUCET@STATS.OX.AC.UK
Department of Statistics, University of Oxford
Oxford OX1 3TG, United Kingdom

Editor: Pierre Alquier

Abstract

Several algorithms involving the Variational Rényi (VR) bound have been proposed to minimize an alpha-divergence between a target posterior distribution and a variational distribution. Despite promising empirical results, those algorithms resort to biased stochastic gradient descent procedures and thus lack theoretical guarantees. In this paper, we formalize and study the VR-IWAE bound, a generalization of the importance weighted auto-encoder (IWAE) bound. We show that the VR-IWAE bound enjoys several desirable properties and notably leads to the same stochastic gradient descent procedure as the VR bound in the reparameterized case, but this time by relying on unbiased gradient estimators. We then provide two complementary theoretical analyses of the VR-IWAE bound and thus of the standard IWAE bound. Those analyses shed light on the benefits or lack thereof of these bounds. Lastly, we illustrate our theoretical claims over toy and real-data examples.

Keywords: variational inference, alpha-divergence, importance weighted auto-encoder, weight collapse, high dimension

1. Introduction

Variational inference methods aim at finding the best approximation to a target posterior density within a so-called variational family of probability densities. This best approximation is traditionally obtained by minimizing the exclusive Kullback–Leibler divergence (Wainwright and Jordan, 2008; Blei et al., 2017), however this divergence is known to have some drawbacks (for instance variance underestimation, see Minka, 2005).

As a result, alternative divergences have been explored (Minka, 2005; Li and Turner, 2016; Bui et al., 2016; Dieng et al., 2017; Li and Gal, 2017; Wang et al., 2018; Daudel et al., 2021, 2023; Daudel and Douc, 2021; Rodríguez-Santana and Hernández-Lobato, 2022), in particular the class of alpha-divergences. This family of divergences is indexed by a scalar α. It provides additional flexibility that can in theory be used to overcome the obstacles associated to the exclusive Kullback–Leibler divergence (which is recovered by letting $\alpha \to 1$).

*. Equal contribution

Among those methods, techniques involving the Variational Rényi (VR) bound introduced in Li and Turner (2016) have led to promising empirical results and have been linked to key algorithms such as the importance weighted auto-encoder (IWAE) algorithm (Burda et al., 2016) in the special case $\alpha = 0$ and the black-box alpha (BB-α) algorithm (Hernandez-Lobato et al., 2016).

Yet methods based on the VR bound are seen as lacking theoretical guarantees. This comes from the fact that they are classified as biased in the community: by selecting the VR bound as the objective function, those methods indeed resort to biased gradient estimators (Li and Turner, 2016; Hernandez-Lobato et al., 2016; Bui et al., 2016; Li and Gal, 2017; Geffner and Domke, 2020, 2021; Zhang et al., 2021; Rodríguez-Santana and Hernández-Lobato, 2022).

Geffner and Domke (2020) have recently provided insights from an empirical perspective regarding the magnitude of the bias and its impact on the outcome of the optimization procedure when the (biased) reparameterized gradient estimator of the VR bound is used. They observe that the resulting algorithm appears to require an impractically large amount of computations to actually optimise the VR bound as the dimension increases (and otherwise seems to simply return minimizers of the exclusive Kullback–Leibler divergence). They postulate that this effect might be due to a weight degeneracy behavior (Bengtsson et al., 2008), but this behavior is not quantified precisely from a theoretical point of view.

In this paper, our goal is to (i) develop theoretical guarantees for VR-based variational inference methods and (ii) construct a theoretical framework elucidating the weight degeneracy behavior that has been empirically observed for those techniques. The rest of this paper is organized as follows:

- In Section 2, we provide some background notation and we review the main concepts behind the VR bound.

- In Section 3, we introduce the VR-IWAE bound. We show in Proposition 1 that this bound, previously defined by Li and Turner (2016) as the expectation of the biased Monte Carlo approximation of the VR bound, can be actually interpreted as a variational bound which depends on an hyperparameter α with $\alpha \in [0, 1)$. In addition, we obtain that the VR-IWAE bound leads to the same stochastic gradient descent procedure as the VR bound in the reparameterized case. Unlike the VR bound, the VR-IWAE bound relies on *unbiased* gradient estimators and coincides with the IWAE bound for $\alpha = 0$, fully bridging the gap between both methodologies.

  We then generalize the approach of Rainforth et al. (2018)—which characterizes the signal-to-noise ratio (SNR) of the reparameterized gradient estimators of the IWAE—to the VR-IWAE bound and establish that the VR-IWAE bound with $\alpha \in (0, 1)$ enjoys better theoretical properties than the IWAE bound (Theorem 1). To further tackle potential SNR difficulties, we also extend the doubly-reparameterized gradient estimator of the IWAE (Tucker et al., 2019) to the VR-IWAE bound (Theorem 2).

- In Section 4, we provide a thorough theoretical study of the VR-IWAE bound. Following Domke and Sheldon (2018), we start by investigating the case where the dimension of the latent space d is fixed and the number of Monte Carlo samples N in the VR-IWAE bound goes to infinity (Theorem 3). Our analysis shows that the hyperparameter α

allows us to balance between an error term depending on both the encoder and the decoder parameters (θ, ϕ) and a term going to zero at a $1/N$ rate. This suggests that tuning α can be beneficial to obtain the best empirical performances.

However, the relevance of such analysis can be limited for a high-dimensional latent space d (Examples 1 and 2). We then propose a novel analysis where N does not grow as fast as exponentially with d (Theorems 4 and 5) or sub-exponentially with $d^{1/3}$ (Theorem 6), which we use to revisit Examples 1 and 2 in Examples 3 and 4 respectively. This analysis suggests that in these regimes the VR-IWAE bound, and hence in particular the IWAE bound, are of limited interest.

- In Section 5, we detail how our work relates to the existing litterature.
- Lastly, Section 6 provides empirical evidence illustrating our theoretical claims for both toy and real-data examples.

2. Background

Given a model with joint distribution $p_\theta(x, z)$ parameterized by θ, where x denotes an observation and z is a latent variable valued in $\mathbb{R}^d$, one is interested in finding the parameter θ which best describes the observations $\mathcal{D} = \{x_1, \dots, x_T\}$. This will be our running example. The corresponding posterior density satisfies:

$$p_\theta(\boldsymbol{z}|\mathcal{D}) \propto \prod_{i=1}^{T} p_\theta(x_i, z_i) \tag{1}$$

with $\boldsymbol{z} = (z_1, \dots, z_T)$, so that the marginal log likelihood reads

$$\ell(\theta; \mathcal{D}) = \sum_{i=1}^{T} \ell(\theta; x_i) \quad \text{with} \quad \ell(\theta; x) := \log p_\theta(x) = \log\left(\int p_\theta(x, z)\mathrm{d}z\right). \tag{2}$$

Unfortunately as this marginal log likelihood is typically intractable, finding θ maximizing it is difficult. Variational bounds are then designed to act as surrogate objective functions more amenable to optimization.

Let $q_\phi(z|x)$ be a variational encoder parameterized by ϕ, common variational bounds are the Evidence Lower BOund (ELBO) and the IWAE bound (Burda et al., 2016):

$$\mathrm{ELBO}(\theta, \phi; x) = \int q_\phi(z|x) \log w_{\theta,\phi}(z; x)\ \mathrm{d}z,$$
$$\ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) = \int\int \prod_{i=1}^{N} q_\phi(z_i|x) \log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)\right) \mathrm{d}z_{1:N}, \quad N \in \mathbb{N}^\star$$

where for all $z \in \mathbb{R}^d$,

$$w_{\theta,\phi}(z; x) = \frac{p_\theta(x, z)}{q_\phi(z|x)}.$$

The IWAE bound generalizes the ELBO (which is recovered for $N = 1$) and acts as a lower bound on $\ell(\theta; x)$ that can be estimated in an unbiased manner. Instead of maximizing $\ell(\theta; \mathcal{D})$ defined in (2), one then considers the surrogate objective

$$\sum_{i=1}^{T} \ell_N^{(\text{IWAE})}(\theta, \phi; x_i)$$

which is optimized by performing stochastic gradient descent steps w.r.t. (θ, ϕ) on it combined to mini-batching. Optimizing this objective w.r.t. ϕ is difficult due to high-variance gradients with low signal-to-noise ratio (Rainforth et al., 2018). To mitigate this problem, reparameterized (Kingma and Welling, 2014; Burda et al., 2016) and doubly-reparameterized gradient estimators (Tucker et al., 2019) have been proposed.

Crucially, stochastic gradient schemes on the IWAE bound (and hence on the ELBO) only resort to unbiased estimators in both the reparameterized (Kingma and Welling, 2014; Burda et al., 2016) and the doubly-reparameterized (Tucker et al., 2019) cases, providing theoretical justifications behind those approaches. In particular, under the assumption that z can be reparameterized (that is $z = f(\varepsilon, \phi; x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$) and under common differentiability assumptions, the reparameterized gradient w.r.t. ϕ of the IWAE bound is given by

$$\frac{\partial}{\partial \phi} \ell_N^{(\text{IWAE})}(\theta, \phi; x) = \int \cdots \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \right) \mathrm{d}\varepsilon_{1:N}$$

and the doubly-reparameterized one by

$$\begin{aligned} &\frac{\partial}{\partial \phi} \ell_N^{(\text{IWAE})}(\theta, \phi; x) \\ &= \int \cdots \int \prod_{i=1}^{N} q(\varepsilon_i) \left(\sum_{j=1}^{N} \left(\frac{w_{\theta,\phi}(z_j; x)}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)} \right)^2 \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi; x); x)|_{\phi'=\phi} \right) \mathrm{d}\varepsilon_{1:N}. \end{aligned} \tag{3}$$

Unbiased Monte Carlo estimators of both gradients are hence respectively given by

$$\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j; x)}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi; x); x) \tag{4}$$

and

$$\sum_{j=1}^{N} \left(\frac{w_{\theta,\phi}(z_j; x)}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k; x)} \right)^2 \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi; x); x)|_{\phi'=\phi},$$

with $\varepsilon_1, \ldots, \varepsilon_N$ being i.i.d. samples generated from q and $z_j = f(\varepsilon_j, \phi; x)$ for all $j = 1 \ldots N$. Maddison et al. (2017) and Domke and Sheldon (2018) in particular established that the variational gap - that is the difference between the IWAE bound and the marginal log-likelihood - goes to zero at a fast $1/N$ rate when the dimension of the latent space d is fixed and the number of samples N goes to infinity.

Another example of variational bound is the Variational Rényi (VR) bound introduced by Li and Turner (2016): it is defined for all $\alpha \in \mathbb{R} \setminus \{1\}$ by

$$\mathcal{L}^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \log \left(\int q_\phi(z|x) \; w_{\theta,\phi}(z;x)^{1-\alpha} \; \mathrm{d}z \right) \tag{5}$$

and it generalizes the ELBO (which corresponds to the extension by continuity of the VR bound to the case $\alpha = 1$, see Li and Turner, 2016, Theorem 1). It is also a lower (resp. upper) bound on the marginal log-likelihood $\ell(\theta; x)$ for all $\alpha > 0$ (resp. $\alpha < 0$).

In the spirit of the IWAE bound optimisation framework, the VR bound is used for variational inference purposes in (Li and Turner, 2016, Section 4.1, 4.2 and 5.2) to optimise the marginal log-likelihood $\ell(\theta, \mathcal{D})$ defined in (2) by considering the global objective function

$$\sum_{i=1}^{T} \mathcal{L}^{(\alpha)}(\theta, \phi; x_i)$$

and by performing stochastic gradient descent steps w.r.t. (θ, ϕ) on it paired up with mini-batching and reparameterization. This VR bound methodology has provided positive empirical results compared to the usual case $\alpha = 1$ and has been widely adopted in the literature (Li and Turner, 2016; Bui et al., 2016; Hernandez-Lobato et al., 2016; Li and Gal, 2017; Zhang et al., 2021; Rodríguez-Santana and Hernández-Lobato, 2022). As discussed in the remark below, this methodology is obviously not limited to the choice of posterior density defined in (1) and is more broadly applicable.

Remark 1 (Black-box alpha energy function) *Let $p_0(z)$ be a prior on a latent variable z valued in $\mathbb{R}^d$ and by $p(x|z)$ the likelihood of the observation x given z, we might consider the posterior density*

$$p(z|\mathcal{D}) \propto p_0(z) \prod_{i=1}^{T} p(x_i|z), \tag{6}$$

leading to the marginal log-likelihood

$$\tilde{\ell}(\mathcal{D}) = \log \left(\int p(\mathcal{D}, z) \mathrm{d}z \right) = \log \left(p_0(z) \prod_{i=1}^{T} p(x_i|z) \mathrm{d}z \right).$$

Here, the latent variable z valued in $\mathbb{R}^d$ is shared across all the observations. Now further assume that the prior density $p_0(z) = \exp(s(z)^T \phi_0 - \log Z(\phi_0))$ has an exponential form, with ϕ_0 and s being the natural parameters and the sufficient statistics respectively and $Z(\phi_0)$ being the normalizing constant ensuring that p_0 is a probability density function.

In order to find the best approximation to the posterior density (6)*, Hernandez-Lobato et al. (2016) offers to minimize the black-box alpha (BB-α) energy function, which is defined by: for all $\alpha \in \mathbb{R} \setminus \{1\}$,*

$$\mathcal{E}(\phi) = \log Z(\phi_0) - \log Z(\tilde{\phi}) - \frac{1}{1-\alpha} \sum_{i=1}^{T} \log \left(\int q_\phi(z) \left(\frac{p(x_i|z)}{f_\phi(z)} \right)^{1-\alpha} \mathrm{d}z \right)$$

where $f_\phi(z) = \exp(s(z)^T\phi)$ *is within the same exponential family as the prior and* $q_\phi(z) = \exp(s(z)^T\tilde{\phi} - \log Z(\tilde{\phi}))$ *with* $\tilde{\phi} = T\phi + \phi_0$ *denoting the natural parameters of* q_ϕ *and* $Z(\tilde{\phi})$ *its normalizing constant. Here, the minimisation is carried out via stochastic gradient descent w.r.t.* ϕ *combined with mini-batching and reparameterization.*

As observed in Li and Gal (2017), minimizing $\mathcal{E}(\phi)$ *w.r.t.* ϕ *is equivalent to maximizing the sum of VR bounds*

$$\sum_{i=1}^{T} \frac{T}{1-\alpha} \log\left(\int q_\phi(z) w_{\theta,\phi}(z;x)^{\frac{1-\alpha}{T}} \mathrm{d}z\right)$$

w.r.t. ϕ*, where this time* $w_{\theta,\phi}(z;x) = p(x_i|z)^T p_0(z)/q_\phi(z)$*.*

However, the stochastic gradient descent scheme originating from having selected the VR bound as the objective function suffers from one important shortcoming: it relies on biased gradient estimators for all $\alpha \notin \{0,1\}$, meaning that there exists no convergence guarantees for the whole scheme. Indeed, Li and Turner (2016) show that the gradient of the VR bound w.r.t. ϕ satisfies

$$\frac{\partial}{\partial\phi}\mathcal{L}^{(\alpha)}(\theta,\phi;x) = \frac{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon,\phi;x);x)\ \mathrm{d}\varepsilon}{\int q(\varepsilon)\ w_{\theta,\phi}(z;x)^{1-\alpha}\ \mathrm{d}\varepsilon},$$

with $z = f(\varepsilon,\phi;x) \sim q_\phi(\cdot|x)$ where $\varepsilon \sim q$. The gradient above being intractable, they approximate it using

$$\sum_{j=1}^{N} \frac{w_{\theta,\phi}(z_j;x)^{1-\alpha}}{\sum_{k=1}^{N} w_{\theta,\phi}(z_k;x)^{1-\alpha}} \frac{\partial}{\partial\phi}\log w_{\theta,\phi}(f(\varepsilon_j,\phi;x);x), \tag{7}$$

where $\varepsilon_1,\ldots,\varepsilon_N$ are i.i.d. samples generated from q and $z_j = f(\varepsilon_j,\phi;x)$ for all $j = 1\ldots N$. The cases $\alpha = 0$ and $\alpha = 1$ recover the stochastic reparameterized gradients of the IWAE bound (4) and of the ELBO (consider (4) with $N = 1$). As a result, we can trace them back to unbiased stochastic gradient descent schemes for IWAE bound and ELBO optimisation respectively. Yet, this is no longer the case when $\alpha \notin \{0,1\}$, hence impeding the theoretical guarantees of the scheme.

In addition, due to the log function, the VR bound itself can only be approximated using biased Monte Carlo estimators, with (Li and Turner, 2016, Section 4.1) using

$$\frac{1}{1-\alpha}\log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(Z_j;x)^{1-\alpha}\right) \tag{8}$$

where $Z_1,\ldots,Z_N$ are i.i.d. samples generated from q_ϕ. Furthermore, while the VR bound and the IWAE bound approaches are linked via the gradient estimator (7), the VR bound does not recover the IWAE bound when $\alpha = 0$.

The next section aims at overcoming the theoretical difficulties regarding the VR bound mentioned above.

3. The VR-IWAE Bound

For all $\alpha \in \mathbb{R} \setminus \{1\}$, let us introduce the quantity

$$\ell_N^{(\alpha)}(\theta, \phi; x) := \frac{1}{1-\alpha} \int \int \prod_{i=1}^{N} q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(z_j; x)^{1-\alpha} \right) \mathrm{d}z_{1:N}, \tag{9}$$

which we will refer to as the *VR-IWAE bound*. Note that for the VR-IWAE bound to be well-defined we will assume that the following assumption holds in the rest of the paper.

(A1) It holds that $0 < p_\theta(x) < \infty$ and the support of $q_\phi(\cdot|x)$ and of $p_\theta(\cdot|x)$ are equal.

We may omit the dependency on x in $z \mapsto q_\phi(z|x)$ and $z \mapsto w_{\theta,\phi}(z; x)$ for notational convenience and we now make two remarks regarding the VR-IWAE bound defined in (9).

- Contrary to the VR bound, VR-IWAE bound (i) can be approximated using an unbiased Monte Carlo estimator and (ii) recovers the IWAE bound by setting $\alpha = 0$. Under common differentiability assumptions, we also have that

  $$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = \mathrm{ELBO}(\theta, \phi; x)$$

  (see Appendix A.1 for details), meaning that the VR-IWAE bound interpolates between the IWAE bound and the ELBO.

- Li and Turner (2016) interpreted the quantity defined in (9) as the expectation of the biased Monte Carlo approximation of the VR bound (8). They established in (Li and Turner, 2016, Theorem 2) some properties on this quantity. In particular, they showed that (i) for all $\alpha \leq 1$ and all $N \in \mathbb{N}^\star$,

  $$\ell_N^{(\alpha)}(\theta, \phi; x) \leq \ell_{N+1}^{(\alpha)}(\theta, \phi; x) \leq \mathcal{L}^{(\alpha)}(\theta, \phi; x)$$

  and (ii) for all $\alpha \in \mathbb{R}$, $\ell_N^{(\alpha)}(\theta, \phi; x)$ approaches the VR bound $\mathcal{L}^{(\alpha)}(\theta, \phi; x)$ as N goes to infinity if the function $z \mapsto w_{\theta,\phi}(z)$ is assumed to be bounded.

Based on the two previous remarks, $\ell_N^{(\alpha)}(\theta, \phi; x)$ seems to be an interesting candidate as a variational bound which generalizes the IWAE bound. We take here another perspective on the quantity $\ell_N^{(\alpha)}(\theta, \phi; x)$ by wanting to frame it as a variational bound with ties to the Rényi's α-divergence variational inference methodology of Li and Turner (2016) and to the IWAE bound, hence the name VR-IWAE bound. We now need to check that the VR-IWAE bound can indeed be used as a variational bound for marginal log-likelihood optimisation in the context of our running example.

3.1 The VR-IWAE Bound as a Variational Bound

As underlined in the following proposition, the VR-IWAE bound is a variational bound for all $\alpha \in [0, 1)$ which enjoys properties akin to those obtained for the IWAE bound (Burda et al., 2016) and which becomes looser as α increases.

Proposition 1 (Properties of the VR-IWAE bound) *The following properties hold for the VR-IWAE bound.*

1. *For all $\alpha \in [0, 1)$ and all $N \in \mathbb{N}^\star$,*

$$\text{ELBO}(\theta, \phi; x) \leq \ell_N^{(\alpha)}(\theta, \phi; x) \leq \ell_{N+1}^{(\alpha)}(\theta, \phi; x) \leq \mathcal{L}^{(\alpha)}(\theta, \phi; x) \leq \ell(\theta; x). \tag{10}$$

   Moreover, if the function $z \mapsto w_{\theta,\phi}(z)$ is bounded, then $\ell_N^{(\alpha)}(\theta, \phi; x)$ approaches the VR bound $\mathcal{L}^{(\alpha)}(\theta, \phi; x)$ as N goes to infinity.

2. *For all $\alpha_1, \alpha_2 \in (0, 1)$ such that $\alpha_1 > \alpha_2$ and all $N \in \mathbb{N}^\star$,*

$$\ell_N^{(\alpha_1)}(\theta, \phi; x) \leq \ell_N^{(\alpha_2)}(\theta, \phi; x) \leq \ell_N^{(\text{IWAE})}(\theta, \phi; x), \tag{11}$$

   where the case of equality is reached if and only if $z \mapsto w_{\theta,\phi}(z)$ is constant for ν-almost all $z \in \mathbb{R}^d$ (with ν denoting the Lebesgue measure).

3. *Further assuming that z can be reparameterized, that is $z = f(\varepsilon, \phi) \sim q_\phi$ where $\varepsilon \sim q$, we have under common differentiability assumptions that*

$$\frac{\partial}{\partial \phi} \ell_N^{(\alpha)}(\theta, \phi; x) = \int \int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi)) \right) \mathrm{d}\varepsilon_{1:N} \tag{12}$$

   and an unbiased estimator of $\partial \ell_N^{(\alpha)}(\theta, \phi; x)/\partial \phi$ is given by

$$\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi)) \tag{13}$$

   where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. samples generated from q and $z_j = f(\varepsilon_j, \phi)$ for all $j = 1 \ldots N$.

The proof of Proposition 1 is deferred to Appendix A.2 and we now comment on Proposition 1. Observe that both the VR and VR-IWAE bounds share the same estimated reparameterized gradient w.r.t. ϕ, that is (7) is exactly (13), hence they lead to the same stochastic gradient descent algorithm. However, when $\alpha \in (0, 1)$, a key difference is that in the VR bound case this estimator is biased while it is unbiased for VR-IWAE bound.

This motivates the VR-IWAE bound as a generalization of the IWAE bound that overcomes the theoretical difficulties of the VR bound, as unbiased gradient estimates provide the convergence of the stochastic gradient descent procedure (under proper conditions on the learning rate). In fact, and as we shall see next, the estimated reparameterized gradient w.r.t. ϕ written in (7) and (13) - that we have now properly justified using the VR-IWAE bound - also enjoys an advantageous signal-to-noise ratio behavior when $\alpha \in (0, 1)$.

3.2 Signal-to-noise Ratio (SNR) Analysis

Rainforth et al. (2018) identified some issues associated to using reparameterized gradient estimators of the IWAE bound. They did so by looking at the signal-to-noise ratio (SNR) of those estimates: their main theorem (Rainforth et al., 2018, Theorem 1) shows that while increasing N leads to a tighter IWAE bound and improves the SNR for learning θ, it actually worsens the SNR for learning ϕ.

Let us now investigate if and how the conclusions of (Rainforth et al., 2018, Theorem 1) extend to the VR-IWAE bound. To this end, we first recall the definition of the SNR used in Rainforth et al. (2018). Given a random vector $X = (X_1, \dots, X_L)$ of dimension $L \in \mathbb{N}^\star$, the SNR may be defined as follows:

$$\mathrm{SNR}[X] = \left(\frac{|\mathbb{E}(X_1)|}{\sqrt{\mathbb{V}(X_1)}}, \dots, \frac{|\mathbb{E}(X_L)|}{\sqrt{\mathbb{V}(X_L)}} \right).$$

Writing $\theta = (\theta_1, \dots, \theta_L)$ and $\phi = (\phi_1, \dots, \phi_{L'})$ and with $L, L' \in \mathbb{N}^\star$, we now consider for all $\ell = 1 \dots L$ and all $\ell' = 1 \dots L'$ the unbiased estimates of the reparameterized gradient of the VR-IWAE bound w.r.t. θ_ℓ and w.r.t. $\phi_{\ell'}$ given by: for all $M, N \in \mathbb{N}^\star$ and all $\alpha \in [0, 1)$,

$$\delta^{(\alpha)}_{M,N}(\theta_\ell) = \frac{1}{(1-\alpha)M} \sum_{m=1}^{M} \frac{\partial}{\partial \theta_\ell} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(f(\varepsilon_{m,j}, \phi))^{1-\alpha} \right), \tag{14}$$

$$\delta^{(\alpha)}_{M,N}(\phi_{\ell'}) = \frac{1}{(1-\alpha)M} \sum_{m=1}^{M} \frac{\partial}{\partial \phi_{\ell'}} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(f(\varepsilon_{m,j}, \phi))^{1-\alpha} \right), \tag{15}$$

where $(\varepsilon_{m,j})_{1 \leq m \leq M, 1 \leq j \leq N}$ are i.i.d. samples generated from q and $z_{m,j} = f(\varepsilon_{m,j}, \phi)$ for all $m = 1 \dots M$ and all $n = 1 \dots N$. Note that the link with the reparameterized gradient estimator (13) from Proposition 1 can be made by considering the case $M = 1$ in (15). We then have the following theorem.

Theorem 1 (SNR analysis) *Let $\alpha \in [0, 1)$ and for all $N \in \mathbb{N}^\star$ and all $j = 1 \dots N$, define $\tilde{w}_{1,j} = w_{\theta,\phi}(f(\varepsilon_{1,j}, \phi))$ and $\hat{Z}_{1,N,\alpha} = N^{-1} \sum_{j=1}^{N} \tilde{w}_{1,j}^{1-\alpha}$. Assume that the eighth moments of $\tilde{w}_{1,1}^{1-\alpha}$, $\partial \tilde{w}_{1,1}^{1-\alpha} / \partial \theta_\ell$ and $\partial \tilde{w}_{1,1}^{1-\alpha} / \partial \phi_{\ell'}$ are finite, where ℓ is an integer between 1 and L and ℓ' is an integer between 1 and L'. Furthermore, assume that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/\hat{Z}_{1,N,\alpha})^4) < \infty$. Lastly, assume that $\partial \mathbb{E}(\tilde{w}_{1,1}^{1-\alpha}) / \partial \theta_\ell \neq 0$ and that*

$$\begin{aligned} &\partial \mathbb{V}(\tilde{w}_{1,1}^{1-\alpha}) / \partial \phi_{\ell'} > 0, \quad \text{if } \alpha = 0 \\ &\partial \mathbb{E}(\tilde{w}_{1,1}^{1-\alpha}) / \partial \phi_{\ell'} \neq 0, \quad \text{if } \alpha \in (0, 1). \end{aligned} \tag{16}$$

Then, under common differentiability assumptions, the SNR of the VR-IWAE bound reparameterized gradient estimates w.r.t θ_ℓ and w.r.t $\phi_{\ell'}$ defined in (14) *and* (15) *respectively satisfy*

$$\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)] = \Theta(\sqrt{MN}) \tag{17}$$

$$\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\phi_{\ell'})] = \begin{cases} \Theta(\sqrt{M/N}) & \text{if } \alpha = 0, \\ \Theta(\sqrt{MN}) & \text{if } \alpha \in (0, 1). \end{cases} \tag{18}$$

The proof of Theorem 1 can be found in Appendix A.3. Theorem 1 states that for $\alpha \in (0,1)$, the SNR for learning the generative network (θ) and for learning the inference network (ϕ) both improve as N increases, unlike the IWAE bound case $\alpha = 0$ where the second SNR worsens as N increases. This provides theoretical support suggesting that taking $\alpha > 0$ in the VR-IWAE bound may help to ensure a good training signal, thus leading to improved empirical performances compared to the IWAE bound.

In the following, we investigate another way to provide gradient estimators of the VR-IWAE bound with an advantageous SNR behavior in practice.

3.3 Doubly-reparameterized Gradient for the VR-IWAE Bound

To remedy the SNR issue identified in Rainforth et al. (2018), Tucker et al. (2019) proposed a new estimator of the gradient of the IWAE bound (3) under the name doubly-reparameterized gradient estimator. As written in the theorem below, the doubly-reparameterized gradient estimator of the IWAE bound (3) in fact generalizes to the case $\alpha \in (0,1)$.

Theorem 2 (Generalized doubly-reparameterized gradient) *Under common differentiability assumptions and assuming that z can be reparameterized, that is $z = f(\varepsilon, \phi) \sim q_\phi$ where $\varepsilon \sim q$, we have that: for all $\alpha \in [0,1]$,*

$$\frac{\partial}{\partial \phi} \ell_N^{(\alpha)}(\theta, \phi; x) = \int \int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N h_j(\alpha) \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi))|_{\phi'=\phi} \right) \mathrm{d}\varepsilon_{1:N}, \tag{19}$$

with $z_j = f(\varepsilon_j, \phi)$ for all $j = 1 \dots N$ and

$$h_j(\alpha) = \alpha \, \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} + (1-\alpha) \, \left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}} \right)^2.$$

An unbiased estimator of $\partial \ell_N^{(\alpha)}(\theta, \phi; x)/\partial \phi$ is then given by

$$\sum_{j=1}^N h_j(\alpha) \frac{\partial}{\partial \phi} \log w_{\theta,\phi'}(f(\varepsilon_j, \phi))|_{\phi'=\phi} \tag{20}$$

where $\varepsilon_1, \dots, \varepsilon_N$ are i.i.d. samples generated from q and $z_j = f(\varepsilon_j, \phi)$ for all $j = 1 \dots N$.

The proof of Theorem 2 is deferred to Appendix A.4. One can then check that we recover the usual doubly-reparameterized gradient estimator of the IWAE bound (resp. the ELBO) when $\alpha = 0$ (resp. $\alpha = 1$). Like the reparameterized gradient estimator (13), which we have studied in Section 3.2 as it corresponds to the special case $M = 1$ in Theorem 1, this second (doubly-reparameterized) gradient estimator too may lead to improved empirical performances. From there, large-scale learning occurs by using that Proposition 1 implies

$$\ell(\theta; \mathcal{D}) = \sum_{i=1}^T \ell(\theta; x_i) \geq \sum_{i=1}^T \mathcal{L}^{(\alpha)}(\theta, \phi; x_i) \geq \sum_{i=1}^T \ell_N^{(\alpha)}(\theta, \phi; x_i)$$

and by following the training procedure for the IWAE bound. Indeed, we have access to an unbiased estimator of the lower bound of the full data set $\sum_{i=1}^T \ell_N^{(\alpha)}(\theta, \phi; x_i)$ (as well

as an unbiased estimator of its reparameterized/doubly-reparameterized gradient) using mini-batching. Seeking to maximize the objective function $\sum_{i=1}^{T} \mathcal{L}^{(\alpha)}(\theta, \phi; x_i)$ by optimising $\sum_{i=1}^{T} \ell_N^{(\alpha)}(\theta, \phi; x_i)$ in fact amounts to seeking to minimize a specific Rényi's α-divergence with a mean-field assumption on the variational approximation (see Remark 2 for detail).

Remark 2 *Define $\boldsymbol{z} = (z_1, \ldots, z_T)$ and $q_\phi(\boldsymbol{z}) = \prod_{i=1}^{T} q_\phi(z_i)$. Then, for all $\alpha \in (0, 1)$:*

$$\begin{aligned}
\sum_{i=1}^{T} \mathcal{L}^{(\alpha)}(\theta, \phi; x_i) &= \sum_{i=1}^{T} \frac{1}{1-\alpha} \log \left(\int q_\phi(z) w_{\theta,\phi}(z; x_i)^{1-\alpha} \mathrm{d}z \right) \\
&= \sum_{i=1}^{T} \frac{1}{1-\alpha} \log \left(\int q_\phi(z_i) w_{\theta,\phi}(z_i; x_i)^{1-\alpha} \mathrm{d}z_i \right) \\
&= \frac{1}{1-\alpha} \log \left(\int \int \prod_{i=1}^{T} q_\phi(z_i) \prod_{j=1}^{T} w_{\theta,\phi}(z_j; x_j)^{1-\alpha} \mathrm{d}z_{1:T} \right) \\
&= \frac{1}{1-\alpha} \log \left(\int q_\phi(\boldsymbol{z}) w_{\theta,\phi}(\boldsymbol{z}; \mathcal{D})^{1-\alpha} \mathrm{d}\boldsymbol{z} \right)
\end{aligned}$$

where $p_\theta(\mathcal{D}, \boldsymbol{z}) = \prod_{i=1}^{T} p(x_i, z_i)$ and $w_{\theta,\phi}(\boldsymbol{z}; \mathcal{D}) = p_\theta(\mathcal{D}, \boldsymbol{z}) / q_\phi(\boldsymbol{z})$. Observe that the last equality is a VR bound, meaning that maximizing the global objective function $\sum_{i=1}^{T} \mathcal{L}^{(\alpha)}(\theta, \phi; x_i)$ is equivalent to minimizing the Rényi's α-divergence between the two probability distributions with associated probability densities $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z}|\mathcal{D})$ respectively w.r.t. the Lebesgue measure. Hence, this approach belongs to alpha-divergence variational inference methods with the particularity that it makes a mean-field assumption on the variational approximation $q_\phi(\boldsymbol{z})$.

At this stage, we have formalized and motivated the VR-IWAE bound. We now want to get an understanding of its theoretical properties.

4. Theoretical Study of the VR-IWAE Bound

The starting point of our approach is to exploit the fact that prior theoretical works study the particular case $\alpha = 0$ (corresponding to the IWAE bound) when the dimension of the latent space $\dim(z) = d$ is fixed and the number of samples N goes to infinity.

4.1 Behavior of the VR-IWAE Bound when d is Fixed and N goes to Infinity

A quantity that has been of interest to assess the quality of the IWAE bound is the *variational gap*, which is defined as the difference between the IWAE bound and the marginal log-likelihood:

$$\Delta_N(\theta, \phi; x) := \ell_N^{(\mathrm{IWAE})}(\theta, \phi; x) - \ell(\theta; x) = \int \int \prod_{i=1}^{N} q_\phi(z_i) \log \left(\frac{1}{N} \sum_{j=1}^{N} \overline{w}_{\theta,\phi}(z_j) \right) \mathrm{d}z_{1:N} \quad (21)$$

where for all $z \in \mathbb{R}^d$

$$\overline{w}_{\theta,\phi}(z) := \frac{w_{\theta,\phi}(z)}{\mathbb{E}_{Z \sim q_\phi}\left(w_{\theta,\phi}(Z)\right)} = \frac{w_{\theta,\phi}(z)}{p_\theta(x)},$$

so that $\overline{w}_{\theta,\phi}(z_1), \dots, \overline{w}_{\theta,\phi}(z_N)$ correspond to the relative weights. The analysis of the variational gap (21), first performed in Maddison et al. (2017) and then refined in Domke and Sheldon (2018), investigated the case where $\dim(z) = d$ is fixed and N goes to infinity. Informally, they obtained in their Theorem 3 that the variational gap behaves as follows

$$\Delta_N(\theta, \phi; x) = -\frac{\gamma_0^2}{2N} + o\left(\frac{1}{N}\right)$$

with γ_0 denoting the variance of the relative weights, that is

$$\gamma_0^2 := \mathbb{V}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}(Z)).$$

This result suggests that using N is very beneficial to reduce the variational gap, as it goes to zero at a fast $1/N$ rate. It motivates a study - in a regime where d is fixed and N goes to infinity - of the more general variational gap defined for all $\alpha \in [0, 1)$ by

$$\Delta_N^{(\alpha)}(\theta, \phi; x) := \ell_N^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x).$$

The following result generalizes (Domke and Sheldon, 2018, Theorem 3) to the VR-IWAE bound.

Theorem 3 *Let $\alpha \in [0, 1)$. Then, it holds that*

$$0 < \mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha}) < \infty. \tag{22}$$

Further assume that there exists $\beta > 0$ such that

$$\mathbb{E}_{Z \sim q_\phi}(|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1|^{2+\beta}) < \infty, \tag{23}$$

where we have defined $\overline{w}_{\theta,\phi}^{(\alpha)}(z) = w_{\theta,\phi}(z)^{1-\alpha} / \mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha})$ for all $z \in \mathbb{R}^d$. Lastly, assume that the following condition holds

$$\limsup_{N \to \infty} \mathbb{E}(1/R_{\alpha,N}) < \infty, \tag{24}$$

where, for all $N \in \mathbb{N}^\star$, $R_{\alpha,N} = N^{-1} \sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha}$ and $Z_1, \dots, Z_N$ are i.i.d. samples generated according to q_ϕ. Then, denoting $\gamma_\alpha^2 = (1-\alpha)^{-1} \mathbb{V}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$, we have:

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right). \tag{25}$$

The proof of this result is deferred to Appendix B.1 and we now aim at interpreting Theorem 3, starting with the conditions (23) and (24).

4.1.1 Conditions (23) and (24)

A first remark is that the conditions (23) and (24) stated in Theorem 3 exactly generalize the ones from (Domke and Sheldon, 2018, Theorem 3), which are recovered by setting $\alpha = 0$. This then prompt us to investigate in the following proposition how restrictive the conditions (23) and (24) are as a function of α.

Proposition 2 *Let $\alpha_1, \alpha_2 \in [0, 1)$ with $\alpha_1 > \alpha_2$. Then, the two following assertions hold.*

1. *If (23) holds with $\alpha = \alpha_2$, then (23) holds with $\alpha = \alpha_1$.*
2. *If (24) holds with $\alpha = \alpha_2$, then (24) holds with $\alpha = \alpha_1$.*

The proof of this result is deferred to Appendix B.2. It notably relies the fact that the condition (24) is equivalent to the statement that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}(1/R_{\alpha,N}) < \infty$, which follows from Lemma 4 in Appendix A.3 with $k = 1$. Notice that this provides an interesting equivalent condition to (24) that might be easier to check empirically.

Proposition 2 then states that the conditions (23) and (24) with $\alpha = \alpha_1$ are at worse as restrictive as the case $\alpha = \alpha_2$, where $\alpha_1 > \alpha_2$. Putting this into perspective with Domke and Sheldon (2018), the conditions (23) and (24) when $\alpha > 0$ are hence not more restrictice than the conditions presented in (Domke and Sheldon, 2018, Theorem 3) for the more usual IWAE bound case $\alpha = 0$. In fact, one would even be inclined to think that those conditions become easier to satisfy as α increases, motivating once again the use of $\alpha \in (0, 1)$ in practice to be in the conditions of application of Theorem 3.

4.1.2 Interpreting (25)

Under the assumptions of Theorem 3, (25) states: for all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right).$$

The variational gap $\Delta_N^{(\alpha)}(\theta, \phi; x)$ is hence composed of two main terms:

- A term going to zero at a $1/N$ rate that depends on γ_α^2. Here γ_α^2 is controlled thanks to (23), as (23) implies that $\mathbb{V}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)) < \infty$ or equivalently that $\gamma_\alpha^2 < \infty$.

- An error term $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$. This term decreases away from zero as α increases due to the fact that $\mathcal{L}^{(\alpha)}(\theta, \phi; x)$ decreases away from its upper bound $\ell(\theta; x)$ as α increases (see for example Li and Turner, 2016, Theorem 1). It is equal to zero when $\alpha = 0$ or when the posterior and the encoder distributions are equal to one another.

  Unless $\alpha = 0$ or the posterior and encoder distributions are matching, the error term $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ hence maintains a dependency in (θ, ϕ) in the variational gap even as N goes to infinity. This is coherent with Theorem 1, in the sense that the case $\alpha \in (0, 1)$ might ensure a better learning of both θ and ϕ in practice compared to the case $\alpha = 0$ (as the latter does not keep a dependency in ϕ as N goes to infinity).

Since the error term $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ is decreasing away from zero as α increases and the term going to zero at a $1/N$ rate depends on the behavior of γ_α^2 (with γ_α^2 going to 0 as α goes to 1, see Lemma 5 of Appendix B.3), there might then be a tradeoff to achieve when choosing α in order to obtain the best empirical performances.

To the best of our knowledge, and by appealing to the link between the VR-bound and the VR-IWAE bound methodologies established in Section 3.1, Theorem 3 is the first result shedding light via (25) on how the quantity γ_α^2 alongside with the error term

$\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ may play a role to guarantee the success of gradient-based methods involving the VR-bound. While the result obtained in Theorem 3 is encouraging and might further motivate the use of $\alpha \in (0, 1)$ in practice, one may seek to identify potential limitations of Theorem 3.

4.1.3 Limitations of Theorem 3

To investigate the limitations of Theorem 3, let us provide below two insightful examples in which all the terms appearing in (25) are tractable.

Example 1 *Let $\sigma > 0$, $S_1, \dots, S_N$ be i.i.d. normal random variables and assume that the distribution of the relative weights $\overline{w}_{\theta,\phi}(z_1), \dots, \overline{w}_{\theta,\phi}(z_N)$ is log-normal of the form*

$$\log \overline{w}_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d} S_i, \quad i = 1 \dots N, \tag{26}$$

where the relationship between mean and variance ensures that the relative weights have expectation 1. Then, we can apply Theorem 3: for all $\alpha \in [0, 1)$,

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right)$$

with

$$\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) = -\frac{\alpha\sigma^2 d}{2} \quad \text{and} \quad \gamma_\alpha^2 = \frac{\exp\left[(1-\alpha)^2\sigma^2 d\right] - 1}{1 - \alpha}.$$

In particular, we can write the weights under the form (26) with $\sigma = 1$ by setting $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$, $q_\phi(z|x) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, $\theta = 0 \cdot \boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, where $\boldsymbol{I}_d$ is the d-dimensional identity matrix and $\boldsymbol{u}_d$ the d-dimensional vector whose coordinates are all equal to 1.

The proof of Example 1 is deferred to Appendix B.4 and we now comment on Example 1. A first comment is that as α increases, the error term $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ worsens linearly with α while γ_α^2 decreases with α, which supports our claim that there might exist an optimal α that balances between the two terms appearing in the variational gap as a rule of thumb.

Furthermore, the variance of the relative weights and more generally γ_α^2 is exponential with d. This means that the analysis of Domke and Sheldon (2018)—that we extended to $\alpha \in [0, 1)$ in Theorem 3—may not capture what is happening in some high-dimensional scenarios as we may never use N large enough in high-dimensional settings for the asymptotic regime of Theorem 3 to kick in. We now present our second example.

Example 2 *We consider the linear Gaussian example from Rainforth et al. (2018), that is $p_\theta(z) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$, $p_\theta(x|z) = \mathcal{N}(x; z, \boldsymbol{I}_d)$ with $\theta \in \mathbb{R}^d$, and $q_\phi(z|x) = \mathcal{N}(z; Ax + b, 2/3\ \boldsymbol{I}_d)$ with $A = \mathrm{diag}(\tilde{a})$ and $\phi = (\tilde{a}, b) \in \mathbb{R}^d \times \mathbb{R}^d$. Here, the optimal parameter values $(\theta^\star, \phi^\star)$ are given by $\theta^\star = T^{-1} \sum_{t=1}^T x_t$ and $\phi^\star = (a^\star, b^\star)$ with $a^\star = 1/2\boldsymbol{u}_d$ and $b^\star = \theta^\star/2$ (see Rainforth et al., 2018, Appendix B). Furthermore, the true marginal likelihood and true posterior density are given by $p_\theta(x) = \mathcal{N}(x; \theta, 2\boldsymbol{I}_d)$ and $p_\theta(z|x) = \mathcal{N}(z; (\theta + x)/2, 1/2\ \boldsymbol{I}_d)$ respectively. Then, we can apply Theorem 3: for all $\alpha \in [0, 1)$,*

$$\Delta_N^{(\alpha)}(\theta, \phi; x) = \mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) - \frac{\gamma_\alpha^2}{2N} + o\left(\frac{1}{N}\right),$$

with

$$\mathcal{L}^{(\alpha)}(\theta,\phi;x)-\ell(\theta;x)=\frac{d}{2}\left[\log\left(\frac{4}{3}\right)+\frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right]-\frac{3\alpha}{4-\alpha}\left\|Ax+b-\frac{\theta+x}{2}\right\|^2$$
$$\gamma_\alpha^2=\frac{1}{1-\alpha}\left[(4-\alpha)^d(15-6\alpha)^{-\frac{d}{2}}\exp\left(\frac{24(1-\alpha)^2}{(5-2\alpha)(4-\alpha)}\left\|Ax+b-\frac{\theta+x}{2}\right\|^2\right)-1\right].$$

The proof of Example 2 is deferred to Appendix B.5. To interpret Example 2, observe that the case of optimality is particularly telling in this example, since when $(\theta,\phi)=(\theta^\star,\phi^\star)$ it holds that $\gamma_\alpha^2=(1-\alpha)^{-1}[(4-\alpha)^d(15-6\alpha)^{-d/2}-1]$ and γ_α^2 is thus exponential in d despite the parameters (θ,ϕ) being optimal for the setting considered.

Hence, and in line with our conclusions for Example 1, the relevance of Theorem 3 can be limited for a high-dimensional latent space d. This calls for an in-depth study of the variational gap as both d and N go to infinity.

4.2 Behavior of The VR-IWAE Bound when both d and N go to Infinity

To better capture what is happening to the VR-IWAE bound in high-dimensional scenarios, we now let $d,N\to\infty$ in the variational gap

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x):=\ell_{N,d}^{(\alpha)}(\theta,\phi;x)-\ell_d(\theta;x),$$

where we have emphasized notationally the dependence on d in the VR-IWAE bound (9), the log-likelihood (2) and in the variational gap. We will consider the two cases:

$$\text{(i)}\quad d,N\to\infty\quad\text{with}\quad\frac{\log N}{d}\to 0,$$
$$\text{(ii)}\quad d,N\to\infty\quad\text{with}\quad\frac{\log N}{d^{1/3}}\to 0,$$

that is, N grows slower than exponentially with d as in (i) or slower than sub-exponentially with $d^{1/3}$ as in (ii). As we shall see, those two cases will rely on a different set of assumptions each in order to carry out the analysis. In both scenarios, we will prove that a single importance weight dominates all the others, which strongly impacts the variational gap. To this end, let us rewrite the variational gap $\Delta_{N,d}^{(\alpha)}(\theta,\phi;x)$ under a more convenient form. Writing $\overline{w}_i=\overline{w}_{\theta,\phi}(z_i)$ for all $i=1\dots N$, we first re-order the weights $\overline{w}_1,\dots,\overline{w}_N$ as

$$\overline{w}^{(1)}<\overline{w}^{(2)}<\dots<\overline{w}^{(N-1)}<\overline{w}^{(N)},$$

where we have made the assumption that the weights have no tie almost surely. Now denoting by $q_\phi^{(N)}$ the density of $\overline{w}^{(N)}$ and defining for all $\alpha\in[0,1)$

$$T_{N,d}^{(\alpha)}=\sum_{j=1}^{N-1}\left(\frac{\overline{w}^{(j)}}{\overline{w}^{(N)}}\right)^{1-\alpha}\tag{27}$$

(we have dropped the dependency in x appearing in $T_{N,d}^{(\alpha)}$ for notational ease here), we then have the following proposition.

Proposition 3 *For all $\alpha \in [0,1)$, the variational gap $\Delta_{N,d}^{(\alpha)}(\theta,\phi;x)$ can be rewritten as*

$$\Delta_{N,d}^{(\alpha)}(\theta,\phi;x) = \Delta_{N,d}^{(\alpha,MAX)}(\theta,\phi;x) + R_{N,d}^{(\alpha)}(\theta,\phi;x) \tag{28}$$

where

$$\Delta_{N,d}^{(\alpha,MAX)}(\theta,\phi;x) = \int q_\phi^{(N)}(\overline{w}^{(N)}) \log\left(\overline{w}^{(N)}\right) \mathrm{d}\overline{w}^{(N)} + \frac{\log N}{\alpha - 1} \tag{29}$$

$$0 \leq R_{N,d}^{(\alpha)}(\theta,\phi;x) \leq \frac{1}{1-\alpha}\mathbb{E}(T_{N,d}^{(\alpha)}). \tag{30}$$

The proof of Proposition 3 can be found in Appendix B.6. To continue the analysis, the key intuition will be that the log weights typically satisfy a central limit theorem (CLT), hence the weights are approximately log-normal as the dimension d increases. One such case for instance arises when the posterior and variational distributions are such that the log weights satisfy

$$\log \overline{w}_i = \sum_{j=1}^{d} X_{i,j}, \quad i = 1 \dots N, \tag{31}$$

where, for all $i = 1 \dots N$, $X_{i,1}, \dots X_{i,d}$ are i.i.d. random variables and $\mathbb{E}(\exp(\sum_{j=1}^{d} X_{i,j})) = 1$ (since the relative weights satisfy $\mathbb{E}(\overline{w}_i) = 1$). Indeed, denoting $\xi_{i,j} = -(X_{i,j} - \mathbb{E}(X_{1,1}))$, $\sigma^2 = \mathbb{V}(\xi_{1,1})$ and $S_i = \sum_{j=1}^{d} \xi_{i,j}/(\sigma\sqrt{d})$, (31) can equivalently be rewritten as

$$\log \overline{w}_i = -\log \mathbb{E}(\exp(-\sigma\sqrt{d}S_1)) - \sigma\sqrt{d}S_i, \quad i = 1 \dots N, \tag{32}$$

where under the assumption that $\sigma^2 < \infty$, S_i converges in distribution to the standard normal distribution by the CLT for all $i = 1 \dots N$. Consequently, the distribution of the weights originating from (32) can be approximated in high-dimensional settings by the log-normal distribution from Example 1, that is

$$\log \overline{w}_i = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \dots N.$$

For this reason, we first show in the following how the rest of the analysis unfolds when the distribution of the weights is assumed to be exactly log-normal. We will then use this analysis as a stepping stone to treat the more general case where the distribution of the weights is approximately log-normal of the form (32).

4.2.1 Log-normal Distribution Assumption for the Weights

Let $S_1, \dots, S_N$ be i.i.d. random variables and let the weights $\overline{w}_1, \dots, \overline{w}_N$ be of the form

$$\log \overline{w}_i = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i, \quad S_i \sim \mathcal{N}(0,1), \quad i = 1 \dots N, \tag{33}$$

that is we consider the case where the distribution of the weights is log-normal. Let $S^{(1)} \leq \dots \leq S^{(N)}$ denote the ordered sequence of $S_1, \dots, S_N$ and recall that $\overline{w}^{(1)} \leq \dots \leq \overline{w}^{(N)}$ denotes the ordered sequence of $\overline{w}_1, \dots, \overline{w}_N$. We then have the following lemma, which provides asymptotic results on the expectation of $S^{(1)}$ as $N \to \infty$.

Lemma 1 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Then,*

$$\mathbb{E}(S^{(1)}) = -\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right). \tag{34}$$

The proof of this lemma can be found in Appendix B.7.1. Intuitively, Lemma 1 will serve as the basis to study the two terms appearing in Equation (28) of Proposition 3, as both $\Delta_{N,d}^{(\alpha,MAX)}(\theta,\phi;x)$ and $\mathbb{E}(T_{N,d}^{(\alpha)})$ depend on $S^{(1)}$ through the relation

$$\log \overline{w}^{(N)} = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S^{(1)}.$$

From there, we can derive the two propositions below.

Proposition 4 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (33)*. Then, for all $\alpha \in [0,1)$,*

$$\lim_{N,d\to\infty} \Delta_{N,d}^{(\alpha,MAX)}(\theta,\phi;x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

Proposition 5 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (33)*. Then, for all $\alpha \in [0,1)$, we have*

$$\lim_{\substack{N,d\to\infty \\ \log N/d\to 0}} \mathbb{E}(T_{N,d}^{(\alpha)}) = 0. \tag{35}$$

The proof of these two propositions are deferred to Appendix B.7.2 and Appendix B.7.3 respectively. Importantly, Proposition 5 implies that the largest weight $\overline{w}^{(N)}$ converges to 1 in probability, meaning that there is a weight collapse when $N, d \to \infty$ with $\log N/d \to 0$ (following the definition of weight collapse given in Bengtsson et al., 2008). By using (35) with $\alpha = 0$, this weight collapse indeed follows from Markov's inequality (in order to get that $T_{N,d}^{(0)}$ converges to 0 in probability) combined with the fact that $\overline{w}^{(N)} = (1 + T_{N,d}^{(0)})^{-1}$.

Building on Proposition 3, Proposition 4 and Proposition 5, we now deduce the following theorem, which describes the asymptotic behavior of the variational gap as $N, d \to \infty$ in the log-normal distribution case for values of α in $[0,1)$.

Theorem 4 (i.i.d. normal random variables) *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (33)*. Then, for all $\alpha \in [0,1)$, we have*

$$\lim_{\substack{N,d\to\infty \\ \log N/d\to 0}} \Delta_{N,d}^{(\alpha)}(\theta,\phi;x) + \frac{d\sigma^2}{2}\left(1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

While Theorem 4 states that increasing N decreases the variational gap $\Delta_{N,d}^{(\alpha)}(\theta,\phi;x)$ for N large enough, it does so by a factor which is negligible compared to the term $-d\sigma^2/2$. This is in sharp contrast to Theorem 3 and more specifically to Example 1, which predicts that for log-normal weights the variational gap decreases in $1/N$ in the fixed d, large N regime.

Contrary to Example 1, the term $-d\sigma^2/2$ does not depend on α here. In fact, by taking the expectation in (33), $\mathrm{ELBO}_d(\theta, \phi; x) - \ell_d(\theta; x) = -d\sigma^2/2$, meaning that the following approximation of the variational gap in the context of Theorem 4 holds: for all $\alpha \in [0, 1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) \approx \mathrm{ELBO}_d(\theta, \phi; x) - \ell_d(\theta; x), \quad \text{as } N, d \to \infty \text{ with } \tfrac{\log N}{d} \to 0.$$

Hence, Theorem 4 shows that in high-dimensional scenarios and under the log-normal distribution assumption (33), we cannot expect to gain much from the VR-IWAE bound unless N grows exponentially with d, in the sense that the improvement is negligible compared to the ELBO. This result holds *for all values of* α *in* $[0, 1)$, thus it holds for the IWAE bound ($\alpha = 0$) as well.

We obtain the following slightly more general result by building on the proof of Theorem 4.

Theorem 5 (General i.i.d. normal random variables) *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy*

$$\log \overline{w}_i = -\frac{B_d^2}{2} - B_d S_i, \quad i = 1 \dots N, \tag{36}$$

and that there exists $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$. Then, for all $\alpha \in [0, 1)$, we have

$$\lim_{\substack{N,d \to \infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + \frac{B_d^2}{2} \left\{ 1 - 2 \, \frac{\sqrt{2 \log N}}{B_d} + \frac{1}{1-\alpha} \, \frac{2 \log N}{B_d^2} + O\left(\frac{\log \log N}{B_d \sqrt{\log N}} \right) \right\} = 0.$$

The proof of Theorem 5 can be found in Appendix B.7.4. We now revisit the Gaussian example given in Example 1 in the context of Theorem 5.

Example 3 *Set $p_\theta(z|x) = \mathcal{N}(z; \theta, \boldsymbol{I}_d)$ and $q_\phi(z) = \mathcal{N}(z; \phi, \boldsymbol{I}_d)$, with $\theta, \phi \in \mathbb{R}^d$. Denoting $B_d = \|\theta - \phi\|$, we can write the weights $\overline{w}_1, \dots, \overline{w}_N$ under the form (36) (see (81) of Appendix B.4). Hence, Theorem 5 applies if there exists $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$. This is for example the case if $\theta = 0 \cdot \boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$ with $\sigma_- = 1$.*

As we shall see next, our conclusion regarding the behavior of the VR-IWAE bound in high-dimensional settings extends to cases where the log-normal assumption does not necessarily hold exactly, that is if we assume instead that (32) holds, where $S_1, \dots S_N$ are i.i.d. random variables whose distribution is close to a normal as $N, d \to \infty$.

4.2.2 Beyond the Log-normal Distribution Assumption

Following (32), let us set

$$\log \overline{w}_i = -\log \mathbb{E}(\exp(-\sigma \sqrt{d} S_1)) - \sigma \sqrt{d} S_i, \quad i = 1 \dots N, \tag{37}$$

where the i.i.d. random variables $S_1, \dots, S_N$ are defined as follows:

$$S_i = \frac{1}{\sigma \sqrt{d}} \sum_{j=1}^{d} \xi_{i,j}, \quad i = 1 \dots N. \tag{38}$$

The assumption (A2) below ensures that $S_1, \dots S_N$ have a distribution that is close to a normal as $N, d \to \infty$, so that (33) is recovered in the limit.

(A2) For all $i = 1 \dots N$,

(a) $\xi_{i,1}, \dots, \xi_{i,d}$ are i.i.d. random variables which are absolutely continuous with respect to the Lebesgue measure and satisfy $\mathbb{E}(\xi_{i,1}) = 0$ and $\mathbb{V}(\xi_{i,1}) = \sigma^2 < \infty$.

(b) There exists $K > 0$ such that:

$$|\mathbb{E}(\xi_{i,1}^k)| \leq k! K^{k-2} \sigma^2, \quad k \geq 3.$$

Here, the condition (A2)b corresponds to the well-known Bernstein condition. Paired up with (A2)a, this condition permits us to appeal to classical limit theorems for large deviations in order to enlarge the so-called zone of normal convergence beyond the CLT (Petrov, 1995; Saulis and Statulevičius, 2000). This enables us to establish preliminary results which are used to prove the results of Section 4.2.2 we will now present (we refer to Appendix B.8.2 for the statement of those preliminary results). We first provide the equivalent of Lemma 1 in the more general context of (38) and under (A2).

Lemma 2 *Assume* (A2). *Let* $S_1, \dots, S_N$ *be as in* (38). *Then, as* $N, d \to \infty$, *with* $\frac{\log N}{d^{1/3}} \to 0$, (34) *holds.*

The proof of this result is deferred to Appendix B.8.3. Notice that we are now assuming that N grows slower than sub-exponentially with $d^{1/3}$ in Lemma 2. The following two propositions give results akin to those obtained in Proposition 4 and Proposition 5.

Proposition 6 *Assume* (A2). *Let* $S_1, \dots, S_N$ *be as in* (38). *Further assume that the weights* $\overline{w}_1, \dots, \overline{w}_N$ *satisfy* (37). *Then, setting*

$$a := \log \mathbb{E}(\exp(-\xi_{1,1})), \tag{39}$$

we have that $a > 0$ *and that for all* $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d \to \infty \\ \log N / d^{1/3} \to 0}} \Delta_{N,d}^{(\alpha, MAX)}(\theta, \phi; x) + da \left\{ 1 - \frac{\sigma}{a} \sqrt{\frac{\log N}{d}} + O\left(\frac{\log \log N}{\sqrt{d \log N}} \right) \right\} = 0.$$

Proposition 7 *Assume* (A2). *Let* $S_1, \dots, S_N$ *be as in* (38). *Further assume that the weights* $\overline{w}_1, \dots, \overline{w}_N$ *satisfy* (37). *Then, for all* $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d \to \infty \\ \log N / d^{1/3} \to 0}} \mathbb{E}(T_{N,d}^{(\alpha)}) = 0.$$

The proof of Proposition 6 and Proposition 7 can be found in Appendix B.8.4 and Appendix B.8.5 respectively.

Remark 3 *The log-normal case corresponds to setting* $a = \sigma^2/2$ *in Proposition 6 (this can be checked using the definition of* a *in* (39) *combined with* (94) *from the proof of Proposition 6 in Appendix B.8.4). Contrary to Proposition 4, the* $(1 - \alpha)^{-1}(d\sigma^2)^{-1} 2 \log N$ *term is now subsumed by the final* $O(\log \log N / \sqrt{d \log N})$ *term in Proposition 6, which comes from the fact that Proposition 6 makes the additional assumption* $\log N / d^{1/3} \to 0$ *as* $N, d \to \infty$. *Hence, Proposition 4 and Proposition 6 agree with each other in the log-normal case.*

Proposition 3, Proposition 6 and Proposition 7 lead to the theorem below, which characterizes the asymptotics of the variational gap as $N, d \to \infty$ for $\alpha \in [0, 1)$ in the more general case where the distribution of the weights is approximately log-normal according to (37).

Theorem 6 (i.i.d. random variables) *Assume* (A2). *Let* $S_1, \dots, S_N$ *be as in* (38). *Further assume that the weights* $\overline{w}_1, \dots, \overline{w}_N$ *satisfy* (37) *and let* $a > 0$ *be defined as in* (39). *Then, for all* $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da \left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0.$$

We have thus obtained that, under the assumptions of Theorem 6, the VR-IWAE bound is of limited interest for all values of $\alpha \in [0, 1)$ unless N grows at least sub-exponentially with $d^{1/3}$. In fact, by taking the expectation in the expression of the log-weights, we have that $\mathrm{ELBO}_d(\theta, \phi; x) - \ell_d(\theta; x) = -da$ (using for example (95) from the proof of Proposition 6 in Appendix B.8.4). Hence, the following approximation of the variational gap holds in the context of Theorem 6: for all $\alpha \in [0, 1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) \approx \mathrm{ELBO}_d(\theta, \phi; x) - \ell_d(\theta; x), \quad \text{as } N, d \to \infty \text{ with } \tfrac{\log N}{d^{1/3}} \to 0.$$

Since the weights are assumed to be *approximately* log-normal this time as opposed to Section 4.2.1, the condition that N should grow at least exponentially with d to avoid a weight collapse effect has now been replaced by the less restrictive yet still stringent condition that N should grow at least sub-exponentially with $d^{1/3}$.

As described below, the assumptions on the distribution of the weights—that is, on the ratio between the posterior and the variational distributions—appearing in Theorem 6 are met for the linear Gaussian setting from Example 2.

Example 4 *We consider the linear Gaussian setting from Rainforth et al. (2018) that we recalled in Example 2. Denoting* $\lambda = \left\|\frac{\theta+x}{2} - Ax - b\right\|/\sqrt{d}$, *the weights can be written in the form of* (37) *with* $\sigma^2 = 1/18 + 8/3\lambda^2$ *and we also have* $a = \lambda^2 + 1/6 + 1/2\log(3/4)$. *As a result, we can apply Theorem 6 if* (A2) *holds. This is for example the case at optimality when* $(\theta, \phi) = (\theta^\star, \phi^\star)$ *(and the derivation details for this example can be found in Appendix B.8.6).*

Example 4 states that we are in the conditions of application of Theorem 6 when the parameters (θ, ϕ) are optimal, with corresponding optimal posterior density $p_{\theta^\star}(z|x) = \mathcal{N}(z; (\theta^\star + x)/2, 1/2\boldsymbol{I}_d)$ and optimal variational density $q_{\phi^\star}(z|x) = \mathcal{N}(z; (\theta^\star + x)/2, 2/3\boldsymbol{I}_d)$.

This example showcases how, in some instances where the variational family is not large enough to contain the target density, there can be a weight collapse phenomenon as d increases that severely impacts the VR-IWAE bound, even when the parameters (θ, ϕ) are set to be the optimal ones for the problem considered. This concludes our theoretical study of the VR-IWAE bound, which sheds lights on the conditions behind the success or failure of this bound. In the next section, we describe how our theoretical results relate to the existing literature.

5. Related Work

Alpha-divergence variational inference. Our work provides the theoretical grounding behind VR-bound gradient-based schemes (Hernandez-Lobato et al., 2016; Bui et al., 2016; Li and Turner, 2016; Dieng et al., 2017; Li and Gal, 2017; Zhang et al., 2021; Rodríguez-Santana and Hernández-Lobato, 2022). It also unifies the VR and IWAE bound methodologies (Burda et al., 2016; Rainforth et al., 2018; Tucker et al., 2019; Domke and Sheldon, 2018; Maddison et al., 2017) and serves as a foundation for improving on both methodologies.

Proof techniques. Several of our theoretical results generalize known findings from the literature in order to build the VR-IWAE bound methodology and to characterize its asymptotics. Some of our proofs are straightforwardly derived from existing ones, such as the proofs of Theorems 2 and 3 (which are established by directly adapting the proofs written in Tucker et al. (2019) and in Domke and Sheldon (2018) respectively). However, a number of our proof techniques differs significantly from/alter parts of known proofs (see Appendix C for details). Lastly, the derivations made in Section 4.2 for the asymptotics of the VR-IWAE bound when $N, d \to \infty$ are, to the best of our knowledge, the first of their kind.

Importance sampling. Common variational bounds and their gradients can often be expressed in terms of the importance weights $w_{\theta,\phi}$ (with our novel VR-IWAE bound being no exception to that rule). As such, the success of gradient-based variational inference has been known to depend on the behavior of the importance weights and there has been a growing interest in understanding this behavior through the use of insights and tools from the importance sampling (IS) literature (Maddison et al., 2017; Domke and Sheldon, 2018; Dhaka et al., 2021; Geffner and Domke, 2021). In particular, it is well-known that IS can perform poorly in high dimensions unless the target and reference/proposal distributions are close.

Picklands III (1975) for instance showed that, under commonly satisfied assumptions, the right tail of the importance weights distribution approximates a generalized Pareto distribution, that is, a heavy-tailed distribution with three parameters (u, σ, k) and moments of order up to $\lfloor 1/k \rfloor$. This behavior is typical in high dimensions and it makes IS fail, as the IS estimators are dominated by the few largest terms. Leveraging this result, Dhaka et al. (2021) considered the case of black-box variational inference and viewed the importance weights as approximately drawn from a generalized Pareto distribution with tail index k. The importance weights taken to an exponent $1 - \alpha$ are then approximately distributed according to a generalized Pareto distribution with tail index $(1 - \alpha)k$ and they deduced that the estimates should be more stable as α increases towards 1 due to lighter tails.

The analysis from Dhaka et al. (2021) goes hand in hand with our findings, as (i) Theorems 1 and 3 predict improvements in terms of SNR and variance as α increases, at the cost of an increasing bias and (ii) our results from Section 4.2 show that, as d increases, the VR-IWAE bound fails regardless of the value of $\alpha \in [0, 1)$ and provides negligible improvements compared to the ELBO ($\alpha = 1$). However, one main specificity of our work is our precise characterization of how the distribution of the importance weights impacts the tightness of the VR-IWAE bound. Specifically, Theorem 3 generalizes Domke and Sheldon (2018) to the VR-IWAE bound, while the results from Section 4.2 provide the first theoretical

justification behind the empirical findings from Geffner and Domke (2021) regarding the impact of weight collapse on the tightness of variational bounds.

The next section is devoted to illustrating the theoretical claims we have made thus far over toy and real-data experiments.

6. Numerical Experiments

In this section, our goal is to verify the validity of the theoretical results we established over several numerical experiments, starting with a Gaussian example in which the distribution of the weights is exactly log-normal.

6.1 Gaussian Example

We consider the Gaussian example described in Example 3, for which the weights $\overline{w}_1, \dots, \overline{w}_N$ can be written under the form (36) with $B_d = \|\theta - \phi\|$, meaning that the distribution of the weights is log-normal. On the one hand, Theorem 3 predicts that for all $\alpha \in [0, 1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = -\frac{\alpha B_d^2}{2} - \frac{\exp\left[(1-\alpha)^2 B_d^2\right] - 1}{2(1-\alpha)N} + o\left(\frac{1}{N}\right) \tag{40}$$

(this follows from a straightforward adaptation of Example 1). On the other hand, Theorem 5 tells us that if there exists $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$, then: for all $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + \frac{B_d^2}{2}\left\{1 - 2\,\frac{\sqrt{2\log N}}{B_d} + \frac{1}{1-\alpha}\,\frac{2\log N}{B_d^2} + O\left(\frac{\log\log N}{B_d\sqrt{\log N}}\right)\right\} = 0. \tag{41}$$

We now want to check the validity of the two asymptotic results above. To do so, we need to be able to approximate the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$, which can be done using the unbiased Monte Carlo (MC) estimator given for all $N \in \mathbb{N}^\star$ by

$$\frac{1}{1-\alpha} \log\left(\frac{1}{N}\sum_{j=1}^{N} \overline{w}_{\theta,\phi}(Z_j)^{1-\alpha}\right),$$

with $Z_1, \dots, Z_N$ being i.i.d. samples generated according to q_ϕ. As for the approximation returned by Theorem 3, we will represent it according to (40) through functions of the form

$$c_1 \mapsto -\frac{\alpha B_d^2}{2} - \frac{\exp\left[(1-\alpha)^2 B_d^2\right] - 1}{2(1-\alpha)N} + \frac{c_1}{N} \tag{42}$$

and for the approximation returned by Theorem 5, we will represent it according to (41) through functions of the form

$$c_2 \mapsto -\frac{B_d^2}{2} + B_d\sqrt{2\log N} + \frac{\log N}{\alpha - 1} + \frac{c_2 B_d \log\log N}{\sqrt{\log N}}. \tag{43}$$

We first consider the case where $\theta = 0 \cdot \boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$. In that setting, $B_d = \sqrt{d}$ and we have that (i) the $1/N$ term from (40) is exponential in $(1-\alpha)^2 d$ and (ii) we are in the conditions of application of Theorem 5 by setting $\sigma_- = 1$.

Consequently, for this choice of (θ, ϕ) and *regardless of the value of* $\alpha \in [0, 1)$, we are expecting Theorem 5 to capture the behavior of the variational gap as d and N increase in such a way that $\log N/d$ decreases. This is indeed what we observe in Figure 1, in which we let $d \in \{10, 100, 1000\}$, $\alpha \in \{0., 0.2, 0.5\}$, $N \in \{2^j \ : \ j = 1 \dots 9\}$ and we compare the behavior of the variational gap to the behavior predicted by Theorem 5 through curves of the form (43).

Unsurprisingly, although valid in low dimensions for a proper choice of α, the analysis of Theorem 3 requires an unpractical amount of samples N to properly capture the behavior of the variational gap as d increases (additional plots providing the comparison with Theorem 3 are made available in Appendix D.1 for the sake of completeness).

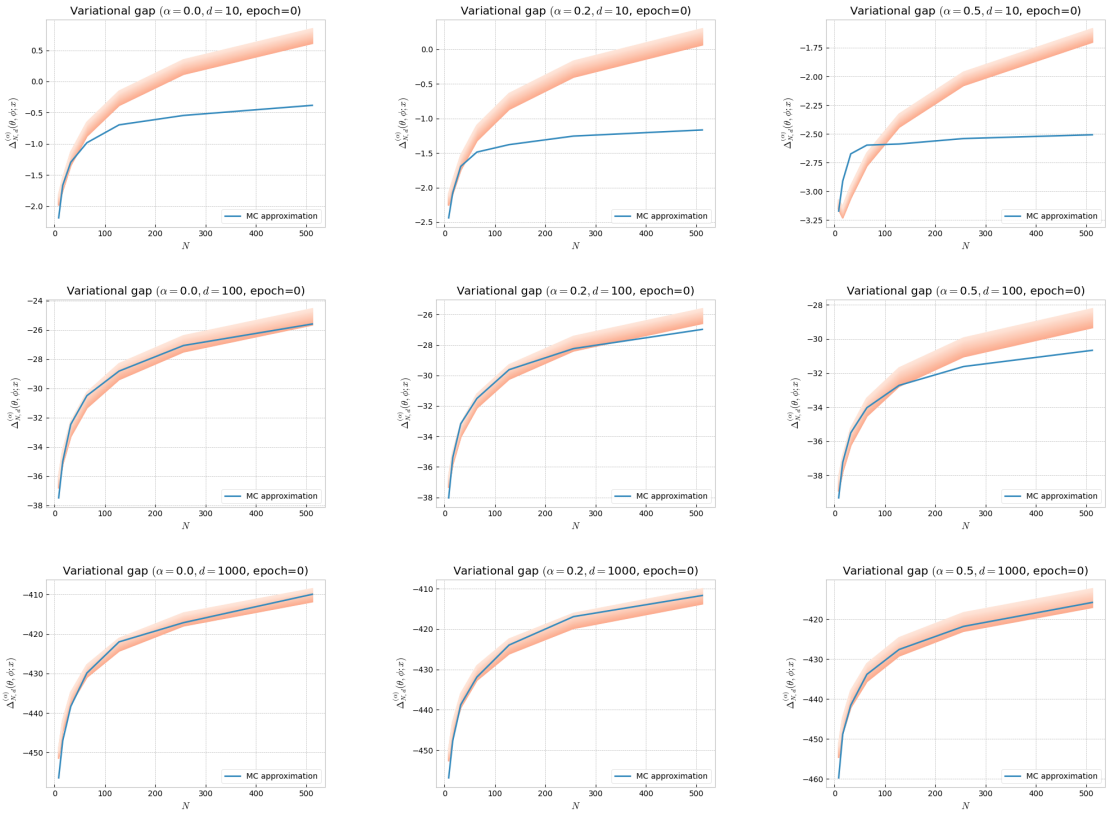

Figure 1: Plotted in blue is the MC estimate of the variational gap $\Delta^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) for the toy example described in Section 6.1 as a function of N, for varying values of (α, d) and with $(\theta, \phi) = (0 \cdot \boldsymbol{u_d}, \boldsymbol{u}_d)$ so that $B_d = \sqrt{d}$. Plotted in orange are curves of the form (43) with tailored values of c_2.

We next train the parameter ϕ in order to measure the impact of the training procedure on the validity of our asymptotic results. Here, this impact is reflected in the quantity B_d

through the simple relation $B_d = \|\theta - \phi\|$. In case the training is successful, $B_d/\sqrt{d}$ is then anticipated to decrease from 1 to 0 (having set $\theta = 0 \cdot \boldsymbol{u}_d$ and initialized with $\phi = \boldsymbol{u_d}$).

Hence, as the training progresses, we will be less and less able to find $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$, which will contradict the assumption we make in Theorem 5. At the same time, the $1/N$ term from (40) will decrease thanks to its dependency in B_d, meaning that (40) may become a better approximation than (41) during the training procedure.

This behavior is empirically confirmed in Figure 2 (and we also check in Figure 14 of Appendix D.1 that $B_d/\sqrt{d}$ indeed goes from 1 to 0 during the training procedure). In those plots, the parameter ϕ was optimised via stochastic gradient descent using the reparameterized gradient estimator (13) with $N = 100$ and we set $\alpha = 0.2$ and $d = 1000$ (and a similar trend can be observed for other values of α and d).

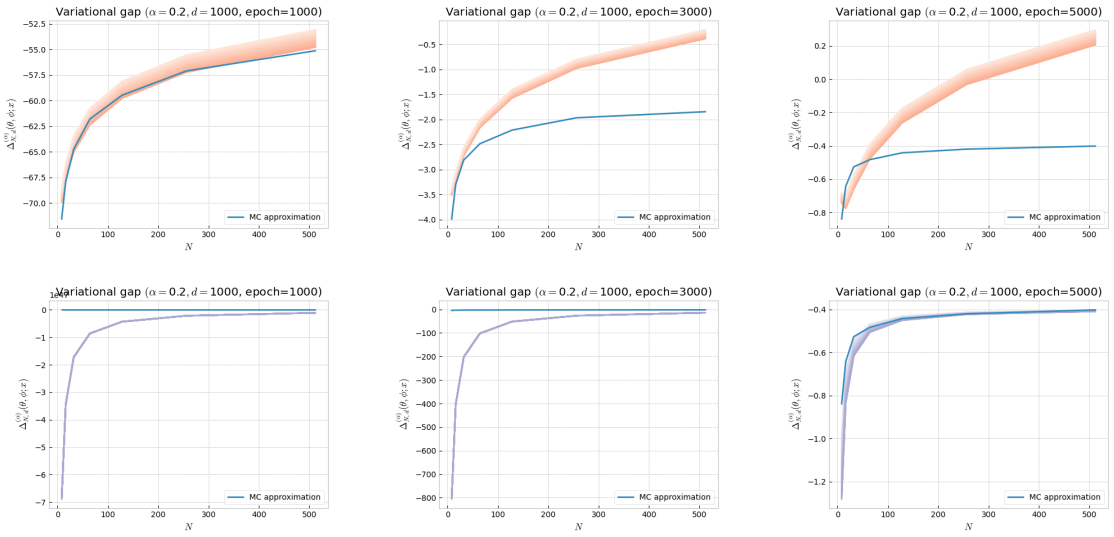

Figure 2: Plotted in blue is the MC estimate of the variational gap $\Delta^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) at epochs $\{1000, 3000, 5000\}$ for the toy example described in Section 6.1 as a function of N, for $\alpha = 0.2$ and $d = 1000$. Plotted in orange (resp. in purple) are curves of the form (43) with tailored values of c_2 (resp. of the form (42) with tailored values of c_1).

The main insight we get from our first numerical experiment is then that: as the dimension d increases and N does not grow faster than exponentially with d, we should not expect much empirically from the VR-IWAE bound as a lower bound to the marginal log-likelihood when the distribution of the weights is log-normal. This is true unless the encoder and decoder distributions become very close to one another, in which case Theorem 3 does apply instead of Theorem 5.

Thus, while this limitation of the VR-IWAE bound holds for all $\alpha \in [0, 1)$, it may be mitigated by (i) proposing successful training procedures (further shedding light on the importance of finding gradient estimators with good SNR properties) and (ii) selecting suitable

variational families which can capture the complexity within the target posterior density. Furthermore, the analysis provided by Theorem 3 may also apply in lower dimensional settings, under the condition that the variance term appearing in Theorem 3 is well-behaved and that the value of α is properly tuned. We next present a second numerical experiment, where this time the weights are not exactly log-normal.

6.2 Linear Gaussian Example

We are interested in the linear Gaussian example from Rainforth et al. (2018), which we already highlighted in Examples 2 and 4. The data set $\mathcal{D} = \{x_1, \dots, x_T\}$ is generated by sampling $T = 1024$ datapoints from $\mathcal{N}(0, 2\boldsymbol{I}_d)$ and we will consider three initializations for the parameters (θ, ϕ) involving a Gaussian perturbation of standard deviation σ_{perturb} of the ground truth values $(\theta^\star, \phi^\star)$:

(i) $\sigma_{\text{perturb}} = 0.5$: the parameters are initialized far from $(\theta^\star, \phi^\star)$,
(ii) $\sigma_{\text{perturb}} = 0.01$: the parameters are initialized close to $(\theta^\star, \phi^\star)$,
(iii) $\sigma_{\text{perturb}} = 0.$: the parameters are equal to $(\theta^\star, \phi^\star)$.

The first two initializations follow from Rainforth et al. (2018) and should notably permit us to approximately characterize the behavior of the linear model before and after training.

Our first step is to check that, as written in Example 4, the distribution of the weights is approximately log-normal as d increases for the initializations above. To do so, we randomly select a datapoint x, draw $N = 1000000$ weight samples in dimension $d = \{20, 100, 1000\}$ for $\sigma_{\text{perturb}} \in \{0.5, 0.01, 0.\}$, before plotting for each d a histogram of the resulting log-weight distribution as well as a Q-Q plot to test the normality assumption of those log-weights.

The results are shown on Figure 3 and we see that while the log-normality phenomenon happens in dimension $d = 100$ when a large perturbation is being considered, even a small perturbation to no perturbation at all can induce some log-normality of the weights as d further increases, which is in line with the theory (and similar plots can be observed for other randomly selected datapoints).

We next want to test the validity of our asymptotic results. On the one hand, Theorem 3 predicts that: for all $\alpha \in [0, 1)$,

$$\Delta_{N,d}^{(\alpha)}(\theta, \phi; x) = \mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \ell_d(\theta; x) - \frac{\gamma_{\alpha,d}^2}{2N} + o\left(\frac{1}{N}\right), \tag{44}$$

where $\mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \ell_d$ and $\gamma_{\alpha,d}^2$ can be analytically computed using Example 2 (and we have emphasized the dependency in d in each of those terms). On the other hand, Theorem 6 predicts under (A2) that: for all $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha)}(\theta, \phi; x) + da\left(1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right) = 0,$$

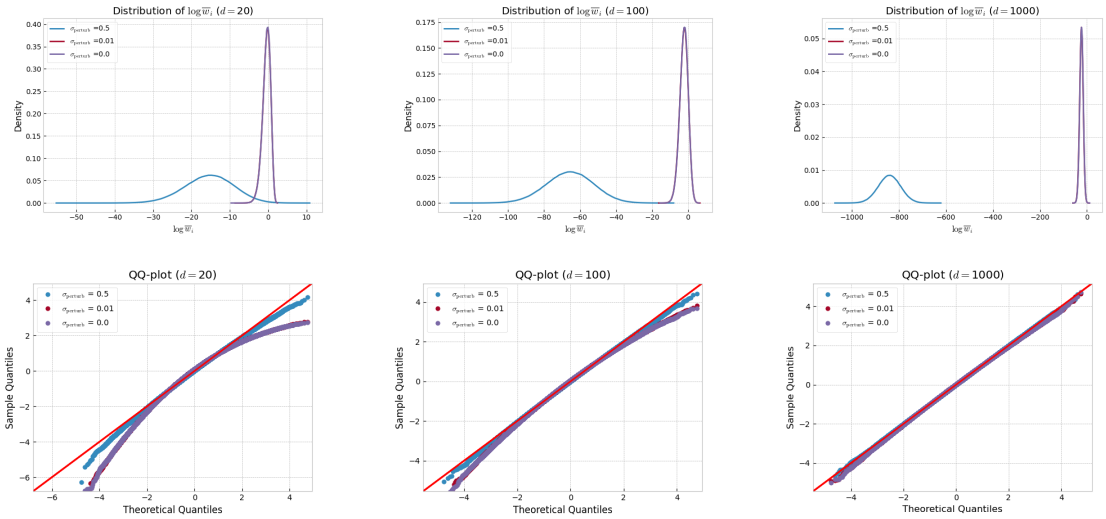

Figure 3: Plotted is the distribution of $\log \overline{w}_i$ and the corresponding QQ-plot for the linear Gaussian example described in Section 6.2 for a randomly selected datapoint x, varying values of d and three different initializations of the parameters (θ, ϕ).

where σ^2 and a can be computed analytically according to Example 4. Hence, to check whether these results apply, we want to look at functions of the form

$$\text{(Theorem 3)} \quad c_1 \mapsto \mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \ell_d(\theta; x) - \frac{\gamma_{\alpha,d}^2}{2N} + \frac{c_1}{N} \tag{45}$$

$$\text{(Theorem 6)} \quad c_2 \mapsto -da + \sqrt{d}\sigma\sqrt{2\log N} + \frac{c_2\sqrt{d}\log\log N}{\sqrt{\log N}} \tag{46}$$

and see how well they approximate the behavior of the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$. Based on Example 4, we are expecting the regime predicted by Theorem 6 to apply as d increases if N does not grow faster than $d^{1/3}$ and this is indeed what we observe in Figure 4.

While this process is noticeably quicker for an initialization that is far from the optimum ($\sigma_{\text{perturb}} = 0.5$), all three initializations considered here eventually exhibit the behavior predicted by Theorem 6 as d further increases. As already mentioned in Section 4.2.2, this sends the important message that the VR-IWAE bound can strongly deteriorate as d increases due to a mismatch between the targeted density and its variational approximation, even though the parameters themselves are optimal.

As for Theorem 3, we obtain that this theorem applies in low to medium dimensions when the value of α is well-chosen and/or the parameters are close to being optimal, but fails as d increases unless we use an unpractical amount of samples N (see Appendix D.2.1).

We now want to get insights regarding the training of the VR-IWAE bound in practice. We follow the methodology used in Rainforth et al. (2018), which looked into the convergence

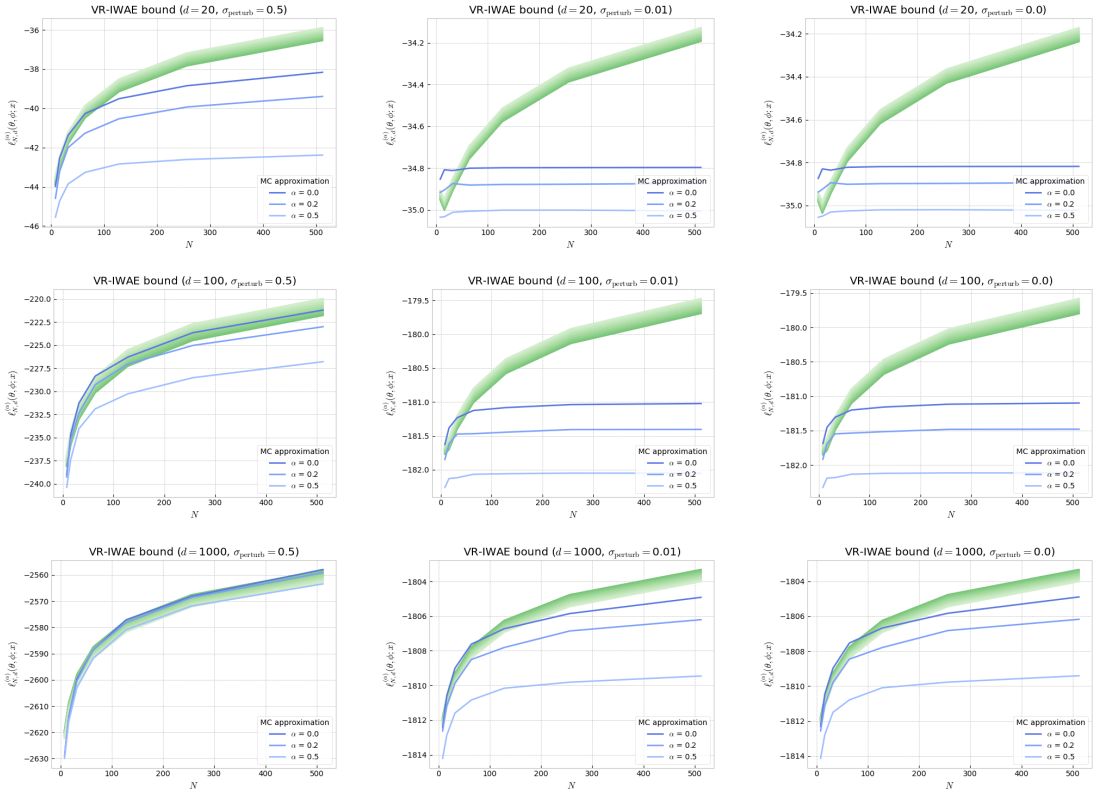

Figure 4: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d) and three different initializations of (θ, ϕ). Plotted in green are curves of the form (46) with tailored values of c_2.

of the SNR for the numerical example considered here in the specific case of the IWAE bound ($\alpha = 0$). Our goal is thus to check whether we can observe the SNR advantages when $\alpha > 0$ predicted by Theorem 1 in the reparameterized case.

Let us decompose θ as $(\theta_\ell)_{1 \leq \ell \leq d}$ and ϕ as $(\phi_{\ell'})_{1 \leq \ell' \leq d+1}$. We then look at the reparameterized estimated gradients of the VR-IWAE bound $(\delta^{(\alpha)}_{1,N}(\theta_\ell))_{1 \leq \ell \leq d}$ and $(\delta^{(\alpha)}_{1,N}(\phi_{\ell'}))_{1 \leq \ell' \leq d+1}$ defined in (14) and (15) respectively as a function of N, for varying values of α, varying values of d and for the two initializations $\sigma_{\text{perturb}} = 0.01$ and $\sigma_{\text{perturb}} = 0.5$. The results are shown in Figure 5 (resp. Figure 6) and they have been obtained by randomly selecting 10 indexes ℓ ranging between 1 and d and averaging over the resulting $\text{SNR}(\delta^{(\alpha)}_{1,N}(\theta_\ell))$ values (resp. by randomly selecting 10 indexes ℓ' ranging between 1 and $d+1$ and averaging over the resulting $\text{SNR}(\delta^{(\alpha)}_{1,N}(\phi_{\ell'}))$ values). Theoretical lines have also been added to Figures 5 and 6 in order to reflect the asymptotic regimes predicted by Theorem 1.

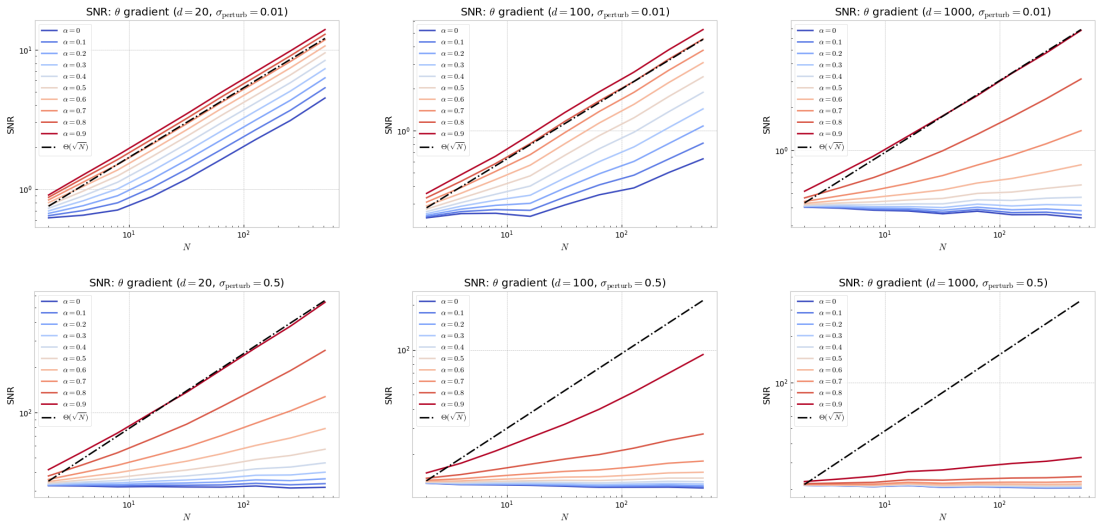

Figure 5: Plotted is the SNR of the generative network (θ) gradients in the reparameterized case (computed over 1000 MC samples) for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

Observe that, in the favourable setting of low to medium dimensions with a small perturbation near the optimum (that is $d \in \{20, 100\}$ with $\sigma_{\text{perturb}} = 0.01$), the asymptotic rates predicted by Theorem 1 for the SNR match the observed rates. In particular, the SNR of the inference network gradients vanishes for $\alpha = 0$ while it does not for $\alpha > 0$, which showcases the potential benefits of using the VR-IWAE bound with $\alpha > 0$ instead of the IWAE bound. More generally, increasing α increases the SNR of both the generative and the inference networks, with what seems to be a monotonic increase with α.

However, the improvement in SNR for both the generative and inference networks becomes less pronounced as we get further away from the optimum ($\sigma_{\text{perturb}} = 0.5$) and/or increase d ($d = 1000$). We relate this behavior to the weight collapse effect established in Theorem 6 and anticipate that observing the asymptotic rates predicted by Theorem 1 requires an unpractical amount of samples N as d increases, regardless of the value of $\alpha \in [0, 1)$. Note that the use of doubly-reparameterized gradient estimators for ϕ mitigates the decay in SNR (see Figure 16 of Appendix D.2.2).

Lastly, the behavior of the VR-IWAE bound as well as the SNR behavior of its gradient estimators are not the only way to measure the success of gradient-based methods involving the VR-IWAE bound. For example, we observe that while increasing α does not lower the Mean Squared Error (MSE) for log-likelihood estimation, it can be useful in lowering the MSE of the θ gradient estimates (see Figures 17 and 18 of Appendix D.2.2). We now move on to our third and final numerical experiment, in which we examine a real-data scenario.

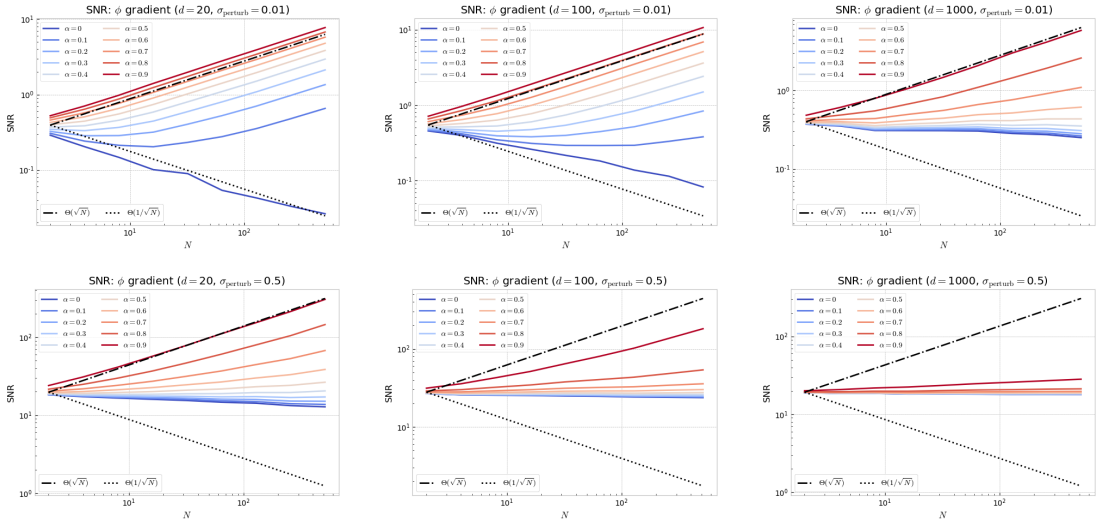

Figure 6: Plotted is the SNR of the inference network (ϕ) gradients in the reparameterized case (computed over 1000 MC samples) for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

6.3 Variational Auto-encoder

We consider the case of a variational auto-encoder (VAE) model designed to generate MNIST digits with a d-dimensional latent space, where $p_\theta(z)$ is a fixed standard Gaussian distribution, $p_\theta(x|z)$ is a product over the output dimensions of independent Bernoulli random variables with logits $\pi_\theta(z)$, $q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sigma_\phi(x))$ and the functions $\pi_\theta(z)$ and $(\mu_\phi(x), \sigma_\phi(x))$ are parameterized by neural networks. More precisely, both the encoding and decoding networks are MLPs with two hidden layers of size 200 and tanh nonlinearities.

We first want to investigate whether the distribution of the weights appears to become log-normal in this setting as the dimension of the latent space d increases.

To verify this claim empirically, we randomly select a datapoint x in the test set and for $d \in \{5, 10, 50, 100, 1000, 5000\}$ we randomly generate some model parameters (θ, ϕ), before drawing $N = 1000000$ (unnormalized) weight samples. For each d, we then normalize the weights and plot a histogram of the resulting log-weight distribution, alongside with a QQ-plot to test the normality assumption of those log-weights. The results are shown in Figure 7 and they illustrate the fact that the weights tend to become log-normal as d increases (and similar plots can be obtained for other randomly selected datapoints and other initializations of the parameters (θ, ϕ)).

From there, we want to check the validity of our asymptotic results. To do so, a first comment is that, regardless of the distribution of the weights, Theorem 3 predicts the

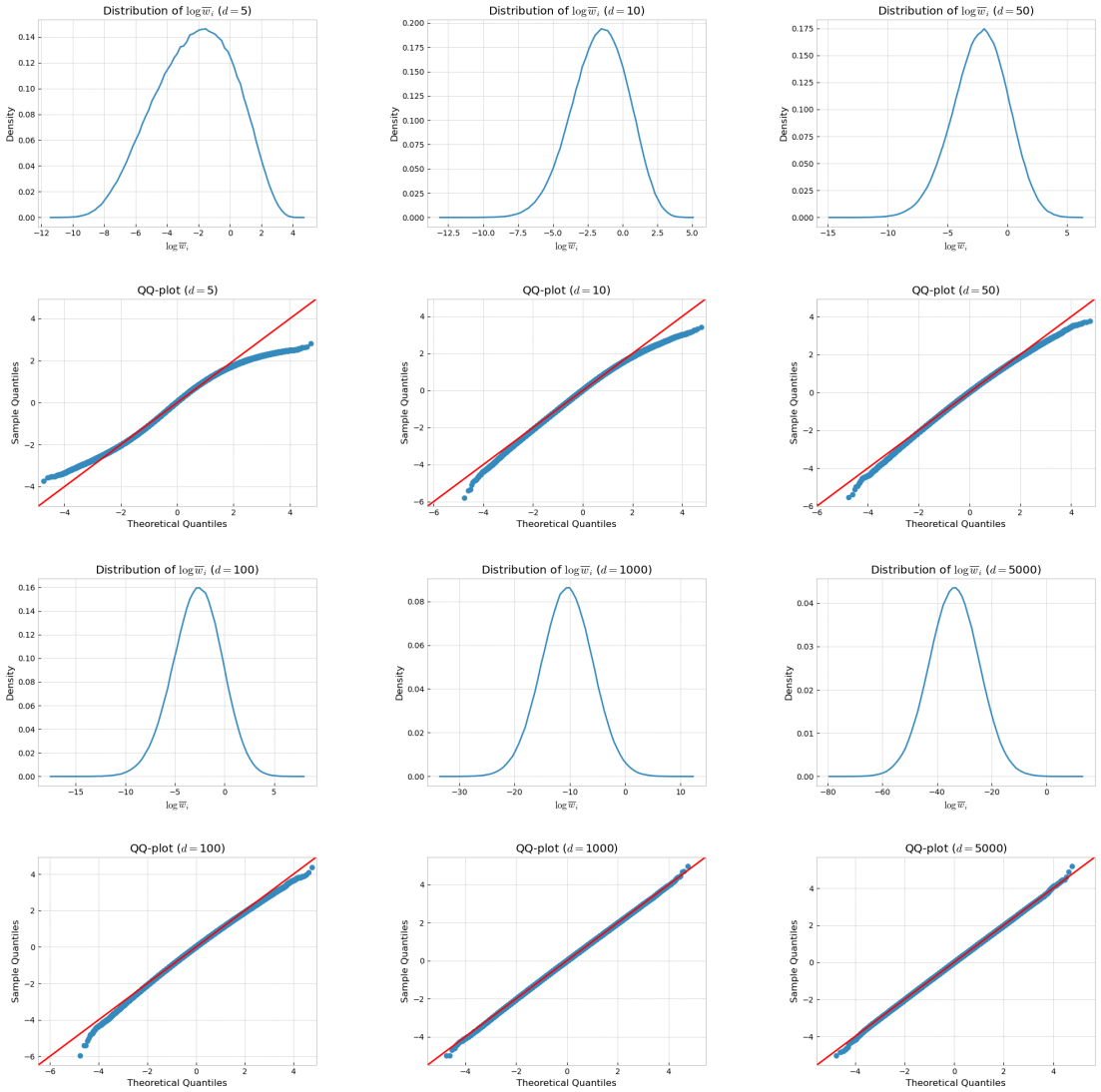

Figure 7: Plotted is the distribution of $\log \overline{w}_i$ and the corresponding QQ-plot for the VAE in Section 6.3, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of d.

following: for all $\alpha \in [0, 1)$,

$$\ell_{N,d}^{(\alpha)}(\theta, \phi; x) = \mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \frac{\gamma_{\alpha,d}^2}{2N} + o\left(\frac{1}{N}\right), \tag{47}$$

where, $\ell_{N,d}^{(\alpha)}(\theta, \phi; x)$ denotes the VR-IWAE bound, $\mathcal{L}_d^{(\alpha)}(\theta, \phi; x)$ the VR-bound and $\gamma_{\alpha,d}^2 = (1-\alpha)^{-1}\mathbb{V}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))$ (and we have emphasized the dependency in d in each of those

terms). If we further make the assumption that the weights are of the form (37) (which appears to approximately be the case as the dimension d increases as per Figure 7), then Theorem 6 predicts under (A2) that: for all $\alpha \in [0, 1)$,

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \ell_{N,d}^{(\alpha)}(\theta, \phi; x) - \mathrm{ELBO}_d(\theta, \phi; x) - \sqrt{d}\sigma\sqrt{2\log N} + O\left(\frac{\sqrt{d}\log\log N}{\sqrt{\log N}}\right) = 0. \quad (48)$$

Here we have emphasized the dependency in d in $\mathrm{ELBO}_d(\theta, \phi; x)$ and we have also used the fact that $\mathrm{ELBO}_d(\theta, \phi; x) - \ell_d(\theta; x) = -da$ (as previously stated, this follows from taking the expectation in (95) from the proof of Proposition 6 in Appendix B.8.4). Hence, to check whether these results apply, we want to look at functions of the form

$$\text{(Theorem 3)} \quad c_1 \mapsto \mathcal{L}_d^{(\alpha)}(\theta, \phi; x) - \frac{\gamma_{\alpha,d}^2}{2N} + \frac{c_1}{N} \quad (49)$$

$$\text{(Theorem 6)} \quad c_2 \mapsto \mathrm{ELBO}_d(\theta, \phi; x) + \sqrt{d}\sigma\sqrt{2\log N} + \frac{c_2\sqrt{d}\log\log N}{\sqrt{\log N}} \quad (50)$$

and see how well they approximate the behavior of the VR-IWAE bound $\ell_{N,d}^{(\alpha)}(\theta, \phi; x)$. Although the functions above contain unknown terms, those terms can all be estimated: the VR bound can be estimated using MC sampling as in (8) and so can the ELBO. As for σ, it can be estimated from the sample standard deviation of the log-weights (and $\gamma_{\alpha,d}^2$ can be estimated in a similar fashion).

Note as a side remark that we are considering the VR-IWAE bound as the quantity of interest whose behavior shall be mimicked by (47) or (48) (through (49) or (50)). Indeed, while we were working with the variational gap in our previous numerical experiments, computing this quantity requires us to estimate both the VR-IWAE bound and the log-likelihood here, which would have incurred an additional source of randomness that we have been able to avoid in both (47) and (48).

Based on Figure 7, we expect two situations to arise at this stage: (i) the asymptotic regime suggested by Theorem 3 captures the behavior of the VR-IWAE bound in low to medium dimensions and (ii) the asymptotic regime predicted by Theorem 6 is accurate as d increases and N does not grow faster than $d^{1/3}$.

This is exactly what we observe in Figures 8 and 9, in which σ is estimated with the 1000000 (unnormalized) weight samples used to build Figure 7 and so are the VR bound and the ELBO (additional plots are also available in Figures 19 and 20 of Appendix D.3). In particular, we see in Figure 8 that the asymptotic regime of Theorem 3 mimics the behavior of the VR-IWAE bound in low to medium dimensions as long as $\gamma_{\alpha,d}^2$ does not grow too quickly with d (and we already observe a mismatch between the two for $\alpha = 0$ and $d = 100$).

Nevertheless, the VR-IWAE bound ends up straying away from the behavior predicted by Theorem 3 as d increases unless N becomes impractically large (with the particularity that this process happens slower as α increases). We then see on Figure 9 that, as d increases to reach high-dimensional settings so that the ratio $\log N/d^{1/3}$ becomes small for the values of N considered here, the behavior predicted by Theorem 6 starts to emerge.

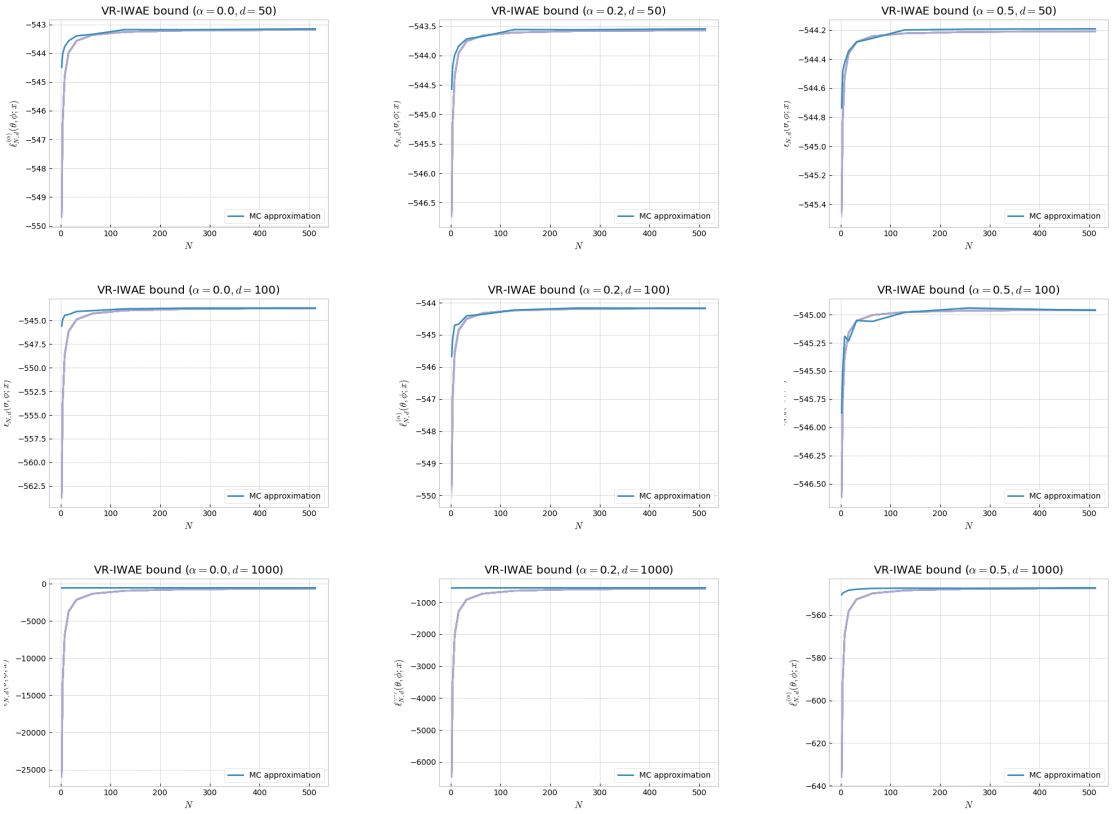

Figure 8: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell_{N,d}^{(\alpha)}(\theta, \phi; x)$ (averaged over 100 MC samples) for the VAE in Section 6.3, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d). Plotted in purple are curves of the form (49) with tailored values of c_1.

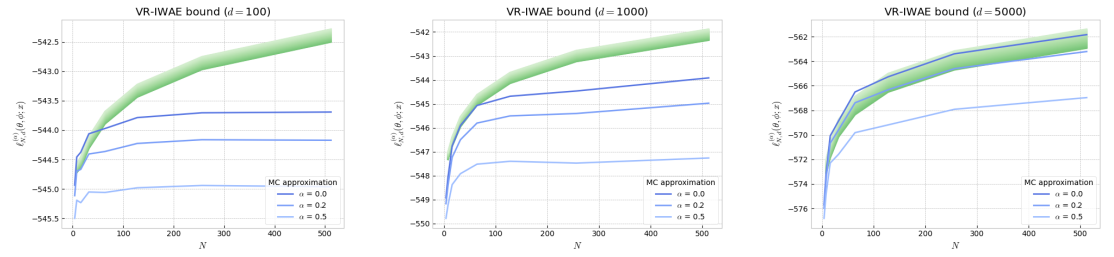

Figure 9: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell_{N,d}^{(\alpha)}(\theta, \phi; x)$ (averaged over 100 MC samples) for the VAE in Section 6.3, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d). Plotted in green are curves of the form (50) with tailored values of c_2.

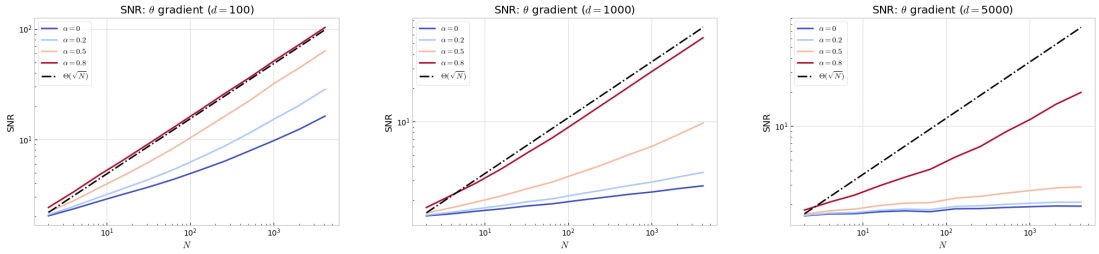

Figure 10: Plotted is the SNR of the generative network (θ) gradients in the reparameterized case (computed over 10000 MC samples) for the VAE in Section 6.3 as a function of N, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d).

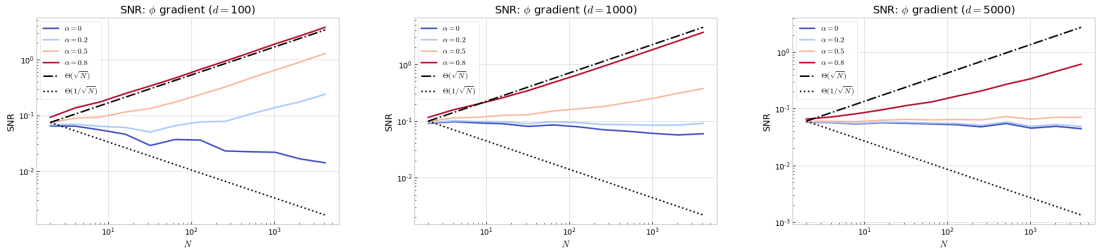

Figure 11: Plotted is the SNR of the inference network (ϕ) gradients in the reparameterized case (computed over 10000 MC samples) for the VAE in Section 6.3 as a function of N, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d).

We now look into the training of the VR-IWAE bound and more specifically into the SNR in medium to high dimensions at initialization, since this scenario corresponds to situations where the VR-IWAE bound seems to resemble more and more the behavior predicted by Theorem 6 (as observed in Figure 9). Following the methodology from the previous subsection, the results are presented in Figures 10 and 11, in which we have plotted the SNR for the generative network and for the inference network respectively in the reparameterized case alongside theoretical lines that reflect the asymptotic regimes predicted by Theorem 1.

As already observed in Section 6.2, the SNR benefits from setting $\alpha > 0$ and the asymptotic rates predicted by Theorem 1 do not capture the SNR behavior as d increases (unless N is unpractically large and/or we appeal to higher values of α). Furthermore, and as we can see in Figure 12, resorting to doubly-reparameterized estimators improves the SNR.

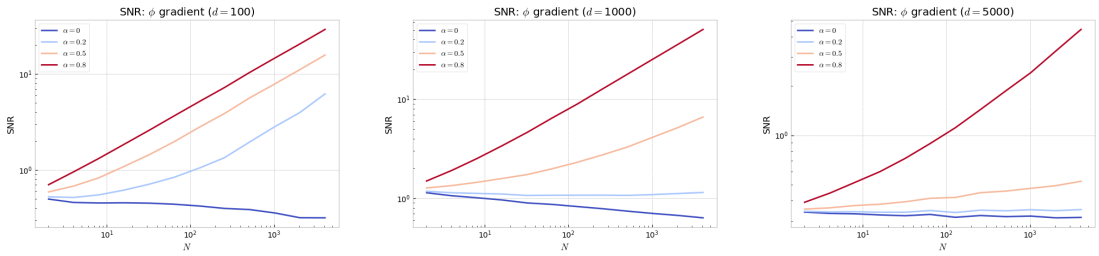

Figure 12: Plotted is the SNR of the inference network (ϕ) gradients in the doubly-reparameterized case (computed over 10000 MC samples) for the VAE in Section 6.3 as a function of N, for a randomly selected datapoint x in the test set, randomly generated model parameters (θ, ϕ) and varying values of (α, d).

We thus confirmed that approximately log-normal weights can arise in real data scenarios as d increases and that our theoretical study provides a useful framework to capture the impact of the weights on the VR-IWAE bound as a function of N, d and α. In line with our empirical findings for the SNR, we also obtained that the asymptotic rates predicted by Theorem 1 match the observed rates in low to medium dimensions and we postulated that the weight collapse occuring in the VR-IWAE bound as d increases may deteriorate the SNR too.

One aspect that remains unexplored empirically is the role of M in the VR-IWAE bound methodology, and in particular the interplay between M and N in the learning outcome. Indeed, the total number of samples needed per iteration in the gradient descent procedure is $N \times M$, with M being responsible for the usual $1/M$ variance reduction in gradient estimators such as (14). Intuitively, and following a similar line of reasoning as for α, we expect to see a bias-variance tradeoff between increasing M or N while keeping $M \times N$ fixed (we refer to Appendix D.3 for details).

Furthermore, if the weight collapse appearing in the VR-IWAE bound as d increases ends up badly impacting the associated gradient descent, our results indicate that practitioners should either (i) set $N = 1$ and allocate the maximum computational budget to M, which in fact corresponds to setting $\alpha = 1$ in the VR-IWAE bound, (ii) find more suitable variational approximations that can capture the complexity within the posterior density or (iii) resort to/construct better gradient estimates (e.g. doubly-reparameterized gradient estimators).

7. Conclusion

In this paper, we formalized the VR-IWAE bound, a variational bound depending on an hyperparameter $\alpha \in [0, 1)$ which generalizes the standard IWAE bound ($\alpha = 0$). We showed that the VR-IWAE bound provides theoretical guarantees behind various VR bound-based schemes proposed in the alpha-divergence variational inference literature and identified other additional desirable properties of this bound.

We then provided two complementary analyses of the variational gap, that is of the difference between the VR-IWAE bound and the marginal log-likelihood. The first analysis shed light on how α may play a role in reducing the variational gap. We then proposed a second analysis to better capture the behavior of the variational gap in high-dimensional scenarios, establishing that the variational gap suffers in this case from a damaging weight collapse phenomenon for all $\alpha \in [0, 1)$. Lastly, we illustrated our theoretical results over several toy and real-data examples.

Overall, our work provides foundations for improving the IWAE and VR methodologies and we now state potential directions of research to extend it. Firstly, one may investigate whether the weight collapse behavior applies beyond the cases we highlighted. Looking into how this weight collapse affects the gradient descent procedures associated to the VR-IWAE bound could be a second direction of research. Thirdly, and in order to improve on the VR-IWAE bound methodology beyond the weight collapse phenomenon, one may seek to further build on the fact that the VR-IWAE bound is the theoretically-sound extension of the IWAE bound that originates from the alpha-divergence variational inference methodology.

Acknowledgments

We would like to thank the action editor and the reviewers for helpful comments and suggestions on the paper. Kamélia Daudel and Arnaud Doucet acknowledge support of the UK Defence Science and Technology Laboratory (Dstl) and Engineering and Physical Research Council (EPSRC) under grant EP/R013616/1. This is part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative. Joe Benton was supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1). Arnaud Doucet also acknowledges support from the EPSRC grant EP/R034710/1.

Appendix A. Deferred Proofs of Section 3

A.1 Extension of $\ell_N^{(\alpha)}$ to the Case $\alpha = 1$

We prove that under common differentiability assumptions, the following limit holds:

$$\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = \text{ELBO}(\theta, \phi; x).$$

Proof Setting $f(\alpha) = \int \int \prod_{i=1}^N q_\phi(z_i) \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(z_j)^{1-\alpha} \right) \mathrm{d}z_{1:N}$, the bound $\ell_N^{(\alpha)}(\theta, \phi; x)$ can be rewritten as

$$\ell_N^{(\alpha)}(\theta, \phi; x) = -\frac{f(\alpha) - f(1)}{\alpha - 1}$$

and hence, $\lim_{\alpha \to 1} \ell_N^{(\alpha)}(\theta, \phi; x) = -f'(1)$. We then get the desired result by observing that

$$f'(\alpha) = \int \int \prod_{i=1}^N q_\phi(z_i) \left(\frac{\frac{1}{N} \sum_{j=1}^N - \log(\bar{w}_{\phi,\theta}(z_j)) \bar{w}_{\phi,\theta}(z_j)^{1-\alpha}}{\frac{1}{N} \sum_{j=1}^N \bar{w}_{\phi,\theta}(z_j)^{1-\alpha}} \right) \mathrm{d}z_{1:N}$$

and letting $\alpha \to 1$ in the quantity above. ■

A.2 Proof of Proposition 1

Proof of Proposition 1 The results for the case $\alpha = 0$ follow from Burda et al. (2016) and we focus on the case $\alpha \in (0, 1)$ in the proof below.

1. One the one hand, (Li and Turner, 2016, Theorem 1) implies that: for all $\alpha \in (0, 1)$,

$$\mathcal{L}^{(\alpha)}(\theta, \phi; x) \leq \ell(\theta; x). \tag{51}$$

   On the other hand, we obtain from (Li and Turner, 2016, Theorem 2) that:

   - For all $N \in \mathbb{N}^\star$ and all $\alpha < 1$

$$\ell_N^{(\alpha)}(\theta, \phi; x) \leq \ell_N^{(N+1)}(\theta, \phi; x) \leq \mathcal{L}^{(\alpha)}(\theta, \phi; x)$$

     which gives (10) when paired with (51).
   - If the function $z \mapsto w_{\theta,\phi}(z)$ is bounded, then $\ell_N^{(\alpha)}(\theta, \phi; x)$ approaches the VR bound $\mathcal{L}^{(\alpha)}(\theta, \phi; x)$ as N goes to infinity.

2. Let $1 > \alpha_1 > \alpha_2 > 0$. Then, the functions $u \mapsto u^{\frac{1-\alpha_1}{1-\alpha_2}}$ and $u \mapsto u^{1-\alpha_2}$ are concave for all $u > 0$ and hence Jensen's inequality implies

$$\begin{aligned}
\ell_N^{(\alpha_1)}(\theta, \phi; x) &= \frac{1}{1-\alpha_1} \int \int \prod_{i=1}^N q_\phi(z_i|x) \log \left(\frac{1}{N} \sum_{j=1}^N \left[w_{\theta,\phi}(z_i)^{1-\alpha_2} \right]^{\frac{1-\alpha_1}{1-\alpha_2}} \right) \mathrm{d}z_{1:N} \\
&\leq \ell_N^{(\alpha_2)}(\theta, \phi; x) \\
&\leq \ell_N^{(0)}(\theta, \phi; x)
\end{aligned}$$

   The desired result (11) follows by using that $\ell_N^{(0)}(\theta, \phi; x) = \ell_N^{(\mathrm{IWAE})}(\theta, \phi; x)$. As for the case of equality, it is obtained as the case of equality of Jensen's inequality.

3. Under the reparameterization trick,

$$\ell_N^{(\alpha)}(\theta, \phi; x) = \frac{1}{1-\alpha} \int \int \prod_{i=1}^N q(\varepsilon_i) \log \left(\frac{1}{N} \sum_{j=1}^N w_{\theta,\phi}(f(\varepsilon_j, \phi))^{1-\alpha} \right) \mathrm{d}\varepsilon_{1:N}$$

   leading, under common differentiability assumptions, to

$$\frac{\partial}{\partial \phi} \ell_N^{(\alpha)}(\theta, \phi; x) = \int \int \prod_{i=1}^N q(\varepsilon_i) \left(\sum_{j=1}^N \frac{w_{\theta,\phi}(f(\varepsilon_j, \phi))^{-\alpha} \frac{\partial}{\partial \phi} w_{\theta,\phi}(f(\varepsilon_j, \phi))}{\sum_{k=1}^N w_{\theta,\phi}(f(\varepsilon_k, \phi))^{1-\alpha}} \right) \mathrm{d}\varepsilon_{1:N}.$$

   The desired result (12) is then obtained using the REINFORCE trick

$$\frac{\partial}{\partial \phi} w_{\theta,\phi}(f(\varepsilon_j, \phi)) = w_{\theta,\phi}(f(\varepsilon_j, \phi)) \frac{\partial}{\partial \phi} \log w_{\theta,\phi}(f(\varepsilon_j, \phi))$$

   and the unbiased estimator (13) follows immediately.

■

A.3 Proof of Theorem 1

The proof of Theorem 1 is based on the proof of the corresponding result in (Rainforth et al., 2018, Theorem 1) (arxiv version of 5 Mar 2019). First, we prove the following useful lemma, which is an extension of (Rainforth et al., 2018, Lemma 1).

Lemma 3 *Suppose we have random variables $X_{i,j}$ for all $i = 1 \dots r$ and $j = 1 \dots N$ satisfying*

(i) $\mathbb{E}\left(X_{i,j}\right) = 0$ *for all* $i = 1 \dots r$ *and* $j = 1 \dots N$*;*

(ii) $\mathbb{E}\left(|X_{i,j}|^r\right) < \infty$ *for all* $i = 1 \dots r$ *and* $j = 1 \dots N$*;*

(iii) for each $i = 1 \dots r$*, the random variables* $X_{i,1}, \dots, X_{i,N}$ *are i.i.d.;*

(iv) for each $i = 1 \dots r$ *and* $j = 1 \dots N$*, the random variables* $X_{i,j}$ *and* $\{X_{i',j'}\}_{i'=1\dots r;\, j' \neq j}$ *are independent.*

Then

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} X_{1,j}\right) \dots \left(\frac{1}{N}\sum_{j=1}^{N} X_{r,j}\right)\right] = \begin{cases} O(N^{-r/2}) & \text{if } r \text{ is even,} \\ O(N^{-(r+1)/2}) & \text{if } r \text{ is odd.} \end{cases}$$

Proof We have

$$\left(\frac{1}{N}\sum_{j=1}^{N} X_{1,j}\right) \dots \left(\frac{1}{N}\sum_{j=1}^{N} X_{r,j}\right) = \frac{1}{N^r} \sum_{\{A_1,\dots,A_t\}} \sum_{(j_1,\dots,j_t)} \left(\prod_{i\in A_1} X_{i,j_1}\right) \dots \left(\prod_{i\in A_t} X_{i,j_t}\right) \quad (52)$$

where the first sum is over all partitions $\{A_1, \dots, A_t\}$ of the set $\{1, \dots, r\}$ (so that t is an integer between 1 and r for each partition) and the second sum is over all tuples $(j_1, \dots, j_t)$ with each element being a distinct integer between 1 and N (e.g. for the partition $\{A_1\}$ with $A_1 = \{1, \dots, r\}$ and $t = 1$, the second sum reduces to the sum over $j_1 = 1 \dots N$ and for the partition $\{A_1, \dots, A_r\}$ with $A_p = \{p\}$ for all $p = 1 \dots r$ and $t = r$, the second sum corresponds to the sum over all permutations over the subsets of length r of $(1, \dots, N)$).

Now consider the case where $|A_p| = 1$ for some integer p between 1 and t in a certain partition $\{A_1, \dots, A_t\}$. Without any loss of generality, we let $|A_1| = 1$. Then, by the independence of condition (iv) followed by condition (i), we have

$$\begin{aligned} \mathbb{E}\left[\left(\prod_{i\in A_1} X_{i,j_1}\right) \dots \left(\prod_{i\in A_t} X_{i,j_t}\right)\right] &= \mathbb{E}\left[X_{i^*,j_1}\right] \mathbb{E}\left[\left(\prod_{i\in A_2} X_{i,j_2}\right) \dots \left(\prod_{i\in A_t} X_{i,j_t}\right)\right] \\ &= 0 \end{aligned}$$

where i^* is the single element of A_1. Hence, we can restrict the sum over $\{A_1, \dots, A_t\}$ to only consider partitions where every partition has at least two elements. Furthermore, by the generalized Hölder's inequality (i.e. given the r random variables $X_1, \dots, X_r$, it holds that $\mathbb{E}(|\prod_{p=1}^r X_p|) \leq \prod_{p=1}^r \mathbb{E}(|X_p|^r)^{1/r}$) and conditions (ii) and (iii), we also have

$$\mathbb{E}\left[\left|\left(\prod_{i\in A_1} X_{i,j_1}\right) \dots \left(\prod_{i\in A_t} X_{i,j_t}\right)\right|\right] \leq \prod_{i=1}^{r} \mathbb{E}\left(|X_{i,1}|^r\right)^{1/r} < \infty,$$

where we have used that the product on the l.h.s. of the equation above contains exactly r terms since $\{A_1, \ldots, A_t\}$ is a partition of $\{1, \ldots, r\}$. Putting this together with (52) yields:

$$\left| \mathbb{E}\left[\left(\frac{1}{N} \sum_{j=1}^{N} X_{1,j} \right) \ldots \left(\frac{1}{N} \sum_{j=1}^{N} X_{r,j} \right) \right] \right| \leq \frac{1}{N^r} \sum_{\substack{\{A_1,\ldots,A_t\} \\ \text{all } |A_p| \geq 2}} \sum_{(j_1,\ldots,j_t)} \left(\prod_{i=1}^{r} \mathbb{E}\left(|X_{i,1}|^r \right)^{1/r} \right)$$

$$\leq \frac{1}{N^r} \sum_{\substack{\{A_1,\ldots,A_t\} \\ \text{all } |A_p| \geq 2}} N^t \left(\prod_{i=1}^{r} \mathbb{E}\left(|X_{i,1}|^r \right)^{1/r} \right).$$

Finally, note that (i) any partition $\{A_1, \ldots, A_t\}$ of $\{1, \ldots, r\}$ where each part has size at least 2 can have at most $\lfloor r/2 \rfloor$ parts, so $t \leq \lfloor r/2 \rfloor$ and (ii) we can crudely bound the number of partitions by r^r. Hence

$$\left| \mathbb{E}\left[\left(\frac{1}{N} \sum_{j=1}^{N} X_{1,j} \right) \ldots \left(\frac{1}{N} \sum_{j=1}^{N} X_{r,j} \right) \right] \right| \leq \frac{r^r}{N^{r - \lfloor r/2 \rfloor}} \left(\prod_{i=1}^{r} \mathbb{E}\left(|X_{i,1}|^r \right)^{1/r} \right)$$

$$= \begin{cases} O(N^{-r/2}) & \text{if } r \text{ is even} \\ O(N^{-(r+1)/2}) & \text{if } r \text{ is odd.} \end{cases}$$

■

We next prove a second lemma.

Lemma 4 *Let k be a positive integer. Set $R_{\alpha,N} = N^{-1} \sum_{i=1}^{N} w_{\theta,\phi}(Z_i)^{1-\alpha}$, where $Z_1, \ldots, Z_N$ are i.i.d. samples generated according to q_ϕ. Then, the condition*

$$\limsup_{N \to \infty} \mathbb{E}\left((1/R_{\alpha,N})^k \right) < \infty \tag{53}$$

is equivalent to the statement that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/R_{\alpha,N})^k) < \infty$.

Proof of Lemma 4 Fix a positive integer $N \geq 2$. For all $x \in [0, 1)$, we have by convexity of the function $x \mapsto (1-x)^{-k}$ that

$$\left(\frac{1}{1-x} \right)^k \geq \left(\frac{N}{N-1} \right)^k + k \left(\frac{N}{N-1} \right)^{k+1} \left(x - \frac{1}{N} \right).$$

It follows that if $x_1, \ldots, x_N \in (0, 1)$ are such that $\sum_{i=1}^{N} x_i = 1$, then

$$\frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{1-x_i} \right)^k \geq \left(\frac{N}{N-1} \right)^k.$$

Given $\alpha \in [0, 1)$ and N positive reals $w_1, \ldots, w_N$, we may set $x_i = w_i^{1-\alpha} / \left(\sum_{i=1}^{n} w_i^{1-\alpha} \right)$ in the above to get

$$\frac{1}{N} \sum_{i=1}^{N} \left(\frac{N-1}{\sum_{\substack{j=1 \\ j \neq i}}^{N} w_j^{1-\alpha}} \right)^k \geq \left(\frac{N}{\sum_{j=1}^{N} w_j^{1-\alpha}} \right)^k.$$

Now consider setting $w_i = w_{\theta,\phi}(Z_i)$ where the Z_i are i.i.d. samples generated according to q_ϕ. We see that the r.h.s. of the above expression is distributed as $(1/R_{\alpha,N})^k$, while each term in the sum on the l.h.s. is distributed as $(1/R_{\alpha,N-1})^k$. We conclude that

$$\mathbb{E}\left((1/R_{\alpha,N-1})^k\right) \geq \mathbb{E}\left((1/R_{\alpha,N})^k\right)$$

for all $N \geq 2$. Hence $\mathbb{E}\left((1/R_{\alpha,N})^k\right)$ is decreasing in N, so $\limsup_{N\to\infty} \mathbb{E}\left((1/R_{\alpha,N})^k\right) < \infty$ if and only if there exists some N such that $\mathbb{E}\left((1/R_{\alpha,N})^k\right) < \infty$. ■

We now move on to the proof of Theorem 1.

Proof of Theorem 1 We use the following shorthand notation

$$\begin{aligned}
&\tilde{w}_{m,j} = w_{\theta,\phi}(f(\varepsilon_{m,j},\phi)), \quad m = 1\dots M,\ j = 1\dots N \\
&Z_\alpha = \mathbb{E}_{\varepsilon\sim q}(w_{\theta,\phi}(f(\varepsilon,\phi))^{1-\alpha})
\end{aligned}$$

and we also recall the notation

$$\hat{Z}_{1,N,\alpha} = \frac{1}{N}\sum_{j=1}^{N} \tilde{w}_{1,j}^{1-\alpha}.$$

We will first prove that

$$\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)] = \sqrt{M}\frac{\left|\sqrt{N}\frac{\partial Z_\alpha}{\partial\theta_\ell} - \frac{Z_\alpha}{2\sqrt{N}}\frac{\partial}{\partial\theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^{3/2}}\right)\right|}{\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell} - \frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right)} + O\left(\frac{1}{N}\right)} \tag{54}$$

$$\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\phi_{\ell'})] = \sqrt{M}\frac{\left|\sqrt{N}\frac{\partial Z_\alpha}{\partial\phi_{\ell'}} - \frac{Z_\alpha}{2\sqrt{N}}\frac{\partial}{\partial\phi_{\ell'}}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^{3/2}}\right)\right|}{\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\phi_{\ell'}} - \frac{\partial\log Z_\alpha}{\partial\phi_{\ell'}}\right]^2\right)} + O\left(\frac{1}{N}\right)}. \tag{55}$$

As the two expressions above follow the same form, it is in fact enough to only prove (54). We will do so by studying the asymptotic variance and expected value of $\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell) := (1-\alpha)\delta^{(\alpha)}_{M,N}(\theta_\ell)$ separately, before combining them to deduce (54).

- **Study of $\mathbb{V}(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell))$.**

We start from the identity

$$\frac{\partial\log\hat{Z}_{1,N,\alpha}}{\partial\theta_\ell} = \frac{\partial\log Z_\alpha}{\partial\theta_\ell} + \frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right),$$

which can for example be verified by using the following identity (which is a version of the Taylor expansion to first order with an explicit form for the remainder)

$$\log(1+x) = x - \int_0^x \frac{t}{1+t}dt,$$

substituting $x = (\hat{Z}_{1,N,\alpha} - Z_\alpha)/Z_\alpha$, differentiating with respect to θ_ℓ and using the chain rule where necessary. Hence,

$$\begin{aligned}
M \cdot \mathbb{V}\left(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell)\right) &= \mathbb{V}\left(\tilde{\delta}^{(\alpha)}_{1,N}(\theta_\ell)\right) = \mathbb{V}\left(\frac{\partial \log(\hat{Z}_{1,N,\alpha})}{\partial \theta_\ell}\right) \\
&= \mathbb{V}\left(\frac{\partial \log Z_\alpha}{\partial \theta_\ell} + \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \\
&= \mathbb{V}\left(\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \\
&= \mathbb{V}\left(\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \\
&\quad + 2\,\mathrm{Cov}\left(\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right), \left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \\
&\quad + \mathbb{V}\left(\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right)
\end{aligned}$$

Furthermore, observe that

$$\begin{aligned}
\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) &= \frac{1}{N}\sum_{j=1}^{N} \frac{\partial}{\partial \theta_\ell}\left(\frac{\tilde{w}_{1,j}^{1-\alpha} - Z_\alpha}{Z_\alpha}\right) \\
&= \frac{1}{N}\sum_{j=1}^{N} \frac{Z_\alpha \frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial \theta_\ell} - \tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial \theta_\ell}}{Z_\alpha^2}. \qquad (56)
\end{aligned}$$

As a result,

$$\mathbb{E}\left[\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right] = \frac{\mathbb{E}[\frac{\partial(\tilde{w}_{1,1}^{1-\alpha})}{\partial \theta_\ell}] - \frac{\partial Z_\alpha}{\partial \theta_\ell}}{Z_\alpha} = 0, \qquad (57)$$

where we have used that under common differentiability assumptions $\mathbb{E}[\partial(\tilde{w}_{1,1}^{1-\alpha})/\partial\theta_\ell] = \partial Z_\alpha/\partial\theta_\ell$, that is we can interchange the order of integration and differentiation. Consequently, we can simplify the expression of $M \cdot \mathbb{V}\left(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell)\right)$ to obtain that

$$\begin{aligned}
M \cdot \mathbb{V}\left(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell)\right) &= \mathbb{E}\left(\left[\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right]^2\right) \\
&\quad + 2\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \left[\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right]^2\right) \\
&\quad + \mathbb{V}\left(\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right) \cdot \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right) \qquad (58)
\end{aligned}$$

We now control the three terms in the r.h.s. of (58) separately.

1. *First term in the r.h.s. of* (58). To control the first term in the r.h.s. of (58), we use (56) to get that

$$\begin{aligned}\mathbb{E}\left(\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right) &= \frac{1}{N^2}\mathbb{E}\left(\sum_{j=1}^N\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)\\ &= \frac{1}{N^2}\mathbb{E}\left(\sum_{j=1}^N\left[\frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2}\right]^2\right)\\ &= \frac{1}{NZ_\alpha^4}\mathbb{E}\left(\left[Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}\right]^2\right)\\ &= \frac{1}{NZ_\alpha^4}\mathbb{E}\left(\left[\tilde{w}_{1,1}^{-\alpha}\left\{(1-\alpha)Z_\alpha\frac{\partial\tilde{w}_{1,1}}{\partial\theta_\ell}-\tilde{w}_{1,1}\frac{\partial Z_\alpha}{\partial\theta_\ell}\right\}\right]^2\right),\end{aligned} \tag{59}$$

where the cross-terms disappeared due to the independence of the $(\varepsilon_{1,j})_{1\leq j\leq N}$ paired up with (57).

2. *Second term in the r.h.s of* (58). We deal with the cross term in (58) by splitting it into two parts

$$\begin{aligned}&\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)\\ &\quad=\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\cdot\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)\\ &\quad+\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\cdot\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)\right).\end{aligned} \tag{60}$$

Using the expression of $\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)$ given in (56) and the fact that

$$\hat{Z}_{1,N,\alpha}-Z_\alpha=\frac{1}{N}\sum_{j=1}^N\left(\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha\right), \tag{61}$$

we set: for all $j=1\dots J$,

$$\begin{aligned}X_{1,j}&=\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha,\\ X_{2,j}=X_{3,j}&=\frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2}\end{aligned}$$

and we can then apply Lemma 3 with $r=3$ (by noting in particular that the required moments are finite under our assumptions): we thus obtain that

$$\mathbb{E}\left(\left(\hat{Z}_{1,N,\alpha}-Z_\alpha\right)\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)=O\left(\frac{1}{N^2}\right)$$

which controls the first term in the r.h.s. of (60). The second term in the r.h.s. of (60) can be bounded as follows

$$\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)\right)$$
$$\leq \mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)^2\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^4\right)^{1/2}\mathbb{E}\left(\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)^2\right)^{1/2}. \quad (62)$$

By taking this time: for all $j = 1 \dots N$,

$$X_{1,j} = X_{2,j} = \tilde{w}_{1,j}^{1-\alpha} - Z_\alpha,$$
$$X_{3,j} = X_{4,j} = X_{5,j} = X_{6,j} = \frac{Z_\alpha \frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell} - \tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2},$$

in Lemma 3 with $r = 6$ (and noting once again that the required moments are finite under our assumptions), we see that the first term in the r.h.s. of (62) is $O(N^{-3/2})$. As for the second term of the r.h.s. of (62), Cauchy-Schwarz implies that

$$\mathbb{E}\left(\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)^2\right) \leq \mathbb{E}\left(\frac{1}{\hat{Z}_{1,N,\alpha}^4}\right)^{1/2}\mathbb{E}\left(\left(Z_\alpha-\hat{Z}_{1,N,\alpha}\right)^4\right)^{1/2}.$$

Now note that under our assumptions, Lemma 4 can be applied with $k = 4$ so that (53) with $k = 4$ holds and controls the first term in the r.h.s. above. Furthermore, applying Lemma 3 with $r = 4$ and for all $j = 1 \dots N$,

$$X_{1,j} = X_{2,j} = X_{3,j} = X_{4,j} = \tilde{w}_{1,j}^{1-\alpha} - Z_\alpha$$

yields

$$\mathbb{E}\left(\left(Z_\alpha-\hat{Z}_{1,N,\alpha}\right)^4\right) = O\left(\frac{1}{N^2}\right).$$

We can then conclude that

$$\mathbb{E}\left(\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)^2\right)^{1/2} = O\left(\frac{1}{N^{1/2}}\right), \quad (63)$$

and so the r.h.s. of (62) is bounded above by $O(N^{-2})$. It follows that the second term in the r.h.s. of (60) is $O(N^{-2})$ too and we can conclude that

$$\mathbb{E}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right) = O\left(\frac{1}{N^2}\right) \quad (64)$$

that is the second term of the r.h.s. of (58) is $O(N^{-2})$.

3. *Third term of the r.h.s. in* (58). For the third term of the r.h.s. in (58), note that

$$\begin{aligned}
&\mathbb{V}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right)\\
&\leq\mathbb{E}\left(\left[\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)\\
&\leq\mathbb{E}\left(\left[\left(\hat{Z}_{1,N,\alpha}-Z_\alpha\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^4\right)^{1/2}\mathbb{E}\left(\frac{1}{\hat{Z}_{1,N,\alpha}^4}\right)^{1/2}
\end{aligned}$$

where the final line follows from Cauchy–Schwarz. As a result, using (56) and (61), taking for all $j=1\dots N$

$$\begin{aligned}
X_{1,j}=X_{2,j}=X_{3,j}=X_{4,j}&=\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha\\
X_{5,j}=X_{6,j}=X_{7,j}=X_{8,j}&=\frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2},
\end{aligned}$$

and since the required moments are finite under our assumptions, Lemma 3 with $r=8$ implies that

$$\mathbb{E}\left(\left[\left(\hat{Z}_{1,N,\alpha}-Z_\alpha\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^4\right)=O\left(\frac{1}{N^4}\right).$$

Combined with (53) with $k=4$ (which holds under our assumptions by Lemma 4 with $k=4$), this implies that

$$\mathbb{V}\left(\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{\hat{Z}_{1,N,\alpha}}\right)\cdot\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right)=O\left(\frac{1}{N^2}\right).\tag{65}$$

Putting (58), (59), (64) and (65) together, we see that

$$\begin{aligned}
\mathbb{V}\left(\tilde{\delta}_{M,N}^{(\alpha)}(\theta_\ell)\right)&=\frac{1}{MNZ_\alpha^4}\mathbb{E}\left(\left[\tilde{w}_{1,1}^{-\alpha}\left\{(1-\alpha)Z_\alpha\frac{\partial\tilde{w}_{1,1}}{\partial\theta_\ell}-\tilde{w}_{1,1}\frac{\partial Z_\alpha}{\partial\theta_\ell}\right\}\right]^2\right)+O\left(\frac{1}{MN^2}\right)\\
&=\frac{1}{MNZ_\alpha^2}\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell}-\frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right)+O\left(\frac{1}{MN^2}\right)
\end{aligned}$$

and it follows that

$$\begin{aligned}
\sqrt{\mathbb{V}(\tilde{\delta}_{M,N}^{(\alpha)}(\theta_\ell))}&=\frac{1}{\sqrt{MN}Z_\alpha}\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell}-\frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right)+O\left(\frac{1}{N}\right)}\\
&=\frac{1}{\sqrt{MN}Z_\alpha}\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell}-\frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right)}\sqrt{1+O\left(\frac{1}{N}\right)}\\
&=\frac{1}{\sqrt{MN}Z_\alpha}\left(\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial\log\tilde{w}_{1,1}}{\partial\theta_\ell}-\frac{\partial\log Z_\alpha}{\partial\theta_\ell}\right]^2\right)}+O\left(\frac{1}{N}\right)\right).
\end{aligned}\tag{66}$$

- **Study of $\mathbb{E}(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell))$.**

We start from the identity

$$\frac{\partial \log \hat{Z}_{1,N,\alpha}}{\partial \theta_\ell} = \frac{\partial \log Z_\alpha}{\partial \theta_\ell} + \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \frac{1}{2}\frac{\partial}{\partial \theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right]^2\right) + \left(\frac{(\hat{Z}_{1,N,\alpha} - Z_\alpha)^2}{Z_\alpha \cdot \hat{Z}_{1,N,\alpha}}\right)\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right),$$

which can for example be proved using the following identity (which is a version of the Taylor expansion to second order with an explicit form for the remainder)

$$\log(1+x) = x - \frac{x^2}{2} + \int_0^x \frac{t^2}{1+t}\mathrm{d}t,$$

substituting $x = (\hat{Z}_{1,N,\alpha} - Z_\alpha)/Z_\alpha$, differentiating with respect to θ_ℓ and using the chain rule where necessary. It follows that

$$\begin{aligned}
\mathbb{E}\left(\tilde{\delta}^{(\alpha)}_{M,N}(\theta_\ell)\right) &= \mathbb{E}\left(\tilde{\delta}^{(\alpha)}_{1,N}(\theta_\ell)\right) = \mathbb{E}\left(\frac{\partial \log \hat{Z}_{1,N,\alpha}}{\partial \theta_\ell}\right) \\
&= \mathbb{E}\Bigg(\frac{\partial \log Z_\alpha}{\partial \theta_\ell} + \frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right) - \frac{1}{2}\frac{\partial}{\partial \theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right]^2\right) \\
&\qquad + \left(\frac{(\hat{Z}_{1,N,\alpha} - Z_\alpha)^2}{Z_\alpha \cdot \hat{Z}_{1,N,\alpha}}\right)\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\Bigg) \\
&= \frac{\partial \log Z_\alpha}{\partial \theta_\ell} - \frac{1}{2}\mathbb{E}\left(\frac{\partial}{\partial \theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right]^2\right)\right) + R_2(\hat{Z}_{1,N,\alpha}) \\
&= \frac{\partial \log Z_\alpha}{\partial \theta_\ell} - \frac{1}{2N}\frac{\partial}{\partial \theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}^{1-\alpha}_{1,1})}{Z_\alpha^2}\right] + R_2(\hat{Z}_{1,N,\alpha}) \qquad (67)
\end{aligned}$$

where we denote

$$R_2(\hat{Z}_{1,N,\alpha}) = \mathbb{E}\left(\left(\frac{(\hat{Z}_{1,N,\alpha} - Z_\alpha)^2}{Z_\alpha \cdot \hat{Z}_{1,N,\alpha}}\right)\frac{\partial}{\partial \theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha} - Z_\alpha}{Z_\alpha}\right)\right)$$

and where we have used (56), (57), (61) and the fact that under common differentiability assumptions, we have that

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial}{\partial\theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right]^2\right)\right) &= 2\mathbb{E}\left(\left[\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right]\frac{\partial}{\partial\theta_\ell}\left(\left[\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right]\right)\right) \\
&= \frac{2}{N^2}\mathbb{E}\left(\sum_{j=1}^{N}\frac{\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha}{Z_\alpha}\cdot\frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2}\right) \\
&= \frac{1}{N}\mathbb{E}\left(\frac{\partial}{\partial\theta_\ell}\left(\left[\frac{\tilde{w}_{1,j}^{1-\alpha}-Z_\alpha}{Z_\alpha}\right]^2\right)\right) \\
&= \frac{1}{N}\frac{\partial}{\partial\theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right]
\end{aligned}$$

(here the cross-terms disappear due to the independence of the $(\varepsilon_{1,j})_{1\leq j\leq N}$ paired up with (57)). Notice then that we can split up $R_2(\hat{Z}_{1,N,\alpha})$ as

$$\begin{aligned}
R_2(\hat{Z}_{1,N,\alpha}) = \mathbb{E}&\left(\left(\frac{(\hat{Z}_{1,N,\alpha}-Z_\alpha)^2}{Z_\alpha^2}\right)\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right) \\
&+\mathbb{E}\left(\left(\frac{(\hat{Z}_{1,N,\alpha}-Z_\alpha)^2}{Z_\alpha^2}\right)\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right) \quad (68)
\end{aligned}$$

The first term in (68) can be bounded by applying Lemma 3 with $r=3$ and for all $j=1\dots N$,

$$\begin{aligned}
X_{1,j} = X_{2,j} &= \tilde{w}_{1,j}^{1-\alpha}-Z_\alpha, \\
X_{3,j} &= \frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2},
\end{aligned}$$

noting the required moments are finite under our assumptions, so that

$$\mathbb{E}\left(\left(\hat{Z}_{1,N,\alpha}-Z_\alpha\right)^2\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right)=O\left(\frac{1}{N^2}\right).$$

The second term in (68) can be bounded using Cauchy-Schwarz as follows

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{(\hat{Z}_{1,N,\alpha}-Z_\alpha)^2}{Z_\alpha^2}\right)\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right) \\
&\leq \mathbb{E}\left(\left(\frac{(\hat{Z}_{1,N,\alpha}-Z_\alpha)^2}{Z_\alpha^2}\right)^2\left[\frac{\partial}{\partial\theta_\ell}\left(\frac{\hat{Z}_{1,N,\alpha}-Z_\alpha}{Z_\alpha}\right)\right]^2\right)^{1/2}\mathbb{E}\left(\left(\frac{Z_\alpha}{\hat{Z}_{1,N,\alpha}}-1\right)^2\right)^{1/2}
\end{aligned}$$

The second term is $O(N^{-1/2})$ by (63), while the first can be bounded by $O(N^{-3/2})$ using Lemma 3 with $r=6$ and for all $j=1\dots N$:

$$\begin{aligned}
X_{1,j} = X_{2,j} = X_{3,j} = X_{4,j} &= \tilde{w}_{1,j}^{1-\alpha}-Z_\alpha, \\
X_{5,j} = X_{6,j} &= \frac{Z_\alpha\frac{\partial(\tilde{w}_{1,j}^{1-\alpha})}{\partial\theta_\ell}-\tilde{w}_{1,j}^{1-\alpha}\frac{\partial Z_\alpha}{\partial\theta_\ell}}{Z_\alpha^2}.
\end{aligned}$$

Hence by combining with (67), we have

$$\mathbb{E}\left(\tilde{\delta}^{(\alpha)}_{M,N}\right) = \frac{\partial \log Z_\alpha}{\partial \theta_\ell} - \frac{1}{2N}\frac{\partial}{\partial \theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^2}\right) \tag{69}$$

- **Deducing** $\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)]$**.**

Finally, putting (66) and (69) together, we get

$$\begin{aligned}\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)] &= \frac{\left|\frac{\partial \log Z_\alpha}{\partial \theta_\ell} - \frac{1}{2N}\frac{\partial}{\partial \theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^2}\right)\right|}{\frac{1}{\sqrt{MN}Z_\alpha}\left(\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial \log \tilde{w}_{1,1}}{\partial \theta_\ell} - \frac{\partial \log Z_\alpha}{\partial \theta_\ell}\right]^2\right)} + O\left(\frac{1}{N}\right)\right)} \\ &= \sqrt{M}\frac{\left|\sqrt{N}\frac{\partial Z_\alpha}{\partial \theta_\ell} - \frac{Z_\alpha}{2\sqrt{N}}\frac{\partial}{\partial \theta_\ell}\left[\frac{\mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})}{Z_\alpha^2}\right] + O\left(\frac{1}{N^{3/2}}\right)\right|}{\sqrt{\mathbb{E}\left(\tilde{w}_{1,1}^{2(1-\alpha)}\left[(1-\alpha)\frac{\partial \log \tilde{w}_{1,1}}{\partial \theta_\ell} - \frac{\partial \log Z_\alpha}{\partial \theta_\ell}\right]^2\right)} + O\left(\frac{1}{N}\right)}\end{aligned}$$

which is exactly (54). Since we have assumed that $\frac{\partial Z_\alpha}{\partial \theta_\ell}$ is non-zero and since this term corresponds to the leading order term, we then deduce that

$$\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\theta_\ell)] = \Theta(\sqrt{MN})$$

and we thus recover (17).

Similarly, (55) holds for $\mathrm{SNR}[\delta^{(\alpha)}_{M,N}(\phi_{\ell'})]$ and we obtain the desired result (18) by splitting the cases $\alpha \in (0,1)$ and $\alpha = 0$. In the former, we have $\partial Z_\alpha/\partial \phi_{\ell'} = \partial \mathbb{E}(\tilde{w}_{1,1}^{1-\alpha})/\partial \phi_{\ell'} \neq 0$ by (16), so the leading order term is $\Theta(\sqrt{MN})$, while in the latter case we have $\partial Z_\alpha/\partial \phi_{\ell'} = 0$ while $\partial \mathbb{V}(\tilde{w}_{1,1}^{1-\alpha})/\partial \phi_{\ell'} > 0$ and so the leading order term is $\Theta(\sqrt{M/N})$ ■

A.4 Proof of Theorem 2

Proof of Theorem 2 Recall from Proposition 1 that

$$\frac{\partial}{\partial \phi}\ell_N^{(\alpha)}(\theta, \phi; x) = \int\int \prod_{i=1}^N q(\varepsilon_i)\left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial \phi}\log w_{\theta,\phi}(f(\varepsilon_j, \phi))\right)\mathrm{d}\varepsilon_{1:N}.$$

We will now follow the reasoning of Tucker et al. (2019). To do so, we expand the total derivative of $\ell_N^{(\alpha)}$ with respect to ϕ by using that

$$\frac{\partial}{\partial \phi}\log w_{\theta,\phi}(f(\varepsilon_j, \phi)) = -\frac{\partial}{\partial \phi}\log q_\phi(f(\varepsilon_j, \phi'))|_{\phi'=\phi} + \frac{\partial}{\partial \phi}f(\varepsilon_j, \phi)\ \frac{\partial}{\partial z_j}\log w_{\theta,\phi}(z_j)$$

which gives

$$\begin{aligned}\frac{\partial}{\partial\phi}\ell_N^{(\alpha)}(\theta,\phi;x) &= -\int\int\prod_{i=1}^N q(\varepsilon_i)\left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(f(\varepsilon_j,\phi'))|_{\phi'=\phi}\right)\mathrm{d}\varepsilon_{1:N}\\ &\quad+\int\int\prod_{i=1}^N q(\varepsilon_i)\left(\sum_{j=1}^N \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}f(\varepsilon_j,\phi)\ \frac{\partial}{\partial z_j}\log w_{\theta,\phi}(z_j)\right)\mathrm{d}\varepsilon_{1:N}\\ &:= -A+B. \end{aligned}\tag{70}$$

Notice now that

$$\begin{aligned}A &= \sum_{j=1}^N\int\int\prod_{i=1}^N q(\varepsilon_i)\left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(f(\varepsilon_j,\phi'))|_{\phi'=\phi}\right)\mathrm{d}\varepsilon_{1:N}\\ &= \sum_{j=1}^N\int\int\prod_{i=1}^N q_\phi(z_i)\left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(z_j')|_{z_j'=z_j}\right)\mathrm{d}z_{1:N},\end{aligned}\tag{71}$$

where we have set $z_j' = f(\varepsilon_j,\phi')$. Observe in addition that for all $j=1\dots N$, the reparameterization trick implies:

$$\begin{aligned}&\int q_\phi(z_j)\ \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(z_j')|_{z_j'=z_j}\ \mathrm{d}z_j\\ &\qquad\qquad=\int q(\varepsilon_j)\ \frac{\partial}{\partial z_j}\left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\right)\frac{\partial}{\partial\phi}f(\varepsilon_j,\phi)\ \mathrm{d}\varepsilon_j\end{aligned}$$

and hence

$$\begin{aligned}&\int q_\phi(z_j)\ \frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\frac{\partial}{\partial\phi}\log q_\phi(z_j')|_{z_j=z_j'}\ \mathrm{d}z_j\\ &=(1-\alpha)\int q(\varepsilon_j)\left[\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}-\left(\frac{w_{\theta,\phi}(z_j)^{1-\alpha}}{\sum_{k=1}^N w_{\theta,\phi}(z_k)^{1-\alpha}}\right)^2\right]\frac{\partial}{\partial\phi}f(\varepsilon_j,\phi)\frac{\partial}{\partial z_j}\log w_{\theta,\phi}(z_j)\,\mathrm{d}\varepsilon_j.\end{aligned}$$

The desired equality (19) is then obtained by combining the last equality above with (70) and (71). ■

Appendix B. Deferred Proofs and Results of Section 4

B.1 Proof of Theorem 3

Proof of Theorem 3 For convenience in the proof, let us first introduce the notation

$$R_\alpha = w_{\theta,\phi}(Z)^{1-\alpha}$$

with $Z\sim q_\phi$ and let us observe that under (A1) we have that $\mathbb{E}(R_\alpha)>0$. Furthermore, (A1) and Jensen's inequality applied to the concave function $u\mapsto u^{1-\alpha}$ yield

$$\mathbb{E}(R_\alpha)\le p_\theta(x)^{1-\alpha}<\infty,\tag{72}$$

meaning that (22) holds. Now decompose the variational gap into the two following terms:

$$\Delta_N^{(\alpha)}(\theta,\phi;x) = \left[\ell_N^{(\alpha)}(\theta,\phi;x) - \mathcal{L}^{(\alpha)}(\theta,\phi;x)\right] + \left[\mathcal{L}^{(\alpha)}(\theta,\phi;x) - \ell(\theta;x)\right].$$

To get the desired result (25), we only need to study the behavior of the term inside the first bracket. This will be done via an adaptation of the proof of (Domke and Sheldon, 2018, Theorem 3) to our more general framework, which is provided here for the sake of completeness. We write

$$\ell_N^{(\alpha)}(\theta,\phi;x) - \mathcal{L}^{(\alpha)}(\theta,\phi;x) = \frac{1}{1-\alpha}\mathbb{E}\left(\log\left(1+\delta_{\alpha,N}\right)\right)$$

where for all $z \in \mathbb{R}^d$,

$$\delta_{\alpha,N} = \frac{R_{\alpha,N}}{\mathbb{E}(R_\alpha)} - 1 \in (-1,\infty).$$

The second-order Taylor expansion of $\log\left(1+\delta_{\alpha,N}\right)$ gives

$$\log\left(1+\delta_{\alpha,N}\right) = \delta_{\alpha,N} - \frac{1}{2}\delta_{\alpha,N}^2 + \int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x.$$

Now using that $\mathbb{E}(\delta_{\alpha,N}) = 0$ and that $\mathbb{E}(\delta_{\alpha,N}^2) = \mathbb{V}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z))/N$, we deduce

$$\ell_N^{(\alpha)}(\theta,\phi;x) - \mathcal{L}^{(\alpha)}(\theta,\phi;x) = -\frac{\gamma_\alpha^2}{2N} + \frac{1}{1-\alpha}\mathbb{E}\left(\int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x\right).$$

All that is left to prove is then that

$$\lim_{N\to\infty} N\left|\mathbb{E}\left(\int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x\right)\right| = 0. \tag{73}$$

By (Domke and Sheldon, 2018, Lemma 7), we have that for all $\varepsilon > 0$ and all $\beta \in (0,1]$ there exist positive constants C_ε and D_β such that

$$\left|\int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x\right| \leq C_\varepsilon \left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}} + D_\beta |\delta_{\alpha,N}|^{2+\beta}$$

and as a result

$$N\left|\mathbb{E}\left(\int_0^{\delta_{\alpha,N}} \frac{x^2}{1+x}\mathrm{d}x\right)\right| \leq C_\varepsilon N\mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}}\right) + D_\beta N\mathbb{E}\left(|\delta_{\alpha,N}|^{2+\beta}\right). \tag{74}$$

Recall that under our assumptions, there exists $\beta > 0$ such that (23) holds. Without loss of generality, one can assume that $\beta \in (0,1]$. [Indeed, assuming that $\beta > 1$, we can find $0 < \beta' \leq 1 < \beta$ so that

$$\mathbb{E}_{Z\sim q_\phi}(|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1|^{2+\beta'}) < \infty,$$

which follows from Jensen's inequality applied to the concave function $u \mapsto u^{(2+\beta')/(2+\beta)}$ and from (23).] Let us now show that the two terms in (74) go to 0 as $N \to \infty$ for a suitable choice of ε ($\varepsilon = \beta/3$).

- First term of (74). Observe first that Hölder's inequality with $p = (1+\varepsilon)/\varepsilon$ and $q = 1+\varepsilon$ implies the following:

$$\mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}}\right) \leq \mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|\right)^{\frac{\varepsilon}{1+\varepsilon}} \mathbb{E}\left(|\delta_{\alpha,N}|^{2+3\varepsilon}\right)^{\frac{1}{1+\varepsilon}}.$$

From there, we deduce that

$$\limsup_{N\to\infty} N\mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}}\right) \leq \limsup_{N\to\infty} \mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|\right)^{\frac{\varepsilon}{1+\varepsilon}} \limsup_{N\to\infty}\left[N\mathbb{E}\left(|\delta_{\alpha,N}|^{2+3\varepsilon}\right)^{\frac{1}{1+\varepsilon}}\right],$$

having used that for any two sequences of non-negative real numbers $(a_N)_{N\in\mathbb{N}^*}$ and $(b_N)_{N\in\mathbb{N}^*}$, $\limsup_{N\to\infty}(a_N b_N) \leq \limsup_{N\to\infty} a_N \cdot \limsup_{N\to\infty} b_N$.

We then obtain that the first limit is bounded by a constant by appealing to (24). [Indeed, (24) means that for sufficiently large N, $\mathbb{E}(1/R_{\alpha,N})$ is bounded by a constant, and hence so is $\mathbb{E}(|1/(1+\delta_{\alpha,N})|)$ by combining the boundedness of $\mathbb{E}(1/R_{\alpha,N})$ with (72)].

As for the second limit, (Domke and Sheldon, 2018, Lemma 5) with $s = 2+3\varepsilon \geq 2$ and $U_i = \overline{w}^{(\alpha)}_{\theta,\phi}(Z_i) - 1$ implies that there exists a constant $B_\varepsilon > 0$ such that

$$\mathbb{E}\left(|\delta_{\alpha,N}|^{2+3\varepsilon}\right) \leq B_\varepsilon N^{-(2+3\varepsilon)/2} \mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}^{(\alpha)}_{\theta,\phi}(Z) - 1\right|^{2+3\varepsilon}\right).$$

Setting $\varepsilon = \beta/3$, we can rewrite the term on the r.h.s. as

$$B_{\beta/3} N^{-(2+\beta)/2} \mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}^{(\alpha)}_{\theta,\phi}(Z) - 1\right|^{2+\beta}\right),$$

leading in particular to the inequality

$$\mathbb{E}\left(|\delta_{\alpha,N}|^{2+\beta}\right) \leq B_{\beta/3} N^{-(2+\beta)/2} \mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}^{(\alpha)}_{\theta,\phi}(Z) - 1\right|^{2+\beta}\right). \tag{75}$$

Hence, by (23) and since $N^{-(2+\beta)/2} = o(N^{-1})$, we obtain

$$\limsup_{N\to\infty}\left[N\mathbb{E}\left(|\delta_{\alpha,N}|^{2+3\varepsilon}\right)^{\frac{1}{1+\varepsilon}}\right] = 0 \quad \text{when } \varepsilon = \beta/3.$$

As a consequence

$$\limsup_{N\to\infty} N\mathbb{E}\left(\left|\frac{1}{1+\delta_{\alpha,N}}\right|^{\frac{\varepsilon}{1+\varepsilon}} |\delta_{\alpha,N}|^{\frac{2+3\varepsilon}{1+\varepsilon}}\right) = 0 \quad \text{when } \varepsilon = \beta/3. \tag{76}$$

- Second term of (74). Using (75) combined with (23) and since $N^{-(2+\beta)/2} = o(N^{-1})$, we deduce:

$$\lim_{N\to\infty} D_\beta N \mathbb{E}\left(|\delta_{\alpha,N}|^{2+\beta}\right) = 0. \tag{77}$$

Combining (74) with (76) and (77) yields (73) and the proof is concluded. ■

B.2 Proof of Proposition 2

Proof of Proposition 2 We prove the two assertions separately.

1. Assume that (23) holds with $\alpha = \alpha_2$, that is, there exists $\beta > 0$ such that

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}_{\theta,\phi}^{(\alpha_2)}(Z) - 1\right|^{2+\beta}\right) < \infty$$

or equivalently using (22) with $\alpha = \alpha_2$ and setting $a_2 := \mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha_2})$ so that $a_2 \in (0,\infty)$,

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - a_2\right|^{2+\beta}\right) < \infty. \tag{78}$$

We now want to prove that (78) implies (23) with $\alpha = \alpha_1$. Using that $|u^\eta - 1| \leq |u-1|$ for all $u \geq 0$ and all $\eta \in (0,1)$, we have: for all $z \in \mathbb{R}^d$,

$$\left|\overline{w}_{\theta,\phi}^{(\alpha_1)}(z) - 1\right| \leq \left|\frac{w_{\theta,\phi}(z)^{1-\alpha_2}}{\mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha_1})^{\frac{1-\alpha_2}{1-\alpha_1}}} - 1\right|$$

where we have set $\eta = (1-\alpha_1)/(1-\alpha_2)$. Hence,

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|\overline{w}_{\theta,\phi}^{(\alpha_1)}(Z) - 1\right|^{2+\beta}\right) \leq \tilde{a}_1^{-1}\mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - \tilde{a}_1\right|^{2+\beta}\right), \tag{79}$$

where $\tilde{a}_1 = a_1^{(1-\alpha_2)/(1-\alpha_1)}$ with $a_1 := \mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha_1})$. Note in particular that a_1 belongs to $(0,\infty)$ as a consequence of (22) with $\alpha = \alpha_1$ and thus so does $\tilde{a}_1$. Now observe that, setting $p = 2+\beta > 1$, Minkowski's inequality implies that

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - \tilde{a}_1\right|^p\right)^{\frac{1}{p}} \leq \mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - a_2\right|^p\right)^{\frac{1}{p}} + \mathbb{E}_{Z\sim q_\phi}\left(|a_2 - \tilde{a}_1|^p\right)^{\frac{1}{p}}$$

that is

$$\mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - \tilde{a}_1\right|^{2+\beta}\right)^{\frac{1}{2+\beta}} \leq \mathbb{E}_{Z\sim q_\phi}\left(\left|w_{\theta,\phi}(Z)^{1-\alpha_2} - a_2\right|^{2+\beta}\right)^{\frac{1}{2+\beta}} + |a_2 - \tilde{a}_1|$$

We then deduce that (23) holds with $\alpha = \alpha_1$ by combining (79) with the inequality above and the fact that (i) $\tilde{a}_1^{-1} < \infty$, (ii) $\mathbb{E}_{Z\sim q_\phi}(|w_{\theta,\phi}(Z)^{1-\alpha_2} - a_2|^{2+\beta})^{1/(2+\beta)} < \infty$ by (78) and (iii) $|a_2 - \tilde{a}_1| < \infty$.

2. Assume (24) holds for $\alpha = \alpha_2$. Then by Lemma 4 with $k = 1$ we may pick N such that $\mathbb{E}(1/R_{\alpha_2,N}) < \infty$. Observe now that

$$\begin{aligned}\mathbb{E}\left(1/R_{\alpha_1,N}\right) &= \mathbb{E}\left(\frac{N}{\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha_1}}\right) \\ &\leq \mathbb{E}\left(\frac{N}{\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha_1}} \,\middle|\, w_{\theta,\phi}(Z_i) \leq p_\theta(x) \text{ for all } i = 1,\ldots,N\right)\end{aligned}$$

where we have used that $N/(\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha_1})$ is a decreasing function of each $w_{\theta,\phi}(Z_i)$, with $w_{\theta,\phi}(Z_i)$ being independent random variables. Since $\alpha_1 > \alpha_2$, it follows that

$$\begin{aligned}\mathbb{E}\left(1/R_{\alpha_1,N}\right) &\leq \mathbb{E}\left(\frac{N p_\theta(x)^{\alpha_1-\alpha_2}}{\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{1-\alpha_2}} \,\middle|\, w_{\theta,\phi}(Z_i) \leq p_\theta(x) \text{ for all } i = 1\ldots N\right) \\ &\leq p_\theta(x)^{\alpha_1-\alpha_2}\mathbb{E}\left(\frac{1}{R_{\alpha_2,N}} \,\middle|\, w_{\theta,\phi}(Z_i) \leq p_\theta(x) \text{ for all } i = 1\ldots N\right) \\ &\leq p_\theta(x)^{\alpha_1-\alpha_2}\frac{\mathbb{E}(1/R_{\alpha_2,N})}{\mathbb{P}\left(w_{\theta,\phi}(Z) \leq p_\theta(x)\right)^N} \\ &< \infty\end{aligned}$$

since $\mathbb{E}(w_{\theta,\phi}(Z)) = p_\theta(x)$ which implies that $\mathbb{P}\left(w_{\theta,\phi}(Z) \leq p_\theta(x)\right) > 0$. We see that there exists a choice of N for which $\mathbb{E}\left(1/R_{\alpha_1,N}\right) < \infty$, from which (24) follows by Lemma 4 with $k = 1$.

■

B.3 Behavior of γ_α^2

Lemma 5 *Let $\alpha \in [0, 1)$. Then, under common integrability and differentiability assumptions*

$$\lim_{\alpha\to 1} \gamma_\alpha^2 = 0.$$

Proof By definition of γ_α^2, we have that

$$\gamma_\alpha^2 = \frac{1}{\mathbb{E}(w_{\theta,\phi}^{1-\alpha})^2} \cdot \frac{1}{1-\alpha}\mathbb{E}\left(\left[w_{\theta,\phi}^{1-\alpha} - \mathbb{E}(w_{\theta,\phi}^{1-\alpha})\right]^2\right).$$

On the one hand, we have that $\mathbb{E}(w_{\theta,\phi}^{1-\alpha}) \to 1$ as $\alpha \to 1$ under convenient integrability assumptions. On the other hand, for all $z \in \mathbb{R}^d$,

$$\begin{aligned}&\lim_{\alpha\to 1} \frac{1}{1-\alpha}\left[w_{\theta,\phi}(z)^{1-\alpha} - \mathbb{E}(w_{\theta,\phi}^{1-\alpha})\right]^2 \\ &\quad= \lim_{\alpha\to 1}\left\{2\left[w_{\theta,\phi}(z)^{1-\alpha} - \mathbb{E}(w_{\theta,\phi}^{1-\alpha})\right] \cdot \left[-w_{\theta,\phi}(z)^{1-\alpha}\log w_{\theta,\phi}(z) + \mathbb{E}\left(w_{\theta,\phi}^{1-\alpha}\log w_{\theta,\phi}\right)\right]\right\}\end{aligned}$$

so that under convenient differentiability assumptions,

$$\lim_{\alpha \to 1} \frac{1}{1-\alpha} \mathbb{E}\left(\left[w_{\theta,\phi}^{1-\alpha} - \mathbb{E}(w_{\theta,\phi}^{1-\alpha}) \right]^2 \right) = 0$$

thus implying that $\lim_{\alpha \to 1} \gamma_\alpha^2 = 0$. ■

B.4 Proof of Example 1

Proof of Example 1 Using (26), we first deduce that: for all $m \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}_{Z \sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^m\right) &= \mathbb{E}_{S \sim \mathcal{N}(0,1)}\left[\exp\left(-\frac{m\sigma^2 d}{2} - m\sigma\sqrt{d}S \right) \right] \\
&= \exp\left(-\frac{m\sigma^2 d}{2} \right) \mathbb{E}_{S \sim \mathcal{N}(0,1)}\left[\exp\left(-m\sigma\sqrt{d}S \right) \right] \\
&= \exp\left(-\frac{m\sigma^2 d}{2} \right) \exp\left(\frac{m^2\sigma^2 d}{2} \right) \\
&= \exp\left(\frac{m(m-1)\sigma^2 d}{2} \right).
\end{aligned}$$

Therefore, plugging in $m = 1 - \alpha$,

$$\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) = \frac{1}{1-\alpha} \log \mathbb{E}_{Z \sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^{1-\alpha} \right) = -\frac{\alpha\sigma^2 d}{2},$$

which gives the desired result for $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$. In addition: for all $m \in \mathbb{R}$,

$$\mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z)^m) = \exp\left(\frac{m(m-1)\sigma^2 d}{2} \right) p_\theta(x)^m. \tag{80}$$

Now note that (26) can be rewritten as: for all $i = 1 \dots N$,

$$\log w_{\theta,\phi}(z_i) = -\frac{\sigma^2 d}{2} - \sigma\sqrt{d}S_i + \log p_\theta(x), \quad S_i \sim \mathcal{N}(0,1).$$

Hence, we get that: for all $i = 1 \dots N$,

$$\begin{aligned}
\log \overline{w}_{\theta,\phi}^{(\alpha)}(z_i) &= (1-\alpha) \log w_{\theta,\phi}(z_i) - \log \mathbb{E}_{Z \sim q_\phi}(w_{\theta,\phi}(Z)^{1-\alpha}) \\
&= -(1-\alpha)\sigma\sqrt{d}S_i - \frac{(1-\alpha)^2\sigma^2 d}{2}
\end{aligned}$$

where we have used (80) with $m = 1 - \alpha$. As a result,

$$\begin{aligned}
\mathbb{V}_{Z \sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)) &= \mathbb{V}_{S \sim \mathcal{N}(0,1)}\left(\exp\left\{ -(1-\alpha)\sigma\sqrt{d}S \right\} \exp\left\{ -\frac{(1-\alpha)^2\sigma^2 d}{2} \right\} \right) \\
&= \exp\left((1-\alpha)^2\sigma^2 d \right) \left(\exp\left((1-\alpha)^2\sigma^2 d \right) - 1 \right) \exp\left(-(1-\alpha)^2\sigma^2 d \right) \\
&= \exp\left((1-\alpha)^2\sigma^2 d \right) - 1,
\end{aligned}$$

which yields the desired result for γ_α^2. In addition, we show that (23) and (24) both hold, meaning that we can apply Theorem 3. To see this, note that $\mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^m)$ and $\mathbb{E}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)^m)$ are well-defined and finite for all $\alpha, m \in \mathbb{R}$. Furthermore, observe that by convexity of the function $u \mapsto |u|^{2+\beta}$ with $\beta > 0$, it holds that

$$\mathbb{E}_{Z\sim q_\phi}(|\overline{w}_{\theta,\phi}^{(\alpha)}(Z) - 1|^{2+\beta}) \leq 2^{1+\beta}\left(\mathbb{E}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)^{2+\beta}) + 1\right)$$

thus (23) holds for all $\beta > 0$. Lastly, $\mathbb{E}(1/R_{\alpha,N}) \leq \mathbb{E}(N^{-1}\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{\alpha-1})$ by the HM-AM inequality and the r.h.s. is finite for all $\alpha \in \mathbb{R}$ thus (24) also holds.

In the particular case $p_\theta(z|x) = \mathcal{N}(z;\theta,\boldsymbol{I}_d)$ and $q_\phi(z|x) = \mathcal{N}(z;\phi,\boldsymbol{I}_d)$: for all $i = 1\dots N$,

$$\begin{aligned}
\log \overline{w}_{\theta,\phi}(z_i) &= -\frac{1}{2}\left(\|z_i-\theta\|^2 - \|z_i-\phi\|^2\right)\\
&= -\frac{1}{2}\left(\|\theta\|^2 - \|\phi\|^2 + 2\langle z_i, \phi-\theta\rangle\right)\\
&= -\frac{1}{2}\left(\langle\theta-\phi,\theta+\phi\rangle + 2\langle z_i,\phi-\theta\rangle\right)\\
&= -\frac{1}{2}\left(\langle\theta-\phi,\theta-\phi+2\phi\rangle + 2\langle z_i,\phi-\theta\rangle\right)\\
&= -\frac{1}{2}\left(B_d^2 + 2\langle z_i-\phi,\phi-\theta\rangle\right)\\
&= -\frac{B_d^2}{2} - B_d S_i \quad \text{with } S_i = \tfrac{1}{B_d}\langle z_i-\phi,\phi-\theta\rangle, \qquad (81)
\end{aligned}$$

where we have set $B_d = \|\phi-\theta\|$. Since $z_i - \phi \sim \mathcal{N}(0,\boldsymbol{I}_d)$, it follows that $S_i \sim \mathcal{N}(0,1)$ as required. Lastly, when $\theta = 0\cdot\boldsymbol{u}_d$ and $\phi = \boldsymbol{u}_d$, we have that $B_d = \sqrt{d}$. ■

B.5 Proof of Example 2

Proof Let us first prove that $p_\theta(x) = \mathcal{N}(x;\theta,2\boldsymbol{I}_d)$ and $p_\theta(z|x) = \mathcal{N}(z;(\theta+x)/2, 1/2\ \boldsymbol{I}_d)$. To see this, note that

$$p_\theta(x,z) = \left(\frac{1}{(2\pi)^{d/2}}\right)^2 \exp\left(-\frac{1}{2}\left\{\|z-\theta\|^2 + \|x-z\|^2\right\}\right).$$

As a result, only considering the dependency in z, we have that

$$p_\theta(x,z) \propto \exp\left(-\frac{1}{2}\cdot 2\left\|z - \frac{\theta+x}{2}\right\|^2\right),$$

which implies that $p_\theta(z|x) = \mathcal{N}(z;(\theta+x)/2, 1/2\ \boldsymbol{I}_d)$. Furthermore,

$$\begin{aligned}
p_\theta(x) &= \int p_\theta(x,z)\mathrm{d}z\\
&= \frac{1}{(2\pi\cdot 2)^{d/2}}\int \frac{1}{(2\pi\cdot 1/2)^{d/2}}\exp\left(-\frac{1}{2}\cdot 2\left\|z-\frac{\theta+x}{2}\right\|^2\right)\mathrm{d}z\cdot\exp\left(-\frac{1}{2}\cdot\frac{1}{2}\|\theta-x\|^2\right)\\
&= \mathcal{N}(x;\theta,2\boldsymbol{I}_d).
\end{aligned}$$

Hence, for all $z \in \mathbb{R}^d$

$$\log \overline{w}_{\theta,\phi}(z) = \log\left(\frac{p_\theta(z|x)}{q_\phi(z|x)}\right) = \log\left(\frac{(2\pi \cdot 2/3)^{d/2}}{(2\pi \cdot 1/2)^{d/2}} \exp\left[-\left\|z - \frac{\theta + x}{2}\right\|^2 + \frac{3}{4}\|z - Ax - b\|^2\right]\right)$$

and from there, we can straightforwardly deduce that: for all $i = 1 \dots N$,

$$\log \overline{w}_{\theta,\phi}(z_i) = \frac{d}{2}\log\left(\frac{4}{3}\right) - \left\|z_i - \frac{\theta + x}{2}\right\|^2 + \frac{3}{4}\|z_i - Ax - b\|^2. \tag{82}$$

Using (82) we can write that

$$\begin{aligned}\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) &= \frac{1}{1-\alpha}\log\left(\int q_\phi(z|x)\overline{w}_{\theta,\phi}(z)^{1-\alpha}\mathrm{d}z\right) \\ &= \frac{d}{2}\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log\left(I\right)\end{aligned}$$

where

$$\begin{aligned}I &= \int \frac{1}{(4\pi/3)^{d/2}} \exp\left[-\frac{3}{4}\|z - Ax - b\|^2 + (1-\alpha)\left(-\left\|z - \frac{\theta + x}{2}\right\|^2 + \frac{3}{4}\|z - Ax - b\|^2\right)\right]\mathrm{d}z \\ &= \int \frac{1}{(4\pi/3)^{d/2}} \exp\left[-\frac{3\alpha}{4}\|z - Ax - b\|^2 - (1-\alpha)\left\|z - \frac{\theta + x}{2}\right\|^2\right]\mathrm{d}z.\end{aligned}$$

I is well-defined and finite for all $\alpha < 4$. Completing the square leads to

$$\begin{aligned}I = \left(\frac{3}{4-\alpha}\right)^{d/2} \exp\Bigg(\left(\frac{4}{4-\alpha}\right)\Bigg\|\frac{3\alpha}{4}(Ax + b) \\ + (1-\alpha)\frac{\theta + x}{2}\Bigg\|^2 - \frac{3\alpha}{4}\|Ax + b\|^2 - (1-\alpha)\left\|\frac{\theta + x}{2}\right\|^2\Bigg).\end{aligned}$$

As a result,

$$\begin{aligned}\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x) &= \frac{d}{2}\left[\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right] \\ &+ \frac{4}{(4-\alpha)(1-\alpha)}\left\|\frac{3\alpha}{4}(Ax + b) + (1-\alpha)\frac{\theta + x}{2}\right\|^2 - \frac{3\alpha}{4(1-\alpha)}\|Ax + b\|^2 - \left\|\frac{\theta + x}{2}\right\|^2.\end{aligned}$$

It can then be checked that the second line simplifies to $-\frac{3\alpha}{4-\alpha}\left\|Ax + b - \frac{\theta+x}{2}\right\|^2$, from which we deduce the desired result for $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$. [Notice in particular that $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ is well-defined for all $\alpha \neq 1, \alpha < 4$ with continuous extension at $\alpha = 1$.] On the other hand, we have that

$$\gamma_\alpha^2 = \frac{1}{1-\alpha}\mathbb{V}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}^{(\alpha)}(Z)) = \frac{1}{1-\alpha}\left(\frac{\mathbb{E}_{Z\sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^{2-2\alpha}\right)}{\mathbb{E}_{Z\sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^{1-\alpha}\right)^2} - 1\right).$$

Furthermore, for all $\alpha' \in \mathbb{R} \setminus \{1\}$, it holds that

$$\mathbb{E}_{Z\sim q_\phi}\left(\overline{w}_{\theta,\phi}(Z)^{1-\alpha'}\right) = \exp\left((1-\alpha')\left[\mathcal{L}^{(\alpha')}(\theta, \phi; x) - \ell(\theta; x)\right]\right). \tag{83}$$

Now using (83) with $\alpha' = \alpha$ and $\alpha' = 2\alpha - 1$ and combining with the expression of $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$, we obtain

$$\gamma_\alpha^2 = \frac{1}{1-\alpha}\left(\exp(A) - 1\right)$$

with

$$\begin{aligned} A &= 2(1-\alpha)\left\{\frac{d}{2}\left[\log\left(\frac{4}{3}\right) + \frac{1}{2(1-\alpha)}\log\left(\frac{3}{5-2\alpha}\right)\right] - \frac{3(2\alpha-1)}{5-2\alpha}\left\|Ax + b - \frac{\theta+x}{2}\right\|^2\right\} \\ &\quad - 2(1-\alpha)\left\{\frac{d}{2}\left[\log\left(\frac{4}{3}\right) + \frac{1}{1-\alpha}\log\left(\frac{3}{4-\alpha}\right)\right] - \frac{3\alpha}{4-\alpha}\left\|Ax + b - \frac{\theta+x}{2}\right\|^2\right\} \\ &= \frac{d}{2}\log\left(\frac{(4-\alpha)^2}{5-2\alpha}\right) + \frac{24(1-\alpha)^2}{(5-2\alpha)(4-\alpha)}\left\|Ax + b - \frac{\theta+x}{2}\right\|^2 \end{aligned}$$

from which we deduce the desired result for γ_α^2 [notice in particular that γ_α^2 is well-defined for all $\alpha < 5/2$].

In addition, the assumptions made in Theorem 3 are satisfied for all $\alpha \in [0, 1)$. To see this, set $m = 1 - \alpha'$ in (83) and use that $\mathcal{L}^{(\alpha)}(\theta, \phi; x) - \ell(\theta; x)$ is well-defined for all $\alpha \neq 1, \alpha < 4$ with continuous extension at $\alpha = 1$, so that $\mathbb{E}_{Z\sim q_\phi}(\overline{w}_{\theta,\phi}(Z)^m)$ and thus $\mathbb{E}_{Z\sim q_\phi}(w_{\theta,\phi}(Z)^m)$ and $\mathbb{E}_{Z\sim q_\phi}(\overline{w}^{(\alpha)}_{\theta,\phi}(Z)^m)$ are well-defined and finite for all $m > -3$. Furthermore, the convexity of the function $u \mapsto |u|^{2+\beta}$ with $\beta > 0$ implies that

$$\mathbb{E}_{Z\sim q_\phi}(|\overline{w}^{(\alpha)}_{\theta,\phi}(Z) - 1|^{2+\beta}) \leq 2^{1+\beta}\left(\mathbb{E}_{Z\sim q_\phi}(\overline{w}^{(\alpha)}_{\theta,\phi}(Z)^{2+\beta}) + 1\right)$$

thus (23) holds for all $\beta > 0$. Lastly, $\mathbb{E}(1/R_{\alpha,N}) \leq \mathbb{E}(N^{-1}\sum_{i=1}^N w_{\theta,\phi}(Z_i)^{\alpha-1})$ by the HM-AM inequality and the r.h.s. is finite for all $\alpha > -2$ thus (24) also holds. ■

B.6 Proof of Proposition 3

Proof of Proposition 3 For all $\alpha \in [0, 1)$, we can rewrite the variational gap $\Delta^{(\alpha)}_{N,d}(\theta, \phi)$ as

$$\begin{aligned} \Delta^{(\alpha)}_{N,d}(\theta, \phi; x) &= \frac{1}{1-\alpha}\int\int \prod_{i=1}^N q_\phi(z_i)\log\left(\frac{1}{N}\sum_{j=1}^N \overline{w}_j^{1-\alpha}\right)\mathrm{d}\overline{w}_{1:N} \\ &= \frac{1}{1-\alpha}\int\int \prod_{i=1}^N q_\phi(z_i)\log\left(\frac{1}{N}\sum_{j=1}^N (\overline{w}^{(j)})^{1-\alpha}\right)\mathrm{d}\overline{w}_{1:N} \\ &= \frac{1}{1-\alpha}\left[\int\int \prod_{i=1}^N q_\phi(z_i)\log\left(\frac{1}{N}(\overline{w}^{(N)})^{1-\alpha}\right)\mathrm{d}\overline{w}_{1:N}\right. \\ &\qquad \left. + \int\int \prod_{i=1}^N q_\phi(z_i)\log\left(1 + \sum_{j=1}^{N-1}\left(\frac{\overline{w}^{(j)}}{\overline{w}^{(N)}}\right)^{1-\alpha}\right)\mathrm{d}\overline{w}_{1:N}\right] \\ &= \Delta^{(\alpha,MAX)}_{N,d}(\theta, \phi; x) + R^{(\alpha)}_{N,d}(\theta, \phi; x) \end{aligned}$$

where we have used (29) and where we have set

$$R_{N,d}^{(\alpha)}(\theta,\phi;x) := \frac{1}{1-\alpha}\int\int \prod_{i=1}^{N} q_\phi(z_i) \log\left(1+\sum_{j=1}^{N-1}\left(\frac{\overline{w}^{(j)}}{\overline{w}^{(N)}}\right)^{1-\alpha}\right) \mathrm{d}\overline{w}_{1:N}.$$

All that is left to do is now to prove (30). Observe that by definition of $T_{N,d}^{(\alpha)}$ in (27) and since $\alpha \in [0,1)$, we can write

$$\begin{aligned} 0 \leq R_{N,d}^{(\alpha)}(\theta,\phi;x) &= \frac{1}{1-\alpha}\int\int \prod_{i=1}^{N} q_\phi(z_i) \log\left(1+T_{N,d}^{(\alpha)}\right) \mathrm{d}\overline{w}_{1:N} \\ &\leq \frac{1}{1-\alpha}\int \prod_{i=1}^{N} q_\phi(z_i)\ T_{N,d}^{(\alpha)}\ \mathrm{d}\overline{w}_{1:N} \\ &= \frac{1}{1-\alpha}\mathbb{E}(T_{N,d}^{(\alpha)}), \end{aligned}$$

which concludes the proof. ■

B.7 Deferred Proofs of Section 4.2.1

B.7.1 Proof of Lemma 1

Proof of Lemma 1 First, note that since $S_1,\ldots,S_N$ are i.i.d. normal random variables, so are $-S_1,\ldots,-S_N$. Setting $M_N = \max_{1\leq i\leq N} -S_i$, we also have $S^{(1)} = -M_N$. A standard result (obtained, for example, by combining Theorem 1.1.2 and Example 1.1.7 in de Haan and Ferreira, 2007) is that for all $x \in \mathbb{R}$,

$$\lim_{N\to\infty} P\left(a_N^{-1}\left(M_N - b_N\right) \leq x\right) = \exp(-e^{-x}) \tag{84}$$

with $a_N = 1/\sqrt{2\log N}$ and $b_N = \sqrt{2\log N} - \frac{1}{2}(\log\log N + \log 4\pi)/(\sqrt{2\log N})$. Since $\mathbb{E}(|M_N|) \leq \mathbb{E}(M_N^2)^{1/2} \leq \mathbb{E}(\sum_{i=1}^N S_i^2)^{1/2} \leq N^{1/2} < \infty$ for all N, it follows by (Pickands III, 1968, Theorem 2.1) that

$$\lim_{N\to\infty} a_N^{-1}\left(\mathbb{E}\left(M_N\right) - b_N\right) = \mathbb{E}(U),$$

where U is a Gumbel random variable and $\mathbb{E}(U)$ is given by the Euler–Masceroni constant. Using that $S^{(1)} = -M_N$, we deduce

$$\lim_{N\to\infty} -a_N^{-1}\left(\mathbb{E}(S^{(1)}) + b_N\right) = \mathbb{E}(U).$$

Finally, plugging in the definition of a_N and b_N, we obtain

$$\begin{aligned} \mathbb{E}(S^{(1)}) &= -\sqrt{2\log N} + \frac{\log\log N + \log 4\pi}{2\sqrt{2\log N}} - \frac{\mathbb{E}(U)}{\sqrt{2\log N}} + o\left(\frac{1}{\sqrt{2\log N}}\right) \\ &= -\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right) \end{aligned}$$

and we have thus recovered (34). ■

B.7.2 Proof of Proposition 4

Proof of Proposition 4 First note that

$$\log \overline{w}^{(N)} = -\frac{d\sigma^2}{2} - \sqrt{d}\sigma S^{(1)}.$$

Combining this result with the definition of $\Delta_{N,d}^{MAX}(\theta, \phi; x)$ in (29) yields

$$\Delta_{N,d}^{MAX}(\theta, \phi) = -\frac{d\sigma^2}{2} - \sqrt{d}\sigma \mathbb{E}(S^{(1)}) + \frac{\log N}{\alpha - 1}.$$

Now using (34), we deduce

$$\begin{aligned}\Delta_{N,d}^{MAX}(\theta, \phi) &= -\frac{d\sigma^2}{2} + \sqrt{d}\sigma\left(\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right)\right) + \frac{\log N}{\alpha - 1} \\ &= -\frac{d\sigma^2}{2}\left\{1 - 2\sqrt{\frac{2\log N}{d\sigma^2}} + \frac{1}{1-\alpha}\frac{2\log N}{d\sigma^2} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right\},\end{aligned}$$

which concludes the proof. ■

B.7.3 Proof of Proposition 5

We first prove a useful intermediate lemma regarding the concentration of $S^{(1)}$.

Lemma 6 *Let $S_1, \ldots, S_N$ be i.i.d. normal random variables, set $S^{(1)} = \min_{1 \leq i \leq N} S_i$, and define $I_N = [-4\sqrt{\log N}, -\sqrt{\log N}]$. Then as $N \to \infty$, we have*

$$\mathbb{P}(S^{(1)} \notin I_N) = O\left(\frac{1}{N^4}\right).$$

Proof We control the probability of the events $\{S^{(1)} > -\sqrt{\log N}\}$ and $\{S^{(1)} < -4\sqrt{\log N}\}$ separately. First, note that

$$\begin{aligned}\log \mathbb{P}(S^{(1)} > -\sqrt{\log N}) &= N \log(\overline{\Phi}(-\sqrt{\log N})) \\ &= N \log\left(1 - \overline{\Phi}(\sqrt{\log N})\right) \\ &= -N\overline{\Phi}(\sqrt{\log N})(1 + o(1)) \\ &= -N\frac{\phi(\sqrt{\log N})}{\sqrt{\log N}}(1 + o(1)) \qquad (85)\end{aligned}$$

where in the final line we have used the standard approximation

$$\overline{\Phi}(x) = \frac{\phi(x)}{x}(1 + o(1)) \qquad (86)$$

as $x \to \infty$. We deduce that

$$\mathbb{P}(S^{(1)} > -\sqrt{\log N}) = \exp\left\{-N\frac{N^{-1/2}}{\sqrt{2\pi}\sqrt{\log N}}(1 + o(1))\right\} = O\left(\frac{1}{N^4}\right) \qquad (87)$$

as $N \to \infty$. Second, as $N \to \infty$ we have, by a union bound, that

$$\begin{aligned}\mathbb{P}(S^{(1)} < -4\sqrt{\log N}) &\leq N\Phi(-4\sqrt{\log N}) \\ &= N\frac{\phi(4\sqrt{\log N})}{4\sqrt{\log N}}(1+o(1)) \\ &= \frac{N^{-7}}{4\sqrt{2\pi}\sqrt{\log N}}(1+o(1)) \\ &= O\left(\frac{1}{N^4}\right) \end{aligned} \tag{88}$$

and so the result follows. ■

We now prove Proposition 5 by building on the proof from Snyder et al. (2008) and on Lemma 6.

Proof of Proposition 5 Denote $\sigma_\alpha = (1-\alpha)\sigma$ for all $\alpha \in [0,1)$. A first remark is that, conditional upon $S^{(1)}$, we can think of the sum in (27) as the sum over $N-1$ i.i.d. random variables

$$\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)}) = (N-1)\mathbb{E}\left(\exp\left(-\sigma_\alpha\sqrt{d}(S-S^{(1)})\right)\right)$$

where the expectation is w.r.t. the density of S given by

$$p(z) = \frac{\phi(z)}{\overline{\Phi}(S^{(1)})}\mathbb{I}(z \geq S^{(1)}),$$

with $\phi(z)$ denoting the standard normal density and $\overline{\Phi}(x) = \int_x^\infty \phi(z)\mathrm{d}z$ denoting the normalizing constant. Then,

$$\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)}) = \frac{(N-1)\int_{S^{(1)}}^\infty \exp\left(-\sigma_\alpha\sqrt{d}\left(z-S^{(1)}\right)\right)\phi(z)\mathrm{d}z}{\overline{\Phi}(S^{(1)})}.$$

We can then calculate explicitly

$$\begin{aligned}&\int_{S^{(1)}}^\infty \exp\left(-\sigma_\alpha\sqrt{d}\left(z-S^{(1)}\right)\right)\phi(z)\mathrm{d}z \\ &= \exp(\sigma_\alpha\sqrt{d}S^{(1)} + \sigma_\alpha^2 d/2)\int_{S^{(1)}}^\infty (\sqrt{2\pi})^{-1}\exp\left(-\frac{1}{2}(z+\sigma_\alpha\sqrt{d})^2\right)\mathrm{d}z \\ &= \exp(\sigma_\alpha\sqrt{d}S^{(1)} + \sigma_\alpha^2 d/2)\overline{\Phi}(\sigma_\alpha\sqrt{d}+S^{(1)}).\end{aligned}$$

Denoting $I_N = [-4\sqrt{\log N}, -\sqrt{\log N}]$, on the event $\{S^{(1)} \in I_N\}$, as $N, d \to \infty$ with $\log N/d \to 0$, we have

$$\sigma_\alpha\sqrt{d} + S^{(1)} = \sigma_\alpha\sqrt{d}(1+o_{N,d}(1)),$$

where we use the notation $o_{N,d}(1)$ to denote that the implicit constant, which goes to zero as $N, d \to \infty$ with $\log N/d \to 0$, does not depend on $S^{(1)}$. Using the approximation (86) for $\Phi(x)$ as $x \to \infty$,

$$\overline{\Phi}(\sigma_\alpha\sqrt{d}+S^{(1)}) = \frac{\phi(\sigma_\alpha\sqrt{d}+S^{(1)})}{\sigma_\alpha\sqrt{d}+S^{(1)}}(1+o_{N,d}(1)).$$

Hence, observing that $\exp(\sigma_\alpha\sqrt{d}S^{(1)} + \sigma_\alpha^2 d/2)\phi(\sigma_\alpha\sqrt{d} + S^{(1)}) = \phi(S^{(1)})$, it follows that

$$\int_{S^{(1)}}^{\infty} \exp\left(-\sigma_\alpha\sqrt{d}\left(z - S^{(1)}\right)\right)\phi(z)\mathrm{d}z = \frac{\phi(S^{(1)})}{\sigma_\alpha\sqrt{d}}(1 + o_{N,d}(1))$$

on the event $\{S^{(1)} \in I_N\}$. Using (86), we can also write

$$\Phi(S^{(1)}) = \overline{\Phi}(-S^{(1)}) = \frac{\phi(-S^{(1)})}{-S^{(1)}}(1 + o_{N,d}(1)).$$

Combined with $\phi(-S^{(1)}) = \phi(S^{(1)})$, this allows us to deduce that

$$\begin{aligned}\int_{S^{(1)}}^{\infty} \exp\left(-\sigma_\alpha\sqrt{d}\left(z - S^{(1)}\right)\right)\phi(z)\mathrm{d}z &= \frac{(-S^{(1)})\Phi(S^{(1)})}{\sigma_\alpha\sqrt{d}}(1 + o_{N,d}(1)) \\ &\leq \Phi(S^{(1)})\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1 + o_{N,d}(1)), \qquad (89)\end{aligned}$$

all on the event $\{S^{(1)} \in I_N\}$. Finally, since $S^{(1)} < -\sqrt{\log N}$ implies

$$\overline{\Phi}(S^{(1)}) = 1 + o_{N,d}(1),$$

we have

$$\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)}) \leq (N-1)\Phi(S^{(1)})\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1 + o_{N,d}(1)).$$

We conclude by using the tower law of expectation and splitting according to whether $S^{(1)} \in I_N$, giving

$$\begin{aligned}\mathbb{E}(T_{N,d}^{(\alpha)}) &= \mathbb{E}\left(\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)})\right) \\ &\leq \mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\in I_N\}}(N-1)\Phi(S^{(1)})\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1 + o_{N,d}(1))\right) + \mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\notin I_N\}}\right) \\ &\leq (N-1)\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}\mathbb{E}\left(\Phi(S^{(1)})\right)(1 + o(1)) + \mathbb{P}(S^{(1)} \notin I_N) \\ &\leq \frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1 + o(1)) + O\left(\frac{1}{N^4}\right) \to 0\end{aligned}$$

where in the final line we have used Lemma 6 and that $\Phi(S^{(1)})$ is distributed as the minimum of N independent uniform random variables on $[0, 1]$ and so $\mathbb{E}(\Phi(S^{(1)})) = \frac{1}{N+1}$.

■

B.7.4 Proof of Theorem 5

First note that Lemma 1 and Lemma 6 are not affected by the change in the distribution of the weights we have made in (36). As for Proposition 4 and Proposition 5, they are modified according to Proposition 8 and Proposition 9 below.

Proposition 8 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (36) *and that there exists $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$. Then, for all $\alpha \in [0, 1)$,*

$$\lim_{N,d\to\infty} \Delta_{N,d}^{(\alpha,MAX)}(\theta, \phi; x) + \frac{B_d^2}{2}\left\{1 - 2\ \frac{\sqrt{2\log N}}{B_d} + \frac{1}{1-\alpha}\ \frac{2\log N}{B_d^2} + O\left(\frac{\log\log N}{B_d\sqrt{\log N}}\right)\right\} = 0.$$

Proof First note that

$$\log \overline{w}^{(N)} = -\frac{B_d^2}{2} - B_d S^{(1)}.$$

Combining this result with the definition of $\Delta_{N,d}^{MAX}(\theta, \phi; x)$ in (29) yields

$$\Delta_{N,d}^{MAX}(\theta, \phi) = -\frac{B_d^2}{2} - B_d \mathbb{E}(S^{(1)}) + \frac{\log N}{\alpha - 1}.$$

Now using (34) of Lemma 1, we deduce

$$\Delta_{N,d}^{MAX}(\theta, \phi) = -\frac{B_d^2}{2} + B_d\left(\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right)\right) + \frac{\log N}{\alpha - 1},$$

which concludes the proof. ■

Proposition 9 *Let $S_1, \dots, S_N$ be i.i.d. normal random variables. Further assume that the weights $\overline{w}_1, \dots, \overline{w}_N$ satisfy* (36) *and that there exists $\sigma_- > 0$ such that $B_d \geq \sigma_- \sqrt{d}$. Then, for all $\alpha \in [0, 1)$, we have*

$$\lim_{\substack{N,d\to\infty \\ \log N/d \to 0}} \mathbb{E}(T_{N,d}^{(\alpha)}) = 0.$$

Proof Conditional upon $S^{(1)}$, we can think of the sum in (27) as the sum over $N - 1$ i.i.d. random variables

$$\mathbb{E}(T_{N,d}^{(\alpha)} | S^{(1)}) = (N-1)\mathbb{E}\left(\exp\left(-(1-\alpha)B_d(S - S^{(1)})\right)\right)$$

where the expectation is w.r.t. the density of S given by

$$p(z) = \frac{\phi(z)}{\overline{\Phi}(S^{(1)})}\mathbb{I}(z \geq S^{(1)}),$$

with $\phi(z)$ denoting the standard normal density and $\overline{\Phi}(x) = \int_x^\infty \phi(z)\mathrm{d}z$ denoting the normalizing constant. Now denoting $\sigma_\alpha = (1-\alpha)\sigma_-$ for all $\alpha \in [0, 1)$ and using that

$$\mathbb{E}(T_{N,d}^{(\alpha)} | S^{(1)}) \leq (N-1)\mathbb{E}\left(\exp\left(-\sigma_\alpha\sqrt{d}(S - S^{(1)})\right)\right)$$

we obtain by following the proof of Proposition 5 that the term on the r.h.s. above goes to 0 as $\log N/d \to 0$ with $N, d \to \infty$. We deduce the desired result by combining this with the fact that $\mathbb{E}(T_{N,d}^{(\alpha)} | S^{(1)}) \geq 0$. ■

The proof of Theorem 5 then follows immediately from Proposition 8 and Proposition 9.

B.8 Deferred Proofs and Results of Section 4.2.2

We start by recalling some useful results from Saulis and Statulevičius (2000) regarding large deviations for sums of independent random variables.

B.8.1 Large Deviations for Sums of Independent Random Variables

The random variable ξ is said to satisfy the assumption (A-ξ) if the following holds.

(A-ξ) There exists $\Delta > 0$ such that $|\Gamma_k(\xi)| \leq \frac{k!}{\Delta^{k-2}}$ for all integer $k \geq 3$, where $\Gamma_k(\xi)$ denotes the k-th cumulant of ξ.

We now state without proof (Saulis and Statulevičius, 2000, Lemma 2.3) and (Saulis and Statulevičius, 2000, Theorem 3.1) in the particular case $\gamma = 0$.

Lemma 7 ((Saulis and Statulevičius, 2000, Lemma 2.3) with $\gamma = 0$) *Let ξ be a random variable with $\mathbb{E}(\xi) = 0$ and $\mathbb{E}(\xi^2) = 1$. Denote by $G(\cdot)$ the cdf of ξ. Assume that (A-ξ) holds and set*

$$\Delta_0 = \frac{\sqrt{2}}{36}\Delta.$$

Then, in the interval $0 \leq x < \Delta_0$, the relations of large deviations

$$1 - G(x) = (1 - \Phi(x)) \exp(P(x)) \left(1 + \theta_1 f(x) \frac{x+1}{\Delta_0}\right)$$
$$G(-x) = \Phi(-x) \exp(P(-x)) \left(1 + \theta_2 f(x) \frac{x+1}{\Delta_0}\right)$$

are valid, with Φ denoting the standard normal distribution. Here, P and f are defined by

$$P(x) = \sum_{k=3}^{\infty} \lambda_k x^k + \theta \left(x/\Delta_0\right)^3$$
$$f(x) = \frac{60(1 + 10\Delta_0^2 \exp\left\{-\left(1 - x/\Delta_0\right)\sqrt{\Delta_0}\right\})}{1 - x/\Delta_0},$$

where $\theta, \theta_1, \theta_2$ are some variables not exceeding 1 in absolute value and where for all $k \geq 3$

$$|\lambda_k| \leq \frac{2}{k} \left(16/\Delta\right)^{k-2}$$

so that

$$P(x) \leq \frac{x^3}{2(x + 8\Delta_0)} \quad \textit{and} \quad P(-x) \geq -\frac{x^3}{3\Delta_0}.$$

Theorem 1 ((Saulis and Statulevičius, 2000, Theorem 3.1) with $\gamma = 0$) *Let $\xi_1, \ldots,$ ξ_d be independent random variables with $\mathbb{E}(\xi_j) = 0$ and $\sigma_j^2 = \mathbb{V}(\xi_j) < \infty$. Set*

$$\boldsymbol{S}_d = \frac{1}{B_d} \left(\xi_1 + \ldots + \xi_d\right),$$

where $B_d^2 = \sum_{j=1}^d \sigma_j^2$. *Assume that there exists* $K > 0$ *such that: for all* $j = 1 \dots d$,

$$|\mathbb{E}(\xi_j^k)| \leq k! K^{k-2} \sigma_j^2, \quad k \geq 3. \tag{90}$$

Then,

$$|\Gamma_k(\boldsymbol{S}_d)| \leq \frac{k!}{\Delta_d^{k-2}}, \quad k \geq 3$$

with

$$\Delta_d = \frac{B_d}{K_d}, \quad \textit{where} \quad K_d = 2 \max \left\{ K, \max_{1 \leq j \leq d} \sigma_j \right\},$$

that is, (A-ξ) *holds with* $\xi = \boldsymbol{S}_d$ *and* $\Delta = \Delta_d$.

B.8.2 Preliminary Results

Building on Lemma 7 and Theorem 1, we can now state some preliminary results that will come in handy when proving the results from Section 4.2.2.

Lemma 8 *Let* $\xi_1, \dots, \xi_d$ *be i.i.d. random variables with* $\mathbb{E}(\xi_1) = 0$ *and* $\sigma^2 = \mathbb{V}(\xi_1) < \infty$. *Set*

$$\boldsymbol{S}_d = \frac{1}{B_d} \left(\xi_1 + \ldots + \xi_d \right),$$

where $B_d = \sigma \sqrt{d}$. *Assume that there exists* $K > 0$ *such that:*

$$|\mathbb{E}(\xi_1^k)| \leq k! K^{k-2} \sigma^2, \quad k \geq 3.$$

Set $\Delta_d = B_d / K_d$ *where* $K_d = 2 \max \{K, \sigma\}$. *Then, as* $d \to \infty$, *there exists an analytic function* P_d *such that the cdf of* $\boldsymbol{S}_d$, *denoted* $G_d(\cdot)$, *satisfies*

$$\begin{aligned} 1 - G_d(x) &= (1 - \Phi(x)) \exp(P_d(x))(1 + o(1)) \\ G_d(-x) &= \Phi(-x) \exp(P_d(-x))(1 + o(1)) \end{aligned}$$

uniformly for all $x \geq 0$ *and* $x = o(\sqrt{d})$. *Here,* Φ *denotes the standard normal distribution,* P_d *is such that*

$$P_d(x) = \sum_{k=3}^{\infty} \lambda_{k,d} x^k$$

with

$$|\lambda_{k,d}| \leq A (c / \sqrt{d})^{k-2}, \quad k \geq 3$$

for some constants $A, c > 0$.

Proof Observe first that $\boldsymbol{S}_d$ satisfies $\mathbb{E}(\boldsymbol{S}_d) = 0$ and $\mathbb{E}(\boldsymbol{S}_d^2) = 1$ with $B_d = \sigma \sqrt{d}$ and $K_d = 2 \max(K, \sigma)$. Furthermore, (A-ξ) holds with $\xi = \boldsymbol{S}_d$ and $\Delta = \Delta_d$ by Theorem 1. Then,

we can apply Lemma 7 with $\xi = \boldsymbol{S}_d$ and $\Delta = \Delta_d$ to obtain that in the interval $0 \le x < \Delta_{0,d}$, the relations of large deviations

$$1 - G_d(x) = (1 - \Phi(x)) \exp(P(x)) \left(1 + \theta_1 f(x) \frac{x+1}{\Delta_{0,d}}\right),$$
$$G_d(-x) = \Phi(-x) \exp(P(-x)) \left(1 + \theta_2 f(x) \frac{x+1}{\Delta_{0,d}}\right)$$

are valid. Here, $\Delta_{0,d} = \frac{\sqrt{2}}{36}\Delta_d$ and P, f are defined by

$$P(x) = P_d(x) + \theta \left(x/\Delta_{0,d}\right)^3$$
$$f(x) = \frac{60(1 + 10\Delta_{0,d}^2 \exp\left\{-\left(1 - x/\Delta_{0,d}\right)\sqrt{\Delta_{0,d}}\right\})}{1 - x/\Delta_{0,d}}$$
$$P_d(x) = \sum_{k=3}^{\infty} \lambda_{k,d} x^k,$$

where $\theta, \theta_1, \theta_2$ are some variables not exceeding 1 in absolute value and

$$|\lambda_{k,d}| \le \frac{2}{k}(16/\Delta_d)^{k-2} \le A(c/\sqrt{d})^{k-2}, \quad k \ge 3$$

for some constants $A, c > 0$. Under the assumption $x = o(\sqrt{d})$, $P(x) = P_d(x) + o(1)$, $f(x)\frac{x+1}{\Delta_{0,d}} = o(1)$ and we can thus deduce that as $d \to \infty$ the relations of large deviations become

$$1 - G_d(x) = (1 - \Phi(x)) \exp(P_d(x)) \left(1 + o(1)\right)$$
$$G_d(-x) = \Phi(-x) \exp(P_d(-x)) \left(1 + o(1)\right)$$

uniformly for all $x \ge 0$ and $x = o(\sqrt{d})$. ■

The corollary below then follows from Lemma 8.

Corollary 1 *Under the assumptions of Lemma 8, as $d \to \infty$,*

$$1 - G_d(x) = (1 - \Phi(x))(1 + o(1))$$
$$G_d(-x) = \Phi(-x)(1 + o(1))$$

uniformly for all $x \ge 0$ and $x = o(d^{1/6})$.

Proof Since we consider the case $x \ge 0$ and $x = o(d^{1/6})$, we can apply Lemma 8 to get: as $d \to \infty$,

$$1 - G_d(x) = (1 - \Phi(x)) \exp(P_d(x))(1 + o(1))$$
$$G_d(-x) = \Phi(-x) \exp(P_d(-x))(1 + o(1))$$

where P_d is defined in Lemma 8. In addition, using successively that (i) $|\lambda_{k,d}| \leq A(c/\sqrt{d})^{k-2}$ by Lemma 8 (ii) $x = o(\sqrt{d})$ and (iii) $x^3 = o(\sqrt{d})$, we have that:

$$\begin{aligned}
|P_d(x)| &\leq \sum_{k=3}^{\infty} |\lambda_{k,d}| x^k \\
&\leq Acx^3 d^{-1/2} \sum_{k=3}^{\infty} \left(cxd^{-1/2}\right)^{k-3} \\
&\leq Acx^3 d^{-1/2}(1 + o(1)) \\
&= o(1).
\end{aligned}$$

Similarly, $|P_d(-x)| = o(1)$ and consequently,

$$\begin{aligned}
1 - G_d(x) &= (1 - \Phi(x))(1 + o(1)) \\
G_d(-x) &= \Phi(-x)(1 + o(1))
\end{aligned}$$

uniformly for all $x \geq 0$ and $x = o(d^{1/6})$. ■

We also prove the following concentration result, which parallels the corresponding result Lemma 6 from the exact log-normal case and which will be useful in subsequent proofs.

Lemma 9 *Let $S_1, \ldots, S_N$ be i.i.d. distributed according to* (38)*, set $S^{(1)} = \min_{1 \leq i \leq N} S_i$ and define $I_N = [-4\sqrt{\log N}, -\sqrt{\log N}]$. Then as $N, d \to \infty$ with $\log N / d^{1/3} \to 0$, we have*

$$\mathbb{P}(S^{(1)} \notin I_N) = O\left(\frac{1}{N^4}\right).$$

Proof The proof follows the same structure as the proof of Lemma 6, using Corollary 1 to relate the approximately log-normal case to the exact case.

We control the probability of the events $\{S^{(1)} > -\sqrt{\log N}\}$ and $\{S^{(1)} < -4\sqrt{\log N}\}$ separately. First, since $\sqrt{\log N} = o(d^{1/6})$ as $N, d \to \infty$ with $\log N / d^{1/3} \to 0$, by Corollary 1 we have

$$\begin{aligned}
\log \mathbb{P}(S^{(1)} > -\sqrt{\log N}) &= N \log\left(1 - G_d(-\sqrt{\log N})\right) \\
&= N \log\left(1 - (1 + o(1))\overline{\Phi}(\sqrt{\log N})\right) \\
&= -N \frac{\phi(\sqrt{\log N})}{\sqrt{\log N}}(1 + o(1))
\end{aligned}$$

using the same method as in (85). Following (87), we deduce that

$$\mathbb{P}(S^{(1)} > -\sqrt{\log N}) = O\left(\frac{1}{N^4}\right)$$

Second, we can write

$$\begin{aligned}
\mathbb{P}(S^{(1)} < -4\sqrt{\log N}) &\leq N G_d(-4\sqrt{\log N}) \\
&= N\overline{\Phi}(4\sqrt{\log N})(1 + o(1)) \\
&= O\left(\frac{1}{N^4}\right)
\end{aligned}$$

using the same method as in (88), from which the result follows. ■

B.8.3 Proof of Lemma 2

Proof of Lemma 2 The idea of the proof will be to relate it to the case where $S_1, \ldots, S_N$ are exactly normally distributed, which was proved in Section B.7.1. Recall that we have $S_1, \ldots, S_N$ with cdf G_d and $S^{(1)} = \min_{1 \leq i \leq N} S_i$. Also, let $\tilde{S}_1, \ldots, \tilde{S}_N$ be auxiliary i.i.d. standard Gaussian random variables, and set $\tilde{S}^{(1)} = \min_{1 \leq i \leq N} S_i$.

By the assumption that the $\xi_{i,j}$ are absolutely continuous with respect to the Lebesgue measure, G_d is continuous and hence we can construct $S_1, \ldots, S_N$ and $\tilde{S}_1, \ldots, \tilde{S}_N$ on a common probability space by drawing N uniform random variables $U_1, \ldots, U_N \sim U[0,1]$ and setting $S_i = G_d^{-1}(U_i)$, $\tilde{S}_i = \Phi^{-1}(U_i)$. We then have that $S^{(1)} = G_d^{-1}(U^{(1)})$ and $\tilde{S}^{(1)} = \Phi^{-1}(U^{(1)})$.

From Lemma 1 we know that

$$\mathbb{E}(\tilde{S}^{(1)}) = -\sqrt{2 \log N} + O\left(\frac{\log \log N}{\sqrt{\log N}}\right)$$

so it suffices to prove that

$$\mathbb{E}(|\tilde{S}^{(1)} - S^{(1)}|) = O\left(\frac{\log \log N}{\sqrt{\log N}}\right). \tag{91}$$

Letting $I_N = [-4\sqrt{\log N}, -\sqrt{\log N}]$, we will split the above expectation according to whether $\tilde{S}^{(1)} \in I_N$.

- Assuming first that $\tilde{S}^{(1)} \in I_N$, so that $\tilde{S}^{(1)} = o(d^{1/6})$, if we let $h \in \mathbb{R}$ be an arbitrary real satisfying $h = o_{N,d}(1)$, then using Corollary 1 we can write

$$\begin{aligned} G_d(\tilde{S}^{(1)} + h) &= \Phi(\tilde{S}^{(1)} + h)(1 + o_{N,d}(1)) \\ &= -\frac{\phi(\tilde{S}^{(1)} + h)}{\tilde{S}^{(1)} + h}(1 + o_{N,d}(1)) \\ &= \Phi(\tilde{S}^{(1)})\frac{\phi(\tilde{S}^{(1)} + h)}{\phi(\tilde{S}^{(1)})}(1 + o_{N,d}(1)) \\ &= U^{(1)} \exp\left\{-h\tilde{S}^{(1)} - h^2/2\right\}(1 + o_{N,d}(1)) \end{aligned}$$

  and so it follows by the continuity of G_d that there is a choice of h, satisfying $h = O_{N,d}(1/\sqrt{\log N})$, such that $G_d(\tilde{S}^{(1)} + h) = U^{(1)}$. We conclude that

$$|\tilde{S}^{(1)} - S^{(1)}| \leq O_{N,d}\left(\frac{1}{\sqrt{\log N}}\right)$$

  and so

$$\mathbb{E}\left(|\tilde{S}^{(1)} - S^{(1)}| \mathbb{1}_{\{\tilde{S}^{(1)} \in I_N\}}\right) \leq O\left(\frac{1}{\sqrt{\log N}}\right). \tag{92}$$

- On the other hand, we may also write

$$\begin{aligned} \mathbb{E}\left(|S_1| \mathbb{1}_{\{\tilde{S}^{(1)} \notin I_N\}}\right) &\leq \mathbb{E}\left(|S_1| \mathbb{1}_{\{|S_1| \geq N^2\}}\right) + \mathbb{E}\left(|S_1| \mathbb{1}_{\{|S_1| < N^2\} \cap \{\tilde{S}^{(1)} \notin I_N\}}\right) \\ &\leq \frac{1}{N^2}\mathbb{E}(|S_1|^2) + N^2 \mathbb{P}\left(\tilde{S}^{(1)} \notin I_N\right) \\ &\leq O\left(\frac{1}{N^2}\right) \end{aligned}$$

where we have used $\mathbb{E}(|S_1|^2) = 1$ and Lemma 6 to bound the second term.

The same result also holds with $\tilde{S}_1$ in place of S_1 (e.g. by considering taking the ξ_i to be i.i.d. Gaussians), and so we see that

$$\mathbb{E}\left(|\tilde{S}^{(1)} - S^{(1)}|\mathbb{1}_{\{\tilde{S}^{(1)} \notin I_N\}}\right) \leq \sum_{i=1}^{N} \mathbb{E}\left(|\tilde{S}_i - S_i|\mathbb{1}_{\{\tilde{S}^{(1)} \notin I_N\}}\right) = O\left(\frac{1}{N}\right) \tag{93}$$

Combining (92) and (93) yields (91) and the proof is concluded. ■

B.8.4 Proof of Proposition 6

Proof of Proposition 6 First, note that since the weights satisfy (37), we may write

$$\log \overline{w}^{(N)} = -\log\left(\mathbb{E}(\exp(-\sigma\sqrt{d}S_1))\right) - \sigma\sqrt{d}S^{(1)}.$$

In addition, using the definition of S_1 written in (38), that is

$$S_1 = \frac{1}{\sigma\sqrt{d}} \sum_{j=1}^{d} \xi_{1,j},$$

where the $\xi_{1,1}, \ldots, \xi_{1,d}$ are i.i.d. random variables, we have that

$$\begin{aligned}\mathbb{E}(\exp(-\sigma\sqrt{d}S_1)) &= \prod_{j=1}^{d} \mathbb{E}(\exp(-\xi_{1,j})) \\ &= \left(\mathbb{E}(\exp(-\xi_{1,1}))\right)^d.\end{aligned}$$

Thus,

$$-\log\left(\mathbb{E}(\exp(-\sigma\sqrt{d}S_1))\right) = -d\log\mathbb{E}(\exp(-\xi_{1,1})) = -da \tag{94}$$

By Jensen's inequality applied to the strictly convex function $u \mapsto -\log(u)$, we have that

$$a < \mathbb{E}(\xi_{1,1}) = 0.$$

Hence,

$$\log \overline{w}^{(N)} = -da - \sigma\sqrt{d}S^{(1)} \tag{95}$$

with $a > 0$. Following the proof of Proposition 4 in Appendix B.7.2, we can then conclude by combining (95) with the definition of $\Delta_{N,d}^{MAX}(\theta, \phi)$ in (29). Indeed,

$$\Delta_{N,d}^{MAX}(\theta, \phi; x) = -da - \sigma\sqrt{d}\mathbb{E}(S^{(1)}) + \frac{\log N}{\alpha - 1}$$

and using (34), we deduce:

$$\Delta_{N,d}^{MAX}(\theta,\phi;x) = -da + \sigma\sqrt{d}\left(\sqrt{2\log N} + O\left(\frac{\log\log N}{\sqrt{\log N}}\right)\right) + \frac{\log N}{\alpha-1}$$
$$= -da\left\{1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + \frac{1}{1-\alpha}\frac{\log N}{da} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right\},$$

Hence, using now that $\frac{1}{1-\alpha}\frac{\log N}{da} = O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)$ under the assumption $\log N/d^{1/3} \to 0$, we can deduce

$$\lim_{\substack{N,d\to\infty \\ \log N/d^{1/3}\to 0}} \Delta_{N,d}^{(\alpha,MAX)}(\theta,\phi) + da\left\{1 - \frac{\sigma}{a}\sqrt{\frac{2\log N}{d}} + O\left(\frac{\log\log N}{\sqrt{d\log N}}\right)\right\} = 0,$$

which yields the desired result. ■

B.8.5 Proof of Proposition 7

Proof of Proposition 7 The proof will build on the proof of Proposition 5. As in that proof, we denote $\sigma_\alpha = (1-\alpha)\sigma$ for all $\alpha \in [0,1)$ and observe that, conditional upon $S^{(1)}$, we can think of the sum in (27) as the sum over $N-1$ i.i.d. random variables

$$\mathbb{E}(T_{N,d}^{(\alpha)}|S^{(1)}) = (N-1)\mathbb{E}\left(\exp\left(-\sigma_\alpha\sqrt{d}(S - S^{(1)})\right)\right)$$

where the expectation is w.r.t. the density of S given by

$$p(z) = \frac{g_d(z)}{\overline{G_d}(S^{(1)})}\mathbb{I}(z \geq S^{(1)}),$$

with g_d denoting the pdf of S_1 and $\overline{G_d}(x) = \int_x^\infty g_d(z)\mathrm{d}z$ for all $x \in \mathbb{R}$, that is

$$\mathbb{E}(T_{N,d}|S^{(1)}) = (N-1)\frac{\int_{S^{(1)}}^\infty \exp(-\sigma_\alpha\sqrt{d}(z - S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G_d}(S^{(1)})}.$$

We are thus required to show that

$$(N-1)\mathbb{E}\left[\frac{\int_{S^{(1)}}^\infty \exp(-\sigma_\alpha\sqrt{d}(z - S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G_d}(S^{(1)})}\right] \to 0 \tag{96}$$

as $N, d \to 0$ with $\log N/d^{1/3} \to 0$. First, we show that contributions due to extreme values of $S^{(1)}$ are negligible. To see this, note that

$$\left|\frac{\int_{S^{(1)}}^\infty \exp(-\sigma_\alpha\sqrt{d}(z - S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G_d}(S^{(1)})}\right| \leq 1,$$

so that Lemma 9 implies

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\notin I_N\}}\frac{\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G}_d(S^{(1)})}\right] \leq (N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\notin I_N\}}\right) \to 0.$$

Hence it suffices to show that

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\frac{\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))g_d(z)\mathrm{d}z}{\overline{G}_d(S^{(1)})}\right] \to 0.$$

Note that by Corollary 1, we have $\overline{G}_d(S^{(1)}) \geq \overline{G}_d(0) = 1-\Phi(0)(1+o(1))$ on the event $\{S^{(1)}\in I_N\}$ as $N, d \to \infty$ with $\log N/d^{1/3} \to 0$, so $\overline{G}_d(S^{(1)})$ is uniformly bounded below. It thus suffices to prove

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))g_d(z)\mathrm{d}z\right] \to 0.$$

We will in fact show that

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\phi(z)\mathrm{d}z\right] \to 0 \tag{97}$$

and

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\left|\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\left(g_d(z)-\phi(z)\right)\mathrm{d}z\right|\right] \to 0. \tag{98}$$

- **Proof of** (97)**.** Following the proof of Proposition 5, we see that (89) holds whenever $S^{(1)}\in I_N$. Restricting to the event $\{S^{(1)}\in I_N\}$ and taking expectations over $S^{(1)}$, we get

$$\begin{aligned}&(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\phi(z)\mathrm{d}z\right]\\ &\leq (N-1)\frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\Phi(S^{(1)})\right](1+o(1)).\end{aligned}$$

  Since $S^{(1)} = o(d^{1/6})$, Corollary 1 implies that

$$\Phi(S^{(1)}) = G_d(S^{(1)})(1+o_{N,d}(1)),$$

  where the uniformity over x in the statement of the theorem implies that the implicit constant is independent of $S^{(1)}$. Restricting to $\{S^{(1)}\in I_N\}$ and taking expectations again, noting that $G_d(S^{(1)})$ is distributed as the minimum of N uniform random variables on $[0,1]$, we see

$$\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\Phi(S^{(1)})\right] = \mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}G_d(S^{(1)})\right](1+o(1)) \leq \frac{1}{N+1}(1+o(1)). \tag{99}$$

  We conclude that

$$(N-1)\mathbb{E}\left[\mathbb{1}_{\{S^{(1)}\in I_N\}}\int_{S^{(1)}}^{\infty}\exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\phi(z)\mathrm{d}z\right] \leq \frac{4\sqrt{\log N}}{\sigma_\alpha\sqrt{d}}(1+o(1)) \to 0,$$

  proving (97).

- **Proof of** (98)**.** Two applications of integration by parts give

$$\int_{S^{(1)}}^{\infty} \exp\{-\sigma_\alpha\sqrt{d}(z-S^{(1)})\}g_d(z)\mathrm{d}z = -G_d(S^{(1)}) + \int_{S^{(1)}}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z-S^{(1)})\}G_d(z)\mathrm{d}z$$

and

$$\int_{S^{(1)}}^{\infty} \exp\{-\sigma_\alpha\sqrt{d}(z-S^{(1)})\}\phi(z)\mathrm{d}z = -\Phi(S^{(1)}) + \int_{S^{(1)}}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z-S^{(1)})\}\Phi(z)\mathrm{d}z.$$

It follows that

$$\begin{aligned}&\left|\int_{S^{(1)}}^{\infty} \exp(-\sigma_\alpha\sqrt{d}(z-S^{(1)}))\left(g_d(z)-\phi(z)\right)\mathrm{d}z\right| \\ &\leq \left|\Phi(S^{(1)})-G_d(S^{(1)})\right| + \left|\int_{S^{(1)}}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z-S^{(1)})\}(G_d(z)-\Phi(z))\mathrm{d}z\right|.\end{aligned}$$

We now deal with each of these terms separately. On the event $\{S^{(1)} \in I_N\}$, we know $\Phi(S^{(1)}) = G_d(S^{(1)})(1+o_{N,d}(1))$ so we have

$$\left|\Phi(S^{(1)})-G_d(S^{(1)})\right| \leq o_{N,d}(1)G_d(S^{(1)})$$

and so by restricting to $\{S^{(1)} \in I_N\}$ and taking expectations

$$(N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\in I_N\}}\left|\Phi(S^{(1)})-G_d(S^{(1)})\right|\right) \leq (N-1)\cdot o(1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)}\in I_N\}}G_d(S^{(1)})\right) \to 0.$$

along the same lines as (99).

We split the second term as an integral from $S^{(1)}$ to 0 and an integral from 0 to ∞ and we write:

$$\left|\int_{S^{(1)}}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z-S^{(1)})\}(G_d(z)-\Phi(z))\mathrm{d}z\right| \leq A_1 + A_2$$

with

$$\begin{aligned}A_1 &= \int_{S^{(1)}}^{0} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z-S^{(1)})\}\left|G_d(z)-\Phi(z)\right|\mathrm{d}z \\ A_2 &= \int_{0}^{\infty} \sigma_\alpha\sqrt{d}\exp\{-\sigma_\alpha\sqrt{d}(z-S^{(1)})\}\mathrm{d}z.\end{aligned}$$

Again, we bound each term individually. For A_2, assuming that $S^{(1)} \in I_N$ we have the bound

$$A_2 = \exp\{\sigma_\alpha\sqrt{d}S^{(1)}\} \leq \exp\{-\sigma_\alpha\sqrt{d\log N}\} \leq \exp\{-\sigma_\alpha(\log N)^2\}$$

for sufficiently large N, d with $\log N/d^{1/3} \to \infty$ and so $(N-1)\mathbb{E}(\mathbb{1}_{\{S^{(1)} \in I_N\}} A_2) \to 0$. To bound A_1, note that by Corollary 1

$$|G_d(z) - \Phi(z)| \leq o_{N,d}(1)\Phi(z)$$

for all $z \in [S^{(1)}, 0]$ so long as $S^{(1)} \in I_N$, and hence under this assumption we can write

$$A_1 \leq o_{N,d}(1) \int_{S^{(1)}}^{0} \sigma_\alpha \sqrt{d} \exp\{-\sigma_\alpha \sqrt{d}(z - S^{(1)})\}\Phi(z)\mathrm{d}z.$$

Changing the upper limit from 0 to ∞, which can only weaken the bound, and then integrating by parts gives

$$\begin{aligned}
&\leq o_{N,d}(1) \int_{S^{(1)}}^{\infty} \sigma_\alpha \sqrt{d} \exp\{-\sigma_\alpha \sqrt{d}(z - S^{(1)})\}\Phi(z)\mathrm{d}z \\
&\leq o_{N,d}(1) \left\{ \Phi(S^{(1)}) + \int_{S^{(1)}}^{\infty} \exp\{-\sigma_\alpha \sqrt{d}(z - S^{(1)})\}\phi(z)\mathrm{d}z \right\}.
\end{aligned}$$

We conclude that

$$\begin{aligned}
(N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)} \in I_N\}} A_1\right) \leq o(1) \Big\{ &(N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)} \in I_N\}} \Phi(S^{(1)})\right) \\
&+ (N-1)\mathbb{E}\left(\mathbb{1}_{\{S^{(1)} \in I_N\}} \int_{S^{(1)}}^{\infty} \exp\{-\sigma_\alpha \sqrt{d}(z - S^{(1)})\}\phi(z)\mathrm{d}z\right) \Big\}
\end{aligned}$$

The former term tends to zero by (99) and the latter tends to zero by (97). We thus see that (98) holds, completing the proof.

■

B.8.6 Proof of Example 4

Proof of Example 4 Recall that by (82): for all $i = 1 \dots N$,

$$\log \overline{w}_i = \frac{d}{2} \log\left(\frac{4}{3}\right) - \left\| z_i - \frac{\theta + x}{2} \right\|^2 + \frac{3}{4} \|z_i - Ax - b\|^2.$$

We want to show that if $z_i \sim q_\phi(\cdot|x) = \mathcal{N}(Ax + b, 2/3\ \boldsymbol{I}_d)$, then $\log \overline{w}_i$ can be written in the form of (37). For this purpose, denote $\mathbf{1} = (1, \dots, 1)^T$ and observe that there exists an orthogonal matrix U such that $U\left(\frac{\theta+x}{2} - Ax - b\right) = \lambda \mathbf{1}$. We can then sample $z_i \sim \mathcal{N}(Ax + b, 2/3\ \boldsymbol{I}_d)$ by setting $z_i = U^{-1} y_i + Ax + b$ where $y_i \sim \mathcal{N}(0, 2/3\ \boldsymbol{I}_d)$. With this parameterization, (82) becomes

$$\begin{aligned}
\log \overline{w}_i &= -\left\| U^{-1} y_i + Ax + b - \frac{\theta + x}{2} \right\|^2 + \frac{3}{4} \|U^{-1} y_i\|^2 + const. \\
&= -\|y_i - \lambda \mathbf{1}\|^2 + \frac{3}{4} \|y_i\|^2 + const. \\
&= -\sum_{j=1}^{d} \left\{ (y_{ij} - \lambda)^2 - \frac{3}{4} y_{ij}^2 \right\} + const.
\end{aligned}$$

where *const*. denotes a fixed constant which depends only on d, θ, A, b and x.

Let us now set $\zeta_{ij} = (y_{ij} - \lambda)^2 - 3y_{ij}^2/4$ and $\xi_{i,j} = \zeta_{ij} - \mathbb{E}(\zeta_{ij})$. Since $y_i \sim \mathcal{N}(0, 2/3\boldsymbol{I}_d)$ it follows that $\xi_{i,1}, \ldots, \xi_{i,d}$ are i.i.d. random variables with $\mathbb{E}(\xi_{i,j}) = 0$. Now defining $\sigma^2 = \mathbb{V}(\xi_{1,1}) < \infty$, we have that σ^2 can be computed analytically by observing that

$$\mathbb{E}(\zeta_{ij}) = \mathbb{E}\left((y_{ij} - \lambda)^2\right) - \mathbb{E}\left(\frac{3}{4}y_{ij}^2\right) = \frac{1}{4}\mathbb{E}(y_{ij}^2) + \lambda^2 = \frac{1}{6} + \lambda^2$$

from which we can deduce that

$$\begin{aligned}\sigma^2 &= \mathbb{E}\left([\zeta_{ij} - \mathbb{E}(\zeta_{ij})]^2\right) = \mathbb{E}\left(\left[\frac{1}{4}y_{ij}^2 - 2\lambda y_{ij} - \frac{1}{6}\right]^2\right) \\ &= \frac{1}{16}\mathbb{E}(y_{ij}^4) + \left(4\lambda^2 - \frac{1}{12}\right)\mathbb{E}(y_{ij}^2) + \frac{1}{36} \\ &= \frac{1}{18} + \frac{8}{3}\lambda^2 < \infty.\end{aligned}$$

Hence, (37) holds by defining S_i as in (38) (noting that the constant terms must match since $\overline{w}_i$ is normalised and so has expected value 1). In addition, we can also analytically compute the quantity a defined in (39) by noting that

$$\begin{aligned}\mathbb{E}\left(\exp\left(-\xi_{1,1}\right)\right) &= \int_{-\infty}^{\infty} \exp\left(-\left[\frac{1}{4}u^2 - 2\lambda u - \frac{1}{6}\right]\right) \cdot \frac{1}{\sqrt{2\pi \cdot \frac{2}{3}}} e^{-\frac{3}{4}u^2}\, \mathrm{d}u \\ &= \sqrt{\frac{3}{4}} \exp\left(\lambda^2 + \frac{1}{6}\right)\end{aligned}$$

so that

$$a = \lambda^2 + \frac{1}{6} + \frac{1}{2}\log\left(\frac{3}{4}\right).$$

Finally, we check that (A2) holds in the case where $(\theta, \phi) = (\theta^\star, \phi^\star)$. For this choice of parameters, $\lambda = 0$, hence σ is independent of d. Furthermore, $\xi_{i,j}$ is clearly absolutely continuous with respect to the Lebesgue measure, and the distribution of $\xi_{i,j}$ is independent of d. It follows that (A2)a holds using our previous observations.

To now check (A2)b, we let $k \geq 3$. In that case, $u \mapsto |u|^k$ is convex and for all real-valued u_1, u_2 and u_3, we have that:

$$\begin{aligned}\left|\frac{1}{3}u_1 + \frac{1}{3}u_2 + \frac{1}{3}u_3\right|^k &\leq \left|\frac{1}{3}|u_1| + \frac{1}{3}|u_2| + \frac{1}{3}|u_3|\right|^k \\ &\leq \frac{1}{3}\left(|u_1|^k + |u_2|^k + |u_3|^k\right)\end{aligned}$$

so that, setting $u_1 = (y_{ij} - \lambda)^2$, $u_2 = -\frac{3}{4}y_{ij}^2$ and $u_3 = -\mathbb{E}(\zeta_{ij})$, it holds that

$$\left|(y_{ij} - \lambda)^2 - \frac{3}{4}y_{ij}^2 - \mathbb{E}(\zeta_{ij})\right|^k \leq 3^{k-1}\left((y_{ij} - \lambda)^{2k} + \left(\frac{3}{4}y_{ij}\right)^{2k} + |\mathbb{E}(\zeta_{ij})|^k\right).$$

Using a similar argument applied to $(y_{ij} - \lambda)^{2k}$, we then deduce that

$$\left|(y_{ij} - \lambda)^2 - \frac{3}{4}y_{ij}^2 - \mathbb{E}(\zeta_{ij})\right|^k \leq 3^{k-1}\left(2^{2k-1}\left(y_{ij}^{2k} + \lambda^{2k}\right) + \left(\frac{3}{4}y_{ij}\right)^{2k} + |\mathbb{E}(\zeta_{ij})|^k\right).$$

Hence,

$$\begin{aligned}
\mathbb{E}\left(|\xi_{i,1}|^k\right) &= \mathbb{E}\left(\left|(y_{ij}-\lambda)^2 - \frac{3}{4}y_{ij}^2 - \mathbb{E}\left(\zeta_{ij}\right)\right|^k\right) \\
&\leq 3^{k-1}\left(2^{2k-1}\left(\mathbb{E}(y_{ij}^{2k}) + \lambda^{2k}\right) + \frac{3^{2k}}{4^{2k}}\mathbb{E}(y_{ij}^{2k}) + \left|\mathbb{E}\left(\zeta_{ij}\right)\right|^k\right) \\
&\leq 3^{k-1}\left((2^{2k-1} + \frac{3^{2k}}{4^{2k}})\cdot(2k-1)!!\cdot\frac{2^k}{3^k} + 2^{2k-1}\lambda^{2k} + \left|\mathbb{E}\left(\zeta_{ij}\right)\right|^k\right) \\
&\leq 3^{k-1}\left((2^{2k-1} + \frac{3^{2k}}{4^{2k}})\cdot\frac{2^{2k}}{3^k}\cdot k! + 2^{2k-1}\lambda^{2k} + \left(\frac{1}{6}+\lambda^2\right)^k\right)
\end{aligned}$$

where we have used that the $(2k)$-th moment of a standard Gaussian random variable is $(2k-1)!!$. Finally, since $\lambda = 0$ in our case, we obtain that

$$\mathbb{E}\left(|\xi_{i,1}|^k\right) \leq k! K^{k-2}\sigma^2$$

for some sufficiently large choice of K which is independent of d, and (A2) thus holds. ■

Remark 4 *We have obtained that* (A2) *holds when (θ, ϕ) are equal to the optimal parameters. If we now assume that x, θ and ϕ are initially drawn from Gaussian distributions with bounded covariance matrices (like it is the case in Rainforth et al., 2018), we anticipate that* (A2) *should approximately hold even for values of the parameters other than the optimal choice. Notice indeed that in that case we inuitively expect $\left\|\frac{\theta+x}{2} - Ax - b\right\| = O(\sqrt{d})$ for most values of x, θ and ϕ. It follows that we should expect $\lambda = O(1)$ and $\sigma = \Theta(1)$ in practice as $d \to \infty$, and so* (A2) *should approximately hold.*

Appendix C. Futher Details Regarding Related Proof Techniques

As mentioned in Section 5, a number of our proof techniques differs significantly from/alter parts of known proofs, which in some cases impacts the corresponding theoretical results. Namely,

- **Theorem 1**. The proof of this result is based on the proof for the case $\alpha = 0$ written in the arxiv version of 5 Mar 2019 of (Rainforth et al., 2018, Theorem 1), which was the latest version available to us. Nevertheless, and contrary to Rainforth et al. (2018), we (i) use an explicit form for the remainder term in Taylor's theorem rather than the mean value form of the remainder, allowing us to get more precise control on the magnitude of the remainder and its gradients, and (ii) we consequently rely on Lemma 3, which is a non-immediate extension of (Rainforth et al., 2018, Lemma 1).

  This significantly impacts the proof technique and as a result, the main difference in terms of assumptions compared to (Rainforth et al., 2018, Theorem 1) is that, for a given $\alpha \in [0, 1)$, we are requiring the eighth moments of $\tilde{w}_{1,1}^{1-\alpha}$, $\partial\tilde{w}_{1,1}^{1-\alpha}/\partial\theta_\ell$ and $\partial\tilde{w}_{1,1}^{1-\alpha}/\partial\phi_{\ell'}$ to be finite in Theorem 1, where (Rainforth et al., 2018, Theorem 1) asked for the fourth moments to be finite with $\alpha = 0$. In addition, we need to further assume that there exists some $N \in \mathbb{N}^\star$ for which $\mathbb{E}((1/\hat{Z}_{1,N,\alpha})^4) < \infty$.

- **Proposition 5**. The proof of this result mostly mirrors the proof written in (Snyder et al., 2008, Section 4.a) which considers the case $\alpha = 0$ and is used in the context of particle filtering. The main difference is that we require an additional lemma (Lemma 6 of Appendix B.7.3) to provide us with a more precise concentration result for $S^{(1)}$. This lemma will also have other uses in the rest of the paper. This does not result in any change of assumptions compared to (Snyder et al., 2008, Section 4.a).

- **Proposition 7**. The proof of this result is arguably the one that required the most alterations out of the three discussed here. It borrows some ideas from the proofs written in the context of particle filtering in Li et al. (2005); Bengtsson et al. (2008); Li et al. (2005), which all aimed at establishing that $\mathbb{E}(T_{N,d}^{(0)}) \to 0$ in the approximate log-normal case under some conditions on N and d. However, we are significantly more thorough in our control of the error terms, which we bound precisely with the aid of two results from Saulis and Statulevičius (2000) and Lemma 9, rather than simply working under convergence in probability.

  In terms of assumptions, it is closest to Li et al. (2005), except for the fact that our result relies on a Bernstein condition while Li et al. (2005) uses Cramer's conditions. Both are in fact equivalent in the i.i.d. setting, and we choose to use Bernstein condition as we believe it might make it easier to generalize our result beyond the i.i.d. case using the results from Saulis and Statulevičius (2000) recalled in Appendix B.8.1.

Appendix D. Additional Numerical Experiments and Derivation Details

D.1 Gaussian Example from Section 6.1

Figure 13 empirically confirms that the asymptotic regime predicted by Theorem 3 does not reflect what is happening in reality in the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$ when the dimension d increases, N is small and the distribution of the weight is log-normal.

Note that we only plotted the variational gap $\Delta_{N,d}^{(\alpha)}(\theta, \phi; x)$ for the cases $d = \{10, 100\}$ in the figure above. This is due to the fact that when $d = 1000$, computing γ_α^2 the $1/N$ term returns an overflow, further illustrating the limitations of the approach from Theorem 3 in the specific setting considered here. Note also that since the variance term is exponential in $(1-\alpha)^2 d$, increasing α does play a role in decreasing γ_α^2 so that the asymptotic regime predicted by Theorem 3 applies in lower dimensions (e.g. $d = 10$ with $\alpha = 0.5$).

D.2 Linear Gaussian Example from Section 6.2

D.2.1 Empirical Experiments for Theorem 3 in the Context of Section 6.2

Figure 15 empirically confirms that we need an unpractical amount of samples N for the asymptotic regime predicted by Theorem 3 to capture the behavior of the variational gap as d increases when $\sigma_{\text{perturb}} = 0$. Note that similar plots and conclusions can be obtained for $\sigma_{\text{perturb}} \in \{0.01, 0.5\}$. Those are not given here for the sake of conciseness.

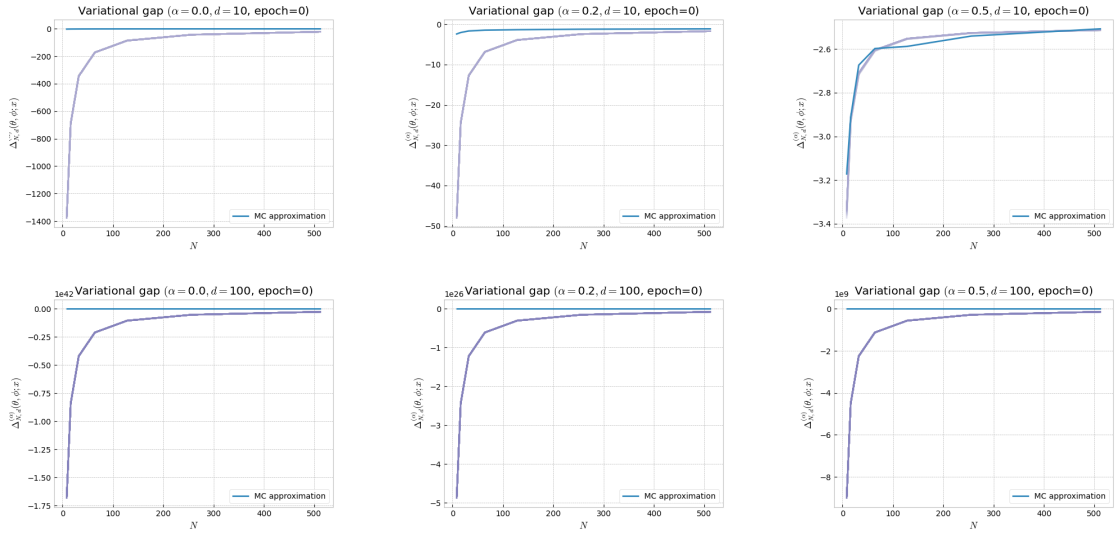

Figure 13: Plotted in blue is the MC estimate of the variational gap $\Delta^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) for the toy example described in Section 6.1 as a function of N, for varying values of (α, d) and with $(\theta, \phi) = (0 \cdot \boldsymbol{u_d}, \boldsymbol{u}_d)$ so that $B_d = \sqrt{d}$. Plotted in purple are curves of the form (42) with tailored values of c_1.

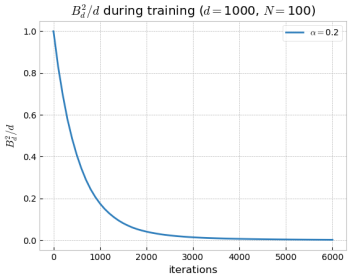

Figure 14: Evolution of B_d^2/d during the training of the ϕ parameter for the toy example described in Section 6.1.

D.2.2 Additional Experimental Results for Section 6.2

We provide some additional results in the context of Section 6.2 regarding the signal-to-noise ratio (SNR) in the doubly-reparameterized case and the Mean Squared Error (MSE) for the VR-IWAE bound and its θ, ϕ gradients.

- **SNR in the doubly-reparameterized case.** In line with Tucker et al. (2019), we observe in Figure 16 that using the doubly-reparameterized gradient estimator for ϕ increases the SNR when $\alpha = 0$. We in fact see that the SNR is increased for all values

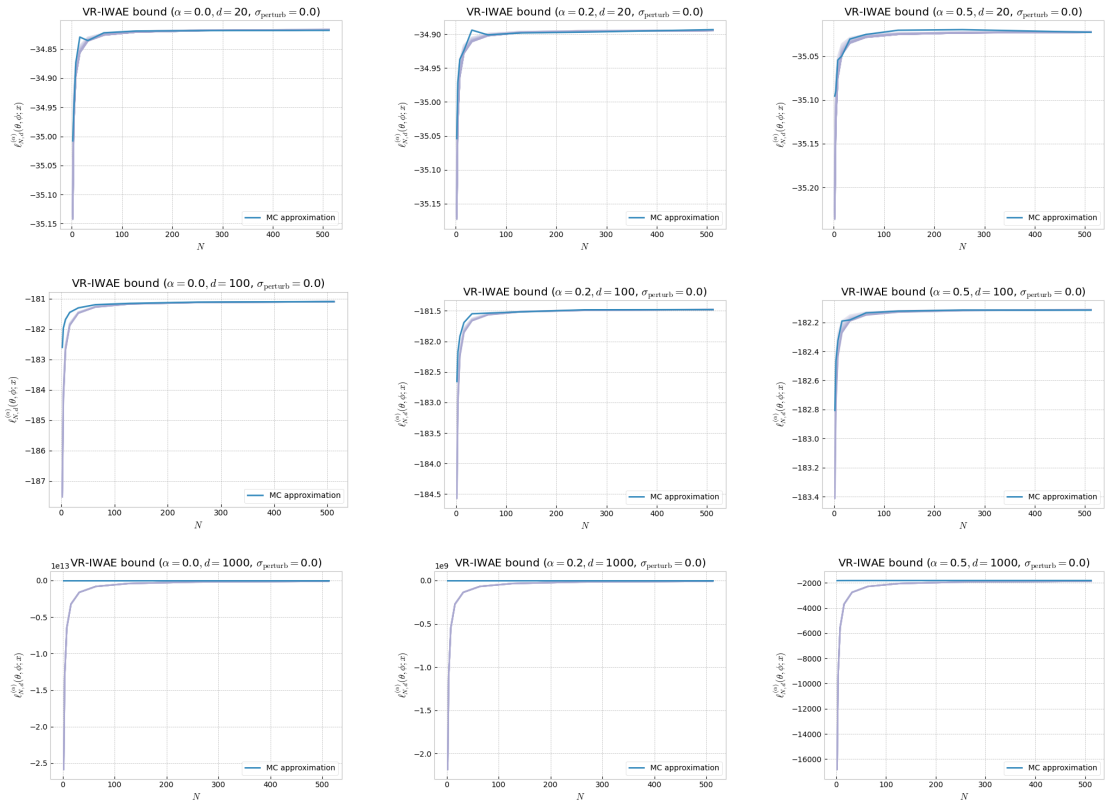

Figure 15: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 1000 MC samples) for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d). Plotted in purple are curves of the form (45) with tailored values of c_1.

of α, extending the conclusions from Tucker et al. (2019) to $\alpha \in [0, 1)$ in the example considered here.

However, as we get further away from the optimum ($\sigma_{\text{perturb}} = 0.5$) and/or increase the dimension ($d = 1000$), we observe that it still remains challenging to obtain an increasing SNR for the ϕ gradients for small values of α, even when using doubly-reparameterized gradient estimators.

- **MSE for the VR-IWAE bound and its θ, ϕ gradients.** We observe on Figures 17 and 18 that while increasing α does not lower the MSE of the VR-IWAE estimator

$$\frac{1}{1-\alpha} \log \left(\frac{1}{N} \sum_{j=1}^{N} w_{\theta,\phi}(Z_j; x)^{1-\alpha} \right)$$

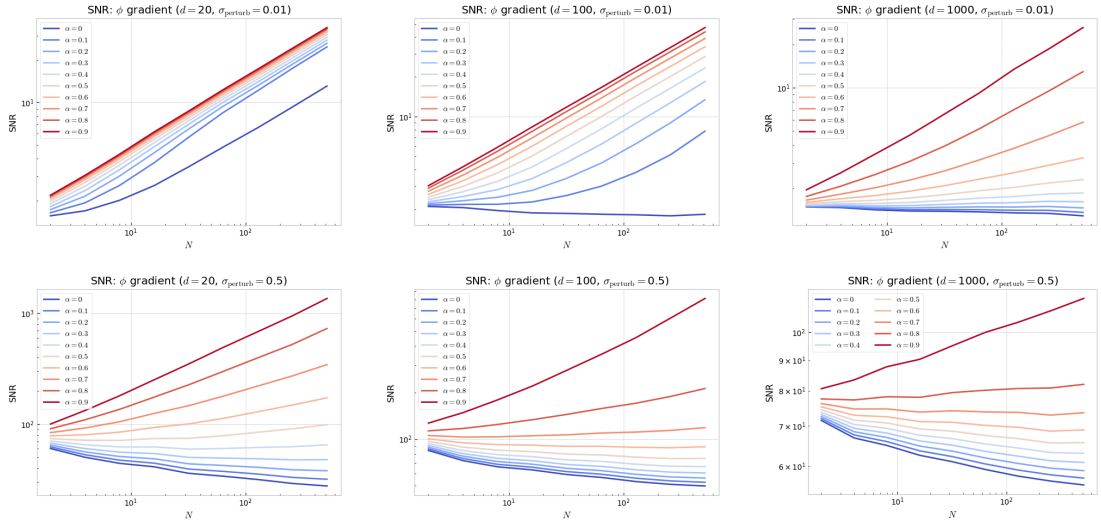

Figure 16: Plotted is the SNR of the inference network (ϕ) gradients in the doubly-reparameterized case (computed over 1000 MC samples) for the linear Gaussian example in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

for log-likelihood estimation, it can be useful in lowering the MSE of its θ gradients

$$\frac{1}{1-\alpha}\nabla_\theta \log\left(\frac{1}{N}\sum_{j=1}^{N} w_{\theta,\phi}(Z_j;x)^{1-\alpha}\right)$$

compared to the θ gradients of the true log-likelihood $\nabla_\theta \ell_d(\theta;x)$.

In the low perturbation regime ($\sigma_{\text{perturb}} = 0.01$) and in medium to high dimensions ($d = 100, 1000$), we indeed see in Figure 17 that every tested value of $\alpha > 0$ achieves lower θ gradient MSE than $\alpha = 0$ for low values of N. As we increase to $N = 2^9$, the value of α achieving the lowest MSE is $\alpha = 0.3$ for $d = 100$, and $\alpha = 0.8$ for $d = 1000$. This sheds light on a bias-variance tradeoff between low bias at $\alpha = 0$ and low variance at $\alpha = 1$, and is in line with the findings of Theorem 3.

In the high perturbation regime ($\sigma_{\text{perturb}} = 0.5$), we see in Figure 17 that the choice of α appears to make less of a difference, especially when the dimension d is high. This suggests that bias reduction may be more important when the inference distribution $q_\phi(z|x)$ is far from the optimum.

D.3 Variational Auto-encoder from Section 6.3

We present additional results for the VAE example discussed in Section 6.3.

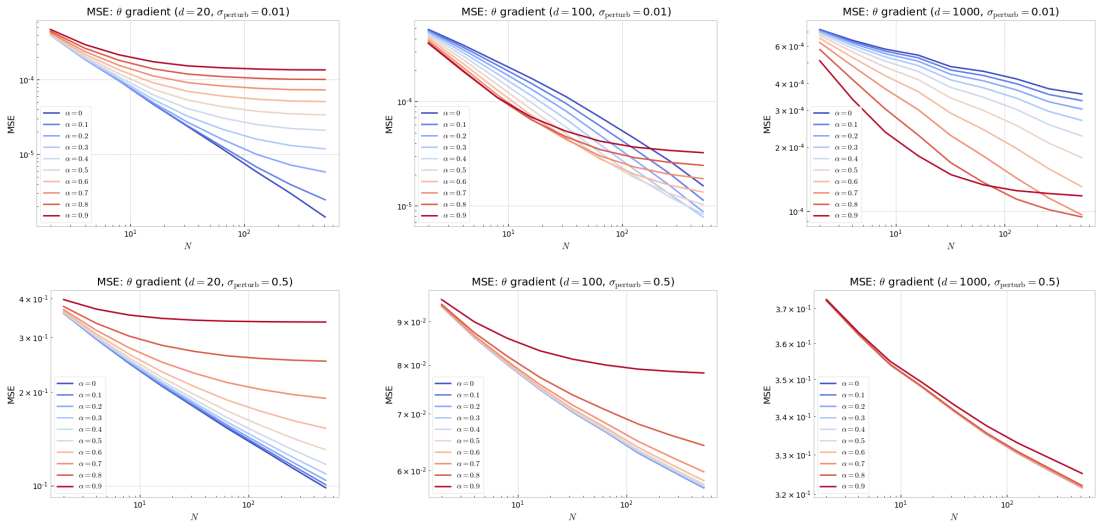

Figure 17: Plotted is the MSE of the generative network (θ) gradients (computed over 1000 MC samples) compared to the log-likelihood gradients for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

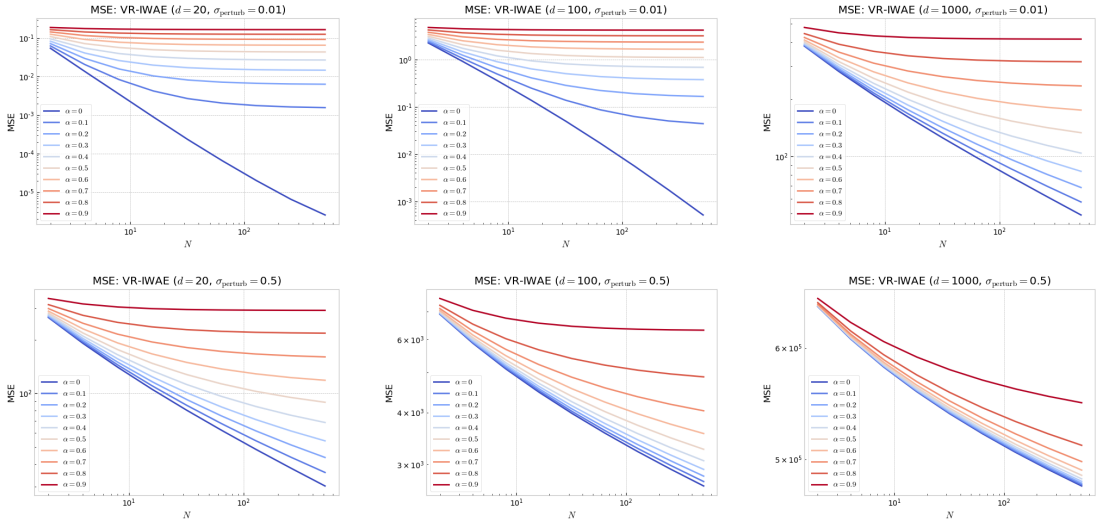

Figure 18: Plotted is the MSE of the VR-IWAE estimate (computed over 1000 MC samples) compared to the log-likelihood gradients for the linear Gaussian example described in Section 6.2 as a function of N, for varying values of (α, d), a randomly selected datapoint x and 10 different initializations of the parameters (θ, ϕ).

D.3.1 Complementary Plots for the VR-IWAE Bound

Figures 19 and 20 provide additional plots to Figures 8 and 9 reinforcing the conclusions drawn in Section 6.3.

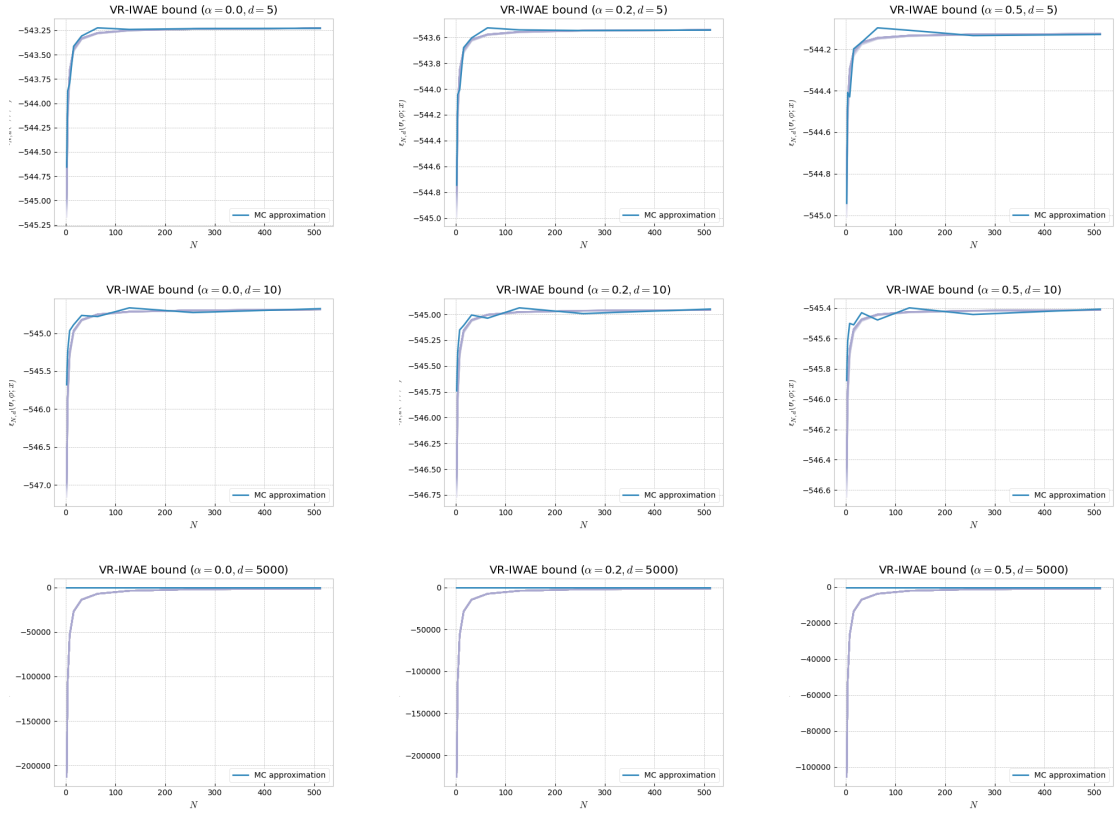

Figure 19: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 100 MC samples) for the VAE in Section 6.3, for a randomly selected datapoint x in the testing set, randomly generated model parameters (θ, ϕ) and varying values of (α, d). Plotted in purple are curves of the form (49) with tailored values of c_1.

D.3.2 Impact of α and of M on Empirical Performances

We discuss here the impact of α and of M on the empirical performances of the VR-IWAE bound metholodogy in the reparameterized and doubly-reparameterized cases.

- **Impact of α on the empirical performances.** We investigate how the choice of α impacts the Negative Log Likelihood (NLL) after training the VAE with the VR-IWAE

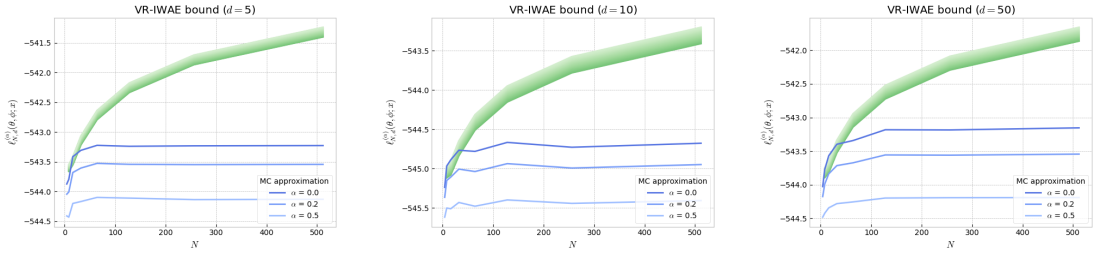

Figure 20: Plotted in blue is the MC estimate of the VR-IWAE bound $\ell^{(\alpha)}_{N,d}(\theta, \phi; x)$ (averaged over 100 MC samples) for the VAE in Section 6.3, for a randomly selected datapoint x in the testing set, randomly generated model parameters (θ, ϕ) and varying values of (α, d). Plotted in green are curves of the form (50) with tailored values of c_2.

bound. The NLL can indeed be used to evaluate the empirical performances of VAEs (since a lower NLL corresponds to a higher likelihood of the data under the VAE model, which indicates better training of the generative network θ). Furthermore, although the NLL is intractable, following Burda et al. (2016) it can be approximated using the negative IWAE bound with $N = 5000$.

We plot in Figure 21 the NLL estimate on the MNIST test set as a function of α after training VAEs on the MNIST training set using either the reparameterized ("rep") or the doubly-reparameterized ("drep") gradient estimators of the VR-IWAE objective with $N = 10, 100$ and $d = 50$. Here, all the models are trained for 1000 epochs using the Adam optimizer with learning rate $1e-3$ and batch size 100.

We observe that the doubly-reparameterized gradient estimator generally achieves better NLL results than the reparameterized one when α is fixed. In addition, for both cases the value of α achieving the best NLL performance lies in the middle of $(0, 1)$, around $\alpha = 0.5$. In line with Theorem 3, this suggests that there is a bias-variance tradeoff to consider when choosing α, and that the best setting can lie between the standard IWAE ($\alpha = 0$, low bias) and ELBO ($\alpha = 1$, low variance) objectives, with the optimal choice of α being dependent on the data set, model architecture, as well as the stochastic gradient descent procedure used for training.

- **Impact of M on the empirical performances.** We investigate how the choice of M and N affects the training of VAE when $M \times N$ is fixed in the VR-IWAE bound methodology. We plot in Figure 22 the NLL on the MNIST test set after training the VAE on the MNIST training set for 1000 epochs, with $M \times N = 100$, $d = 50$ and $\alpha \in \{0, 0.2\}$.

  We observe a bias-variance tradeoff that is similar to the analysis above for the impact of α. Indeed, the cases $M = 100$ and $M = 1$ have a particular meaning in the plots of Figure 22: $M = 100$ corresponds to the ELBO with maximum computational

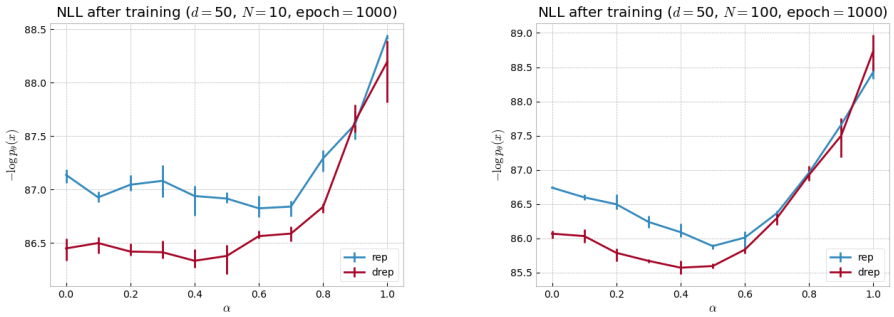

Figure 21: Plotted is the Negative Log Likelihood (NLL) estimate on the test set of the MNIST data set as described in Section 6.3 as a function of α, after training on the train set for 1000 epochs with $N \in \{10, 100\}$. The error bars are computed over 3 trials with different network initialisations and seeds during training.

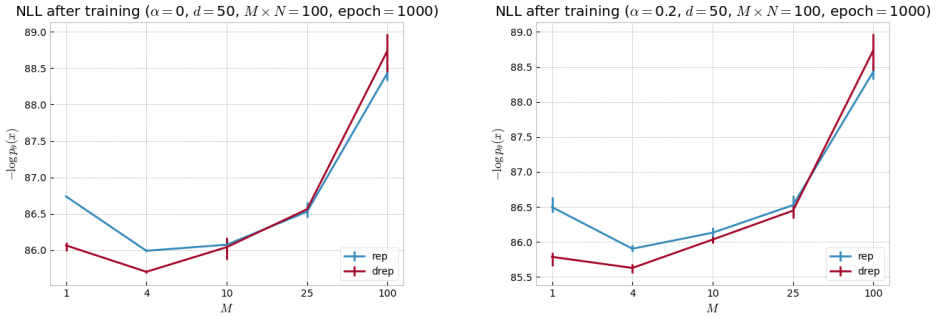

Figure 22: Plotted is the Negative Log Likelihood (NLL) estimate on the test set of the MNIST data set as described in Section 6.3 as a function of M while fixing $M \times N = 100$, after training on the train set for 1000 epochs with $\alpha = 0, 0.2$. The error bars are computed over 3 trials with different network initialisations and seeds during training.

budget for M (i.e. $\alpha = 1$, low variance), while $M = 1$ corresponds to the VR-IWAE bound with maximum computational budget for N (i.e. low bias, with the lowest bias being achieved for $\alpha = 0$). Here, the best value of test NLL is obtained for $\alpha = 0.2, M = 4, N = 25$ among our tested combinations. Note that one potential advantage of tuning α instead of M and N is that α resides on a one-dimensional continuous interval, whereas M and N are integer values and so their choice is more limited (that is, they can be more difficult to tune for a given computational budget).

References

Thomas Bengtsson, Peter Bickel, and Bo Li. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and Statistics: Essays in honor of David A. Freedman*, pages 316–334. Institute of Mathematical Statistics, 2008.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773.

Thang D. Bui, Daniel Hernández-Lobato, José Miguel Hernández-Lobato, and Yingzhen Li. Black-box α-divergence for deep generative models. In *NIPS Workshop on Approximate inference*, 2016.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations (ICLR)*, 2016.

Kamélia Daudel and Randal Douc. Mixture weights optimisation for alpha-divergence variational inference. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4397–4408. Curran Associates, Inc., 2021.

Kamélia Daudel, Randal Douc, and François Roueff. Monotonic alpha-divergence minimisation for variational inference. *Journal of Machine Learning Research*, 24(62):1–76, 2023.

Kamélia Daudel, Randal Douc, and François Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *The Annals of Statistics*, 49(4):2250 – 2270, 2021. doi: 10.1214/20-AOS2035.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2007. ISBN 9780387344713.

Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael Riis Andersen, Jonathan H Huggins, and Aki Vehtari. Challenges and opportunities in high-dimensional variational inference. volume 34, pages 7787–7798. Neural information processing systems foundation, 3 2021.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ-upper bound minimization. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4470–4479. Curran Associates, Inc., 2018.

Tomas Geffner and Justin Domke. Empirical evaluation of biased methods for alpha divergence minimization. *3rd Symposium on Advances in Approximate Bayesian Inference*, 2020.

Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3650–3659. PMLR, 18–24 Jul 2021.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernández-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *International Conference on Machine Learning*, pages 1511–1520. PMLR, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Bo Li, Thomas Bengtsson, and Peter Bickel. Curse-of-dimensionality revisited: collapse of importance sampling in very large scale systems. Technical Report 696, Department of Statistics, University of California at Berkeley, 2005.

Yingzhen Li and Yarin Gal. Dropout inference in Bayesian neural networks with alpha-divergences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2052–2061. PMLR, 06–11 Aug 2017.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6573–6583. Curran Associates, Inc., 2017.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

Valentin V. Petrov. *Limit Theorems of Probability Theory*. 06 1995. ISBN 9780198534990.

James Pickands III. Moment Convergence of Sample Extremes. *The Annals of Mathematical Statistics*, 39(3):881 – 889, 1968. doi: 10.1214/aoms/1177698320.

James Picklands III. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1 1975. doi: 10.1214/AOS/1176343003.

Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference*

on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 4277–4285. PMLR, 10–15 Jul 2018.

Simón Rodríguez-Santana and Daniel Hernández-Lobato. Adversarial α-divergence minimization for bayesian approximate inference. *Neurocomputing*, 471:260–274, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.09.076.

Leonas Saulis and Vytautas Statulevičius. *Limit Theorems on Large Deviations*, pages 185–266. 01 2000. ISBN 978-3-642-08170-5. doi: 10.1007/978-3-662-04172-7_5.

Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629 – 4640, 2008. doi: https://doi.org/10.1175/2008MWR2529.1.

George Tucker, Dieterich Lawson, Shixiang Shane Gu, and Chris J. Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. ISSN 1935-8237. doi: 10.1561/2200000001.

Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Ruqi Zhang, Yingzhen Li, Christopher De Sa, Sam Devlin, and Cheng Zhang. Meta-learning divergences for variational inference. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 4024–4032. PMLR, 13–15 Apr 2021.

Statement of Authorship

Title of paper:	Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics
Publication status:	Published
Publication details:	Kamélia Daudel, Joe Benton, Yuyang Shi, Arnaud Doucet. *Journal of Machine Learning Research*, 24(243):1–83, 2023.

Student Confirmation

Student name:	Joe Benton
Contribution to the paper:	Kamélia Daudel lead the project and, along with Arnaud Doucet, provided the overarching vision for the work, including introducing the VR-IWAE bound (Section 3.1) and predicting the phenomenon of weight collapse. I provided the analysis of the signal-to-noise ratio (Section 3.2) and the proofs in the analysis of the VR-IWAE bound in the case where both d and N go to infinity (Section 4.2). The experiments (Section 6) were implemented by Yuyang Shi.

Signature:  Date: 29/10/2023

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: George Deligiannidis, Associate Professor of Statistics

Supervisor comments: None

Signature:  Date: 31/10/2023

7 Conclusion

In this final chapter, we summarize the main contributions of this thesis and indicate some outstanding open questions which may provide the basis for further research.

7.1 Contributions

This thesis has focused on two central questions. First, under what conditions do current generative modeling methods provide good approximations to the data distribution? Second, can we extend the existing methods to new settings? We have addressed these questions in the particular contexts of diffusion models and importance weighted autoencoders.

First, we have shown how to extend diffusion models beyond real vector spaces to a much wider class of state spaces using the DMM framework. This opened up a broad range of applications, including to discrete spaces, to manifolds and to distributions over measures. Furthermore, we demonstrated that DMMs perform competitively in practice for each of these applications. In the case of discrete spaces, we also developed several optimizations that significantly improved performance, including using tau-leaping to simulate the reverse process, developing a predictor-corrector scheme to increase the accuracy of samples, and finding good choices for the forward rate matrix and the parameterization of the reverse process.

Second, we derived new bounds on the approximation error of diffusion models. We improved on state-of-the-art KL error bounds for SDE-based diffusion models using techniques inspired by stochastic localization. In addition, we provided the first bounds for the more general flow matching framework which hold in the fully deterministic setting even for data supported on a submanifold of $\mathbb{R}^d$. Both bounds converge polynomially in the relevant problem parameters, and the former holds without smoothness assumptions on the data distribution. The fast convergence rates of these bounds go some way towards explaining why diffusion models can efficiently learn to approximate the data manifold well even in high dimensions.

Finally, we turned our attention to the importance weighted autoencoder. In an attempt to address some of the limitations of the IWAE, we introduced the VR-IWAE bound, which extends both the IWAE and variational Rényi methods. We explained how the VR-IWAE allows one to interpolate between these methods, trading off between mode-covering and mode-seeking behaviour. We also explained how it improves on the signal-to-noise properties of the IWAE bound. Finally, we explored the phenomenon of weight collapse in these bounds, and demonstrated theoretically and empirically how it limits the effectiveness of both the original IWAE bound and our extension the VR-IWAE.

7.2 Open Questions

Despite the progress made during this thesis there remain many important and exciting open questions surrounding the topics discussed. We highlight some of the most interesting ones here.

Can we find additional applications of DMMs? While we presented several example applications of DMMs in Chapter 2, the scope for possible applications is likely much greater. For example, our framework could in principle incorporate state spaces which have both continuous and discrete components, or some infinite-dimensional spaces with finite representations (though care must be taken to ensure the relevant assumptions of local compactness and separability hold). Alternatively,

we could consider modeling distributions which can be naturally embedded in $\mathbb{R}^d$, but using alternative noising processes which might better respect the underlying structure in the data or encode some prior, perhaps derived from domain-specific knowledge, on the data distribution. For example, some works have explored using noising processes specifically designed for learning categorical data (Richemond et al., 2022) or data defined on manifolds with a boundary (Lou et al., 2023).

Are there ways to improve the efficiency of training or sampling DMMs? Sampling from diffusion models is often expensive, as we must typically take many sequential steps to simulate the reverse process, requiring many neural network function evaluations. There has been much work to speed up the sampling of diffusion models on $\mathbb{R}^d$, including using the probability flow ODE (Song et al., 2021b), using more efficient methods to simulate the reverse ODE (Jolicoeur-Martineau et al., 2021; Zhang et al., 2023), or using alternative noising processes (Song et al., 2021a; Lipman et al., 2023). However, much of this machinery does not transfer easily to our more general framework for DMMs.

We would like to find methods for more efficient training and sampling of general DMMs, or particular sub-classes of DMMs other than those defined using SDEs on $\mathbb{R}^d$. For example, our work using tau-leaping for discrete-space diffusion models in Chapter 3 can be viewed as one instance of such an optimization. Extensions of the probability flow ODE to the manifold setting are another (De Bortoli et al., 2022). Other possibilities might include extensions of the probability flow method to other state spaces, extensions of the flow matching framework, or higher-order sampling methods for more general Markov processes.

Is generalized score matching effective at learning distributions on state spaces other than $\mathbb{R}^d$? In Chapter 2, we introduced generalized score matching and explained how it could be used to find an approximation φ to a data distribution q_0 whose density was known only up to a normalizing constant. It is known that score matching can be effective in learning unnormalized probability distributions on $\mathbb{R}^d$ (Hyvärinen, 2005). However, we did not study the empirical effectiveness of

generalized score matching on other state spaces. It would be interesting to assess the viability of generalized score matching on alternative state spaces, and compare it to other extensions of score matching which have been proposed (Hyvärinen, 2007; Lyu, 2009; Yu et al., 2022).

What are the inductive biases imparted by diffusion models? All of the results in this thesis concerning the effectiveness of diffusion models have been framed in terms of a bound on the distance between the learned approximation and the true data distribution. However, optimal performance according to this criterion would mean perfectly reproducing the empirical data distribution – such a generative model would be useless! Instead, the value of a good generative model is that it can interpolate reasonably between the observed samples and generate datapoints which are not in the training dataset but are nevertheless perceptually convincing.

That diffusion models are able to do this with such success indicates that there is some aspect of their design which gives them a favourable inductive bias and allows them to detect the data manifold without overfitting. As yet, we lack a proper understanding of this behaviour, which seems essential for explaining the empirical success of these methods. Initial work by Pidstrigach (2022) has highlighted conditions under which diffusion models learn to completely recover the data manifold or memorize the data distribution. However, practical applications of diffusion models operate far away from the regime discussed in that work, so the broader question still remains very much open.

Does the approximation error of flow matching converge polynomially for non-smooth distributions? Our convergence results for flow matching in Chapter 4 require the assumption that $\alpha_t X_0 + \beta_t X_1$ is λ-regular for all $t \in [0, 1]$. Conversely, existing results for the general flow matching setting which do not assume smoothness of the data distributions instead require some stochasticity in the sampling procedure to smooth the resulting marginals (Albergo et al., 2023a). However, we suspect that the approximation error of fully deterministic flow matching should converge polynomially under much weaker assumptions than

those made in Chapter 4, for example requiring only a small amount of Gaussian smoothing of the target distribution (as in Chapter 5). As evidence, we note that Chen et al. (2023d) indicate the error of the deterministic probability flow ODE with the true score converges polynomially under minimal smoothness assumptions, and that Li et al. (2023) derive polynomial error bounds for the probability flow ODE with a score approximation under the additional assumption that we can control the difference between the derivatives of the true and approximate scores. We therefore ask whether it may be possible to replace the assumption of λ-regularity by a weaker or more natural assumption and still obtain polynomial error bounds, or alternatively to provide a more natural interpretation of what it means for a distribution to be λ-regular.

How does weight collapse in the VR-IWAE bound influence the gradient descent procedure? In Chapter 6, we showed that the VR-IWAE bound suffers from weight collapse in high dimensions. However, as Rainforth et al. (2018) indicate, it is not necessarily the case that a tight variational bound is preferable for learning the data distribution. Indeed, it may be the case that the VR-IWAE is an effective generative model despite weight collapse. To investigate this further, we would want to study how weight collapse affects the gradient decent procedure. For example, is it the case that a single weight dominates the gradient updates, and if so, what implications does this have for the effectiveness of the learning procedure?

Bibliography

Albergo, Michael S, Nicholas M Boffi, and Eric Vanden-Eijnden (2023a). "Stochastic Interpolants: A Unifying Framework for Flows and Diffusions". In: *arXiv preprint arXiv:2303.08797*.

Albergo, Michael S and Eric Vanden-Eijnden (2023b). "Building Normalizing Flows with Stochastic Interpolants". In: *International Conference on Learning Representations*.

Anderson, Brian D O (1982). "Reverse-time diffusion equation models". In: *Stochastic Processes and their Applications* 12.3, pp. 313–326.

Anil, Rohan et al. (2023). "PaLM 2 Technical Report". In: *arXiv preprint arXiv:2305.10403*.

Austin, Jacob, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg (2021). "Structured Denoising Diffusion Models in Discrete State-Spaces". In: *Advances in Neural Information Processing Systems*.

Ben-Hamu, Heli, Samuel Cohen, Joey Bose, Brandon Amos, Aditya Grover, Maximilian Nickel, Ricky T Q Chen, and Yaron Lipman (2022). "Matching Normalizing Flows and Probability Paths on Manifolds". In: *International Conference on Machine Learning*.

Bengtsson, Thomas, Peter Bickel, and Bo Li (2008). "Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems". In: *Probability and Statistics: Essays in honor of David A. Freedman*. Vol. 2. Institute of Mathematical Statistics, pp. 316–335.

Block, Adam, Youssef Mroueh, and Alexander Rakhlin (2022). "Generative Modeling with Denoising Auto-Encoders and Langevin Sampling". In: *arXiv preprint arXiv:2002.00107*.

Brown, Tom B et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*.

Burda, Yuri, Roger Grosse, and Ruslan Salakhutdinov (2016). "Importance Weighted Autoencoders". In: *International Conference on Learning Representations*.

Cattiaux, Patrick, Giovanni Conforti, Ivan Gentil, and Christian Léonard (2022). "Time reversal of diffusion processes under a finite entropy condition". In: *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*.

Chen, Hongrui, Holden Lee, and Jianfeng Lu (2023a). "Improved Analysis of Score-based Generative Modeling: User-Friendly Bounds under Minimal Smoothness Assumptions". In: *International Conference on Machine Learning*.

Chen, Ricky T Q, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud (2018). "Neural Ordinary Differential Equations". In: *Advances in Neural Information Processing Systems*.

Chen, Sitan, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim (2023b). "The probability flow ODE is provably fast". In: *arXiv preprint arXiv:2305.11798*.

Chen, Sitan, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang (2023c). "Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions". In: *International Conference on Learning Representations*.

Chen, Sitan, Giannis Daras, and Alexandros G Dimakis (2023d). "Restoration-Degradation Beyond Linear Diffusions: A Non-Asymptotic Analysis For DDIM-Type Samplers". In: *International Conference on Machine Learning*.

Conforti, Giovanni, Alain Durmus, and Marta Gentiloni Silveri (2023). "Score diffusion models without early stopping: finite Fisher information is all you need". In: *arXiv preprint arXiv:2308.12240*.

De Bortoli, Valentin (2022). "Convergence of denoising diffusion models under the manifold hypothesis". In: *Transactions on Machine Learning Research*.

De Bortoli, Valentin, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet (2022). "Riemannian Score-Based Generative Modeling". In: *Advances in Neural Information Processing Systems*.

De Bortoli, Valentin, James Thornton, Jeremy Heng, and Arnaud Doucet (2021). "Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling". In: *Advances in Neural Information Processing Systems*.

Domke, Justin and Daniel Sheldon (2018). "Importance Weighting and Variational Inference". In: *Advances in Neural Information Processing Systems*.

Gillespie, Daniel T (2001). "Approximate accelerated stochastic simulation of chemically reacting systems". In: *The Journal of Chemical Physics* 115.4, pp. 1716–1733.

Goodfellow, Ian J, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*.

Grathwohl, Will, Ricky T Q Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud (2019). "FFJORD: Free-form Continuous Dynamics for Scalable Reverse Generative Models". In: *International Conference on Learning Representations*.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*.

Hoogeboom, Emiel, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling (2021). "Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions". In: *Advances in Neural Information Processing Systems*.

Huang, Chin-Wei, Milad Aghajohari, Avishek Joey Bose, Prakash Panangaden, and Aaron Courville (2022). "Riemannian Diffusion Models". In: *Advances in Neural Information Processing Systems*.

Huang, Chin-Wei, Jae Hyun Lim, and Aaron Courville (2021). "A Variational Perspective on Diffusion-Based Generative Models and Score Matching". In: *Advances in Neural Information Processing Systems*.

Hyvärinen, Aapo (2005). "Estimation of Non-Normalized Statistical Models by Score Matching". In: *Journal of Machine Learning Research* 6.24, pp. 695–709.

— (2007). "Some extensions of score matching". In: *Computational Statistics and Data Analysis* 51, pp. 2499–2512.

Ingraham, John, Vikas K Garg, Regina Barzilay, and Tommi Jaakkola (2019). "Generative models for graph-based protein design". In: *Advances in Neural Information Processing Systems*.

Jolicoeur-Martineau, Alexia, Ke Li, Rémi Piché-Taillefer, Tal Kachman, and Ioannis Mitliagkas (2021). "Gotta Go Fast When Generating Data with Score-Based Models". In: *arXiv preprint arXiv:2105.14080*.

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila (2020). "Analyzing and Improving the Image Quality of StyleGAN". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119.

Kingma, Diederik P and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations*.

Kong, Zhifeng and Wei Ping (2021). "On Fast Sampling of Diffusion Probabilistic Models". In: *ICML 2021 Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.

Le, Matthew et al. (2023). "Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale". In: *arXiv preprint arXiv:2306.15687*.

Lee, Holden, Jianfeng Lu, and Yixin Tan (2022). "Convergence for score-based generative modeling with polynomial complexity". In: *Advances in Neural Information Processing Systems*.

— (2023). "Convergence of score-based generative modeling for general data distributions". In: *International Conference on Algorithmic Learning Theory*.

Li, Gen, Yuting Wei, Yuxin Chen, and Yuejie Chi (2023). "Towards Faster Non-Asymptotic Convergence for Diffusion-Based Generative Models". In: *arXiv preprint arXiv:2306.09251*.

Li, Yingzhen and Richard E Turner (2016). "Rényi Divergence Variational Inference". In: *Advances in Neural Information Processing Systems*.

Lipman, Yaron, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le (2023). "Flow Matching for Generative Modeling". In: *International Conference on Learning Representations*.

Liu, Xingchao, Chengyue Gong, and Qiang Liu (2023). "Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow". In: *International Conference on Learning Representations*.

Liu, Xingchao, Lemeng Wu, Mao Ye, and Qiang Liu (2022). "Let us Build Bridges: Understanding and Extending Diffusion Generative Models". In: *arXiv preprint arXiv:2208.14699*.

Lou, Aaron and Stefano Ermon (2023). "Reflected Diffusion Models". In: *International Conference on Machine Learning*.

Lyu, Siwei (2009). "Interpretation and Generalization of Score Matching". In: *Uncertainty in Artificial Intelligence*.

Maddison, Chris J., Dieterich Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Whye Teh (2017). "Filtering Variational Objectives". In: *Advances in Neural Information Processing Systems*.

Minka, Thomas P (2005). "Expectation Propagation for Approximate Bayesian Inference". In: *Technical Report MSR-TR-2005-173*.

Montanari, Andrea (2023). "Sampling, Diffusions, and Stochastic Localization". In: *arXiv preprint arXiv:2305.10690*.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu (2016a). "WaveNet: A Generative Model for Raw Audio". In: *arXiv preprint arXiv:1609.03499*.

Oord, Aaron van den, Nal Kalchbrenner, and Koray Kavukcuoglu (2016b). "Pixel Recurrent Neural Networks". In: *International Conference on Machine Learning*.

Pidstrigach, Jakiw (2022). "Score-Based Generative Models Detect Manifolds". In: *Advances in Neural Information Processing Systems*.

Popov, Vadim, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov (2021). "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech". In: *International Conference on Machine Learning*.

Rainforth, Tom, Adam R Kosiorek, Tuan Anh Le, Chris J Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh (2018). "Tighter Variational Bounds are Not Necessarily Better". In: *International Conference on Machine Learning*.

Ramesh, Aditya, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen (2022). "Hierarchical Text-Conditional Image Generation with CLIP Latents". In: *arXiv preprint arXiv:2204.06125*.

Rényi, Alfréd (1961). "On Measures of Entropy and Information". In: *Berkeley Symposium on Mathematical Statistics and Probability* 4.1, pp. 547–561.

Rezende, Danilo Jimenez and Shakir Mohamed (2015). "Variational Inference with Normalizing Flows". In: *International Conference on Machine Learning*.

Richemond, Pierre H, Sander Dieleman, and Arnaud Doucet (2022). "Categorical SDEs with Simplex Diffusion". In: *arXiv preprint arXiv:2210.14784*.

Rozen, Noam, Aditya Grover, Maximilian Nickel, and Yaron Lipman (2021). "Moser Flow: Divergence-based Generative Modeling on Manifolds". In: *Advances in Neural Information Processing Systems*.

Saharia, Chitwan et al. (2022). "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding". In: *Advances in Neural Information Processing Systems*.

Saulis, Leonas and Vytautas Statulevičius (1991). *Limit Theorems for Large Deviations*. Springer Science & Business Media.

Sohl-Dickstein, Jascha, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli (2015). "Deep Unsupervised Learning Using Nonequilibrium Thermodynamics". In: *International Conference on Machine Learning*.

Song, Jiaming, Chenlin Meng, and Stefano Ermon (2021a). "Denoising Diffusion Implicit Models". In: *International Conference on Learning Representations*.

Song, Yang and Stefano Ermon (2019). "Generative Modeling by Estimating Gradients of the Data Distribution". In: *Advances in Neural Information Processing Systems*.

Song, Yang, Sahaj Garg, Jiaxin Shi, and Stefano Ermon (2020). "Sliced Score Matching: A Scalable Approach to Density and Score Estimation". In: *Uncertainty in Artificial Intelligence*.

Song, Yang, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole (2021b). "Score-Based Generative Modeling through Stochastic Differential Equations". In: *International Conference on Learning Representations*.

Trippe, Brian L, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola (2023). "Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem". In: *International Conference on Learning Representations*.

Tucker, George, Chris J Maddison, Dieterich Lawson, and Shixiang Gu (2019). "Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives". In: *International Conference on Learning Representations*.

Vahdat, Arash and Jan Kautz (2020). "NVAE: A Deep Hierarchical Variational Autoencoder". In: *Advances in Neural Information Processing Systems*.

Vincent, Pascal (2011). "A Connection Between Score Matching and Denoising Autoencoders". In: *Neural Computation* 23.7, pp. 1661–1674.

Yang, Kaylee Yingxi and Andre Wibisono (2022). "Convergence in KL and Rényi Divergence of the Unadjusted Langevin Algorithm Using Estimated Score". In: *NeurIPS 2022 Workshop on Score-Based Methods*.

Yu, Shiqing, Mathias Drton, and Ali Shojaie (2022). "Generalized Score Matching for General Domains". In: *Information and Inference: A Journal of the IMA* 11.2, pp. 739–780.

Zhang, Qinsheng and Yongxin Chen (2023). "Fast Sampling of Diffusion Models with Exponential Integrator". In: *International Conference on Learning Representations*.