Let's Say We Are Not Sure:
Model Uncertainty in Finance from a Machine Learning Perspective



Yannick Limmer

St. Hilda's College

University of Oxford

Report submitted for

Doctor of Philosophy

Trinity 2025

Acknowledgements

I would like to thank Blanka Horvath for her constant support, insightful guidance, and for granting me the freedom to pursue my ideas. Her advice has consistently been both precise and generous, extending well beyond the strictly academic.

I thank Johannes Muhle-Karbe, whose early encouragement and mentorship were instrumental in setting me on this path.

I have also been fortunate to collaborate with Anastasis Kratsios, whose mathematical clarity and deep understanding of the geometry of learning have been a constant inspiration.

On the practitioner side, I am grateful to Hans Bühler for sharing his insight, passion, and time as a co-author. I also thank Gordon Lee, Ben Wood, and Greg Sidier for their valuable perspectives and encouragement along the way.

I am grateful to all my peers I have met during these years at the Mathematical Institute and the Oxford-Man Institute (OMI), in particular Marcello, Fayçal, Patsy, Alvarito, Harrison, Jonathan, Milena, and Leandro, who made OMI a truly unique place to work.

I also wish to thank Anthony Ledford, Jen Desmond, and Vanessa Wilkins for their wonderful work at the OMI.

Beyond Oxford, Owen, Magnus, Phil, Marco, Zack, Niklas, and Raeid provided valuable perspective and support that shaped this PhD in less obvious, but no less important, ways.

A special line is reserved for Niklas Weber —
with whom it all began in our very first lecture, and without whom this journey would have looked very different.

To the boys —
Rigney, Luke, Saxon, Gabe, Tara, Paolo, Throbber, Crab, Hamish, Will Randell, and Bodo.

To my family —
Mum, Dad, Linus, and Noel, for their love and unwavering belief in me.

Abstract

In contemporary finance, the pervasive challenge of model uncertainty stems from the empirical complexity of markets, whose heteroscedastic, regime-switching, and path-dependent dynamics resist faithful encapsulation within any single stochastic formulation. Confronting this challenge, this dissertation pursues a set of complementary investigations, each employing and adapting modern machine learning methods to characterise, exploit, and attenuate such uncertainty. The principal contributions are as follows.

First, we reformulate worst-case hedging as a two-player game between a trading strategy and a generator that perturbs the data-generating process. The resulting concept, termed *Robust Hedging GAN*, produces strategies that dominate both static and pathwise worst-case benchmarks, supported theoretically through a novel duality with coherent risk measures and validated empirically on S&P 500 options.

Second, adopting a probabilistic view of model risk, we introduce a family of *uncertainty measures* that rank distributions over candidate models. Building on this, we propose a variance-controlled subsampling scheme and a CVaR-SGD algorithm that enable risk-aware stochastic optimisation in high dimensions while preserving out-of-sample guarantees. Extensive simulations and a case study on limit-order-book data demonstrate consistent gains over mixture-of-experts and Bayesian ensemble baselines.

Third, to accommodate dynamic belief updating about latent states, we design the *Filterformer*, an attention-guided extension of the Deep Kalman Filter. We establish that it universally approximates the optimal nonlinear filter on compact horizons and prove its Wasserstein-2 stability under misspecified priors. Numerical experiments on stochastic volatility and robotic-sensor fusion tasks corroborate its accuracy and robustness.

Taken together, these contributions advance the theoretical foundations and practical methodology of robust financial machine learning, offering principled tools for decision-making under model uncertainty.

Contents

1 Introduction **1**

1.1 Model Uncertainty in Financial Markets 1

1.2 Robust Hedging with Generative Adversarial Networks 3

1.3 Supercharging Model Distribution Methods With Uncertainty Measures 9

1.4 The Filterformer: Attention-Based Stochastic Filtering 14

2 Adversarial Deep Hedging **25**

2.1 Introduction 25

2.2 Preliminaries 28

2.3 Robust Hedging GAN 36

2.4 Numerical Results 40

2.5 Conclusion 52

3 Uncertainty Measures **55**

3.1 Introduction 55

3.2 Robust strategies based on uncertainty measures 58

3.3 Comparing strategies in a Gaussian i.i.d. context 61

3.4 Subsampling: An Agnostic Choice of Model Distribution 66

3.5 Overcoming Memory Constraints: A CVaR-Stochastic Gradient Descent 70

3.6 Conclusion 76

4 Deep Filters Can Filter **81**

4.1 Introduction 81

4.2 The Setting 86

4.3 The Filterformer: An Attention-Based Deep Filter 92

4.4 Main Result 95

4.5 Conclusion 98

References **101**

Appendices

A Appendix of Chapter 2 **125**
A.1 Proofs of Section 2.2.1 125
A.2 Rough path theory and distances of stochastic processes 127
A.3 Setting of Herrmann, Muhle-Karbe, and Seifried 131
A.4 Out-of-sample Testing 132
A.5 Inverse Calibration in the Black-Scholes model 135

B Appendix of Chapter 3 **137**
B.1 Basics on Risk Measures 137
B.2 Results for Section 3.3 138
B.3 Results for Section Section 3.5 146

C Proofs of Chapter 4 **151**
C.1 Isometric Copies of Every Compact Riemannian Manifold in the Path Space — Proof of Proposition 4.2.1 151
C.2 Stability of the Optimal Filter — Proof of Proposition 4.4.2 152
C.3 Stable and Lossless Encoding — Proof of Proposition 4.3.1 165
C.4 Approximability of Locally Lipschitz Maps By Filterformers — Proof of Proposition 4.4.3 172
C.5 An Auxiliary Lemma — I. Pinelis 180

Le doute n'est pas un état bien agréable, mais l'assurance est un état ridicule.

— Voltaire

1 Introduction

Contents

1.1 Model Uncertainty in Financial Markets 1
1.2 Robust Hedging with Generative Adversarial Networks 3
1.2.1 Worst-Case Methods and Their Extensions 4
1.2.2 Robust Hedging GAN 5
1.2.3 Related Work 7
1.3 Supercharging Model Distribution Methods With Uncertainty Measures 9
1.3.1 Uncertainty Measure Method 10
1.3.2 CVaR Stochastic Gradient Descent 11
1.3.3 Subsampling as a Generator of Model Ambiguity 12
1.3.4 Related Work 13
1.4 The Filterformer: Attention-Based Stochastic Filtering 14
1.4.1 Filtering Model Uncertainty 14
1.4.2 Deep Filtering: Balancing Expressiveness and Tractability . . 17
1.4.3 Universal Approximation of the Filterformer 20
1.4.4 Alternative Applications of (Deep) Filters 21

1.1 Model Uncertainty in Financial Markets

The pursuit of generating ever more realistic market models has been at the forefront of quantitative finance ever since the advent of the Black-Scholes model, yet even the most sophisticated models remain approximations of an elusive "true" market reality [37, 80]. In this context, model uncertainty—encompassing aspects such as parameter

misestimation, overfitting of highly flexible models, and the broader challenge of model-misspecification—has increasingly been recognized as a fundamental obstacle in tasks ranging from portfolio allocation and derivative pricing to risk management [81, 123, 174].

Over the decades, the dominant approach in financial modelling has been to restrict analysis to a pre-specified model class and calibrate parameters against historical data [262, 263]. While early models were chosen to balance explainability with mathematical tractability [134, 146], recent advances in numerical methods and machine learning have shifted the emphasis towards predictive power [43]. This is exemplified by the emergence of neural-network-based market generators that can produce synthetic data mirroring observed market dynamics with considerable fidelity [50, 54, 155, 176, 228].

However, the flexibility of these models does not eliminate the inherent uncertainty in financial systems. Whether stemming from the heteroscedastic and non-stationary nature of market dynamics [80, 136, 149, 195] or from the risks of overfitting high-dimensional models to limited data [13, 30], the inability to perfectly capture the underlying probability measure (denoted by $\mathbb{P}^\star$) forces practitioners to rely on empirical approximations (denoted by $\hat{\mathbb{P}}$), often with significant implications for risk management [95, 225]. In many cases, this challenge is framed in terms of *Knightian uncertainty* or model ambiguity, where agents must evaluate potential payoffs under a spectrum of plausible models [57, 216].

In this thesis, we investigate how robustification techniques—originally developed to counteract the limitations of both parametric and distributionally robust frameworks [21, 78, 205, 246, 287]—can be leveraged within the realm of optimal control problems in finance. By integrating insights from traditional financial modelling with novel machine learning methods, we aim to provide robust, data-driven solutions that not only improve predictive performance but also mitigate the risks associated with model misestimation and overfitting.

Approaches to model uncertainty. The methods developed to address model uncertainty in this thesis can be broadly categorized into two approaches: worst-case methods and model-distribution methods. Worst-case methods operate under an adversarial perspective, optimizing performance against the most unfavorable model within a predefined

uncertainty set. In contrast, model-distribution methods take a probabilistic viewpoint, considering a distribution over potential models rather than a single adversarial worst-case scenario. In both cases, we introduce data-driven approaches that enhance the agent's ability to adapt to uncertain environments, leveraging modern machine learning techniques to refine decision-making under uncertainty. These constitute the key methodological developments of this thesis, which we subsequently support with related theoretical results. An overview of uncertainty methods mentioned below is given in Table 1.1.

Abbrv.	Objective	Global	Directional	Adjustable	Agnostic	Complexity[1]
[M]	$\int_{\mathscr{P}} J(a;\mathbb{P})\mathfrak{P}(d\mathbb{P})$	✓	×	×	n.a.	$\mathcal{O}(n)$
[WC]	$\inf_{\mathbb{P}\in\mathscr{P}} J(a;\mathbb{P})$	✓	✓	×	n.a.	$\mathcal{O}(n\log n)$
[BWC]	$\inf_{\mathbb{P}\in B(\mathbb{P}')} J(a;\mathbb{P})$	×	✓	✓	×	$\mathcal{O}(n\log n)$
[PWC]	$\inf_{\mathbb{P}\in\mathscr{P}} J(a;\mathbb{P})+\alpha(\mathbb{P})$	✓	✓	✓	×	$\mathcal{O}(n\log n)$
[UM]	$U_{\mathbb{P}\sim\mathfrak{P}}(J(a;\mathbb{P}))$	✓	✓	✓	✓	$\mathcal{O}(n^2)$

Table 1.1: Overview of methods for model uncertainty. See Section 1.2.1 and Section 1.3 for details.

1.2 Robust Hedging with Generative Adversarial Networks

Building on the categorization of uncertainty methods outlined above, we now focus on the central methodological contribution of Chapter 2 of this thesis: the *robust hedging GAN*, a general solver for (penalized) worst-case methods. Incorporating deep learning into the robustification process, makes the approach more computationally tractable, therefore significantly more appealing for practitioners. The goal is to enable robust decision-making under model uncertainty in complex, high-dimensional, and data-intensive settings, without relying on specific assumptions of the model class employed. While the method is developed and analyzed in the context of a hedging problem, its applicability extends well beyond this use case.

[1]Purely symbolic, $\mathcal{O}(n)$ requires sampling from one measure, $\mathcal{O}(n\log n)$ requires sampling from one measure plus a search on measure space, $\mathcal{O}(n^2)$ requires sampling of measures from a distribution plus sampling from the generated measures.

1.2.1 Worst-Case Methods and Their Extensions

To contextualize this contribution, note that most problems in quantitative finance can be described by optimizing an objective function $J : \mathcal{A} \times \mathscr{P} \to \mathbb{R}$, which evaluates the "goodness" of a strategy on a model, i.e.

$$\arg\max_{a \in \mathcal{A}} J(a; \mathbb{P}), \tag{1.1}$$

where $\mathcal{A}$ is the space of permissible actions and the measure $\mathbb{P} \in \mathscr{P}$ represents a model from the set of all models $\mathscr{P}$.

Worst-case method [WC]. The *worst-case* method evaluates performance under the most adverse model in a given class. Originating from the work of [120], where authors derived this approach by relaxing the axioms underpinning expected utility theory [227, 258], which was further refined in [101, 138, 209, 234]. Whilst offering conceptual clarity and robustness, it suffers from notable limitations. Applying caution uniformly across all models, the worst-case method often leads to overly conservative or trivial strategies [20, 35, 115, 224]. Computationally, its reliance on a global minimization across the model space $\mathscr{P}$ can be prohibitively expensive or unfeasible.

Bounded worst-case method [BWC]. To alleviate the rigidity of [WC], bounded worst-case methods constrain attention to a narrower, context-specific subset of models $\mathscr{P}$. This refinement—used in financial control by, e.g. [21, 287]—offers better tractability and reduces conservatism. However, it inherits a core flaw from [WC]: all models within the bounded region are treated equally, regardless of empirical plausibility. This implies, agents will have to balance between choosing a small uncertainty region and therefore work with a very local framework; or including more models into consideration, resulting again in overly conservative results. Consequently, it introduces an additional layer of subjectivity by requiring the practitioner to define the boundaries of plausibility ex ante.

Penalized worst-case method [PWC]. A more nuanced extension of the worst-case approach is to attach penalties to implausible models, as for instance suggested in [28, 131, 145, 172, 284]. Instead of optimizing against the worst possible model unconditionally, [PWC] incorporates a penalty function $\alpha(\mathbb{P})$ into the objective, which reduces the influence of less likely models. The resulting optimization problem takes the form

$$\arg\max_{a\in\mathcal{A}} \inf_{\mathbb{P}\in\mathscr{P}} J(a;\mathbb{P}) + \alpha(\mathbb{P}),$$

where $\alpha(\mathbb{P})$ expresses a preference for more plausible models, often via a divergence or distance from a reference measure.

In financial contexts, when $J(a;\mathbb{P})$ corresponds to a monetary risk measure of the profit and loss from strategy a, the penalty $\alpha(\mathbb{P})$ can be interpreted in monetary terms. Recall that $J(a;\mathbb{P})$ is considered as the amount of money that an agent needs to put aside to make the investment acceptable from a risk point of view. The penalty then reduces this required capital buffer for models deemed less likely to be the true model.

Remark 1.2.1 (Relation to risk measures). The [PWC] framework is closely related, though not equivalent, to the application of monetary risk measures. For instance, when $J(a,\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[u(X(a))]$ and α satisfies certain conditions, the duality of risk measures (cf. [22, 111]) yields

$$\inf_{\mathbb{P}\in\mathscr{P}} \left(\mathbb{E}_{\mathbb{P}}[u(X)] + \alpha_\rho^{\min}(\mathbb{P})\right) = -\rho(u(X)).$$

However, the presence of a utility function alters the objective, and similar dual results do not exist for other choices such as variance or CVaR. Furthermore, the implied penalty

$$\alpha^{\min}(\mathbb{P}) = \sup_{Y\in L^0} \left(\mathbb{E}_{\mathbb{P}}[-Y] - \rho(Y)\right)$$

is defined via a supremum and hence often impractical. ◁

1.2.2 Robust Hedging GAN

At its core, the robust hedging GAN reformulates the classical robust control problem as a two-player game between an adversarial market environment and a hedging agent. Drawing inspiration from generative adversarial networks (GANs) [47, 84, 127, 218, 244],

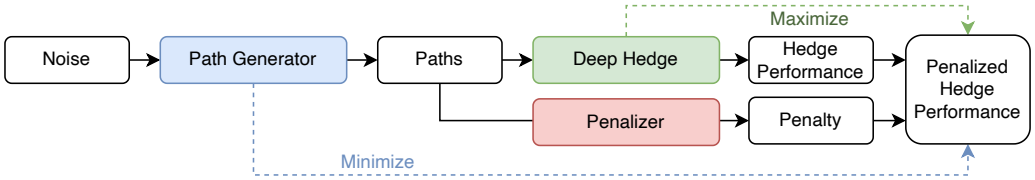


Figure 1.1: The Robust Hedging GAN architecture. Random noise is transformed by the path generator into asset price paths. The deep hedge module then applies a hedging strategy to these paths. Simultaneously, the penalizer assesses how realistic the generated paths are, penalizing unrealistic ones to buffer against misleading hedging performance. The deep hedge aims to maximize the combined performance metric, while the generator seeks to minimize it.

we reinterpret the generator as an adversarial "mother nature" who proposes challenging market scenarios, while the discriminator acts as a robust hedger, adapting her strategy to withstand adverse conditions. The generator produces synthetic paths of asset prices from noise inputs, and the discriminator optimizes hedging strategies—represented by a trainable neural network inspired by the deep hedging framework [50]—based on these generated paths. Training proceeds in an adversarial fashion: the generator updates its outputs to minimize the hedger's (penalized[2]) utility, while the hedger adjusts her strategy to maximize it. The result is a strategy that is robust not against a fixed set of worst-case models, but against a dynamically learned and data-informed class of plausible market evolutions. See Figure 1.1 for an overview of the architecture.

By leveraging deep learning, we make worst-case hedging computationally feasible for complex market environments, addressing the intractability of classical approaches while preserving their robustification properties. Note that our approach can be used for arbitrarily complex models (including market generators using high-parametric machine learning techniques), therefore achieving quasi-distributional robustness [214, 287]. It is evident that the framework is still limited to the class of generators used, however, this class can be sufficiently large to include "quasi" all relevant measures.

Scope of the robust hedging GAN. The robust hedging GAN integrates seamlessly with all three worst-case paradigms discussed above. If the generator is unrestricted, it recovers the [WC] approach. Constraining it to a specific parametric class yields a

[2]The term *penalized utility* might be misleading here, as the penalizer *increases* the resulting utility for unlikely paths, implying a penalty to the generator.

[BWC] setup, while including a penalty term in the generator's objective mirrors the logic of [PWC]. In this way, the GAN framework provides a flexible and extensible platform for adversarial robustification under model uncertainty. Moreover, it does so in a computationally viable manner, opening up robust control to high-dimensional applications previously out of reach for traditional methods.

Penalizing models in practice. In some low-parametric models, introducing a penalty for model uncertainty is relatively straightforward. For instance, the distance between two Black-Scholes models can be simply defined as the L^2-distance of their volatility parameters. However, for more complex models or data-driven market generators, defining meaningful distances between models remains a challenge. To address this, we propose to use a data-driven penalization approach in these scenarios (as suggested in [228]), based on the *signature Maximum Mean Discrepancy (signature-MMD)* [67], which is agnostic to the specific model class used, as it entirely operates on generated paths and has proven effective in practical settings [5]. We demonstrate numerically that this approach allows us to achieve meaningful robustness results in more complex settings, such as the Heston model.

1.2.3 Related Work

Our contribution builds on deep hedging [50] while introducing adversarial robustification that remains model-agnostic and data-driven. In contrast to recent "robust deep hedging" formulations [118, 119, 156, 205], we operationalize robustness via adversarial path generation and penalization. Notably, [205] quantify parameter uncertainty by restricting attention to a bounded set endowed with a *uniform* prior. We, instead, avoid encoding prior beliefs as distributions and permit the agent to tune uncertainty aversion directly; despite the different modeling stance, we adopt their calibration for out-of-sample checks.

Adversarial deep learning approaches. In [160, 287], authors implement Wasserstein robustness by transforming profit-and-loss distributions with a neural map. Our route targets model uncertainty at its core; we modify the *path generator* and impose Wasserstein-

type penalties on generated paths. This cleanly decouples model uncertainty from policy design and recovers analogous structures when the penalty targets a Wasserstein ball.

A closely related and complementary line of work is [178], where authors develop a GAN-based robust utility framework, recover optimal strategies in multi-asset frictionless markets under logarithmic utility, and derive pathwise constraints for adversarial market measures. Their perspective and ours are highly synergistic: both exploit adversarially trained generators to stress test decision rules, while differing in emphasis—our formulation targets hedging performance across data-driven scenarios, whereas their focus is on utility maximization with structural constraints on the adversary.

Also noteworthy is the contribution of [62], who employ machine learning techniques—most notably Gaussian process surrogates—within an optimal control framework to address problems of utility maximization and hedging; with results providing clear evidence of performance gains in these settings. Nevertheless, the scope of model uncertainty accommodated in their framework is more limited than that captured by our approach—restricted to the mixture of measure approach (see Section 1.3).

Asymptotic vs. scalable non-asymptotic schemes. The article [21] examines Wasserstein uncertainty and emphasise the distinction between parameter and distributional robustness, but their analysis focuses on *asymptotically small* balls. Building on the economic framework of [208], [145] adopt asymptotic and local-volatility assumptions. Our framework admits scalable, non-asymptotic robustness without these locality requirements and subsumes these settings in our empirical study.

Outside of finance, efficient algorithms for distributionally robust optimization have been developed (e.g. f-divergence DRO and related surveys) [223, 245, 265], alongside optimal-transport approaches tailored to Wasserstein uncertainty [38, 219]. These lines inform our computational design and regularization choices.

Robust market generation and neural SDEs. Within data-driven market simulation, robustness for neural SDE generators was introduced in [119]; related market generators include [54, 76, 77, 176], and we build on concepts from [228], particularly in approximating the SigW_1-metric/SigMMD. While we use hedging as the primary example in

this section, note that our techniques can also be applied to other optimization tasks prone to model uncertainty (c.f. [34, 58, 59]).

1.3 Supercharging Model Distribution Methods With Uncertainty Measures

While (penalized) worst-case methods are theoretically designed to be global, they are often bound to local applications in practice. The complexity of model spaces makes exhaustive optimization infeasible, meaning solutions are inevitably found through local search procedures [178, 198, 287]. This fundamental limitation arises because identifying the true worst-case model in high-dimensional spaces is computationally intractable, forcing worst-case approaches to operate within a constrained region of the model space. Chapter 3 of this thesis explores the use of *uncertainty measures* to establish a framework that overcomes these issues, while handling model uncertainty in a practical and data-driven way [51].

Mixture method [M]. Instead of focusing solely on one model at a time, we construct a distribution over the space of measures, treating model uncertainty probabilistically rather than deterministically. This places our approach in the above-mentioned class of model-distribution methods. Simply put, we consider all measures in $\mathscr{P}$ simultaneously, each weighted with a corresponding probability. To make this precise, identify the set of measures with a parameter space Θ, i.e. $\mathscr{P} \equiv \{\mathbb{P}_\theta : \theta \in \Theta\}$. Assigning each model a probability is the tantamount to posing a distribution $\theta \sim \mathfrak{P}$ on Θ. If the agent now operates under the mixture of these distributions, i.e. the measure $\bar{\mathbb{P}} \overset{\text{def.}}{=} \int_{\mathscr{P}} \mathbb{P}_\theta \mathfrak{P}(d\theta)$, we refer to its optimization henceforth as *mixture method*.

While the mixture approach is not a robustification method per se but rather an extension of the existing model, it is popular in practical applications due to its methodological simplicity. Using a distribution on the measure space increases dimensionality, yet the curse of dimension can be evaded when using Monte Carlo methods [56, 122], since sampling is done from a single measure. This makes this approach effective for applications like deep hedging [205]. However, the mixture method has its limitations, ranging

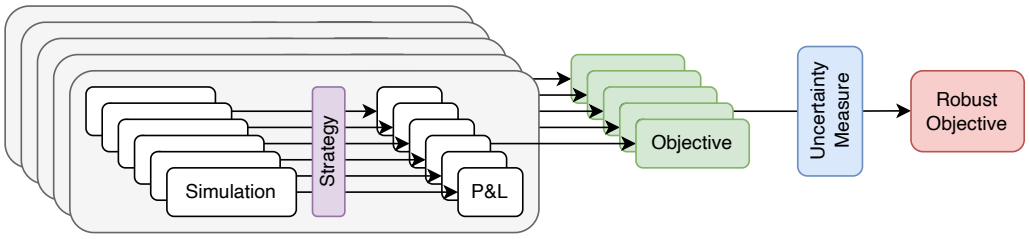


Figure 1.2: Evaluation of a strategy with the uncertainty measure methodology using subsampling. The strategy is tested across multiple simulated market scenarios, producing a profit and loss (P&L) value for each run. These P&L outcomes are used to compute the investment objective defined by the agent. This process is repeated using alternative simulations to generate a distribution of objectives, which are then analyzed using a metric we refer to as the uncertainty measure.

from distortion of model properties, via limited adjustability, to the non-directionality of the approach. The most significant drawback, nonetheless, is that the strategy becomes vulnerable to misestimation of $\mathfrak{P}$ itself (as for instance postulated in [81]), which is to some extent tantamount to deferring model uncertainty to a different domain.

1.3.1 Uncertainty Measure Method

To overcome limitations of the mixture method, we suggest the use of *uncertainty measures [UM]*, building on the theory of [172]. While this approach is closely related to the mixture method—in that it requires constructing a distribution over potential models and assigning a probability to each model being the true one—it differs substantially in the fact that, rather than merely aggregating these models, we evaluate each model independently to derive distinct utility values; and in a subsequent step, we analyse the resulting distribution of utility values. This distribution effectively represents the utility achieved under a randomly selected model, allowing us to define an objective function based on it. Note that this situation is similar to characterizing the "goodness" of a $P\&L$-value, which risk measures have proven to be very effective at. So if U is a risk measure under the assumption that $\theta \sim \mathfrak{P}$, we can write the uncertainty measure objective as $U_{\theta\sim\mathfrak{P}}(J(a;\mathbb{P}_\theta))$.

With that, we coin the term *uncertainty measures*, which are defined as risk measures [10, 108, 109]—however applied to the model space. Recall that a monetary risk-measure J, identifies the amount of money $J(a;\mathbb{P}_\theta)$ an agent needs to put aside to make the

strategy a acceptable under the measure $\mathbb{P}_\theta$ from a risk perspective. In the case of an uncertainty measure, $U_{\theta\sim\mathfrak{P}}(J(a;\mathbb{P}_\theta))$ refers to the amount of money the agent needs to put aside for the strategy to be acceptable for all measures under $\theta \sim \mathfrak{P}$.

In Chapter 3, we benchmark our technique against approaches that optimize a single mixture measure, combining a distribution over models with their respective laws. Our method demonstrates improved robustness by flexibly adjusting to model risk; and that the mixture measure under canonical choices for the model distribution yields comparatively insufficient robustification. This is best seen in a simple and controlled environment, where analytic solutions are available, but we also demonstrate that our findings extend to simulated and empirical experiments.

1.3.2 CVaR Stochastic Gradient Descent

While the use of *uncertainty measures* provides a flexible and principled framework for addressing model risk, it typically incurs a substantial computational cost. As highlighted in Table 1.1, the overall evaluation complexity is expressed as $\mathcal{O}(n^2)$, the highest among the listed approaches. Specifically, obtaining a sufficiently accurate estimate of $J(a, \mathbb{P}_\theta)$ requires averaging evaluations for n samples. In order to compute the uncertainty-aware objective $U_{\theta\sim\mathfrak{P}}(J(a;\mathbb{P}_\theta))$, this procedure must itself be repeated across n sampled models, resulting in quadratic scaling with respect to the number of simulations.

Although modern processing units (such as GPUs) allow for parallel execution and can significantly reduce wall-clock time [272], the available memory often becomes the limiting factor [222, 273]. This bottleneck is particularly prominent in high-dimensional settings or when transitioning from single-period predictions to more complex tasks involving the simulation of entire asset price trajectories. Such scenarios demand evaluations over expansive state spaces and across numerous model variants, quickly exhausting available memory resources.

To mitigate this issue, we introduce a memory-efficient variant of stochastic gradient descent (SGD) [41, 42, 249] tailored to uncertainty-aware optimization under Conditional Value-at-Risk (CVaR). The core idea is to avoid the need to retain all simulation paths simultaneously in memory. Our approach operates by simulating under a single model

instance at a time and retaining only those samples relevant to the tail-risk quantification intrinsic to CVaR. This selective retention significantly alleviates memory pressure while preserving the ability to capture model risk effectively. Moreover, since simulations are independent, this procedure naturally lends itself to distributed computation across multiple devices, aligning well with contemporary deep learning workflows.

1.3.3 Subsampling as a Generator of Model Ambiguity

While our results indicate that the specific choice of model distribution $\mathfrak{P}$ is less critical in the context of uncertainty-measure-based methods—particularly when compared to the mixture approach—selecting a suitable family of models remains of fundamental importance. Indeed, in settings where modern deep solvers are capable of operating in highly expressive model classes, the construction of $\mathfrak{P}$ becomes the central design choice.

Before introducing more sophisticated ambiguity generation schemes—such as filtering-based approaches (see Section 1.4)—we begin by analyzing a simple and computationally efficient method to derive a distribution $\mathfrak{P}$ from a single estimated model $\hat{\mathbb{P}}$. Although limited in generality, this method is practically attractive: it is easy to implement, requires minimal hyperparameter tuning (which is often delicate in competing approaches), and integrates seamlessly with existing modelling pipelines.

The underlying idea can be summarized as follows. Given an estimated model $\hat{\mathbb{P}}$, we draw a relatively small number of samples from it and use these to empirically evaluate the objective function empirically. We refer to such set of samples as a subsample. Since the sample size is deliberately kept small, repeated subsampling results in variability in the empirical objective. By associating each such empirical measure with a distinct model in the support of $\mathfrak{P}$, the repeated subsampling procedure implicitly defines a distribution over models.

This approach is supported by multiple conceptual justifications. First, if $\hat{\mathbb{P}}$ itself is an empirical distribution (e.g. based on the last observed returns of an asset price process), then the described method coincides with the classical *bootstrapping* technique commonly used in statistical inference and portfolio analysis [102, 103]. In this case, a subsample corresponds to a resampling (with replacement) from historical observations.

Second, the idea of subsampling is deeply embedded in the machine learning literature under the paradigm of *mini-batching* [41, 126]. Mini-batch stochastic gradient descent corresponds, in our framework, to applying an expectation-type uncertainty measure over this induced model distribution (see Section 3.5).

To assess the effectiveness of this approach, we provide numerical evidence showing that our subsampling-based ambiguity generation method yields robust performance in practical investment and hedging tasks. Despite its simplicity, the approach performs comparably to more elaborate Bayesian techniques. Moreover, when combined with the deep hedging framework [50] in a controlled simulation environment, our robustification scheme leads to demonstrable improvements in hedging performance.

1.3.4 Related Work

Our work builds upon several strands of research at the intersection of robust optimization, risk-aware decision-making, and data-driven methods in finance.

Risk measures for model ambiguity. Our method uses risk measures as *uncertainty measures*—specifically entropic risk and CVaR [10, 251]—to quantify uncertainty in a model-agnostic fashion. This connects to recent work in risk-sensitive reinforcement learning [68, 69, 275], where coherent risk measures guide decision-making under uncertainty and to the use of CVaR in robust control [157]. Unlike worst-case optimization, our approach adjusts the degree of robustness continuously via uncertainty aversion parameters.

Subsampling. To approximate model uncertainty without specifying priors, we propose *subsampling* (akin to bootstrapping [102, 217]), drawing parallels to resampling techniques widely used in statistics [241, 242], decision-making, and deep learning [41, 44, 229]. This allows us to construct empirical distributions over models efficiently—a key methodological innovation relative to fully Bayesian or mixture approaches. Subsampling methods have also found attention in recent studies, see for instance [74, 240], including in the context of financial model uncertainty [196].

Computational aspects. Our scalable stochastic gradient approach to uncertainty-aware learning draws inspiration from both distributionally robust optimization and recent advances in deep hedging. For literature handling model uncertainty in the context of deep hedging, refer to Section 1.2.3.

Finally, our work relates to the growing literature on *risk-sensitive deep learning* in finance and reinforcement learning [163], where performance metrics explicitly account for tail risks and robustness.

1.4 The Filterformer: Attention-Based Stochastic Filtering

This section introduces the *deep filtering* paradigm that underpins Chapter 4 of this thesis [150]. We proceed in four steps. Section 1.4.1 motivates filtering as a dynamic, Bayesian alternative to static distributions of models introduced in Section 1.3 and we formulate the robust filtering problem for the remainder of this section. Section 1.4.2 addresses the rise of deep filtering methods building on this framework, we introduce in Section 1.4.3 the *Filterformer*—an attention-based extension of the Deep Kalman Filter–and discuss two core theoretical results in Chapter 4: $\mathcal{W}_2$-Lipschitz stability of the robust optimal filter, and a universal approximation theorem showing that the Filterformer can recover the optimal posterior law for a broad class of latent-parameter models. Section 1.4 concludes with a survey of alternative deep-filtering applications.

1.4.1 Filtering Model Uncertainty

In Section 1.3 we introduced *subsampling* as an easily implementable, if somewhat *ad hoc*, mechanism for constructing a distribution of probabilistic models $\mathfrak{P}$ on the model space $\mathscr{P}$. Although attractive for its simplicity, subsampling is not optimized for any statistical or economic criterion; its performance is not necessarily optimal.

As a possible way to enhance performance, Bayesian approaches have proven popular, where an initial belief is encoded as a prior distribution over the model parameters and subsequently updated with incoming data to better reflect the evolving market

conditions. This is for instance demonstrated in [55, 86, 100, 132, 158, 164, 276] and references therein.

However, all of these methods share a common limitation: once the distribution over candidate models has been fixed *all* incoming data is treated *ex post*. Consequently, both methods yield *static* measures of model risk that may lag behind the rapidly changing dynamics of financial markets. This is exemplified in [281], where we show that for control problems, Bayesian updates to a prior distribution over the model space $\mathfrak{P}$ may result in worse performance than quickly adapting the estimated measure $\hat{\mathbb{P}}$ to incoming information without taking model uncertainty into account a priori (see also [33]). This is especially evident when the prior distribution is drastically misspecified.

Filtering has been suggested as the intersection of these approaches [137], as it offers a complementary, *dynamic* viewpoint while retaining a Bayesian framework. Instead of considering a fixed distribution of models that we gain more and more information of, we regard the model parameters $\theta \in \Theta$ as a latent, time-evolving state that must be inferred from a continuous stream of noisy market observations. As before, the agent represents, at each point t in time, her current belief by assigning each model a probability of being the true model, i.e. $\theta_t \sim \mathfrak{P}$. However, this distribution is dependent on the previous observable states of the market and will change with further incoming information. We can model this type of behaviour by postulating θ to be a non-observable stochastic process, which is coupled to the observable stochastic process describing market information Y. The filtering literature often refers to θ as the *signal process* and Y the *measurement process*. Filtering is for instance used in [18, 200, 266, 285] for the calibration of equity market models, in [233] for commodity models, and in [6, 139] for interest rate models. See also [248] for an overview.

Filtering dynamics. To precisely articulate the contribution of this part of this thesis, we now introduce some technical notation for the filtering problem outlined above. On an intuitive level, we model the market state Y to develop depending on three components, (i) observed market states $Y_{[0:t]}$, which can be limited to just the current market state, but also extends to non-Markovian settings, (ii) the unobservable current

model parameters θ_t, as well as (iii) a source of noise W_t, which corresponds to the market state change that cannot be explained by the modelling dynamics. The model parameters θ evolve themselves and can generally depend (individually) on the same three components, however, often only relate to previous model parameters θ_t together with noise W_t in applications.

To make this precise, consider the following continuous-time partially observed system

$$d\theta_t = [a_0(t, Y_{[0:t]}) + a_1(t, Y_{[0:t]})\theta_t]dt + \sum_{i=1}^{2} b_i(t, Y_{[0:t]})dW_t^{(i)}, \tag{1.2}$$

$$dY_t = [A_0(t, Y_{[0:t]}) + A_1(t, Y_{[0:t]})\theta_t]dt + \sum_{i=1}^{2} B_i(t, Y_{[0:t]})dW_t^{(i)}, \tag{1.3}$$

where a_0 and A_0 respectively take values in $\mathbb{R}^{d_\theta}$ and $\mathbb{R}^{d_Y}$, and where $a_1, A_1, b_1, b_2, B_1, B_2$ are matrix-valued of respective dimensions $d_\theta \times d_\theta$, $d_Y \times d_\theta$, $d_\theta \times d_\theta$, $d_\theta \times d_Y$, $d_Y \times d_\theta$, and $d_Y \times d_Y$, and the entries of $a_0, a_1, b_0, b_1, A_0, A_1, B_0, B_1$ are measurable nonanticipative functionals on the measurable space $([0:T] \times C([0:T], \mathbb{R}^{d_Y}), \mathcal{B}_{[0:T]} \times \mathcal{B}_T^{d_Y})$, and where $\mathcal{B}_t^d \overset{\text{def.}}{=} \sigma(C([0:t], \mathbb{R}^d))$ denotes the σ-algebra generated by continuous paths on $[0:t]$ to $\mathbb{R}^d$. We will always assume that the filtration $\mathfrak{F}$ is right-continuous and that the filtered probability space $(\Omega, \mathcal{F}, \mathfrak{F}, \mathbb{P})$ supports independent $\mathfrak{F}$-adapted Brownian motions, $W_\cdot^{(1)} \overset{\text{def.}}{=} (W_t^{(1)})_{0\le t\le T}$ and $W_\cdot^{(2)} \overset{\text{def.}}{=} (W_t^{(2)})_{0\le t\le T}$, of respective dimensions d_θ and d_Y, for positive integers d_θ and d_Y; all technical requirements are detailed in Assumption 4.2.1 below.

In summary, given the coupled dynamics defined by $a_0, a_1, b_1, b_2, A_0, A_1, B_1, B_2$, we obtain two processes, θ, Y, evolving stochastically in time.

Filtering problem. This dynamic modelling allows us to extract a distribution $\mathfrak{P}_t$ at each time t. This is the objective of the so-called *robust stochastic filtering problem*, studied by [72, 73, 82, 88, 89, 186], which is to identify a function $f_t : C([0:t], \mathbb{R}^{d_Y}) \to \mathcal{P}_\theta$, such that

$$\mathfrak{P}_t := f_t((Y_s)_{s\le t}) = \mathbb{P}(\theta_t \in \cdot | \mathcal{F}_t^Y), \tag{1.4}$$

where $\mathbb{P}(\theta_t \in \cdot | \mathcal{F}_t^Y)$ is the conditional law of the d_θ-dimensional model parameter process θ_t given the σ-algebra generated by the previously observed market states $(Y_s)_{s \leq t}$. Therefore, $\mathfrak{P}_t$ is in this setting the filtered posterior over model parameters θ_t, conditional on all prior observed market states.[3]

Remark 1.4.1 (On the generality of the filtering dynamics). The formulation of the filtering dynamics in (1.2) and (1.3) allows for significantly more expressive modelling capabilities compared to traditional linear state-space models. In particular, all coefficients are nonanticipative functionals of the entire past trajectory $Y_{[0:t]}$, which enables both non-Markovian dependencies and the integration of long-memory effects in the evolution of both observable and latent processes.

Nevertheless, it is important to highlight a structural limitation: the signal component θ_t is affine-linear in both equations. This is not a fundamental requirement of the general filtering framework, and more general nonlinear interactions could be accommodated in principle.

In this thesis, however, we deliberately retain this affine-linear structure in the signal process. This choice is dictated by our main theoretical result in Chapter 4, where we show that such a structure is essential to ensure that the solution to the robust filtering problem (1.4) admits a Gaussian representation. The resulting tractability enables a precise characterisation of the conditional distribution $\mathbb{P}(\theta_t \in \cdot \mid \mathcal{F}_t^Y)$, which is key to establishing our approximation results. ◁

1.4.2 Deep Filtering: Balancing Expressiveness and Tractability

The stochastic filtering problem expressed in (1.4) constitutes a classical topic within probability theory and signal processing. Central issues—most notably the existence, uniqueness, and temporal evolution of the conditional distribution of the signal process θ_t given the observation history $Y_{[0:t]}$—were rigorously addressed in foundational works by [267, 268, 270, 271, 290], and have been further developed in recent studies such as [4, 291]. Despite these theoretical advances, the fact that the optimal filter evolves as

[3] Note that we can recover the initially mentioned pure Bayesian paradigm by choosing constant dynamics for the model parameters without noise, i.e. $a(\theta_{[0:t]}, Y_{[0:t]}, dW_t) := \theta_t$.

an infinite-dimensional, measure-valued, and path-dependent stochastic process renders direct numerical implementation effectively infeasible.

An important exception arises in the class of finite-dimensional filters, which correspond to the exceptional scenarios wherein the joint evolution of the signal $\theta_\cdot$ and the observation process $Y_\cdot$ admits an optimal filter characterized by a finite set of parameters. Prominent instances include the *Kalman-Bucy filters* [166, 167], for which explicit closed-form expressions are obtained under the hypotheses of Markovian Gaussian perturbations and affine Ornstein-Uhlenbeck-type dynamics for $\theta_\cdot$ and $Y_\cdot$; the Wonham filter [286], in which both the signal and the observations evolve as finite-state Markov chains; and the Beňes filter [29], predicated on a specially chosen one-dimensional diffusion model. Nevertheless, empirical evidence from financial markets typically reveals nonlinear interactions between latent state variables [80, 112], thereby violating the restrictive structural assumptions underpinning these finite-dimensional formulations.

The intrinsic computational intractability of the general filtering problem motivates the deployment of approximately optimal schemes. One prominent class comprises *particle filters*, which approximate the true posterior by evolving an interacting system of weighted particles and resampling steps to sequentially represent the filter's distribution [92–94]; these methods are, however, restricted to low-dimensional settings [31]. Alternatively, for affine signal-observation models one may derive *linearized relaxations* of the optimal filtering functional, yielding matrix-valued stochastic Riccati equations whose numerical solution furnishes tractable approximations [97, 125]. A further strategy projects the infinite-dimensional optimal filter onto a finite-dimensional manifold of exponential-family distributions, thereby enabling real-time tracking via a finite set of sufficient parameters [8, 45, 46]. While the latter two techniques achieve high fidelity under the assumed affine dynamics of the coupled system $(\theta_\cdot, Y_\cdot)$, and particle methods perform effectively in low-dimensional regimes, a universal methodology for constructing (approximately) optimal infinite-dimensional filters across more general, nonlinear settings remains elusive.

In this thesis, we pursue a deep-learning-based methodology, motivated by the remarkable capacity of deep neural networks to render tractable a host of once intractable

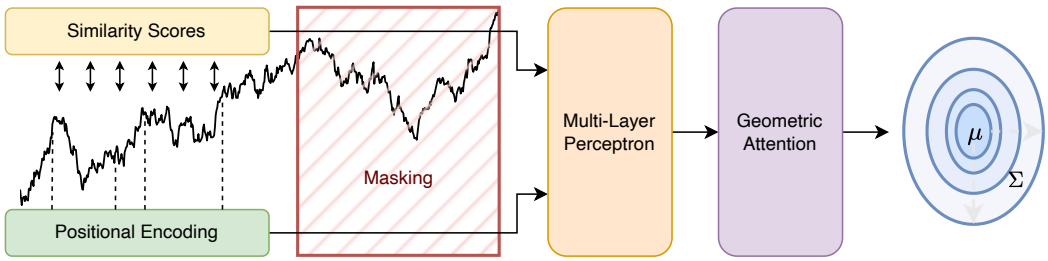


Figure 1.3: Architecture of a Filterformer. The input path is processed in two parallel ways: first, pointwise evaluations at a finite set of selected time points are used to construct the positional encoding; second, similarity scores are computed via distance measurements to reference paths. These features are then passed through a multi-layer perceptron (MLP) and subsequently transformed by a geometric attention mechanism into a distribution-valued output, representing the conditional law of the signal process θ..

numerical challenges. The representational power of such networks in classical supervised and unsupervised settings has spurred their application to stochastic filtering tasks; see, for instance, [27, 48, 63, 141, 144, 177, 253], as well as the formulation of measure-valued deep architectures in [1]. Of particular relevance is the class of *deep Kalman filters* (DKFs) introduced in [185], which underpins the model developed in Chapter 4. At a high level, DKFs employ deep neural networks to map observation histories into Gaussian distributions.

It is important to note that the Bayesian deep learning models—coined DKFs—relate only loosely to the classical Kalman filter (KF) through its *conditionally Gaussian* nature. In fact, the DKF, in contrast to the original KF, is not restricted to linear settings and hence is more broadly applicable.[4] It is this non-linearity which has led to its well-documented success in non-linear settings across a range of domains, from video compression [202, 203] to censoring [152] and mathematical finance [159].

Notwithstanding their heuristic affinity with stochastic filters, a rigorous linkage between DKFs and the canonical filtering problem (1.4) remains largely unexplored. Chapter 4 develops such a connection, offering a theoretical framework that justifies the use of DKFs in this setting.

[4]The expressive power of neural networks would be unnecessary for solving a linear Gaussian (Kalman) filtering problem, which is already well-known to be resolved by the classical *Kalman-Bucy* filter [167].

1.4.3 Universal Approximation of the Filterformer

Building on Section 1.4.1, we will now give an overview of the main theoretical contribution of Chapter 4. In order to achieve the desired approximation results, we propose the *Filterformer* (FF), an attention-based extension of the Deep Kalman Filter (DKF) of [185]. Recall that the model is an approximation of f_t in (1.4), i.e. takes a time stamp t and path $(Y_s)_{s\leq t}$ and maps to a probability distribution, particularly the conditional distribution of the signal given the observed path at time t. The architecture is visualized in Figure 1.3 and corresponds to the following components:

Phase 1. The proposed layer compresses each path $Y_{[0:T]} \in C([0,T],\mathbb{R}^{d_Y})$ to a finite–dimensional vector while remaining *injective* on a broad class of compact sets K—including all piecewise-linear paths, any finite training set, and any K isometric to a closed Riemannian manifold—hence the encoding is *lossless*. The map and its inverse are jointly Lipschitz, providing *stability* in the sense of constructive approximation theory [75, 235], so small numerical errors do not distort predictions. Because the layer learns its parameters directly from data, with no projection onto a fixed Schauder basis, it is *adaptive* to the intrinsic geometry of the chosen K.

Phase 2. The encoded path vectors are fed into a multi-layer perceptron (MLP) that learns task-specific nonlinear combinations—"deep features"—of the data. Thanks to the universal approximation power of ReLU networks [170, 180, 204, 289], the MLP can flexibly adapt to the dynamics in (1.2)-(1.3), capturing the statistics most relevant for filtering. Note that this is the only component inherited from the original DKF framework.

Phase 3. A modified *geometric attention mechanism* [1] then converts the deep features into $\mathcal{N}_{d_\theta}$-valued outputs, providing an approximation to the optimal posterior in (1.4). Concretely, the decoder maps Euclidean features to a point in the geodesic convex hull generated by a finite reference set that covers the image of K under the true filter, thereby producing a probabilistically meaningful prediction.

Main theoretical results. We first show that the robust optimal filter associated with (1.2)-(1.3) is *locally Lipschitz* in the 2-Wasserstein metric $\mathcal{W}_2$ (Proposition 4.4.2), extending the stability theory of [73, 82]. In practical terms, a small perturbation of the observed path produces only a proportionally small change in the posterior distribution of the hidden state, as measured by $\mathcal{W}_2$, which quantifies the minimal "mass-transport cost" between two probability measures.

Building on this regularity, we prove a *new universal approximation theorem*: the *Filterformer* can approximate, to arbitrary accuracy in $\mathcal{W}_2$, any continuous map that sends a compact set of paths to multivariate Gaussian laws (Proposition 4.4.3). Unlike [7], our theorem covers *degenerate* Gaussians (those with singular covariance) and does not rely on an information-geometric distance, but on the standard $\mathcal{W}_2$ metric.

Fast convergence is enabled by the *pathwise-attention encoder*, which embeds any compact path set $K \subset C([0,T],\mathbb{R}^{d_Y})$ into a finite-dimensional Euclidean space via a bi-Lipschitz, lossless map whose dimension does not depend on the target approximation error (Proposition 4.3.1). This geometry-preserving compression bypasses the usual curse of dimensionality [2, 188] and allows the Filterformer to learn accurate filters from high-resolution time series with provably fast rates.

1.4.4 Alternative Applications of (Deep) Filters

Apart from modelling parameter processes for financial models, there are several other filtering applications for filtering, which we will give a brief overview below.

Finance. As streams of financial data are often extremely noisy, stochastic filters are indispensable in situations where latent parameters influence or obscure market factors. Applications include the computation of optimal investment under partial information [36, 187], hedging of credit derivatives [113, 114], and the estimation of volatility from observed intra-day stock prices [19], interest rates [11, 61, 98, 161], spot price models for commodity futures [105, 189, 210, 259, 260], and equilibria under asymmetric information [60]. Indeed, the abilities of stochastic filters to estimate unobservable variables, track market dynamics, and improve the precision of financial models have

made them a staple tool in the quantitative finance toolbox, with several books treating the role of stochastic filtering in finance [32, 87, 107, 247, 283].

Machine learning. Filtering principles have recently found fertile ground *inside* the architectures of modern machine-learning systems, where they offer a mathematically principled alternative to the heuristic gating mechanisms that dominate today's Mixture-of-Experts (MoE) large-language-model (LLM) landscape. In particular, in [257][5] we introduce MoE-F, a filtering-based framework that casts the selection among N pre-trained experts as a finite-state, continuous-time Hidden Markov Model and deploys the classical Wonham-Shiryaev filter to generate time-adaptive, information-efficient expert weights. This perspective stands in sharp contrast to the static or coarsely learned routers employed by contemporary MoE architectures—such as DeepSeek-V2 and its predecessors [90, 91, 130], Switch Transformers [106], Mixtral [162], Gemini [128], DBRX [133], and a host of merger-and-patch approaches [70, 193, 243, 254]. By recasting expert combination as a stochastic filtering problem, MoE-F affords closed-form updates and inherits optimality guarantees from Bayesian filtering theory, and attains substantial empirical gains in news-driven financial-market forecasting, while retaining the favourable computational scaling that has propelled MoE LLMs to the forefront of generative AI.

Other scientific domains. Applications are not only limited to financial mathematics or machine learning, but include a wide variety of scientific domains. In medicine, deep filtering techniques have been employed to model patient trajectories from irregularly sampled electronic health records [184, 185], to extract vital signs from video-based photoplethysmography using LSTM-based filters [40], to improve interpretation of ECG signals [256], and to denoise echocardiographic images via spatiotemporal convolutional autoencoders [274]. In robotics and human-machine interaction, deep Kalman filtering has been used to infer hand kinematics from surface electromyography signals for the control of prosthetic devices [17], while hybrid filtering approaches integrating GNSS and IMU data have advanced autonomous navigation [152]. Applications in computer

[5] This is joint work where I have acted as a second author to support the project with the derivation of theoretical results together with Anastasis Kratsios, the first author is Raeid Saqur.

vision include the reduction of compression artifacts in video data using deep generative temporal models [202]. In systems biology, Kalman filter-based methods have been developed to infer gene regulatory networks from noisy time-series expression data [237], offering robust insights into complex intracellular dynamics. See also [197], where filtering is proposed for model selection in computational biology. Further, in climate and earth system sciences, clustered particle filters have proven effective in managing the curse of dimensionality inherent to data assimilation in high-dimensional dynamical systems [191]. These examples highlight the methodological flexibility and cross-disciplinary impact of filtering frameworks, particularly when enhanced by modern machine learning architectures.

It is not the most intellectual of the species that survives; it is not the strongest that survives; but the species that survives is the one that is able best to adapt and adjust to the changing environment in which it finds itself.

— Leon C. Megginson on Darwin's Origin of Species

2 Adversarial Deep Hedging

Contents

2.1 Introduction 25
2.2 Preliminaries 28
2.2.1 Theoretical foundations 30
2.2.2 The data generating process (the path generator) 32
2.2.3 Deep hedging 35
2.3 Robust Hedging GAN 36
2.4 Numerical Results 40
2.4.1 Comparison to Herrmann, Muhle-Karbe, and Seifried [145] . 41
2.4.2 Out-of-sample test in the Black-Scholes model 43
2.4.3 Market-data induced uncertainty in the Heston model 46
2.4.4 Choice of the uncertainty aversion parameter γ 49
2.4.5 Real market data 50
2.4.6 On applying the robust hedging GAN with an NSDE 50
2.5 Conclusion 52

2.1 Introduction

This chapter presents a numerically efficient, model-agnostic framework for addressing model uncertainty in the penalised worst-case [PWC] setting (see Section 1.2). Our goal is to approximate optimal hedging strategies algorithmically via neural networks, thereby enabling robust decision-making even when the true data-generating process is ambiguous.

The proposed approach fuses three powerful, recently developed tools: deep hedging engines, synthetic market data generation (potentially via generative models), and signature methods on path space. These components are integrated within the structure of a generative adversarial network (GAN). While GANs are typically employed to produce realistic synthetic samples, here the adversarial framework serves a different purpose: the generator explores adverse yet plausible market dynamics, while the discriminator adapts its hedging strategy to remain effective against these stress scenarios. In this way, GAN training is repurposed from purely data generation to the reinforcement of hedging strategies against model uncertainty.

Concretely, we consider the problem of trading an asset $\mathbf{S}^{\mathbb{P}}$—modelled as a stochastic process under some measure $\mathbb{P}$—according to a discrete-time strategy φ over $N \in \mathbb{N}$ trading dates. At each step n, the position $\varphi_{t_{n-1}}$ is held until the next trading time, generating an increment $\varphi_{t_{n-1}}^{\intercal} \Delta \mathbf{S}_{t_n}^{\mathbb{P}}$. The resulting terminal wealth $(\varphi \cdot \mathbf{S}^{\mathbb{P}})_T$ is offset by a derivative payoff $\mathbf{C}_T^{\mathbb{P}} \in \mathcal{X}$, where $\mathcal{X} := \{X : \Omega \to \mathbb{R}\}$, and its performance is evaluated through a utility functional $U^{\mathbb{P}}$. Note that the payoff depends on the selected measure, i.e. $X \sim \mathbb{P}$ and the utility functional can be any mapping $\mathcal{X} \to \mathbb{R}$, as for instance expected utility or a monetary risk measure. In the absence of model uncertainty, the agent solves

$$\min_{\varphi \in \Phi} U^{\mathbb{P}}\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}}\Big),$$

where Φ denotes the set of admissible strategies.

To robustify this problem, we adopt the penalised worst-case formulation

$$\min_{\varphi \in \Phi} \left(\sup_{\mathbb{P} \in \mathcal{P}} U^{\mathbb{P}}\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}}\Big) - \alpha(\mathbb{P}) \right), \tag{2.1}$$

in which the generator selects adverse models $\mathbb{P} \in \mathcal{P}$, the penaliser α downweights implausible alternatives, and the discriminator (hedger) learns to perform optimally under the resulting stress scenarios.

The three components of (2.1) naturally align with the GAN architecture:

Generator: Acts as an adversary ("mother nature"), selecting dynamics $\mathbb{P} \in \mathcal{P}$ that yield the poorest hedging outcome for a given φ. This extends beyond a single reference model, systematically probing worst-case scenarios.

Penaliser: Reduces the impact of models deemed less plausible through the penalty term $\alpha(\mathbb{P})$. The choice of α is discussed in Section 2.2.1; examples include the SigW_1- and SigMMD-distances introduced in Appendix A.2.3.

Discriminator: Represents the robust hedger, choosing $\varphi \in \Phi$ to minimise (2.1) against the adversarially chosen models. In effect, it learns to exploit whatever favourable features remain in adverse market conditions.

Remark 2.1.1 (Interpretation of the robustified objective (2.1)). When the utility operators $U^{\mathbb{P}}$, $\mathbb{P} \in \mathcal{P}$, represent monetary risk measures and $\mathbf{X} \in \mathcal{X}$ denotes a contingent claim, $U^{\mathbb{P}}(\mathbf{X})$ can be interpreted as the minimal cash amount that must be added to $\mathbf{X}$ to render it acceptable from a risk perspective under $\mathbb{P}$. The penalty $\alpha(\mathbb{P})$ then reduces this capital requirement for models $\mathbb{P}$ considered less plausible. In the absence of model uncertainty, optimization is performed solely under a reference model $\mathbb{Q}$. The robust formulation extends this evaluation to all $\mathbb{P} \in \mathcal{P}$, while downweighting the influence of implausible alternatives through α. ◁

Outline. The remainder of this chapter is organized as follows. We begin with a theoretical discussion of the problem, where we repeat the formulation of the problem in a mathematical context and examine properties of the loss function (Section 2.2.1). Next, we revisit the deep hedging framework (Sections 2.2.2 and 2.2.3), before we formulate the robust hedging GAN in detail (Section 2.3). We conclude our analysis with numerical results (Section 2.4), where we first replicate the result of [145] and then compare the out-of-sample performance of our approach against the classic deep hedge. In Appendix A, we include a brief introduction to rough path theory, focussing on signatures and expected signatures, which are then used to derive the SigW_1-distance; as well as some additional material relevant for the construction of the out-of-sample tests.

2.2 Preliminaries

Let $T \in \mathbb{R}_+$ denote the fixed time horizon, and consider discrete trading dates

$$0 = t_0 < t_1 < \ldots < t_{N-1} < T,$$

for some $N \in \mathbb{N}$. The underlying probability framework is given by a filtered measurable space $(\Omega, \mathcal{F}, \mathfrak{F})$, where Ω is the state space, $\mathcal{F}$ is the terminal σ-algebra of measurable events, and $\mathfrak{F} := (\mathcal{F}_t)_{t\in[0,T]}$ is a filtration satisfying the usual conditions and representing the flow of market information over time.

We assume a market with $d \in \mathbb{N}$ hedging instruments whose (discounted) price process is denoted by $\mathbf{S} = (\mathbf{S}_t)_{t\in[0,T]}$, an $\mathbb{R}^d$-valued, $\mathfrak{F}$-adapted stochastic process. The dynamics of $\mathbf{S}$ are considered under different probability measures $\mathbb{P} \in \mathcal{P}$, with $\mathbf{S}^{\mathbb{P}}$ denoting the process under the law $\mathbb{P}$. Throughout, we assume without loss of generality that the risk-free interest rate is zero and that prices are expressed in discounted terms. For simplicity, we also omit transaction costs, although all results in this work can be extended to incorporate proportional transaction costs; see [50] for details.

Trading strategies. Let Φ denote the set of admissible trading strategies, where each $\varphi \in \Phi$ is a discrete-time, $\mathfrak{F}$-adapted process $\varphi := (\varphi_{t_n})_{n\in\{1,\ldots,N\}}$ with values in $\mathbb{R}^d$, representing the holdings in each hedging instrument between trading dates. Further restrictions on Φ (e.g., boundedness or self-financing constraints) can be imposed; see [50] for examples. Given $\varphi \in \Phi$, the discrete stochastic integral

$$(\varphi \cdot \mathbf{S}^{\mathbb{P}})_T := \sum_{n=1}^{N} \varphi_{t_{n-1}}^{\intercal} \Delta\mathbf{S}_{t_n}^{\mathbb{P}}, \quad \Delta\mathbf{S}_{t_n}^{\mathbb{P}} := \mathbf{S}_{t_n}^{\mathbb{P}} - \mathbf{S}_{t_{n-1}}^{\mathbb{P}},$$

represents the cumulative gains from trading over the horizon $[0, T]$.

Hedging in complete markets. Let $\mathcal{X}$ denote the set of $\mathbb{R}$-valued, $\mathcal{F}$-measurable random variables, interpreted as (discounted) contingent claims. If the market is defined under a single measure $\mathcal{P} = \{\mathbb{P}\}$ and is complete, then for any liability $\mathbf{X}^{\mathbb{P}} \in \mathcal{X}$ there exists a strategy $\varphi \in \Phi$ such that

$$-\mathbf{X}^{\mathbb{P}} + p_0 + (\varphi^{t_{n-1}} \cdot \mathbf{S}^{\mathbb{P}})_T \equiv 0, \tag{2.2}$$

where $p_0 \in \mathbb{R}$ is the initial capital, see [111]. In incomplete markets, such replication is generally impossible, and the left-hand side of (2.2) becomes a non-trivial random variable representing the *profit and loss* (P&L) from hedging $\mathbf{X}^{\mathbb{P}}$ with φ.

Risk-based performance evaluation. To assess the performance of $\varphi \in \Phi$, it is necessary to assign a numerical score to the random $\mathrm{P\&L}$. A common approach uses *convex risk measures* [111], which quantify the minimal cash amount that, when added to a position, makes it acceptable from a specified risk perspective (e.g., that of a regulatory body). Other choices include the expectation of $u(\mathrm{P\&L})$ for a mapping $u : \mathbb{R} \to \mathbb{R}$, called expected utility [179, 215, 230]. Our framework includes both methods, we refer to any mapping $U^{\mathbb{P}} : \mathcal{X} \to \mathbb{R}$ as a *utility functional*, indicating that it operates under the measure $\mathbb{P}$. In the case of hedging, we therefore denote the agent's objective (without consideration of model uncertainty) as

$$U^{\mathbb{P}}\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{X}^{\mathbb{P}}\Big). \tag{2.3}$$

Multiple models and model uncertainty. If $\mathcal{P}$ contains more than one probability measure, then different $\mathbb{P} \in \mathcal{P}$ will, in general, yield different capital requirements (2.3). Since the true measure is unknown *ex ante*, the agent must account for this model uncertainty when evaluating hedging performance.

A natural approach is the *worst-case* principle, whereby the agent computes the risk measure under each $\mathbb{P} \in \mathcal{P}$ and takes the maximum. To avoid excessive conservatism, the capital requirement is partially reduced for less plausible models via a *penalty function* $\alpha : \mathcal{P} \to \mathbb{R}$. This leads to the penalised worst-case formulation (2.1) introduced above, where the associated loss function is

$$\mathcal{L} : \Phi \times \mathcal{P} \longrightarrow \mathbb{R}, \quad (\varphi, \mathbb{P}) \longmapsto U^{\mathbb{P}}\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}}\Big) - \alpha(\mathbb{P}), \tag{2.4}$$

with $\mathbf{C}_T^{\mathbb{P}} \in \mathcal{X}$ denoting the discounted payoff of the derivative under $\mathbb{P}$.

2.2.1 Theoretical foundations

In this section, we examine a series of theoretical properties of the loss function $\mathcal{L}$ defined in (2.4). The results presented here are grounded in established principles from the theory of convex risk measures and robust optimisation (see, e.g., [109–111]), adapted to the specific robust hedging framework introduced above, similar as in [50]. Whilst some of these statements follow as natural extensions of classical results, this is not the case for all, and a careful formulation is essential to ensure the mathematical soundness of our setting. For clarity of exposition, all proofs are deferred to Appendix A.1.

Theorem 2.2.1 (Convexity)**.** *Suppose that $U^{\mathbb{P}}$ is convex for all $\mathbb{P} \in \mathcal{P}$. Then the mapping*

$$\varphi \longmapsto \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{L}(\varphi, \mathbb{P})$$

is convex on Φ.

The property in Theorem 2.2.1 aligns with the general literature on optimisation under convex risk measures in incomplete markets (cf. [154, 173, 288]). In this framework, the value of the optimisation problem (2.1) can be interpreted as quantifying the efficiency with which a given contingent claim can be replicated under model uncertainty. This naturally leads to a pricing paradigm known as *indifference pricing*.

To make this connection precise, observe that the robust hedging problem induces a functional $\pi : \mathcal{X} \to \mathbb{R}$ defined by

$$\pi(\mathbf{X}) := \inf_{\varphi\in\Phi} \sup_{\mathbb{P}\in\mathcal{P}} \Big(U^{\mathbb{P}}\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T + \mathbf{X}^{\mathbb{P}}\Big) - \alpha(\mathbb{P})\Big),$$

where $\mathbf{X} \in \mathcal{X}$ denotes the discounted payoff of a contingent claim under each $\mathbb{P} \in \mathcal{P}$. The next result shows that π is itself a convex risk measure.

Theorem 2.2.2. *Let $U^{\mathbb{P}}$ be a convex risk measure for every $\mathbb{P} \in \mathcal{P}$ and suppose that the admissible set Φ is convex. Then π defined above is a convex risk measure on $\mathcal{X}$.*

In particular, $\pi(-\mathbf{C}_T)$ represents the minimal additional cash required—assuming optimal hedging—to render the position acceptable both from a risk and model uncertainty

perspective. The *indifference price* of $\mathbf{C}_T$ is obtained by normalising this cash amount against the capital required for a zero claim:

$$p(\mathbf{C}_T) := \pi(-\mathbf{C}_T) - \pi(0).$$

If a replicating strategy exists, this indifference price coincides with the replication cost.

Lemma 2.2.3. *Suppose there exist $\varphi \in \Phi$ and $p_0 \in \mathbb{R}$ such that*

$$\mathbf{C}_T^{\mathbb{P}} \equiv p_0 + (\varphi \cdot \mathbf{S}^{\mathbb{P}})_T, \quad \forall \mathbb{P} \in \mathcal{P}.$$

Then $p(\mathbf{C}_T) = p_0$.

Remark 2.2.1 (On arbitrage). The absence of arbitrage is not imposed as a structural assumption in our setting. Following the discussion in [50], we exclude markets that are *irrelevant*, in the sense that

$$\inf_{\varphi \in \Phi} \rho\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T\Big) > -\infty,$$

for each $\mathbb{P} \in \mathcal{P}$. Notably, market irrelevance does not require the existence of explicit arbitrage opportunities; it may also arise through forms of statistical arbitrage. Moreover, even in the presence of apparent arbitrage and in the absence of trading frictions, such opportunities may remain unexploitable in practice. ◁

The convexity established in Theorem 2.2.1 concerns the dependence of the loss function on the hedging strategy φ. The second optimisation in (2.1)—the supremum over $\mathbb{P} \in \mathcal{P}$—is generally more intricate to analyse. However, when the risk measure belongs to the class of *optimised certainty equivalents* (OCEs), concavity in the second argument of $\mathcal{L}$ can be established.

Theorem 2.2.4 (Concavity)**.** *Suppose that, for each $\mathbb{P} \in \mathcal{P}$, the risk measure $U^{\mathbb{P}}$ is an optimised certainty equivalent, i.e.*

$$U^{\mathbb{P}}(\mathbf{X}) = \inf_{w \in \mathbb{R}} \Big(w + \mathbb{E}_{\mathbb{P}}\left[u(-\mathbf{X} - w)\right] \Big),$$

where $u : \mathbb{R} \to \mathbb{R}$ is continuous, non-decreasing, and convex. If, in addition, $\alpha : \mathcal{P} \to \mathbb{R}$ is convex, then $\mathbb{P} \mapsto \mathcal{L}(\varphi, \mathbb{P})$ is concave for every $\varphi \in \Phi$.

The class of OCE risk measures is sufficiently rich to encompass several prominent examples, including the *entropic risk measure* and the *average value at risk* (also known as *expected shortfall* or *conditional value at risk*); see [111] for a detailed discussion.

2.2.2 The data generating process (the path generator)

As discussed in the introductory section, the robust hedging framework introduced in Section 2.1 is modular in all three of its components. In particular, the *data generating process* can be chosen from a broad class of stochastic or data-driven models without altering the general structure of the optimisation problem. This flexibility allows the modeller to incorporate diverse modelling paradigms, ranging from classical stochastic models to modern neural network–based generators.

In line with the terminology of generative adversarial networks (GANs), we shall refer to the data generating process as a *path generator*. The term is deliberately chosen to encompass both traditional model-based approaches and modern machine learning architectures, thus generalising the notions of *market generator* and *market simulator* introduced in [54, 176, 228]. In our setting, the role of the generator departs from its classical use in GANs—where the goal is to produce samples resembling observed data—and instead focuses on producing scenarios that *challenge* the hedger according to criteria determined by the modeller.

Formally, a path generator is a mapping

$$\mathcal{G}_\xi : \begin{array}{ccc} \mathbb{R}^{N\times d'} & \longrightarrow & \mathbb{R}^{(N+1)\times d}, \\ (\Delta W_{t_n})_{n=1}^N & \longmapsto & (S_{t_n})_{n=0}^N, \end{array} \tag{2.5}$$

transforming d'-dimensional noise increments (e.g., Brownian motion increments) into time-discretised realisations of an asset price process with $d \leq d'$ tradable assets. The parameter vector $\xi \in \Xi$ specifies the generator's configuration, and Ξ denotes the corresponding parameter space.

For fixed parameters $\xi \in \Xi$, the generator $\mathcal{G}_\xi$ represents a specific model $\mathbb{P} \in \mathcal{P}$. Given a sequence of noise increments $(\Delta W_{t_n}^j)_{n=1}^N$ indexed by $j \in J \subseteq \mathbb{N}$, the mapping $\mathcal{G}_\xi$ produces a collection of sample paths $(\mathbf{S}_t^j)_{t\in\{t_n\}}$ distributed according to $\mathbb{P}$. The modeller determines $\mathcal{P}$ by specifying both the type and the richness of the admissible

model classes, as well as any constraints on their parameters. The dimension d' allows for the inclusion of non-tradable latent factors.

Illustrative classes of path generators. The generality of (2.5) admits a variety of concrete specifications, some of which are summarised below:

1. **Classical stochastic equity market models.** One may take $\mathcal{G}_\xi$ to be a parametric model from the classical literature—such as the Black–Scholes model, the Heston model, or discrete-time autoregressive processes—where ξ corresponds to the model parameters. In this case, "training" the generator reduces to calibrating ξ to observed market data. This choice retains direct interpretability and facilitates benchmarking against established results.

2. **Neural SDE (NSDE)–type generators.** Consider dynamics defined under $\mathbb{P} \in \mathcal{P}$ by

$$d\mathbf{S}_t = \mu^{\mathbb{P}}(t, \mathbf{S}_t)\, dt + \sigma^{\mathbb{P}}(t, \mathbf{S}_t)\, d\mathbf{W}_t^{\mathbb{P}}, \quad \mathbf{S}_0 \in \mathbb{R}^{d'}, \tag{2.6}$$

   where $\mu^{\mathbb{P}} : [0,T] \times \mathbb{R}^{d'} \to \mathbb{R}^{d'}$ and $\sigma^{\mathbb{P}} : [0,T] \times \mathbb{R}^{d'} \to \mathbb{R}^{d' \times d'}$ are continuous functions, and $\mathbf{W}^{\mathbb{P}}$ is a d'-dimensional $\mathbb{P}$-Brownian motion. A time-discretised realisation of (2.6) can be obtained from $(\Delta W_{t_n})_{n=1}^N$ using, for instance, the implicit Euler–Maruyama scheme:

$$S_n = S_{n-1} + \mu^{\mathbb{P}}(t_{n-1}, S_{n-1})\, \Delta t_n + \sigma^{\mathbb{P}}(t_{n-1}, S_{n-1})\, \Delta W_{t_n},$$

   with $\Delta t_n := t_n - t_{n-1}$. In NSDE-type generators, $\mu^{\mathbb{P}}$ and $\sigma^{\mathbb{P}}$ may themselves be parametrized by neural networks with parameter sets Ξ^μ and Ξ^σ, respectively, leading to $\Xi := \Xi^\mu \times \Xi^\sigma \times \mathbb{R}^{d'}$. Parameter constraints (e.g., drift restrictions or volatility positivity) may be imposed to ensure admissibility.

3. **LogSig–RNN generators.** Following [194, 228], path generation can also be based on recurrent neural networks with log-signature features. These methods have been shown to produce high-quality path samples with reduced parameter complexity, thereby offering a data-efficient alternative to more generic architectures.

Parametric formulation. The discussion above motivates a parametric reformulation of the robust hedging problem (2.1), in which the set of models $\mathcal{P}$ is represented by a family of path generators $\{\mathcal{G}_\xi : \xi \in \Xi\}$. We assume that the contingent claim $\mathbf{C}_T$ is determined by a measurable *payoff function*

$$c : \Big([0,T] \to \mathbb{R}^d\Big) \longrightarrow \mathbb{R}, \quad S \longmapsto C_T,$$

mapping an entire asset price path to its terminal payoff.[1]

Defintion 2.2.1 (Parametric robust hedging problem)**.** *Let $\mathcal{P}$ be represented by $\{\mathcal{G}_\xi : \xi \in \Xi\}$, and suppose that the operator U admits a heuristic estimator $\hat{U}$. The parametric version of (2.1) is then given by*

$$\min_{\varphi \in \Phi} \Big(\sup_{\xi \in \Xi} \hat{U}\Big((\varphi \cdot S^I)_T - C_T^I\Big) + \hat{\alpha}(\xi)\Big),$$

where:

- $S^I := \Big\{\mathcal{G}_\xi\Big((\Delta W_{t_n}^i)_{n=1}^N\Big)\Big\}_{i \in I}$ *denotes the set of generated sample paths for parameters $\xi \in \Xi$ and simulated Brownian increments $(\Delta W_{t_n}^i)_{n=1}^N$ indexed by $I \subseteq \mathbb{N}$,*
- $C_T^I := \hat{c}(S^I)$ *is the estimated payoff associated with the generated paths, where $\hat{c}$ is an empirical approximation of c,*
- $\hat{\alpha} : \Xi \to \mathbb{R}$ *is defined by $\hat{\alpha}(\xi) := \alpha(\mathbb{P})$ whenever $\mathcal{G}_\xi$ represents $\mathbb{P} \in \mathcal{P}$.*

Remark 2.2.2 (Concavity of the parametrized loss). The concavity in $\mathbb{P}$ established in Theorem 2.2.4 for the non-parametric formulation does not necessarily carry over to the parametrized setting in Definition 2.2.1. The mapping $\xi \mapsto \hat{U}((\varphi \cdot S^I)_T - C_T^I)$ may exhibit highly non-convex behaviour due to the potentially intricate dependence of the generated paths on ξ. Consequently, the associated optimisation landscape may contain multiple local minima, rendering the identification of the global minimum challenging in practice. Nevertheless, stochastic optimisation techniques—most notably stochastic gradient descent (SGD)—can offer effective means of navigating such landscapes and obtaining satisfactory approximations to the global optimum, even in the presence of significant non-convexity [41, 223, 249]. ◁

[1] A more general class of derivatives could be accommodated by allowing the generator to produce both the asset price process and the associated payoff process.

2.2.3 Deep hedging

We briefly recall the *deep hedging* methodology of [50][2] as it forms an essential building block of the proposed algorithm. The central idea is to approximate the trading strategy

$$\varphi : [0,T] \times \Omega \longrightarrow \mathbb{R}^d,$$

which maps time and current market information to portfolio holdings in d tradable assets, by a deep neural network. The network is denoted by φ^θ, where $\theta \in \Theta$ represents the vector of trainable parameters (weights and biases) belonging to a given parameter space Θ.

To fit θ, the strategy φ^θ is evaluated on a batch of simulated paths

$$S^I := \{S^{(i)}\}_{i\in I}, \quad I \subseteq \mathbb{N},$$

generated under a fixed model $\mathbb{P} \in \mathcal{P}$. For each path $S^{(i)}$, the payoff of the derivative is denoted $C_T^{(i)}$, and we write $C_T^I := \{C_T^{(i)}\}_{i\in I}$. Given an empirical estimator $\hat{U}$ of the utility functional U, the empirical hedging performance of φ^θ is measured by

$$\hat{U}\Big((\varphi^\theta \cdot S^I)_T - C_T^I\Big), \tag{2.7}$$

where $(\varphi^\theta \cdot S^I)_T$ denotes the terminal wealth across all paths in I. The network parameters θ are then updated via backpropagation [126] to *maximize* (2.7), typically using stochastic gradient-based methods.

We adopt this approach for approximating hedging strategies within the robust optimisation framework of (2.1), leading to the following consistency result.

Theorem 2.2.5 (Approximation of the robust hedging problem)**.** *Using the notation of Definition 2.2.1, for every $\epsilon > 0$ there exists a parameter space Θ for neural networks such that*

$$\left| \inf_{\theta\in\Theta} \sup_{\xi\in\Xi} \left(\hat{U}\Big((\varphi^\theta \cdot S^I)_T - C_T^I\Big) + \hat{\alpha}(\xi) \right) - \inf_{\varphi\in\Phi} \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{L}(\varphi, \mathbb{P}) \right| < \epsilon,$$

and $|\varphi^{\theta^\star} - \varphi^\star|_{[0,T]\times\Omega} < \epsilon$ for

$$\theta^\star := \arg\min_{\theta\in\Theta} \sup_{\xi\in\Xi} \left(\hat{U}\Big((\varphi^\theta \cdot S^I)_T - C_T^I\Big) + \hat{\alpha}(\xi) \right), \quad \varphi^\star := \arg\min_{\varphi\in\Phi} \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{L}(\varphi, \mathbb{P}).$$

[2]Note that we are using the basic deep hedging algorithm in this work, extensions include [52, 53, 222].

Proof. This follows from Definition 2.2.1 with $\Phi = \{\varphi^\theta : \theta \in \Theta\}$, combined with the universal approximation property for deep hedging strategies [50, Proposition 4.3], whose applicability is ensured by the convexity established in Theorem 2.2.1. □

Remark 2.2.3 (Convexity of the parametrized loss function). While the unparametrized loss function in (2.1) is convex in the strategy φ under the assumptions of Theorem 2.2.1, this property does not generally hold for the parametrized version

$$\theta \longmapsto \hat{U}\Big((\varphi^\theta \cdot S^I)_T - C_T^I\Big),$$

since neural network models are inherently non-convex [126]. Consequently, the associated optimisation landscape may contain multiple local minima, saddle points, and flat regions, which complicates convergence to the global optimum.

In practice, stochastic optimisation algorithms—such as (mini-batch) stochastic gradient descent—are employed to mitigate the risk of convergence to poor local minima and to exploit the regularising effect of noise in the gradient estimates. Although no general theoretical guarantees exist for convergence of neural networks to a global minimum in polynomial time, empirical evidence and prevailing belief in the deep learning community suggest that sufficiently expressive architectures, trained with appropriate stochastic optimisation schemes, can achieve near-optimal solutions within reasonable computational effort [126, 261]. ◁

2.3 Robust Hedging GAN

In Theorem 2.2.5, the robust hedging problem under model uncertainty (2.1) was reformulated as a parametric optimisation over the generator parameters $\xi \in \Xi$ and the hedging strategy parameters $\theta \in \Theta$. We now propose the *robust hedging GAN* as a practical implementation of this optimisation task, inspired by the architecture of generative adversarial networks and augmented with a penalisation mechanism to incorporate prior beliefs on model plausibility. The structure consists of the following three components (see Figure 2.1):

1. **Generator.** The generator $\mathcal{G}_\xi : \mathbb{R}^{N\times d'} \to \mathbb{R}^{(N+1)\times d}$, parametrized by $\xi \in \Xi$, takes as input a batch of d'-dimensional Brownian motion increments $(\Delta W_{t_n})_{n=1}^N$ and outputs time-discretised paths of the asset price process $\mathbf{S}^{\mathbb{P}}$ under the model $\mathbb{P}$ represented by $\mathcal{G}_\xi$. The dimension d' may exceed d to allow for latent, non-tradable factors. The generator's role is adversarial: it selects dynamics that produce unfavourable hedging outcomes for the current discriminator strategy.

2. **Discriminator.** The discriminator corresponds to the deep hedging strategy φ^θ, with $\theta \in \Theta$ denoting the neural network parameters. For the batch of paths S^I generated by $\mathcal{G}_\xi$, the discriminator computes the terminal profit and loss $(\varphi^\theta \cdot S^I)_T - C_T^I$, where C_T^I is the contingent claim payoff for each path in I. The performance is evaluated using an empirical estimator $\hat{U}$ of the utility functional U, as in (2.7), and the discriminator adapts θ to maximize performance against the most adverse generator.

3. **Penaliser.** The penaliser assigns a model-dependent penalty $\alpha(\mathbb{P})$, expressed in the parametric setting as $\hat{\alpha}(\xi)$ for the generator $\mathcal{G}_\xi$. While any penalty functional can be used, we recommend the *signature maximum-mean discrepancy* SigMMD (or an approximation thereof), which compares the distribution of generated paths to that of a set of reference paths directly in path space (see Appendix A.2). The penalty is computed by generating a batch of paths from $\mathcal{G}_\xi$ and measuring the distance to a batch of reference paths. When only non-path data (e.g., volatility smiles) are available, a reference generator is first calibrated to the observed data, and its output paths are then used as the reference set.[3] In our implementation, the penaliser is fixed (non-trainable), but it could be extended to include trainable parameters, particularly when defined via a supremum over a parameterised family.

The robust hedging GAN is trained in two stages: an initial *pre-training* phase, during which the generator and discriminator are trained separately, followed by an *adversarial training* phase, in which the two components interact adversarially with the penaliser to solve the joint optimisation problem.

[3] In this case, the reference generator can also serve to initialise the adversarial training phase, avoiding additional computational cost.

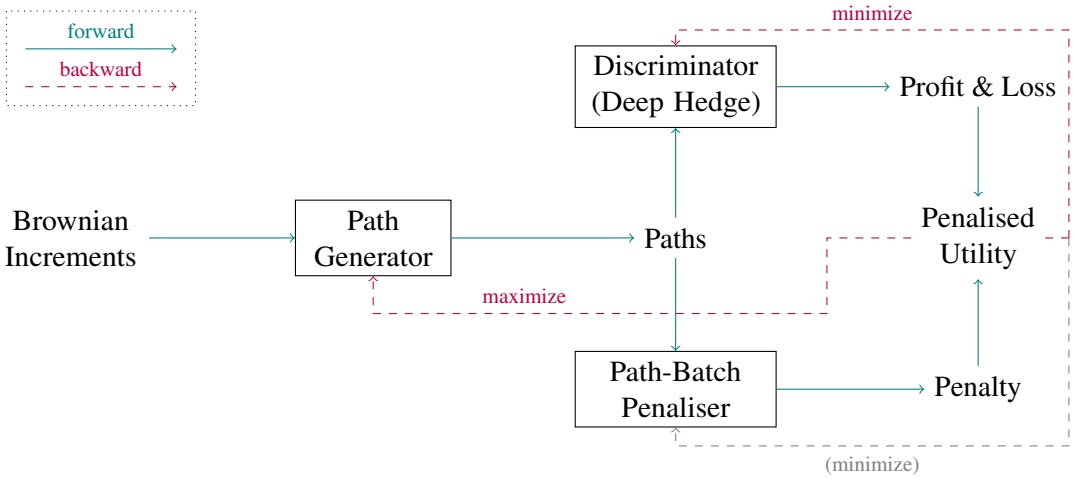


Figure 2.1: Structure of the robust hedging GAN. Green arrows denote the forward pass: white noise is mapped by the generator to candidate trajectories, which feed into the hedger (for strategy implementation) and the penaliser (for plausibility assessment). Red arrows indicate backpropagation: the discriminator is updated to minimise the penalised utility (the combination of the hedge $\mathrm{P\&L}$ and penalty), while the generator is updated to maximize it.

Pre-training. While the eventual goal is to train all components of the robust hedging GAN jointly, it is advantageous to begin by training each component in isolation. This reduces the likelihood that the initial parameter choices for the generator $\xi \in \Xi$ and the discriminator $\theta \in \Theta$ will become trapped in irrelevant local saddle points during the subsequent adversarial phase. The procedure is as follows:

1. *Generator pre-training.* Define and calibrate the path generator $\mathcal{G}_\xi$ to represent a reference model $\mathbb{P} \in \mathcal{P}$. The calibration aligns the generator's parameters ξ with past historical observations or with a market state inferred from derivative prices (e.g., from an implied volatility surface). As described in Section 2.2.2, $\mathcal{G}_\xi$ may correspond to a classical parametric equity market model such as Black–Scholes, Heston, or rough Bergomi, or to a data-driven machine learning–based generator. This choice encodes both our current belief about the prevailing market distribution $\mathbb{P}$ and the admissible set of alternative models in $\mathcal{P}$ reachable under model uncertainty. The parameterisation determines the degrees of freedom available to the generator during adversarial training.

2. *Discriminator pre-training.* Fit the deep hedging strategy φ^θ to the pre-trained generator from Step 1. This is done by repeatedly sampling batches of paths S^I

from $\mathcal{G}_\xi$ and adjusting the network parameters θ to maximize (2.7), i.e.

$$\hat{U}\Big((\varphi^\theta \cdot S^I)_T - C_T^I\Big),$$

where C_T^I denotes the contingent claim payoffs for each simulated path in $I \subseteq \mathbb{N}$.

The outcome of this phase corresponds to a deep hedging solution without explicit consideration of model uncertainty, as in [50].

Following pre-training, the generator $\mathcal{G}_\xi$ may still represent an inaccurate or misspecified model $\mathbb{Q} \neq \mathbb{P}^\star$. This can arise in several scenarios: (i) the market is non-stationary and the true distribution $\mathbb{P}^\star$ has shifted away from the calibrated reference model $\mathbb{Q}$, (ii) the calibration employs simplifying assumptions (e.g., static implied volatilities), (iii) the generator is over-parameterised and overfits to limited data, (iv) the generator class is too restrictive to capture the true market dynamics, or (v) the calibration is imprecise due to insufficient data.

Adversarial training. To address the above issues, we adopt an adversarial training phase in which the generator and discriminator are updated in alternating steps, with the penaliser constraining the generator's exploration.

A single forward pass proceeds as follows: (i) the generator $\mathcal{G}_\xi$ produces a batch of paths S^I; (ii) the discriminator φ^θ applies the hedging strategy to S^I, producing terminal wealth values $(\varphi^\theta \cdot S^I)_T - C_T^I$; (iii) the empirical utility $\hat{U}$ is evaluated; and (iv) in parallel, the generated paths are compared with a batch of reference paths to compute the penalty term $\hat{\alpha}(\xi)$.

The parameters ξ and θ are then updated in two sub-steps:

1. *Generator update.* The generator parameters ξ are updated to *maximize* the loss, i.e. to identify scenarios under which the current strategy φ^θ performs poorly. Economically, this corresponds to moving away from the most likely model towards nearby models in $\mathcal{P}$ that are more challenging for the hedge. The penalty term $\hat{\alpha}(\xi)$ constrains this movement by reducing the loss when ξ deviates substantially from the reference model.

2. *Discriminator update.* The strategy parameters θ are updated to *minimise* the loss, adapting the hedge to the generator's most recent challenge. Robustness is encouraged because the strategy must retain performance against previous generator configurations—otherwise the generator can revert to an earlier state, aided by the penalty function.

This alternating update scheme continues until convergence or until a predetermined number of iterations is reached.[4]

Remark 2.3.1 (Convergence considerations). The convexity and concavity results of Theorems 2.2.1 and 2.2.4 indicate that the non-parametric robust hedging problem has a convex–concave structure. In the parametrized case of Theorem 2.2.5, however, neither convexity nor concavity can be guaranteed; see Remarks 2.2.2 and 2.2.3. Consequently, the training dynamics may converge to local rather than global optima (cf. [175]).

Two practical measures are adopted to mitigate this limitation. First, stochastic gradient-based optimisation is employed, which reduces the likelihood of becoming trapped in poor local optima by introducing gradient noise. Second, the pre-training stage described above provides a favourable initialisation for both generator and discriminator. Even if the final equilibrium is not globally optimal, it can still yield materially improved robustness in the resulting hedging strategy. ◁

2.4 Numerical Results

In this section, we will demonstrate the usage of the *robust hedging GAN*. We will conduct five experiments, (i) we reproduce the results of [145] in a exemplary Black-Scholes setting, (ii) we investigate its out-of-sample performance in an artificial model uncertainty setup within the Black-Scholes model, (iii) move on to an out-of-sample performance test based on the Heston model and motivated by real market data, (iv) test the model on real market data, and (v) demonstrate the application with an NSDE-generator.

[4] An open-source implementation of the robust hedging GAN, including all numerical experiments of this work, is available at `https://github.com/YannickLimmer/adversarial-robust-deep-hedging`.

2.4.1 Comparison to Herrmann, Muhle-Karbe, and Seifried [145]

We begin with a simple case: the Black–Scholes model [37] in the arbitrage-free setting with zero drift, leaving a single parameter, the volatility. The generator takes the form

$$d\mathbf{S}_t^{\mathbb{P}} = \sigma^{\mathbb{P}} \mathbf{S}_t^{\mathbb{P}} \, d\mathbf{W}_t^{\mathbb{P}}, \quad t \in [0, T], \quad \mathbf{S}_0^{\mathbb{P}} \equiv S_0, \tag{2.8}$$

in the notation of (2.6), where $\sigma^{\mathbb{P}} \in \mathbb{R}_+$ and $\mathbf{W}^{\mathbb{P}}$ is a Brownian motion under $\mathbb{P}$. The uncertainty set is

$$\mathcal{P} := \{\mathbb{P} : (2.8) \text{ holds for some } \sigma^{\mathbb{P}} > 0\},$$

and the parameter space is $\Xi := \mathbb{R}_+$.

In this example, the reference model $\mathbb{Q}$ is specified by $\sigma^{\mathbb{Q}} = 0.2$. We allow trading at 18 equally spaced dates with $\Delta t = 5/255$, corresponding to approximately one business week, and set $S_0 = 1$. The hedging problem considers an at-the-money European call option with strike 1 and payoff

$$\mathbf{C}_T = (\mathbf{S}_T - 1)_+.$$

Since [145] employ expected-utility operators, we take

$$U^{\mathbb{P}}(\mathbf{X}) := \mathbb{E}_{\mathbb{P}}\left[\frac{1 - \exp(-\lambda \mathbf{X})}{\lambda}\right],$$

where $\lambda > 0$ is the risk aversion parameter. For the standalone deep hedge, we use a feedforward network with two hidden layers of 128 nodes each, ReLU activation in hidden layers, and an identity output activation.

For the robust hedging GAN, we adopt the penalty[5]

$$\alpha_\varphi(\mathbb{P}) := \frac{1}{\gamma N} U^{\mathbb{P}}(\mathrm{P\&L}^\varphi) \sum_{n=1}^{N} \left(\sigma^{\mathbb{Q}} - \sqrt{\frac{\mathbb{V}_{\mathbb{P}}\left[\log(\Delta \mathbf{S}_{t_n})\right]}{\Delta t_n}}\right)^2.$$

Remark 2.4.1 (Homogeneous penalty). The inclusion of the hedge objective in the penalty, as in [145], has several motivations: (i) invariance of preferences under affine transformations of the utility function; (ii) facilitation of a dynamic programming

[5] In the notation of [145], this corresponds to $f(t, s, y; \sigma) := (\sigma - \bar{\sigma})^2 / T$.

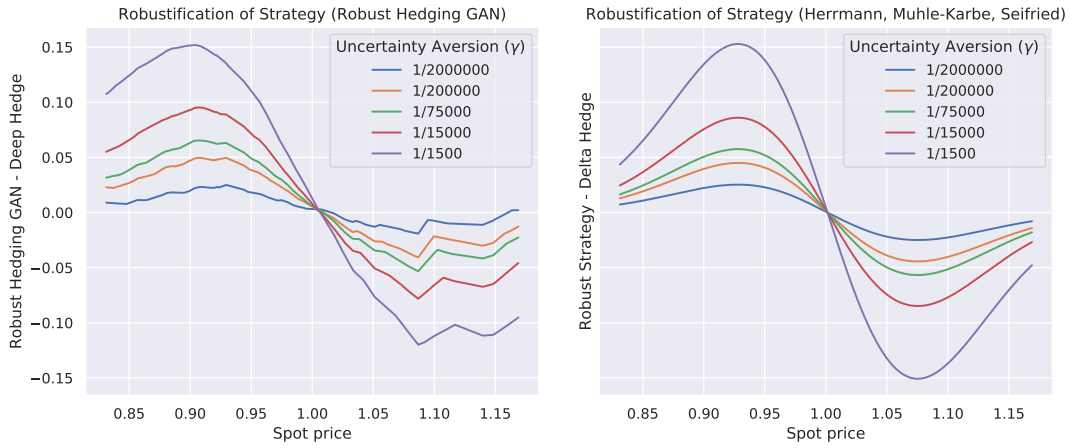


Figure 2.2: Comparison between the robust hedging GAN strategy and the strategy expansion of [145] evaluated for various spot values.

formulation; and (iii) prevention of pathwise dependence of the strategy on the current profit and loss. Our framework can incorporate such dependence, but it is not required. In some contexts, we argue it may be advantageous for the uncertainty analysis to be defined purely in terms of the asset dynamics, independent of the current strategy. ◁

Figure 2.2 compares our robust hedging GAN results with the asymptotic analytical solution of [145]. The two coincide closely, offering insight into the economic nature of the robustification: when the option is out of the money, an upward movement in the stock price increases sensitivity to volatility misspecification. The uncertainty-aware strategy counteracts this by holding more stock than the pure delta hedge, generating additional profit. Conversely, a downward movement reduces this exposure, and the adjustment leads to a small loss relative to the delta hedge—offset by the reduced vulnerability. The in-the-money case exhibits analogous behaviour in the opposite direction. In summary, the robust strategy mitigates exposure to stock price movements into regions of high vulnerability, which in [145] is characterised by high cash gamma.

Remark 2.4.2 (Sources of discrepancy). Small differences between the curves in Figure 2.2 arise from: (i) the relatively simple network architecture (two layers, 128 nodes each) and the fact that the figure plots the difference between two deep hedge outputs, amplifying noise; (ii) the discrete trading frequency (five business days) leading to market incompleteness and deviation from the Black–Scholes delta hedge; and (iii) the analytical

solution's asymptotic nature, based on a quadratic Taylor expansion, which limits its exactness. ◁

Remark 2.4.3 (Relation to further results). Wu and Jaimungal report in [287] similar findings using a different robustification method, modifying the profit and loss directly via a neural network within a Wasserstein ball. Their knock-in call option results are equivalent to our vanilla option results once knocked in. Despite using a different penalty, our outcomes are qualitatively aligned (cf. left panel of Figure 2.2 with Figure 3 in [287]).

The key distinction is that [287] address full distributional robustness, while here we replicate the parameter-uncertainty setting of [145]. A broader discussion of this difference is given in [21]. We conjecture that a sufficiently flexible generator class (see Section 2.2.2) could approximate distributionally robust solutions, a direction explored in Section 2.4.6. ◁

2.4.2 Out-of-sample test in the Black-Scholes model

As a next step, we want to examine the out-of-sample performance of the robust hedging GAN and compare it to the deep hedge. For this, we continue in the above Black-Scholes setting, however define the penaliser as a comparison of the volatility estimates of the generated paths $\sigma^{\mathbb{P}}$ to the volatility estimate of the batch $\sigma^{\mathbb{Q}}$, weighted by $\gamma > 0$. This means,

$$\alpha(\mathbb{P}) := \frac{1}{\gamma N} \sum_{n=1}^{N} \left(\sigma^{\mathbb{Q}} - \sqrt{\mathbb{V}_{\mathbb{P}} \left[\log(\Delta \mathbf{S}_{t_n}) \right]} \right)^2 . \tag{2.9}$$

Example 2.4.1 (Black-Scholes test case). We assume that the derivative, trading frequency, initial asset price, set of eligible parameters Ξ, and estimated parameter $\sigma^{\mathbb{Q}}$ are specified as above.

The hedging objective in this example is chosen as the *entropic risk measure*, i.e.

$$U^{\mathbb{P}}(\mathbf{X}) := \frac{1}{\lambda} \log \mathbb{E}_{\mathbb{P}} \left[\exp(-\lambda \mathbf{X}) \right]$$

for any random variable $\mathbf{X}$, where we fix the risk aversion as $\lambda = 130$.[6] ◁

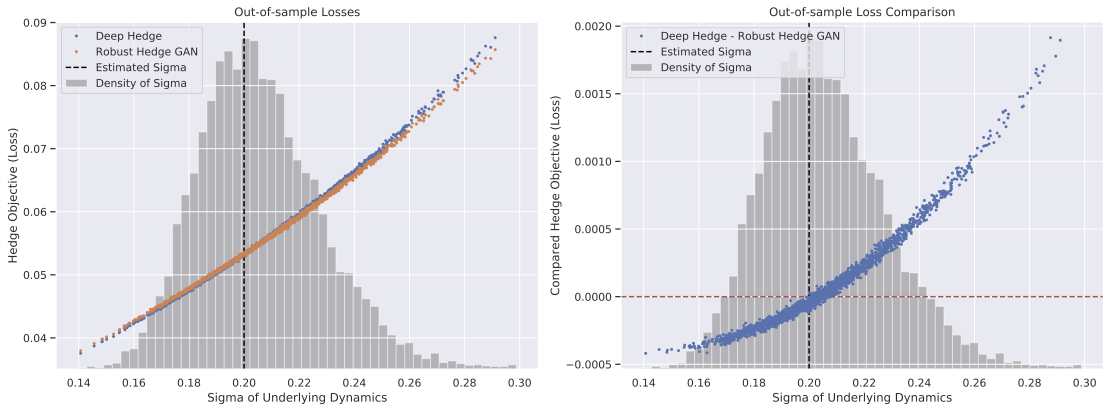


Figure 2.3: Out-of-sample performances (loss) for individual scenarios of the test performed in Example 2.4.1 with $\gamma := 1/75$. For every sampled $\sigma^{\mathbb{P}}$, we compute the performance of the deep hedge and robust hedging GAN on the model $\mathbb{P}$, yielding each a single data point in the plot.

Example 2.4.2 (Black–Scholes out-of-sample test data). We consider an agent who models market dynamics with the driftless Black–Scholes model, calibrating $\sigma^{\mathbb{Q}}$ to the realised volatility over the past 45 business days. Suppose $\sigma^{\mathbb{Q}} = 0.2$. Knowing the sample size, we can invert the estimation procedure to recover the distribution of the true volatility parameter $\sigma^{\mathbb{P}}$, shown in Figure 2.3.

Using this distribution, we test the robust hedging GAN of Example 2.4.1 over a range of "true" volatilities and statistically evaluate its out-of-sample performance. The out-of-sample loss (cf. (A.11)) is computed from $M = 10\,000$ models $\{\mathbb{P}^m\}_{m=1}^M$, with each utility approximated by $10\,000$ simulated paths. ◁

Figure 2.4 shows that, on average, robust hedging GANs reduce out-of-sample loss relative to standalone deep hedges, especially for low uncertainty-aversion (γ small), with the best mean at $\gamma = 1/100$. As $\gamma \downarrow 0$, performance converges to the deep hedge; for larger γ, the cost of robustness outweighs the benefit of excluding poor models.

Remark 2.4.4 (Test hedge). To benchmark these results, we also trained a *test hedge* using the full scenario set: each σ value generates one path, and training is carried out on the mixture $\tilde{\mathbb{P}} = \frac{1}{M} \bigoplus_{m=1}^M \mathbb{P}_m$. This corresponds to Bayesian model averaging [147],

[6]This choice can roughly be motivated by considering mean-volatility, i.e. $U^{\mathbb{P}}(\mathbf{X}) := \mathbb{E}_{\mathbb{P}}[\mathbf{X}] - \frac{\lambda'}{2}\sqrt{\mathbb{V}_{\mathbb{P}}[\mathbf{X}]}$, where we manipulate the expectation to the lower bound of the confidence interval given by $\lambda'/2$ standard deviations. A reasonable choice for the mean-variance risk-aversion is thus $\lambda = \lambda'/V$ with $V^2 \approx \mathbb{V}_{\mathbb{P}}[\mathbf{X}]$. We know that the entropic risk measure coincides with the mean-variance operator for normally distributed random variables.

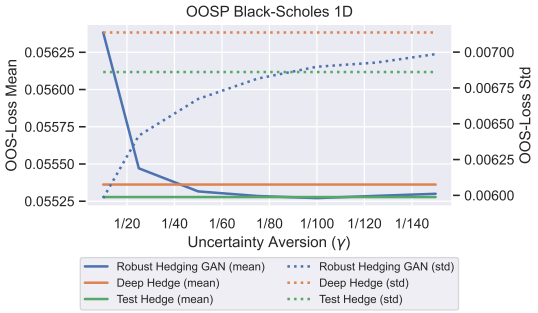


Figure 2.4: Mean and standard deviation of out-of-sample losses for Example 2.4.1 over various γ. This is obtained by taking the mean and standard deviation of the points in Figure 2.3 for each γ.

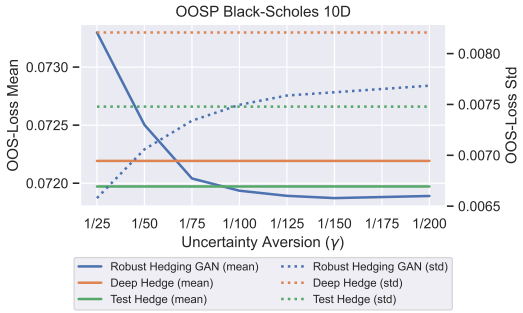


Figure 2.5: Mean and standard deviation of out-of-sample losses for the 10-dimensional case (Example 2.4.3).

discussed critically in [81] and applied to deep hedging in [205]; see also Section 3.3 for more details.

Although this strategy is generally not optimal for either the mean or variance of the out-of-sample loss, it achieves lower loss than the deep hedge. Notably, the robust hedging GAN still outperforms the test hedge in both metrics for $1/\gamma \in \{50, 75\}$. ◁

Variance reductions are more pronounced: standard deviations decrease with increasing γ, reaching zero in the limit $\gamma \to \infty$ (zero strategy), and matching the deep hedge as $\gamma \to 0$. For sufficiently high aversion ($\gamma \geq 1/75$), the robust hedging GAN yields lower variance than even the test hedge.

Figure 2.3 (for $\gamma = 1/75$) shows that the GAN underperforms in low-loss cases ($\sigma^{\mathbb{P}_m} \leq \sigma^{\mathbb{Q}}$) but delivers substantial gains for high-loss scenarios, corresponding to underestimation of volatility—consistent with the well-known result that hedging is harder in more volatile markets [104].

Example 2.4.3 (Multi-dimensional Black–Scholes). We repeat the experiment in a 10-dimensional Black–Scholes setting, with initial prices $S_0 \in \mathbb{R}^{10}$ and covariance $\sigma^{\mathbb{P}} \in \mathbb{R}^{10\times 10}$. The claim is a European basket-maximum call, $\mathbf{C}_T = \left(\max_{i\leq 10} \mathbf{S}_T^{(i)} - 1\right)_+$.

The penalty of (2.9) is extended to a Frobenius norm over the covariance matrix to

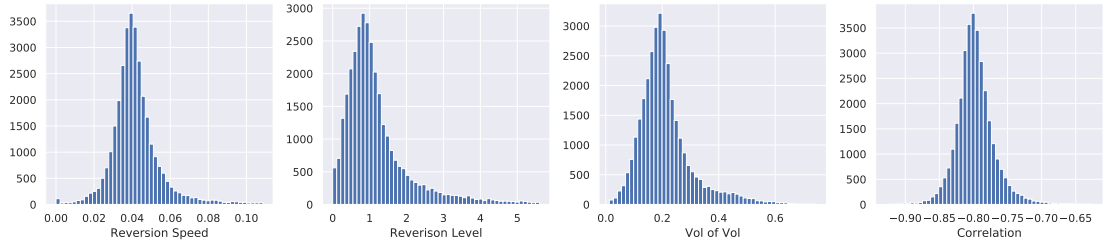


Figure 2.6: Heston parameters motivated by real market option prices as described in Example 2.4.5.

capture correlation uncertainty:

$$\alpha(\mathbb{P}) = \frac{1}{\gamma N} \sum_{n=1}^{N} \sum_{i,j=1}^{10} \left(\sigma_{ij}^{\mathbb{Q}} - \frac{\mathrm{Cov}_{\mathbb{P}}(\mathbf{S}^{(i)}, \mathbf{S}^{(j)})}{\mathbb{V}_{\mathbb{P}}\left[\log(\Delta \mathbf{S}_{t_n}^{(i)})\right] \mathbb{V}_{\mathbb{P}}\left[\log(\Delta \mathbf{S}_{t_n}^{(j)})\right]} \right)^2 .$$

Results in Figure 2.5 are consistent with the one-dimensional case. ◁

Remark 2.4.5 (Ensuring comparability). To ensure fair comparison, both the deep hedge and robust GAN were trained for the same total number of iterations. The deep hedge was trained on 2^{16} simulated paths with increasing batch sizes $\{2^8, 2^{10}, 2^{12}, 2^{14}\}$, five repetitions per size, regenerating paths each full pass. The robust GAN was initialised from the pre-trained deep hedge and trained for $1\,000$ epochs with batch size 2^{16}, matching an additional $1\,000$-epoch retraining of the deep hedge alone. ◁

2.4.3 Market-data induced uncertainty in the Heston model

We now turn to the Heston model [146], whose dynamics are

$$d\mathbf{S}_t^1 = \mathbf{S}_t^1 \sqrt{\mathbf{S}_t^2}\, d\mathbf{W}_t^1, \qquad \mathbf{S}_0^1 = s_0^1, \tag{2.10a}$$

$$d\mathbf{S}_t^2 = \kappa^{\mathbb{P}}\left(\beta^{\mathbb{P}} - \mathbf{S}_t^2\right) dt + \sqrt{\mathbf{S}_t^2}\, \sigma^{\mathbb{P}}\, d\mathbf{W}_t^2, \qquad \mathbf{S}_0^2 = s_0^2, \tag{2.10b}$$

where $\kappa^{\mathbb{P}}, \beta^{\mathbb{P}}, \sigma^{\mathbb{P}} > 0$ and $\rho^{\mathbb{P}} \in [-1, 1]$ is the correlation between $(\mathbf{W}^1, \mathbf{W}^2)$. For simplicity, $s_0^2 = \beta^{\mathbb{P}}$. A model $\mathbb{P} \in \mathcal{P}$ is thus parameterised by

$$\xi := (\kappa^{\mathbb{P}}, \beta^{\mathbb{P}}, \sigma^{\mathbb{P}}, \rho^{\mathbb{P}}) \in \Xi := \mathbb{R}_+^3 \times [-1, 1].$$

Since volatility $\mathbf{S}^2$ is not directly tradable, we introduce the volatility swap

$$\mathbf{S}_t^3 := \mathbb{E}^{\mathbb{P}}\left[\int_0^T \mathbf{S}_s^2\, ds \,\middle|\, \sigma\left((\mathbf{S}_s^1, \mathbf{S}_s^2)_{s \in [0,t]}\right) \right],$$

which admits the closed form

$$\mathbf{S}_t^3 = \int_0^t \mathbf{S}_s^2\, ds + \frac{\mathbf{S}_t^2 - \beta^{\mathbb{P}}}{\kappa^{\mathbb{P}}}\left(1 - e^{-\kappa^{\mathbb{P}}(T-t)}\right) + \beta^{\mathbb{P}}(T-t). \tag{2.11}$$

Trading is permitted in $(\mathbf{S}^1, \mathbf{S}^3)$, while $(\mathbf{S}^1, \mathbf{S}^2)$ inform the hedge and generator fit [50].

Example 2.4.4 (Heston test case). We replicate the Black–Scholes experiment (Example 2.4.1) with the Heston dynamics (2.10), using the same discretisation, payoff (European call), and deep-hedge architecture. The reference model has

$$\xi^{\mathbb{Q}} = (\kappa^{\mathbb{Q}}, \beta^{\mathbb{Q}}, \sigma^{\mathbb{Q}}, \rho^{\mathbb{Q}}) = (1,\ 0.04,\ 0.2,\ -0.8).$$

The robust GAN penalty is the SigMMD (Appendix A.2.3), applied to $(\mathbf{S}^1, \mathbf{S}^2)$ only, with paths time-augmented and lead–lag transformed (Remark A.2.1), signature depth $M = 2$, and scaling $1/\gamma$. ◁

Example 2.4.5 (Heston uncertainty from SPX implied volatilities). To reflect market-data–driven parameter shifts, we calibrate the Heston model daily to SPX end-of-day smiles from 1 January 2010 to 18 March 2019 using the *deep learning volatility* framework [151]. Let $(\xi_m)_{m=1}^{M'} \subset \Xi$ denote the M' daily calibrations.

For $l \in \{1, \dots, 10\}$ and $m \geq 11$, define shifted parameters

$$\xi^{\mathbb{P}_{10m+l}} := \xi^{\mathbb{Q}} + \left(\xi_m - \xi_l\right),$$

yielding $M = 10M'$ scenarios that represent plausible parameter changes over 1–10 business days without recalibration. Minor violations of Ξ are corrected ad hoc (e.g. clipping ρ to $[-1, 1]$, taking σ absolute). The resulting distribution is shown in Figure 2.6. ◁

Results. Figure 2.8 shows that the robust hedging GAN achieves slightly lower mean out-of-sample losses and consistently reduced variance relative to the deep hedge. As in Example 2.4.1, mean losses converge to the deep hedge as $\gamma \downarrow 0$ and rise for large γ, while variances behave inversely, tending to zero in the limit of the zero strategy.

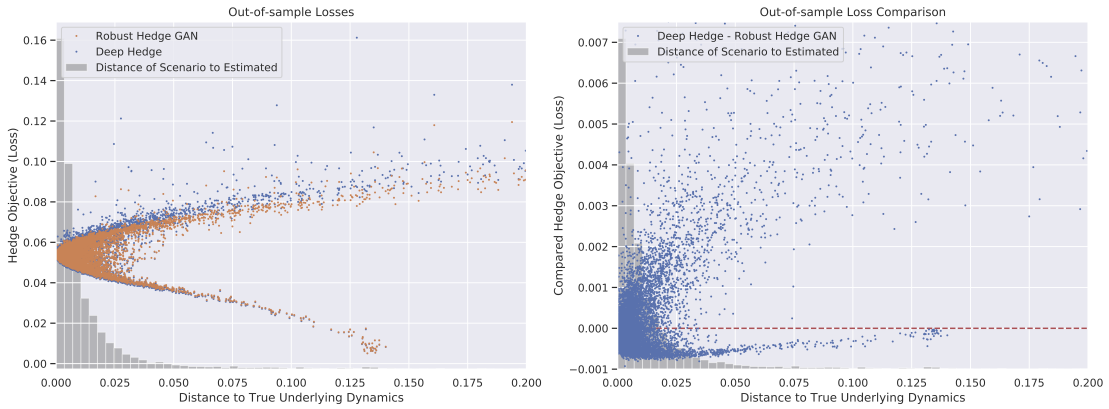


Figure 2.7: Scenario losses in Example 2.4.4 with $\gamma = 17{,}500$. For every sampled set of Heston parameters, we compute the performance of the deep hedge and robust hedging GAN on the corresponding Heston model, yielding each a single data point in the plot.

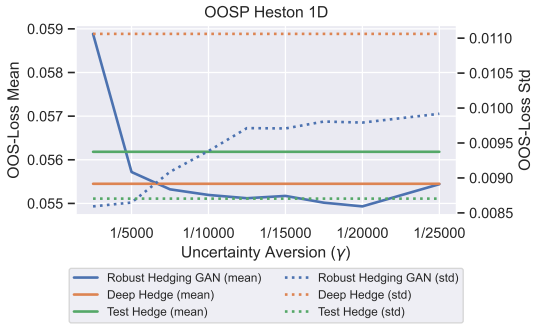


Figure 2.8: Mean and standard deviation of out-of-sample losses for Example 2.4.4 under scenarios from Example 2.4.5, over various γ. This is obtained by taking the average and standard deviation of the data points in Figure 2.8 for each γ.

The GAN also outperforms the artificial test hedge in both metrics for appropriate γ, confirming that even small improvements in mean loss can be significant when coupled with variance reduction.

Scenario-level losses in Figure 2.7 exhibit a parabolic relation with parameter distance: shifts in one direction improve hedging, while the opposite degrades it. The robust GAN concentrates improvements in high-loss regions, at some cost in low-loss cases. In summary, the findings from the Black–Scholes experiments carry over to the more complex and less tractable Heston setting.

2.4.4 Choice of the uncertainty aversion parameter γ

Our analysis does not prescribe a specific method for selecting the uncertainty aversion parameter γ. This choice may be interpreted in two distinct ways. Firstly, γ can be regarded as a subjective decision, determined by the agent in accordance with their individual risk preferences and tolerance towards model misspecification. Alternatively, it can be treated as an optimal tuning parameter within a given market and model uncertainty framework, selected to maximize performance under the prevailing conditions.

To address the latter interpretation—the optimal choice within a fixed uncertainty environment—we propose a procedure analogous to the Heston out-of-sample test of Example 2.4.5. The procedure comprises the following steps:

1. *Specification of uncertainty.* Define the prevailing model uncertainty as a probability distribution over the parameter space Ξ. This distribution encodes the plausible range of parameter values consistent with historical evidence or expert assessment.

2. *Training across γ.* For a set of candidate uncertainty aversion parameters, train a robust hedging GAN in each case. The generator-hedge system is thus optimised to reflect varying degrees of robustness to parameter perturbations.

3. *Out-of-sample evaluation.* Assess the mean and variance of the resulting hedging strategies under the specified uncertainty distribution. Comparing these metrics across γ facilitates the identification of a parameter value yielding a desirable balance between expected performance and stability.

By following this procedure, practitioners can systematically investigate the effect of γ on hedging performance, given a specific model uncertainty description. This offers a principled basis for selecting γ that accounts for prevailing uncertainty and targets robust, reliable outcomes in practice.

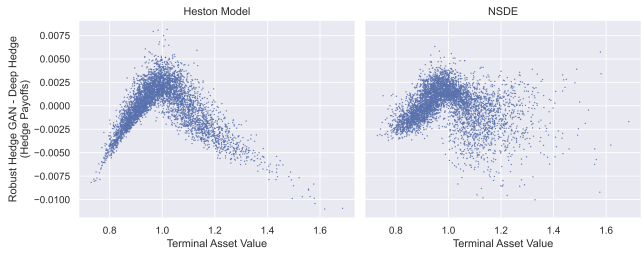


Figure 2.9: Comparing the robust hedging GAN and the deep hedge for both, an NSDE generator and a Heston generator.

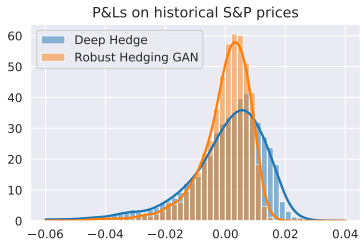


Figure 2.10: Comparing P&Ls of deep hedge and robust hedging GAN on S&P500 market data.

2.4.5 Real market data

We conclude by evaluating our methods on real S&P 500 market data obtained from OptionMetrics[7]. Our aim is to hedge a European call option with the Heston model as the reference generator, targeting low standard deviation of hedging errors. The Heston parameter calibrations from Example 2.4.5 are employed for all tests.

For each trading day in the sample, we train two separate models: a deep hedge and a robust hedging GAN. These are then applied to the realised S&P 500 index path up to the option's expiry. The comparison is illustrated in Figure 2.10. To focus solely on relative hedging performance, we normalise the mean profit-and-loss values to zero, thereby eliminating the effect of option premium considerations.

The results show that both approaches achieve similar mean performances, yet the robust hedging GAN attains a markedly lower out-of-sample objective, measured as the standard deviation of hedging errors. This reduction indicates greater stability and reliability in the robust hedging GAN relative to the standard deep hedge.

We emphasise that the comparative advantage observed here necessarily depends on the accuracy of the Heston calibrations and the fidelity of the market data. Deviations in either can materially affect the relative and absolute performance of the two strategies.

2.4.6 On applying the robust hedging GAN with an NSDE

We conclude our empirical investigation by applying the robust hedging GAN with a neural stochastic differential equation (NSDE) as the path generator. Specifically, we

[7] https://optionmetrics.com/

employ dynamics of the form (2.6) with $d' = 2$, omitting the time component. The neural networks approximating the drift $\mu^{\mathbb{P}}(\cdot)$ and diffusion $\sigma^{\mathbb{P}}(\cdot)$ functions are each initialised with three hidden layers of 36 nodes. To ensure economically and mathematically consistent dynamics, we impose the following structural constraints: the drift component $\mu^{\mathbb{P}}(\cdot)$ in the first dimension is set identically to zero, and the diffusion component $\sigma^{\mathbb{P}}(\cdot)$ is passed through an absolute value operation to enforce non-negativity.

In the first step, we fit the NSDE generator to the estimated Heston model of Example 2.4.4 using the SigMMD distance. Subsequently, we evaluate both the deep hedge and the robust hedging GAN using the NSDE generator and, for comparison, the original Heston generator. In both cases, we employ the same derivative, hedging objective, and parameter values as in Example 2.4.1 and Example 2.4.4. Trading is permitted only in the first dimension of the process, corresponding to the asset price, and, for simplicity, the available information for trading is likewise restricted to this component.[8] The penalty term, in contrast, is computed on both the asset price and the volatility components.

To examine the effect of the robust hedging GAN relative to the deep hedge, we plot the difference in the terminal payoffs of the two strategies against the terminal value of the asset price process—the sole determinant of the derivative's payoff—in Figure 2.9. The payoff differences are generated using paths sampled from the respective reference model.

The results reveal a clear transformation from the deep hedge strategy to the robust hedging GAN strategy when the Heston generator is employed. However, this transformation largely disappears when the NSDE generator is used. This behaviour is consistent with our interpretation of the two generator classes: a structurally constrained path generator such as the Heston model primarily addresses parameter uncertainty, whereas the high-parametric NSDE generator exhibits *quasi-distributional robustness*. The latter arises from the NSDE's universal approximation capability for diffusion processes (cf. Section 2.2.2), which mitigates the effects of parameter misspecification inherent in more restrictive model classes.

[8] This restriction is imposed solely to simplify the analysis and does not affect the general applicability of the approach.

2.5 Conclusion

In this chapter, we introduced the *Robust Hedging GAN*, a novel algorithm for addressing model uncertainty within the penalised worst-case (PWC) framework. By combining deep hedging, synthetic market data generation via flexible path generators, and path-space signature methods, our approach delivers a numerically tractable and model-agnostic approximation to optimal uncertainty-aware hedging strategies.

The integration of Generative Adversarial Network principles enables adversarial training between the hedger and the path generator, producing strategies that are resilient to a wide spectrum of potential model misspecifications. This adversarial set-up forces the strategy to adapt to worst-case perturbations of the reference model, thus improving robustness without being tied to a specific parametric form.

Our numerical experiments in Section 2.4 show that the Robust Hedging GAN can outperform classical deep hedging in various uncertainty scenarios, including Black–Scholes, Heston, and high-dimensional market models, as well as when driven by neural SDE generators. Across these settings, the method consistently reduced out-of-sample loss variance, and—depending on the uncertainty aversion—often improved mean performance. The model-agnostic nature of our approach allows direct application to a broad range of market models and derivative products.

Two key limitations remain. First, the penalty term α must be tailored to each task, with no universal form guaranteeing optimal performance across applications. This selection often requires careful calibration and problem-specific tuning. Second, although the PWC framework is global in theory, in practice the algorithm optimises against locally penalised adversarial models. In high-dimensional model spaces, exhaustive exploration is infeasible, so the training process may fail to identify certain adverse scenarios.

In the next chapter, we turn to model distribution methods, which remove the need for hand-crafted penalty functions and expand the search space for adversarial models. While this comes at a higher computational cost, it overcomes the two principal limitations identified here, and further enhances the robustness properties of uncertainty-aware hedging strategies.

Statement of Authorship

Title of Paper: "Robust Hedging GANs: Towards Automated Robustification of Hedging Strategies"

Publication Status: Published

Publication Details: Yannick Limmer and Blanka Horvath and. "Robust Hedging GANs: Towards Automated Robustification of Hedging Strategies". In: *Applied Mathematical Finance* 31.3 (2024), pp. 164–201. DOI: 10.1080/1350486X.2024.2440661. eprint: https://doi.org/10.1080/1350486X.2024.2440661. URL: https://doi.org/10.1080/1350486X.2024.2440661

Contribution to the paper: I was responsible for all aspects of this work, including the conceptualization, methodological design, formal analysis, implementation, and preparation of the manuscript. My supervisor, Blanka Horvath, provided academic guidance and constructive feedback throughout, and supported the refinement of the writing and presentation in her capacity as an academic advisor.

The hard part is estimating the mean vector; after all, the hedge fund industry employs armies of well-paid professionals whose main job is to come up with an ever improving estimate of the actual return of their tradable assets.

— Hans Buehler

3 Uncertainty Measures

Contents

3.1 Introduction 55
3.2 Robust strategies based on uncertainty measures 58
3.2.1 Plug-in strategy 58
3.2.2 The uncertainty-aware strategy 60
3.3 Comparing strategies in a Gaussian i.i.d. context 61
3.4 Subsampling: An Agnostic Choice of Model Distribution 66
3.5 Overcoming Memory Constraints: A CVaR-Stochastic Gradient Descent 70
3.5.1 Improved precision in high-dimensional investing 72
3.5.2 Deep Learning Example: Deep Hedging of Cliquet Options . 74
3.6 Conclusion 76

3.1 Introduction

Model uncertainty constitutes a fundamental challenge in quantitative finance: the true data-generating process is never observed, and even well-calibrated models are merely approximations. Decisions predicated upon a single estimated probability measure are therefore susceptible to estimation error, which may significantly degrade performance out of sample. Classical robust approaches—such as the (penalized) worst-case methodology—possess theoretical and practical appeal but are ultimately constrained by the

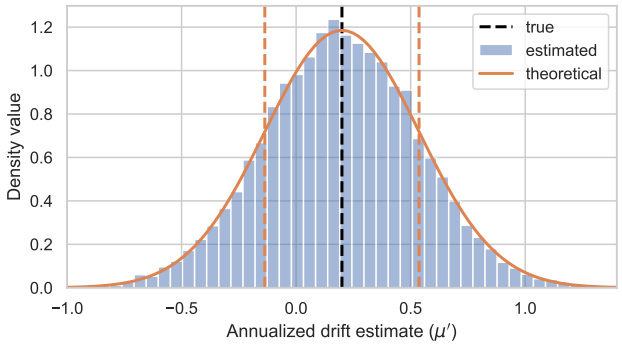


Figure 3.1: Sampling distribution of $\hat{\mu}'$ from $N = 90$ daily returns when the true (annualized) parameters are $\mu' = 20\%$ and $\sigma' = 20\%$. The substantial dispersion reflects the inherent difficulty of drift estimation with limited data.

dimensionality of the model space and the reliance on local search. This chapter advances an alternative, data-driven framework that treats model uncertainty in a probabilistic manner and evaluates the induced distribution of outcomes via a risk-sensitive functional. We refer to this outer functional as an *uncertainty measure* (UM). Its formulation permits a continuous adjustment of robustness through a single aversion parameter, thereby obviating the need for global worst-case assumptions.

Example 3.1.1. To anchor the discussion, consider the problem of estimating the annualized drift μ' of an asset with volatility σ'. Given N daily returns $\{\Delta S_t\}_{t=-N+1}^{0}$, the natural estimator is

$$\hat{\mu}' \;=\; \frac{1}{N\,\Delta t}\sum_{t=-N+1}^{0} \Delta S_t.$$

Suppose that the true parameters are $\mu' = 20\%$ and $\sigma' = 20\%$. With as few as $N = 90$ observations, the sampling error is sufficiently large that $\hat{\mu}'$ frequently fails to recover even the sign of the drift. Figure 3.1 illustrates the sampling distribution of $\hat{\mu}'$ in this setting, based on $10{,}000$ Monte Carlo replications. ◁

This stylised example encapsulates the broader difficulty: when a decision rule—be it a portfolio allocation, hedging strategy, or pricing scheme—is optimised under a single estimated model $\hat{\mathbb{P}}$, its model selection (in this case the drift estimate $\hat{\mu}'$) may be highly sensitive to statistical noise. A central aim of this chapter is to show how the uncertainty measure framework can mitigate this sensitivity. We will return to the

drift estimation example in later sections to illustrate, in a transparent setting, how the proposed methodology stabilises decisions and improves out-of-sample behaviour.

Conceptual basis of the framework. Let $\mathscr{P} \equiv \{\mathbb{P}_\theta : \theta \in \Theta\}$ denote the admissible model class, and let $\theta \sim \mathfrak{P}$ represent a distribution over plausible parameterisations. A traditional approach is to aggregate this distribution into a single *mixture* model $\bar{\mathbb{P}} = \int \mathbb{P}_\theta \, \mathfrak{P}(d\theta)$ and optimise with respect to $\bar{\mathbb{P}}$. Note that in Chapter 2, we used the mixture measure (under the name joint measure) as a benchmark for our numerical analysis of hedging strategies (c.f. Section 2.4). While computationally convenient, this aggregation can distort structural properties and renders the decision entirely dependent upon the accuracy of $\mathfrak{P}$.

The present framework departs from this aggregation step. Instead, each model $\mathbb{P}_\theta$ is evaluated separately, yielding a distribution of objective values $\{J(a; \mathbb{P}_\theta) : \theta \sim \mathfrak{P}\}$ for any admissible strategy a. An uncertainty measure U—for example, an entropic risk measure or the conditional value-at-risk (CVaR)—is then applied to this distribution:

$$\max_{a \in \mathcal{A}} \; U_{\theta \sim \mathfrak{P}}\Big(J(a; \mathbb{P}_\theta)\Big).$$

The choice of U governs the trade-off between expected performance and robustness: moderate aversion weights typical outcomes more heavily, whereas strong aversion emphasises the lower tail and converges towards worst-case optimisation.

Generation of model ambiguity. The construction of $\mathfrak{P}$ is itself a nontrivial task. A fully Bayesian specification may be available in some contexts, but in many practical applications it is computationally burdensome or infeasible. We therefore develop a model-agnostic ambiguity generator based on *subsampling*. Starting from an estimated model $\hat{\mathbb{P}}$ (or directly from empirical data), small subsamples are drawn repeatedly, and the objective is re-evaluated under each subsample. The resulting empirical variation induces a distribution $\mathfrak{P}$ without the need for explicit prior elicitation. This procedure is directly related to classical bootstrap techniques and, in the context of machine learning, to mini-batch sampling; it is compatible with both return-based and path-dependent control problems.

Computational considerations. A straightforward implementation of the uncertainty measure objective appears to require storing all simulated outcomes for all models, which is prohibitive in large-scale settings. To address this, we introduce a memory-efficient stochastic gradient procedure for CVaR optimisation. By processing one model instance at a time and retaining only tail-relevant samples, the method achieves substantial memory savings without compromising statistical efficiency. The independence of simulations across models renders the approach naturally amenable to parallel computation.

Outline. Section 3.2 formalises the uncertainty measure methodology and provides illustrative examples. Section 3.3 continues Section 3.2 and analyses the performance of uncertainty measures analytically. Section 3.4 details the subsampling-based construction of $\mathfrak{P}$ and compares it with Bayesian alternatives, showcasing numerical examples on ARCH/GARCH type models. Section 3.5 presents the memory-efficient CVaR-SGD algorithm and demonstrates high-dimensional and hedging examples.

3.2 Robust strategies based on uncertainty measures

In this section, we introduce *uncertainty measures* and demonstrate their application in addressing model uncertainty. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $\mathcal{A}$ the set of admissible actions. Consider an investment opportunity

$$X : \mathcal{A} \times \Omega \to \mathbb{R},$$

which, for an action $a \in \mathcal{A}$ and an outcome $\omega \in \Omega$, yields the payoff $X(a, \omega)$. In line with standard conventions, we suppress the dependence on ω and regard $X(a)$ as a random variable, i.e. $X(a) \in L^0(\Omega)$, the space of random variables on Ω. We write $X \sim \mathbb{P}$ to indicate that X is distributed according to the probability measure $\mathbb{P}$.

3.2.1 Plug-in strategy

We refer to the *plug-in* strategy (or policy)[1] as the choice of $a \in \mathcal{A}$ which maximises the agent's investment objective J under the assumption that $X(a)$ is distributed according

[1] We use the terms interchangeably: "strategy" is more common in the finance literature, whereas "policy" predominates in the machine learning literature.

to an estimated probability measure $\hat{\mathbb{P}} \in \mathscr{P}$, namely:

$$\arg\max_{a \in \mathcal{A}} J\Big(X(a), \hat{\mathbb{P}}\Big). \tag{3.1}$$

Here, "model" refers to the specification of the problem under a given candidate measure $\hat{\mathbb{P}}$, and we use these terms synonymously. The space $\mathscr{P}$ denotes the set of probability measures on $(\Omega, \mathscr{F})$, and the objective is a function $J : L^0(\Omega) \times \mathscr{P} \to \mathbb{R}$.

Example 3.2.1 (Investing with normally distributed returns). Let the initial stock price be $S_0 \in \mathbb{R}$, and consider an investment until time $t = 1$ with position size $a \in \mathbb{R} =: \mathcal{A}$. The payoff is given by the change in asset price $\Delta S_1 := S_1 - S_0$ (also referred to as return), scaled by the position, i.e. $X(a) := a\,\Delta S_1$. Suppose the agent's objective is to minimise the *entropic risk* $J := \mathrm{Entr}_\lambda$ with risk aversion $\lambda \in \mathbb{R}_+$ (see Equation (B.1)), and that she estimates $\Delta S_1 \sim \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$. The plug-in strategy is then

$$\arg\max_{a \in \mathbb{R}} \mathrm{Entr}_\lambda\Big(a\,\Delta S_1, \hat{\mathbb{P}}\Big) = \frac{\hat{\mu}}{\lambda\,\hat{\sigma}^2},$$

as shown in Lemma B.2.1. ◁

Example 3.2.2 (Hedging). The framework naturally extends to more complex settings such as hedging. Consider a stock price process $S = (S_t)_{t=0}^T$ and a European option with payoff C at maturity T. An admissible action is a predictable process $a = (a_t)_{t=1}^{T-1}$, where a_t is the number of units of stock held between $t-1$ and t. This results in the terminal payoff

$$X(a) := \sum_{t=1}^{T} a_t\,\Delta S_t - C.$$

If the agent assumes that S evolves under an estimated measure $\hat{\mathbb{P}}$ and seeks to minimise the variance of the terminal payoff, the problem becomes

$$\arg\max_{a \in \mathcal{A}} \left\{ -\mathbb{V}^{\hat{\mathbb{P}}}\Big[\sum_{t=1}^{T} a_t\,\Delta S_t - C\Big] \right\},$$

where $\mathbb{V}^{\hat{\mathbb{P}}}$ denotes variance under $\hat{\mathbb{P}}$. ◁

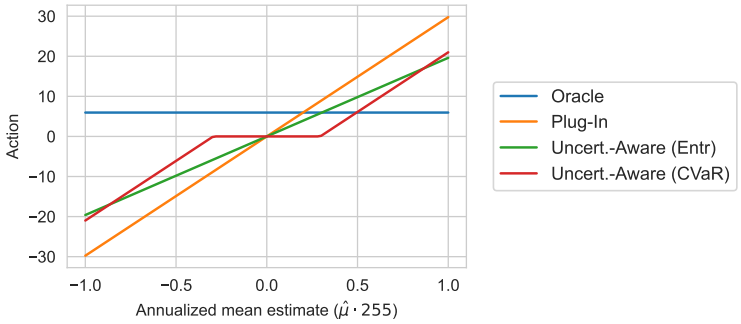


Figure 3.2: Strategies in Examples 3.2.1 and 3.2.3 based on the drift estimated from 90 days of returns. The plug-in strategy is linear in the drift estimate. The uncertainty-aware strategy using Entr remains linear but is less aggressive, as reflected in its smaller slope. Using CVaR introduces an "uncertainty region" where small drift estimates lead to no investment; outside this region, the slope is parallel to that of the plug-in strategy. The oracle strategy is independent of the drift estimate, as it uses the true distribution.

3.2.2 The uncertainty-aware strategy

The plug-in strategy is inherently vulnerable to model misspecification: the optimal decision changes if the model parameters are estimated differently, or if the assumed distributional form is altered. Moreover, there is no explicit accounting for the uncertainty in the estimate $\hat{\mathbb{P}}$ itself.

We now treat the model $\hat{\mathbb{P}}$ as a random variable. Let $(\Theta, \mathscr{F}_\Theta)$ be a measurable space, and let $\mathscr{P}_\Theta$ denote the set of probability measures on $(\Theta, \mathscr{F}_\Theta)$. We index candidate models by $\theta \in \Theta$, writing $\mathbb{P}_\theta$ for the corresponding probability measure. We use $L^0(\Theta)$ for the space of Θ-valued random variables. Our proposal is as follows:

(i) Fix a class of models $(\mathbb{P}_\theta)_{\theta\in\Theta} \subseteq \mathscr{P}$ indexed by the parameter space Θ;

(ii) Treat θ as a random variable in $L^0(\Theta)$ and choose a distribution $\mathfrak{P} \in \mathscr{P}_\Theta$ to encode model uncertainty, so that $\theta \sim \mathfrak{P}$;

(iii) Specify an *uncertainty measure* $U : L^0(\Theta) \times \mathscr{P}_\Theta \to \mathbb{R}$.

This extends (3.1) to the *uncertainty-aware* strategy:

$$\arg\max_{a\in\mathcal{A}} \left\{ U\Big(J(X(a), \mathbb{P}_\theta), \mathfrak{P}\Big)\right\}. \tag{3.2}$$

Example 3.2.3 (Uncertainty-aware investing with normally distributed returns). In the setting of Example 3.2.1, suppose the drift parameter is uncertain, with $\mu \sim \mathcal{N}(\hat{\mu}, \hat{\tau}^2)$ for

$\hat{\tau} \in \mathbb{R}_+$. With $U := \mathrm{Entr}_{\lambda'}$ for uncertainty aversion $\lambda' \in \mathbb{R}_+$ (see Equation (B.2)), the uncertainty-aware strategy becomes

$$\arg\max_{a \in \mathbb{R}} \mathrm{Entr}_{\lambda'}\Big(\mathrm{Entr}_{\lambda}\big(a\,\Delta S_1 \,\big|\, \Delta S_1 \sim \mathcal{N}(\mu, \hat{\sigma}^2)\big) \,\Big|\, \mu \sim \mathcal{N}(\hat{\mu}, \hat{\tau}^2)\Big), \tag{3.3}$$

with analytic representation given in Lemma 3.3.1.

Alternatively, taking $U := \mathrm{CVaR}_\alpha$ (see Equation (B.4)) for $\alpha \in [0, 1]$ yields

$$\arg\max_{a \in \mathbb{R}} \mathrm{CVaR}_{\alpha}\Big(\mathrm{Entr}_{\lambda}\big(a\,\Delta S_1 \,\big|\, \Delta S_1 \sim \mathcal{N}(\mu, \hat{\sigma}^2)\big) \,\Big|\, \mu \sim \mathcal{N}(\hat{\mu}, \hat{\tau}^2)\Big), \tag{3.4}$$

with analytic form in Lemma 3.3.2. Both strategies, together with the plug-in strategy, are illustrated in Figure 3.2. ◁

The transition from the plug-in to the uncertainty-aware framework represents a fundamental shift: rather than committing to a single estimated model, the agent evaluates decisions by incorporating the distribution over possible models. This raises the question of how to select or approximate these model probabilities in a principled way.

To compare the strategies from Examples 3.2.1 and 3.2.3, we will first focus on settings in which canonical statistical methods provide natural estimators for model probabilities, thereby enabling an objective evaluation of uncertainty-aware methods in an out-of-sample context. Finally, we will contrast these strategies with a simple *mixture-of-models* approach to highlight the distinct benefits of uncertainty-aware decision-making over naive model averaging.

We emphasise that the uncertainty measures $\mathrm{Entr}_{\lambda'}$ and CVaR_α are employed to manage *model uncertainty*, whereas the out-of-sample performance comparisons later in the thesis will be evaluated using the original objective Entr_λ.

3.3 Comparing strategies in a Gaussian i.i.d. context

In this section, we analyse, in a simple Gaussian i.i.d. setting, the out-of-sample performance of the strategies introduced in Examples 3.2.1 and 3.2.3. This basic environment not only clarifies why robustification is beneficial, but also illustrates how our approach improves on natural alternatives.

Setting. Let the current time be $t = 0$. We assume that $N \in \mathbb{N}$ historical increments of an asset price process $(S_t)_{t=-N,\dots,0}$, sampled at step size $\delta > 0$, have been observed and are used to estimate the distribution of the next increment ΔS_1. We impose that the increments are i.i.d., i.e.

$$\Delta S_t \overset{d}{=} \Delta S_0 \quad \text{for all } t \in \{-N+1, \dots, 0, 1\}.$$

The *true* distribution is Gaussian,

$$\Delta S_t \sim \mathcal{N}(\mu, \sigma^2) \quad \text{with drift } \mu \in \mathbb{R} \text{ and volatility } \sigma > 0 \text{ per step } \delta.^{2}$$

Throughout, the agent's objective is the entropic criterion Entr_λ with risk aversion $\lambda \in \mathbb{R}_+$ (see Equation (B.1)).

Oracle strategy. If the true parameters were known, the optimal (oracle) position is

$$a_{\text{oracle}} = \frac{\mu}{\lambda\sigma^2},$$

with achieved objective $\mathrm{Entr}_\lambda(a_{\text{oracle}}\,\Delta S_1) = \mu^2/(2\lambda\sigma^2)$ (see Lemma B.2.1).

Plug-in strategy. As μ and σ are unknown, we form estimates from the N past increments. Let

$$\hat{\mu} := \frac{1}{N} \sum_{t=-N+1}^{0} \Delta S_t, \qquad \hat{\sigma^2} := \frac{1}{N-1} \sum_{t=-N+1}^{0} \Delta S_t^2. \tag{3.5}$$

The plug-in position is then

$$a_{\text{plug-in}} := \frac{\hat{\mu}}{\lambda \hat{\sigma^2}} \qquad \text{(cf. Lemma B.2.1 with } \nu = \hat{\mu},\ \tau^2 = \hat{\sigma^2}).$$

These estimators are random. In particular,

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right), \qquad \hat{\sigma^2} \sim \frac{\sigma^2}{N-1}\,\chi^2_{N-1}, \tag{3.6}$$

where the second statement corresponds to the centred (unbiased) variance estimator; in what follows we use $\hat{\sigma^2}$ as a consistent plug-in for σ^2 and, for tractability, first

[2] It is often convenient to express annualized parameters via $\mu = \delta\,\mu'$ and $\sigma = \sqrt{\delta}\,\sigma'$.

consider the simplification $\hat{\sigma^2} \equiv \sigma^2$.[3] Then Lemma B.2.3 yields the out-of-sample performance (OOSP)

$$\mathrm{Entr}_\lambda(a_{\text{plug-in}}\,\Delta S_1) \;=\; \frac{\mu^2}{2\lambda\sigma^2} \;-\; \frac{\mu^2+\sigma^2}{2N\lambda\sigma^2}.$$

For the illustrative parameters $\lambda = 0.84$, $N = 140$, $\sigma = 0.2/\sqrt{255}$, and $\mu = 0.2/255$, the OOSP is *negative*, hence worse than the trivial (zero-investment) strategy.

Uncertainty-aware strategy. To construct the uncertainty-aware policies of Example 3.2.3, the agent needs a distribution over models. A canonical specification takes the point estimates in (3.5) and uses the sampling law in (3.6) to calibrate the dispersion of μ, while keeping σ fixed for analytic transparency. In the notation of Example 3.2.3, we set $\hat{\mu} = \hat{\mu}$ and $\tau^2 = \hat{\sigma^2}/N$, and compute (3.3) and (3.4) in closed form.

Lemma 3.3.1. *The entropic uncertainty-aware policy* (3.3) *admits the representation*

$$a_{\text{u-a}} \;=\; \frac{\hat{\mu}}{\lambda\hat{\sigma^2}\left(1+\frac{\lambda'}{\lambda N}\right)} \;=\; \frac{\lambda N}{\lambda N+\lambda'}\,a_{\text{plug-in}}, \tag{3.7}$$

and, under $\hat{\sigma^2} \equiv \sigma^2$, its out-of-sample performance is

$$\begin{aligned}\mathrm{Entr}_\lambda(a_{\text{u-a}}\,\Delta S_1) &\equiv \mathrm{MeanVar}^{\mathbb{P}}_\lambda(a_{\text{u-a}}\Delta S_1)\\ &= \frac{x_N}{\lambda\sigma^2}\left(\mu^2\left(1-\frac{x_N}{2}\right)-\frac{x_N(\mu^2+\sigma^2)}{2N}\right),\\ x_N &:= \frac{\lambda N}{\lambda N+\lambda'}.\end{aligned} \tag{3.8}$$

Proof. See Appendix B.2. □

Lemma 3.3.2. *For $U = \mathrm{CVaR}_\alpha$, the uncertainty-aware policy* (3.4) *is*

$$a_{\text{u-a}'} \;=\; \mathrm{sign}(\hat{\mu})\,\frac{\left(|\hat{\mu}|-A\sqrt{\hat{\sigma^2}/N}\right)_+}{\lambda\hat{\sigma^2}} \;=\; \left(a_{\text{plug-in}}-\mathrm{sign}(\hat{\mu})\frac{A\sqrt{\hat{\sigma^2}/N}}{\lambda\hat{\sigma^2}}\right)\mathbb{1}_{\{\hat{\mu}^2\geq A^2\,\frac{\hat{\sigma^2}}{N}\}},$$

where $A := \varphi(\Phi^{-1}(\alpha))/\alpha$, with φ and Φ the standard normal density and cdf. Under $\hat{\sigma^2} \equiv \sigma^2$, the OOSP is explicit: $\mathrm{Entr}_\lambda(a_{\text{u-a}'}\Delta S_1) = f(\mu,\sigma,\lambda,\alpha,N)$ for an analytically computable function f.

Proof. See Appendix B.2. □

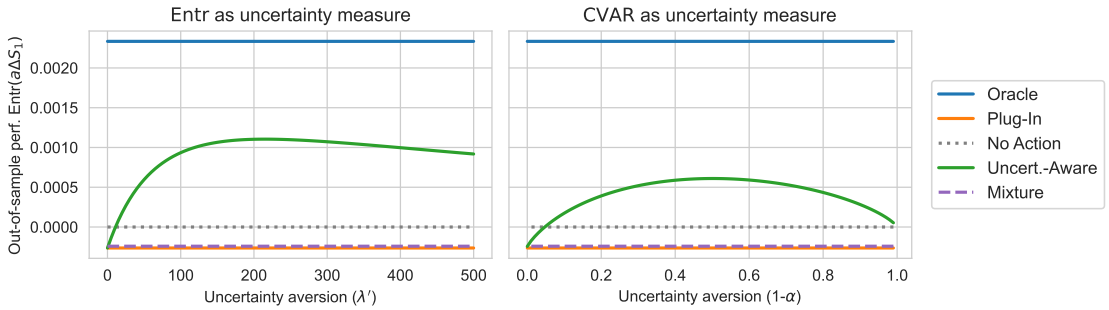


Figure 3.3: Out-of-sample performance of the oracle, plug-in, mixture (defined in Section 3.3), and uncertainty-aware strategies using Entr and CVaR as uncertainty measures. In this example, the plug-in and mixture strategies yield negative OOSP.

Figure 3.3 contrasts the strategies as a function of the robustification parameters λ' and α. The oracle (with perfect information) attains the highest performance. A natural benchmark is whether OOSP exceeds zero; the plug-in strategy fails this test due to estimation error. By contrast, both uncertainty-aware strategies deliver robustness for sufficiently strong robustification, interpolating continuously between the plug-in policy (as $\lambda' \downarrow 0$ or $\alpha \uparrow 1$) and the trivial strategy (as $\lambda' \uparrow \infty$ or $\alpha \downarrow 0$). This is formalised next.

Corollary 3.3.3. *For the policy* (3.3)*,*

$$\lim_{\lambda' \to 0} a_{\mathrm{u-a}} = a_{\text{plug-in}}, \qquad \lim_{\lambda' \to \infty} a_{\mathrm{u-a}} \equiv 0, \qquad \lim_{N \to \infty} a_{\mathrm{u-a}} = a_{\text{oracle}}.$$

Analogously, for (3.4)*,*

$$\lim_{\alpha \to 1} a_{\mathrm{u-a'}} = a_{\text{plug-in}}, \qquad \lim_{\alpha \to 0} a_{\mathrm{u-a'}} \equiv 0, \qquad \lim_{N \to \infty} a_{\mathrm{u-a'}} = a_{\text{oracle}}.$$

Proof. See Appendix B.2. □

Consistent patterns emerge in the optimal robustification (Figure 3.4): stronger signals (large $|\mu|$) call for milder robustification (smaller λ' or larger α); weak signals push the optimum towards extreme protection ($\lambda' \to \infty$ or $\alpha \to 0$), effectively reverting to the zero position; and higher volatility (noisier returns) demands stronger robustification.

[3] This isolates the dominant mean-estimation risk; see also the numerical study below, where variance uncertainty is reintroduced.

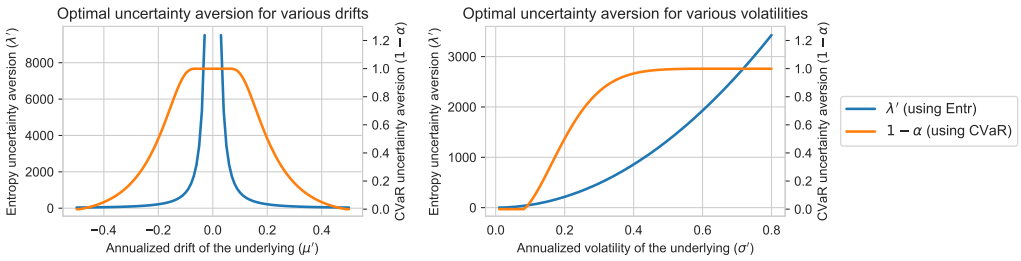


Figure 3.4: Optimal robustification (λ' for Entr and α for CVaR) that maximises OOSP in (3.3) and (3.4) across testing environments with annualized drift and volatility (μ', σ'). Parameters not varied are fixed at $\mu' = 0.2$ and $\sigma' = 0.2$. Markets with lower signal-to-noise (small μ' relative to σ') require stronger robustification.

Comparing to the mixture-of-measures approach. A natural alternative is to *average* models and optimise under the resulting mixture measure; see [83, Definition 6.1] and references therein. Formally, define

$$a_{\text{mix}} := \arg\max_{a \in \mathcal{A}} J\Big(X(a), \int_{\Theta} \mathbb{P}_\theta \, \mathfrak{P}(d\theta)\Big), \tag{3.9}$$

which, in the present setting (and by Lemma B.2.1), yields

$$a_{\text{mix}} = \arg\max_{a \in \mathbb{R}} \Big\{ \text{Entr}_\lambda(a\,\Delta S_1) \;\Big|\; \Delta S_1 \sim \mathcal{N}\big(\hat{\mu},\ \sigma^2 + \sigma^2/N\big) \Big\} \;=\; \frac{\hat{\mu}\, N}{\lambda \sigma^2 (N+1)}. \tag{3.10}$$

Its OOSP is (see Lemma B.2.4)

$$\text{Entr}_\lambda(a_{\text{mix}}\,\Delta S_1) = \frac{\mu^2\, N}{\lambda \sigma^2 (N+1)} - \frac{N(\mu^2 + \sigma^2)}{2\lambda \sigma^2 (N+1)^2}.$$

As illustrated in Figure 3.3, the mixture approach affords some robustness but is insufficient in this setting: it averages over models without explicitly penalizing misspecification or tail risk.

Variance-adjusted plug-in viewpoint. Note that (3.10) coincides with the plug-in position computed under the *inflated* variance $\sigma^2 + \sigma^2/N$. More generally, one may consider the family

$$a_{\text{var-adj}} := \arg\max_{a \in \mathbb{R}} \Big\{ \text{Entr}_\lambda(a\,\Delta S_1) \;\Big|\; \Delta S_1 \sim \mathcal{N}(\hat{\mu},\ \sigma^2 + \tau^2) \Big\},$$

with an ad hoc variance adjustment $\tau^2 \geq 0$. Figure 3.5 plots performance as τ^2 varies: achieving meaningful robustness requires relatively large adjustments, and the optimal τ^2

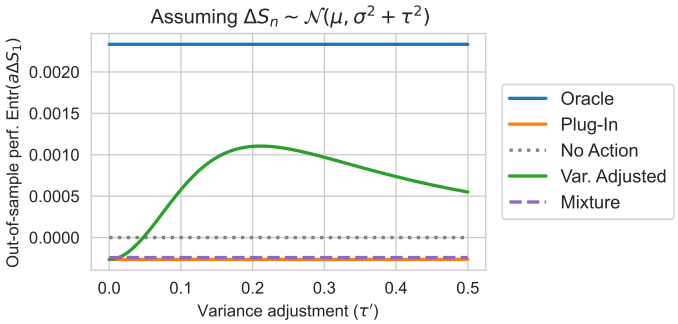


Figure 3.5: Performance of the plug-in strategy with a custom variance adjustment τ^2 in the setting of Example 3.2.1.

bears no simple relation to the actual estimation uncertainty of $\hat{\mu}$. Systematic variance upscaling for risk estimation is studied, e.g., in [238, 239].

Remark 3.3.1 (Relation to the uncertainty-aware strategy). If J is linear in its second argument (the measure) and the uncertainty measure is the *expectation*, then the uncertainty-aware optimisation coincides with the mixture approach:

$$J\Big(X(a), \int_\Theta \mathbb{P}_\theta \mathfrak{P}(d\theta)\Big) = \int_\Theta J\big(X(a), \mathbb{P}_\theta\big)\mathfrak{P}(d\theta).$$

For nonlinear objectives in the measure—such as entropic or mean-variance criteria—this identity fails, and mixture policies generally differ from uncertainty-aware policies. ◁

3.4 Subsampling: An Agnostic Choice of Model Distribution

In many practical applications, obtaining a precise and reliable distribution for model parameters is challenging. In such cases, it is advantageous to employ methods that yield reasonable approximations of the model distribution with minimal modelling assumptions. We propose the use of *subsampling* for this purpose. This method is straightforward to implement, computationally efficient, requires minimal hyperparameter tuning (in contrast to many Bayesian or filtering approaches), and integrates naturally into existing modelling pipelines. Although subsampling is limited in generality, it is often sufficient in practice: when combined with uncertainty measures, the precision of

the model distribution becomes less critical than for the mixture approach, and substantial performance gains can still be realised even with this relatively ad hoc construction.

We now formalise subsampling in the present setting. In many cases, closed-form solutions to the investment problems (3.1) are unavailable and must be approximated via Monte Carlo simulation. Recall that $X(a)$ was previously defined in general as a payoff function depending on the action a and the outcome $\omega \in \Omega$. In most financial applications, however, the asset price process evolves independently of the trading action, so that

$$X(a) = f(a, S), \tag{3.11}$$

where S denotes the relevant underlying process. In what follows we adopt this viewpoint and regard $\mathbb{P}$ as the law of S itself.

An n-*subsample* $(\omega_i)_{i=1}^n \subseteq \Omega$, $n \in \mathbb{N}$, is obtained by drawing n i.i.d. samples from the estimated distribution $\hat{\mathbb{P}}$, resulting in the empirical values $(S(\omega_i))_{i=1}^n$. The associated *subsample measure* assigns equal weight to each point:

$$\mathbb{P}_\theta := \mathcal{U}\Big((\omega_i)_{i=1}^n\Big) = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}, \tag{3.12}$$

where δ_ω is the Dirac measure at ω and $\mathcal{U}$ denotes the uniform distribution over the subsample $\theta := (\omega_i)_{i=1}^n$. This fits naturally into our uncertainty-aware framework by taking $\Theta := \Omega^n$ and $\mathfrak{P} := \hat{\mathbb{P}}^n$.[4]

Connection to the Gaussian i.i.d. example. In the setting of Section 3.3, the above subsampling construction yields, in the large-sample limit, the same strategies as those in Lemmata 3.3.1 and 3.3.2. Indeed, suppose $\hat{\mathbb{P}}$ is the Gaussian estimate $\Delta S_1 \sim \mathcal{N}(\hat{\mu}, \hat{\sigma^2}) \approx \mathcal{N}(\hat{\mu}, \sigma^2)$. Applying (3.12) to $\hat{\mathbb{P}}$, we have $(\Delta S_{1,i})_{i=1}^N \sim \mathcal{N}(\hat{\mu}, \sigma^2)^N$, with sample mean

$$\mathbb{E}_{\hat{\mathbb{P}}}\left[\frac{1}{N}\sum_{i=1}^N \Delta S_{1,i}\right] = \hat{\mu}, \qquad \mathbb{V}_{\hat{\mathbb{P}}}\left[\frac{1}{N}\sum_{i=1}^N \Delta S_{1,i}\right] = \frac{\sigma^2}{N}.$$

Approximating Entr_λ by $\mathrm{MeanVar}_\lambda$ for Gaussian arguments ((B.3)), the subsampling method recovers the theoretical out-of-sample performances derived analytically in Lemmata 3.3.1 and 3.3.2.

[4] This formulation also covers general payoffs X not of the form (3.11), although we do not pursue such abstract settings here.

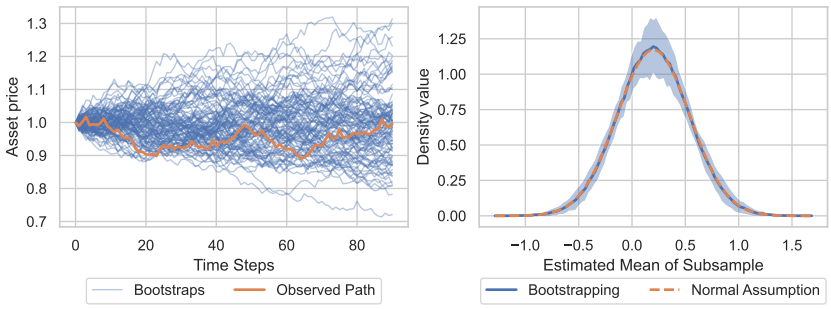


Figure 3.6: Left: bootstrapping a single path by resampling increments (with replacement) to form synthetic paths. Right: distribution of mean estimates from bootstrapping versus those from Gaussian subsampling $\mathcal{N}(\hat{\mu}, \hat{\sigma^2}/N)$. The solid line denotes the average; shading shows one standard deviation.

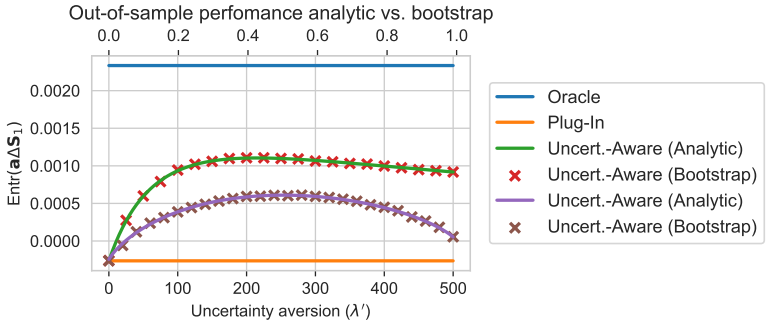


Figure 3.7: Out-of-sample performance of analytic strategies (Lemmata 3.3.1 and 3.3.2) versus bootstrapping increments (equivalent to i.i.d. subsampling under the empirical measure). The close match confirms that the normal approximation in Figure 3.6 is sufficiently accurate in practice.

Bootstrapping vs. subsampling. In the Gaussian example, subsampling draws from the fitted distribution $\hat{\mathbb{P}}$, whereas in bootstrapping one draws directly (with replacement) from the empirical measure of the observed returns. For one-step horizons, this is equivalent to taking $\hat{\mathbb{P}}$ as the empirical distribution $\mathcal{U}\left((\Delta S_t)_{t=-N+1}^{0}\right)$. Bootstrapping is simpler when returns are plausibly i.i.d., but subsampling is more flexible for data with time dependencies, since $\hat{\mathbb{P}}$ can encode richer dynamics.

Practitioners often bootstrap multi-step paths by concatenating resampled historical returns (Figure 3.6, left). This exactly matches our n-subsample definition when $\hat{\mathbb{P}}$ is empirical. The resulting mean estimates are close to those from Gaussian subsampling (Figure 3.6, right), and the corresponding strategy performances differ negligibly (Figure 3.7).

Remark 3.4.1 (Considerations for the estimated measure $\hat{\mathbb{P}}$). If $\hat{\mathbb{P}}$ is taken as the empirical measure, as in naive bootstrapping, the resamples are confined to the N observations.

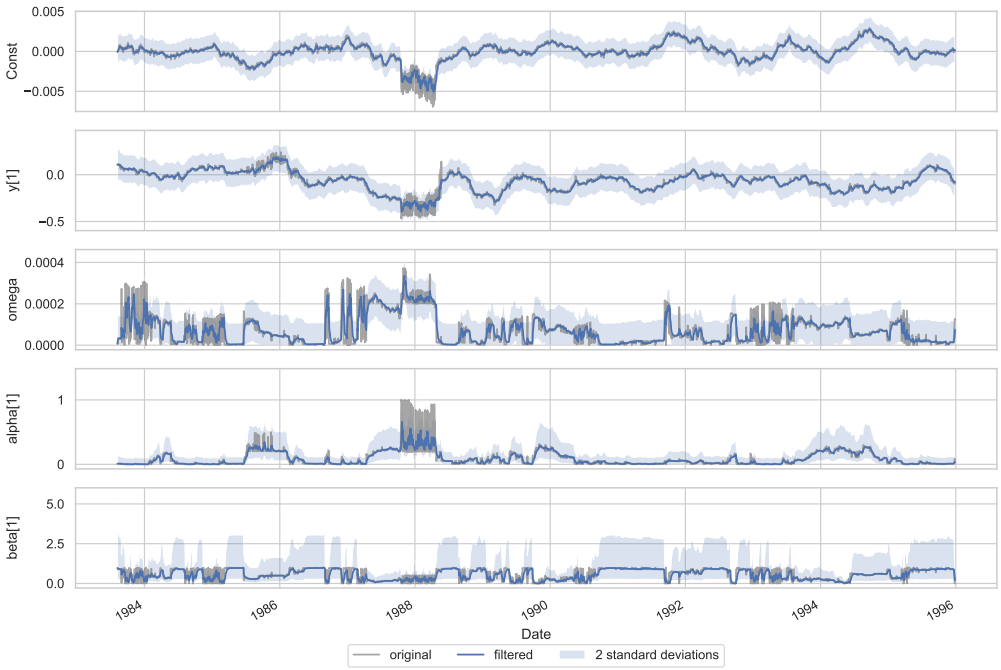


Figure 3.8: ARCH(1)-GARCH(1,1) parameters for the PEP-KO spread estimated via Kalman filtering, showing filtered means and uncertainty bands.

This can underrepresent tail risk and yield unreliable CVaR estimates. Smoothing the centre of the distribution [282] and thickening the tails using EVT techniques [79, 213, 238] can mitigate this, though we leave such refinements for future work. ◁

The effectiveness of any resampling method depends critically on the quality of $\hat{\mathbb{P}}$. Since bootstrapping assumes i.i.d. returns, it erases path-dependent features such as volatility clustering. By contrast, our subsampling procedure can be applied to a fitted dynamic model $\hat{\mathbb{P}}$, preserving such dependencies before applying robustification.

Figure 3.9 shows that "Subsampling + CVaR" substantially improves performance over the plug-in, achieving positive utility for moderate α and outperforming the trivial (zero-investment) benchmark. For comparison, we apply CVaR_α to a model distribution obtained by Kalman filtering [71] the daily AR-GARCH parameter estimates (log-transformed where necessary to maintain constraints). This yields similar performance to subsampling, with the Bayesian filter slightly ahead for some α values. The mixture method applied to the filtered distribution improves over the plug-in but fails to beat the trivial strategy.

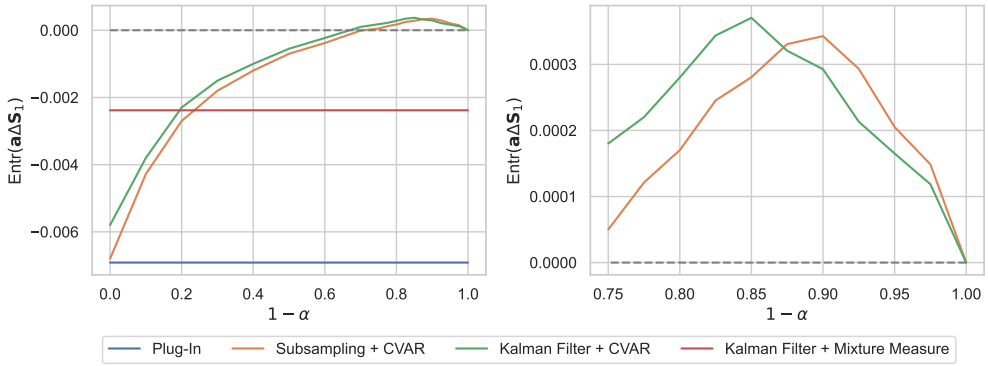


Figure 3.9: OOSP for the PEP-KO spread. "Subsampling + CVaR" uses $n = 30$ and CVaR_α; "Kalman Filter + CVaR" applies CVaR_α to the filtered distribution of AR(1)-GARCH(1,1) models; "Kalman Filter + Mixture" uses (3.9) on the same distribution. Kalman filtering slightly outperforms subsampling with CVaR, but the gain is marginal relative to the added complexity.

These results suggest that, while filtering-based methods can offer incremental gains, the simplicity and competitive performance of subsampling make it an attractive, low-overhead alternative for constructing $\mathfrak{P}$ in uncertainty-aware decision frameworks.

3.5 Overcoming Memory Constraints: A CVaR-Stochastic Gradient Descent

A central computational challenge in applying the uncertainty-aware strategies developed in this chapter is the substantial memory requirement incurred when simulating paths under a distribution of models. Sampling multiple models from a distribution $\mathfrak{P}$ and simulating returns or price paths for each of them leads to a quadratic scaling of computational cost. Although modern GPUs can alleviate the processing time through parallelisation, their memory capacity remains a limiting factor, particularly for large-scale or high-dimensional problems.

This constraint becomes especially acute when moving from single-period return predictions to tasks requiring the simulation of full price paths. Such problems involve larger state spaces and repeated evaluations across numerous model variants, rapidly exhausting available GPU memory.

To address this, we propose a memory-efficient variant of stochastic gradient descent (SGD) tailored to uncertainty-aware optimisation with conditional value-at-risk (CVaR).

The key idea is to simulate paths under a single model at a time rather than in bulk. This design drastically reduces the memory footprint while preserving the ability to capture model risk. Furthermore, the independence of the simulations enables straightforward distribution across multiple compute nodes, making the method well-suited to large-scale deep learning pipelines.

Before presenting the method, recall that standard mini-batch SGD applied to the plug-in objective $J(X(a), \mathbb{P}_\theta)$, with $\mathbb{P}_\theta$ the empirical distribution over a mini-batch $\theta := (\omega_i)_{i=1}^n \sim \hat{\mathbb{P}}^n$, converges—under standard assumptions [261, Theorem 14.8]—to

$$\mathbb{E}_{\theta \sim \hat{\mathbb{P}}^n} \left[J(X(a), \mathbb{P}_\theta) \right].$$

If J is unbiased under empirical evaluation, mini-batch SGD thus optimises the plug-in objective under $\hat{\mathbb{P}}$, implicitly incorporating model uncertainty with the *expectation* as the uncertainty measure.

To incorporate model risk explicitly, we replace the expectation with the CVaR operator, yielding the CVaR-SGD method in Algorithm 1. This algorithm prioritises adverse scenarios by discarding the top α-fraction of outcomes at each iteration.

The algorithm only stores $(1-\alpha) \cdot m$ gradients in memory at any time, reducing requirements proportionally to $(1-\alpha)$. This is particularly valuable for high-dimensional strategy classes $\mathcal{A}$. The computations for different $\mathbb{P}_i$ are independent and may be parallelised or distributed, maintaining efficiency despite reduced memory usage.

From a theoretical perspective, the algorithm produces, in expectation, an unbiased estimator of the gradient of the CVaR-objective by discarding gradients corresponding to the best-performing α-fraction of objectives. This yields the following guarantee:

Theorem 3.5.1. *[CVaR-Stochastic Gradient Descent] Let $\Phi = \{\varphi \in \mathbb{R}^d : \|\varphi\| \leq B\}$, and suppose $\varphi \mapsto J(X(a_\varphi), \mathbb{P})$ is convex and ρ-Lipschitz for all $\mathbb{P} \sim \mathfrak{P}$, i.e.*

$$\|\nabla_\varphi J(X(a_\varphi), \mathbb{P})\| \leq \rho \quad \mathfrak{P}\text{-a.s.}.$$

For $\alpha \in (0,1)$, define

$$\mathcal{L}(\varphi) := \mathrm{CVaR}_\alpha \left(J(X(a_\varphi), \mathbb{P}) \,|\, \mathbb{P} \sim \mathfrak{P} \right).$$

Algorithm 1: CVaR Stochastic Gradient Descent

Input: Initial parameters $\varphi^{(0)}$, learning rate η, batch size m, CVaR level α
Output: Optimised parameters φ^*
$t \leftarrow 0$
while *not converged* **do**
  Initialise sets $\mathcal{J} \leftarrow \varnothing, \mathcal{G} \leftarrow \varnothing$
  Sample integer k with $\mathbb{E}[k/m] = 1 - \alpha$
  for $i = 1$ **to** m **do**
    Sample $\mathbb{P}_i \sim \mathfrak{P}$
    Compute $J \leftarrow J(X(a^{(t)}), \mathbb{P}_i)$
    if $|\mathcal{J}| < k$ **then**
      Insert J into $\mathcal{J}$; compute $g \leftarrow \nabla_\varphi J$ and insert into $\mathcal{G}$
    end
    if $|\mathcal{J}| = k$ ***and*** $J < \max(\mathcal{J})$ **then**
      Replace $\max(\mathcal{J})$ by J and update corresponding gradient in $\mathcal{G}$
    end
  end
  Update:

$$\varphi^{(t+1)} \leftarrow \varphi^{(t)} - \eta \frac{1}{k} \sum_{g \in \mathcal{G}} g$$

  $t \leftarrow t + 1$
end

Let $(\varphi_t)_{t \geq 0}$ be the sequence generated by Algorithm 1 with step size $\eta_t = B/(\rho\sqrt{t})$, and set $\bar{\varphi}_t := t^{-1} \sum_{s=1}^{t} \varphi_s$. Then, for all $t \geq 1$,

$$\mathbb{E}[\mathcal{L}(\bar{\varphi}_t)] - \min_{\varphi \in \Phi} \mathcal{L}(\varphi) \leq \frac{B\rho}{\sqrt{t}}.$$

3.5.1 Improved precision in high-dimensional investing

We illustrate the reduction in memory requirements, and the associated gains in precision, in a controlled high-dimensional extension of Example 3.2.1. Let $d \in \mathbb{N}$ and

$$\Delta S_t \sim \mathcal{N}_d(\mu, \Sigma), \qquad t = -N + 1, \ldots, 0,$$

with unknown $\mu \in \mathbb{R}^d$ and positive-definite $\Sigma \in \mathbb{R}^{d \times d}$. The sample mean

$$\hat{\mu} = \frac{1}{N} \sum_{t=-N+1}^{0} \Delta S_t$$

satisfies $\hat{\mu} \sim \mathcal{N}_d(\mu, \Sigma/N)$. As in Section 3.3, we neglect the estimation error in Σ.

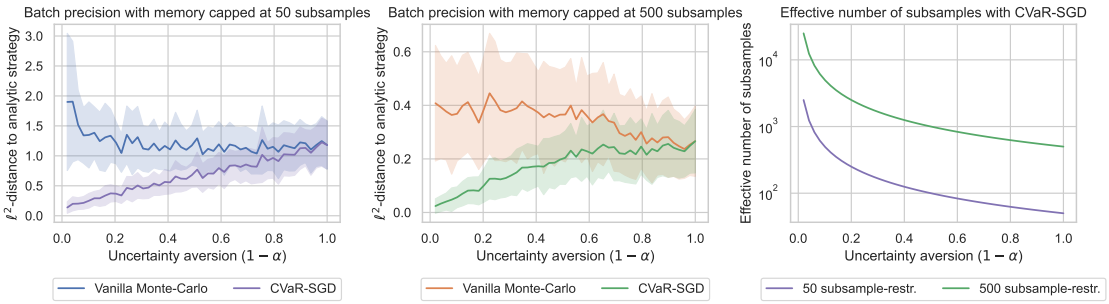


Figure 3.10: Approximation of the analytic strategy in Lemma 3.5.2, using $\hat{\mu}$ and Σ from the last 30 daily returns of the $d = 503$ S&P 500 components. Left/middle: standard GD with $m = 50$ and $m = 500$ subsamples; right: CVaR-SGD at fixed memory cost, enabling an "effective" $m/(1-\alpha)$ subsamples.

Actions are now $a \in \mathbb{R}^d$. With entropic risk objective and risk aversion $\lambda > 0$, the plug-in problem reduces to

$$J(X(a), \hat{\mathbb{P}}) = a^\top \hat{\mu} - \frac{\lambda}{2} a^\top \Sigma a,$$

which is maximised by $a_{\text{plug-in}} = \Sigma^{-1}\hat{\mu}/\lambda$.

Analytic robustification. Using CVaR_α as the uncertainty measure leads to the following closed form:

Lemma 3.5.2. *The* CVaR*-uncertainty-aware strategy*

$$\arg\max_{a \in \mathbb{R}^d} \mathrm{CVaR}_\alpha\Big(\mathrm{Entr}_\lambda(a^\top \Delta S_1) \,\Big|\, \mu \sim \mathcal{N}(\hat{\mu}, \Sigma/N)\Big)$$

is

$$a_{\mathrm{u-a'}} = \frac{1}{\lambda}\left(1 - \frac{\kappa_\alpha}{\sqrt{N}\,\|\Sigma^{-1/2}\hat{\mu}\|}\right)_+ \Sigma^{-1}\hat{\mu},$$

where $\kappa_\alpha = \varphi(\Phi^{-1}(\alpha))/\alpha$. *The zero-investment region* $a_{\mathrm{u-a'}} = 0$ *occurs whenever* $\|\Sigma^{-1/2}\hat{\mu}\| \le \kappa_\alpha/\sqrt{N}$.

Approximation experiment. We extract daily log-returns for $d = 503$ S&P 500 stocks over the 30 trading days preceding 20 May 2025, computing $\hat{\mu}$ and Σ. The analytic CVaR strategy of Lemma 3.5.2 applied to these estimates serves as a benchmark across $\alpha \in [0, 1]$.

Reversion speed	Reversion level	Volatility of volatility	Correlation	Drift	Initial volatility
1.0	0.03	0.2	-0.8	0	0.03

Table 3.1: Plug-in Heston parameters for Section 3.5.2.

We approximate the benchmark via sub-sampling: standard GD with $m = 50$ and $m = 500$ subsamples of size $n = 30$, and CVaR-SGD with $m = 50$ but "effective" $m/(1-\alpha)$ subsamples at the same memory usage. The error metric is the average Euclidean distance to the analytic solution.

Figure 3.10 shows that precision improves markedly with larger m in standard GD, but deteriorates for large α because CVaR discards an α-fraction of simulations. CVaR-SGD reverses this trend by effectively increasing the number of usable subsamples, maintaining high precision even for large α, all while respecting the original memory constraint.

3.5.2 Deep Learning Example: Deep Hedging of Cliquet Options

In this final numerical illustration, we demonstrate that subsampling can be effectively employed to enhance robustness in path-dependent deep learning applications. Specifically, we consider a deep hedging task for a cliquet option over a 120-day trading horizon, with a reset every 40 days. The cliquet payoff is given by

$$C_{\text{cliquet}} = \sum_{i=1}^{3} \max\left(S_{40i} - S_{40(i-1)},\, 0\right),$$

where S_{40i} denotes the underlying asset price at the i-th reset date. The hedger's objective is to minimise the standard deviation of the profit-and-loss (P&L) distribution.

As the plug-in model $\hat{\mathbb{P}}$, we take a Heston model calibrated to the parameters reported in Table 3.1. In our setup, the hedging strategy has access only to the current spot price and the most recent reset price (serving as the prevailing strike).

Vanilla deep hedge. We first train a *vanilla* deep hedge on the plug-in Heston model, following the deep hedging framework of [50]. The hedging policy is parameterised as a neural network

$$\pi_\varphi : \text{ information at time } t \;\mapsto\; \text{asset position at time } t,$$

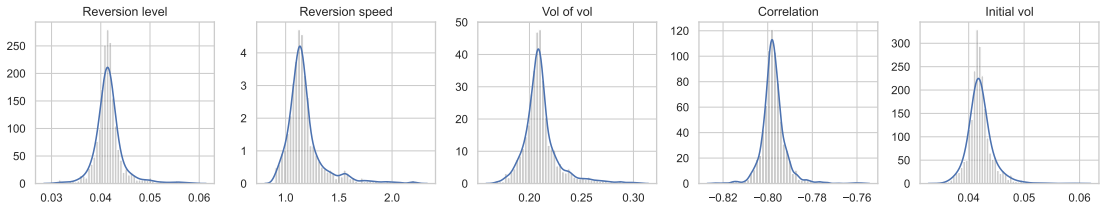


Figure 3.11: Test distribution $\tilde{\mathfrak{P}}$ of Heston parameters, calibrated from empirical one-day changes in S&P500 option calibrations. The centre corresponds to the plug-in parameters in Table 3.1.

where the information process $(h_t)_{t=0}^{120} \subset \mathbb{R}^2$ consists of the current spot price and the last reset price. With Φ denoting the set of all network parameters, the policy class is $\mathcal{A} = \{\pi_\varphi : \varphi \in \Phi\}$. See Section 2.2.3 for more details.

Training proceeds by simulating a large number of spot paths $S = (S_t)_{t=0}^{120}$ under $\hat{\mathbb{P}}$, and minimising the standard deviation of the hedging P&L:

$$\min_{\varphi\in\Phi} \operatorname{Std}_{S\sim\hat{\mathbb{P}}}\left[-C_{\text{cliquet}} + \sum_{t=1}^{T} \pi_\varphi(h_t)\,\Delta S_t\right].$$

Uncertainty-aware deep hedge via subsampling. We then construct uncertainty-aware strategies by replacing the expectation in the objective with a CVaR measure, evaluated via subsampling. Given m subsampled paths $(S^{(i)})_{i=1}^m \sim \hat{\mathbb{P}}^m$, we train

$$\min_{\varphi\in\Phi} \mathrm{CVaR}_\alpha\left[\widehat{\mathrm{Std}}\left(\left[-C_{\text{cliquet}} + \sum_{t=1}^{T} \pi_\varphi(h_t)\,\Delta S_t^{(i)}\right]_{i=1}^{m}\right)\right],$$

where $\widehat{\mathrm{Std}}$ denotes the empirical standard deviation on the subsample. We train such strategies for various uncertainty aversion levels $\alpha \in [0, 1]$ and subsample sizes $m \in \{16, 32, 64\}$.

Experimental setup. To evaluate out-of-sample performance, we test all strategies against a distribution $\tilde{\mathfrak{P}}$ of Heston models representing plausible alternative market dynamics. This distribution is constructed to match the empirical one-day variations in Heston parameters when calibrating to the S&P500 index option chain, and is illustrated in Figure 3.11.

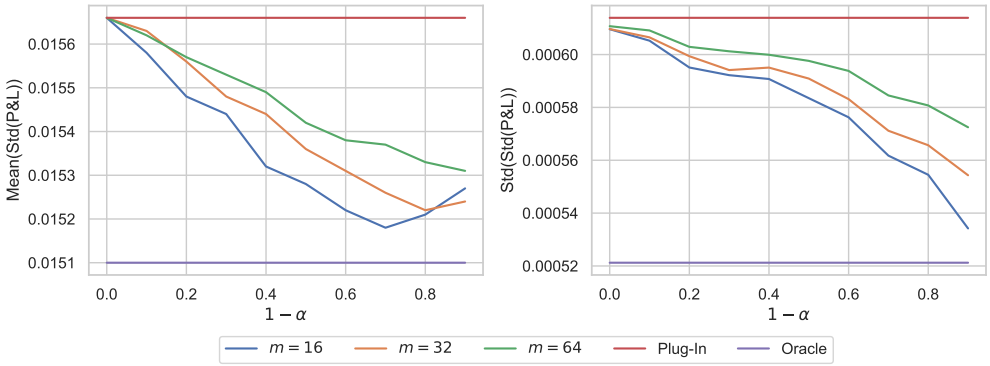


Figure 3.12: Performance across the test distribution $\tilde{\mathfrak{P}}$. Left: mean objective (lower is better). Right: standard deviation of the objective (lower is better). We compare vanilla deep hedge, uncertainty-aware hedges for $m \in \{16, 32, 64\}$, and an oracle hedge trained directly on $\tilde{\mathfrak{P}}$.

Results. The results, shown in Figure 3.12, compare the vanilla deep hedge, its uncertainty-aware variants, and an "oracle" hedge. Uncertainty-aware deep hedges consistently outperform the vanilla hedge across the test distribution, improving both the mean and dispersion of the objective. The strongest gains are observed for small subsample sizes ($m = 16$), indicating that a substantial degree of robustification is most effective. For $m = 32$ and $m = 64$, performance improves markedly as α decreases, suggesting that high uncertainty aversion is most beneficial in these cases.

Oracle benchmark. The oracle hedge is trained directly on the test distribution:

$$\min_{\varphi \in \Phi} \mathbb{E}_{\mathbb{P} \sim \tilde{\mathfrak{P}}} \left[\mathrm{Std}_{S \sim \mathbb{P}} \left[-C_{\text{cliquet}} + \sum_{t=1}^{T} \pi_\varphi(h_t)\, \Delta S_t \right] \right].$$

While this naturally outperforms all other methods—since it uses information unavailable to the agent—the comparison highlights that well-calibrated uncertainty-aware hedges can deliver comparable performance gains without access to this knowledge.

3.6 Conclusion

In this chapter, we have developed and analysed a unified framework for incorporating model uncertainty into decision-making, with a particular focus on quantitative investment and hedging problems. Starting from the classical plug-in approach, we established that estimation error can critically degrade out-of-sample performance, to the extent that naïve

strategies may underperform even the trivial no-investment benchmark. Using a stylised Gaussian i.i.d. setting, we quantified these effects analytically and demonstrated that robustification through explicit uncertainty measures, such as Entr and CVaR, can yield substantial improvements. In this setting, we derived closed-form expressions for the resulting uncertainty-aware policies, allowing a transparent comparison with the plug-in strategy and the mixture-of-measures approach. The analysis revealed that while model averaging offers some resilience, it generally fails to achieve the level of protection afforded by well-calibrated uncertainty measures.

We then introduced *subsampling* as a simple yet effective mechanism for constructing a distribution over candidate models in the absence of reliable prior information. We showed that, in the Gaussian i.i.d. example, subsampling converges to the same theoretical strategies as obtained from the analytical uncertainty-aware formulations, and that its performance closely matches that of the more elaborate bootstrap procedure. Furthermore, we demonstrated that subsampling naturally extends to settings with temporal dependence, and that it can be used in combination with richer model classes—such as autoregressive GARCH dynamics—without loss of practical tractability. An empirical study on historical pairs-trading data illustrated that subsampling can deliver performance gains comparable to those of Bayesian filtering approaches, while being significantly easier to implement.

To address the computational bottlenecks inherent in simulating across large model distributions, particularly in high-dimensional or path-dependent problems, we proposed a memory-efficient stochastic gradient descent method tailored to CVaR objectives. The CVaR-SGD algorithm operates by processing models sequentially and discarding gradients from the best-performing scenarios, yielding an unbiased gradient estimator for the robust objective while reducing memory requirements to a fraction of the naïve approach. We established convergence guarantees under standard convexity and Lipschitz conditions, and demonstrated, in a high-dimensional Gaussian setting, that this approach substantially improves approximation precision for large uncertainty-aversion parameters without increasing memory consumption.

Finally, we applied the proposed methodology to a deep hedging problem for cliquet options under model uncertainty in the underlying Heston dynamics. In this complex,

path-dependent setting, we showed that subsampling-based uncertainty-aware training produces hedging strategies with markedly improved out-of-sample performance, both in terms of mean and variability of the risk measure, when evaluated over a realistic distribution of possible model parameters. The gains were shown to be competitive with those of an oracle strategy that has access to the true model distribution, thereby underscoring the practical relevance of the proposed robustification approach.

Taken together, the results of this chapter provide both theoretical insight and practical tools for integrating model uncertainty into modern data-driven decision frameworks. The analytical derivations clarify the mechanisms by which uncertainty measures improve performance, while the subsampling and CVaR-SGD techniques enable these methods to be deployed in realistic, computationally demanding settings. The combination of tractable uncertainty modelling, computational efficiency, and demonstrable performance gains makes the proposed framework a versatile addition to the toolbox of quantitative risk management and robust optimisation.

Statement of Authorship

Title of Paper: *Uncertainty-Aware Strategies: A Model-Agnostic Framework for Robust Financial Optimization through Subsampling*

Publication Status: Unpublished and unsubmitted work written in a manuscript style

Publication Details: Hans Buehler, Blanka Horvath, Yannick Limmer, and Thorsten Schmidt. *Uncertainty-Aware Strategies: A Model-Agnostic Framework for Robust Financial Optimization through Subsampling*. 2025. arXiv: 2506.07299 [q-fin.CP]. URL: https://arxiv.org/abs/2506.07299

Contribution to the paper: I am the lead author of this work and was responsible for the conceptual development, methodological design, formal analysis, implementation, and writing of the manuscript. Hans Buehler contributed to the conceptualization, particularly in the design of the analytic experiments. Thorsten Schmidt and my supervisor, Blanka Horvath, provided valuable academic guidance, critical feedback and supported writing and editing.

The true mystery of the world is the visible, not the invisible.

— Oscar Wilde

4 Deep Filters Can Filter

Contents

4.1 Introduction 81
4.1.1 The Filterformer Architecture 83
4.1.2 Main Contributions 84
4.1.3 Outline and Notation 85
4.2 The Setting 86
4.2.1 Regularity Conditions on the Partially-Observed System . . . 87
4.2.2 Source Spaces: Regular Compact Subsets of Path Space . . . 88
4.2.3 Target Space: The 2-Wasserstein Space of Probability Measures 91
4.3 The Filterformer: An Attention-Based Deep Filter 92
4.3.1 Phase 1 — Encoding via Pathwise Attention 92
4.3.2 Phase 2 — Transformation via Multi-Layer Perceptrons . . . 94
4.3.3 Phase 3 — Decoding via Geometric Attention 94
4.3.4 The Filterformer Model 95
4.4 Main Result 95
4.4.1 Proof Overview 96
4.5 Conclusion 98

4.1 Introduction

Filtering problems arise whenever one seeks to estimate the latent dynamics of an unobservable process from a continuous stream of noisy measurements. Formally, we consider a latent *signal* process $X_\cdot$, taking values in $\mathbb{R}^{d_X}$ with $d_X \in \mathbb{N}$, and a noisy

observation process $Y_\cdot$, taking values in $\mathbb{R}^{d_Y}$ with $d_Y \in \mathbb{N}$. Both evolve on a filtered probability space $(\Omega, \mathcal{F}, \mathfrak{F}, \mathbb{P})$, where $\mathfrak{F} = (\mathcal{F}_t)_{t\in[0,T]}$ is a right-continuous filtration.[1] The aim of the robust stochastic filtering problem is to identify, for each $t \in (0, T]$, a continuous mapping

$$f_t : C([0,t], \mathbb{R}^{d_Y}) \longrightarrow \mathcal{P}(\mathbb{R}^{d_X}), \qquad y_{[0:t]} \longmapsto \mathbb{P}\Big(X_t \in \cdot \,\Big|\, \mathcal{F}_t^Y\Big), \tag{4.1}$$

where $C([0,t], \mathbb{R}^{d_Y})$ denotes the space of continuous paths $[0,t] \to \mathbb{R}^{d_Y}$ equipped with the uniform norm, and $\mathcal{P}(\mathbb{R}^{d_X})$ the set of Borel probability measures on $\mathbb{R}^{d_X}$ equipped with the weak topology. The continuity—and uniqueness—of f_t distinguishes the robust representation (4.1) from merely measurable versions.

In this chapter we establish, for a broad class of non-linear, non-Markovian partially observed systems, that the robust optimal filter enjoys Lipschitz stability with respect to perturbations of the observation path when accuracy is measured in the 2-Wasserstein metric. Building on this, we introduce the *Filterformer* (FF)—an attention-guided, distribution-valued deep architecture—and prove a universal approximation theorem showing that FFs uniformly approximate the robust filter f_t on compact path sets, with quantitative rates enabled by a lossless pathwise attention encoding.

Filtering and model uncertainty. The robust filtering problem treated here provides a dynamic perspective on model uncertainty that directly complements the discussion in Sections 1.3 and 1.4.1. There, model uncertainty was formulated in terms of a latent, time-evolving parameter process θ, with posterior distribution $\mathfrak{P}_t$ updated sequentially in light of observed market data Y. In this chapter we set $X := \theta$ so that the conditional distribution

$$\mathfrak{P}_t \;=\; \mathbb{P}(\theta_t \in \cdot \mid \mathcal{F}_t^Y)$$

from Section 1.4 coincides with the robust representation

$$f_t(Y_{[0:t]}) \;=\; \mathbb{P}(X_t \in \cdot \mid \mathcal{F}_t^Y).$$

[1] Throughout this chapter X denotes the latent signal; in Section 1.4 we denoted this by θ as the unobservable state representing model parameters. We retain X here to remain consistent with the chapter's technical development.

Accordingly, the present results give rigorous foundation to the filtering perspective on model uncertainty anticipated in Section 1.4.1, where model parameters are regarded as hidden stochastic processes evolving in time.

Deep learning as a resolution to intractability. General optimal filters are infinite-dimensional, path-dependent, and measure-valued, rendering their exact computation infeasible outside special finite-dimensional cases. Classical approximation schemes—such as particle filters, linearisation, or projection methods—either deteriorate rapidly in high dimensions or impose restrictive structural assumptions, see Section 1.4.2 for more details. Deep neural networks, by contrast, offer expressive and task-adaptive classes of functions on path space. The results of this chapter provide rigorous support for their use in stochastic filtering: we show that robust filters are $\mathcal{W}_2$-Lipschitz continuous and that they can be uniformly approximated by our proposed Filterformer architecture, whose pathwise attention mechanism enables efficient and lossless encodings of observation paths.

4.1.1 The Filterformer Architecture

To realise the approximation results developed in this chapter, we propose the *Filterformer* (FF), an attention-based extension of the deep Kalman filter (DKF) of [185]. While the DKF is a conditionally Gaussian latent-variable model loosely inspired by the classical Kalman filter, the FF is explicitly designed to approximate the robust representation f_t in (4.1). That is, the FF takes as input a time t together with an observed path segment $Y_{[0:t]}$ and outputs a probability distribution on $\mathbb{R}^{d_X}$, intended as an approximation of the conditional law of the signal given observations up to time t. The architecture, visualised in Figure 1.3, is composed of three interacting phases:

Phase 1. **Pathwise attention encoder.** The observed path $Y_{[0:T]}$ is compressed into a finite-dimensional vector representation while remaining *injective* on a broad class of compact subsets $K \subset C([0,T],\mathbb{R}^{d_Y})$—including all piecewise-linear paths, any finite training set, and any K isometric to a closed Riemannian manifold. The encoding is therefore *lossless*, and both the encoder and its inverse are Lipschitz, providing *stability* in the sense of constructive approximation theory [75, 235]. Because the encoder learns

directly from data without projection onto a fixed basis, it is also *adaptive* to the intrinsic geometry of K.

Phase 2. **Feature extraction via MLP.** The encoded vectors are processed by a multi-layer perceptron (MLP) which learns nonlinear task-specific combinations of the data. Thanks to universal approximation results for ReLU networks [170, 180, 204, 289], the MLP can flexibly adapt to the dynamics of the system (4.2)-(4.3), extracting the statistics most relevant for filtering. This is the only component directly inherited from the original DKF framework.

Phase 3. **Geometric attention decoder.** Finally, a modified version of the geometric attention mechanism of [1] maps the deep features into $\mathcal{N}_{d_X}$-valued outputs, where $\mathcal{N}_{d_X}$ denotes the set of non-singular Gaussian measures on $\mathbb{R}^{d_X}$ endowed with the $\mathcal{W}_2$ metric. Concretely, the decoder projects Euclidean features to points in a geodesic convex hull generated by a finite reference set covering the image of K under the true filter, thereby producing a probabilistically meaningful approximation of the posterior distribution in (4.1).

In contrast to the original DKF, the FF is not merely heuristic but is accompanied by rigorous guarantees, which we detail below.

4.1.2 Main Contributions

We study partially observed dynamics on $[0,T]$ of the form

$$dX_t = \Big[a_0(t, Y_{[0:t]}) + a_1(t, Y_{[0:t]})\, X_t\Big]\, dt \;+\; \sum_{i=1}^{2} b_i(t, Y_{[0:t]})\, dW_t^{(i)}, \tag{4.2}$$

$$dY_t = \Big[A_0(t, Y_{[0:t]}) + A_1(t, Y_{[0:t]})\, X_t\Big]\, dt \;+\; \sum_{i=1}^{2} B_i(t, Y_{[0:t]})\, dW_t^{(i)}, \tag{4.3}$$

where $a_0 : [0,T] \times C([0,T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{d_X}$ and $A_0 : [0,T] \times C([0,T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{d_Y}$ are vector-valued functionals, and

$$a_1, b_1 : [0,T] \times C([0,T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{d_X \times d_X}, \quad b_2 : [0,T] \times C([0,T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{d_X \times d_Y},$$
$$A_1, B_1 : [0,T] \times C([0,T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{d_Y \times d_X}, \quad B_2 : [0,T] \times C([0,T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{d_Y \times d_Y}.$$

All coefficients are measurable, non-anticipative functionals on the measurable space $([0:T] \times C([0:T], \mathbb{R}^{d_Y}), \mathcal{B}_{[0:T]} \times \mathcal{B}_T^{d_Y})$, where $\mathcal{B}_{[0:T]}$ denotes the Borel σ-algebra on $[0,T]$ and $\mathcal{B}_T^{d_Y}$ the canonical σ-algebra on path space. We assume $(\Omega, \mathcal{F}, \mathfrak{F}, \mathbb{P})$ supports independent $\mathfrak{F}$-adapted Brownian motions $W^{(1)} = (W_t^{(1)})_{t\in[0,T]}$ and $W^{(2)} = (W_t^{(2)})_{t\in[0,T]}$, of respective dimensions d_X and d_Y. All further technical conditions are detailed in Assumption 4.2.1. This class strictly generalises the linear-Gaussian models underlying the Kalman-Bucy filter, whilst retaining the conditional Gaussian structure essential for tractable limit distributions.

The theoretical results of this chapter may be summarised in two principal contributions:

(i) **$\mathcal{W}_2$-Lipschitz stability of the robust filter.** For the partially observed dynamics (4.2)-(4.3), we prove that the robust optimal filter f_t is Lipschitz continuous with respect to perturbations of the observation path when measured in the 2-Wasserstein distance (Proposition 4.4.2). This result quantifies the sensitivity of the conditional law $\mathbb{P}(X_t \in \cdot \mid \mathcal{F}_t^Y)$ to observation noise and establishes the regularity required for stable approximation.

(ii) **Universal approximation by Filterformers.** Building on this stability result, we establish a universal approximation theorem (Proposition 4.4.3, Theorem 4.4.1) for the Filterformer architecture. Specifically, for any compact $K \subset C([0,T], \mathbb{R}^{d_Y})$ and any $\varepsilon > 0$, there exists an FF whose outputs approximate the robust filter uniformly over K to within ε in $\mathcal{W}_2$. The guarantee does not require non-degeneracy of the target Gaussians, and thanks to the pathwise attention encoder, the embedding dimension is independent of the approximation accuracy. This result establishes the FF as a universal surrogate for robust filters in non-linear, non-Markovian systems.

4.1.3 Outline and Notation

Section 4.2 formalizes the basic required regularity conditions on the coupled system (4.2)-(4.3) and it overviews any background material. Section 4.3 rigorously introduces the relevant DKF model. Section 4.4 contains the chapter's main theoretical results, as well as an overview of the proof methodology. Several results of independent interest,

both concerning the approximation capacity of the DKF model and the local Lipschitz regularity of the robust filtering map (1.4) are discussed. Detailed proofs of our results are relegated to Appendix C.

Notation. The following serves as a reference, which records the notation used throughout this chapter. In what follows, $N, I, J \in \mathbb{N}_+$, A is an arbitrary $I \times J$ matrix, $x, x_1, \dots, x_J \in \mathbb{R}^N$, f is an arbitrary real-valued function on $\mathbb{R}$, and $t \in \mathbb{R}$.

1. *Componentwise Composition:* $f \bullet x \overset{\text{def.}}{=} (f(x_n))_{n=1}^N$ for any $N \in \mathbb{N}_+$.
2. *Rectified Linear Unit (ReLU):* $\mathrm{ReLU} : \mathbb{R} \to \mathbb{R}$ given by $\mathrm{ReLU}(t) \overset{\text{def.}}{=} \max\{0, t\}$.
3. *Rowwise Product:* $v \odot X \overset{\text{def.}}{=} (v_i \, X_{i,j})_{i=1,\dots,N,\, j=1,\dots,d}$.
4. *Softmax Function:* $\mathrm{Softmax}(x) \overset{\text{def.}}{=} (e^{x_n} / \sum_{i=1}^N e^{x_i})_{n=1}^N$.
5. *Sparsity:* $\|(A_{i,j})_{i,j=1}^{I,J}\|_0 \overset{\text{def.}}{=} \#\{A_{i,j} \neq 0 : \, i, j = 1, \dots, I\}$.
6. *Vector Concatenation:* $\oplus_{j=1}^N x_j \overset{\text{def.}}{=} (x_1, \dots, x_J)^\top$ is the $J \times N$-matrix whose j^{th} row is x_j.
7. *Vectorization:* $\mathrm{vec}((A_{i,j})_{i,j=1}^{I,J}) \overset{\text{def.}}{=} (A_{1,1}, \dots, , A_{I,1}, \dots, A_{1,J}, \dots, A_{I,J})$.
8. *Euclidean Norm:* $\|A\| = (\sum_{i,j=1}^{I,J} |A_{i,j}|^2)^{1/2}$.
9. *Supremum Norm:* $\|A\|_\infty = \sup\{A_{i,j} : \, i = 1, ..., I, j = 1, ..., J\}$.

4.2 The Setting

This section formalises the framework in which our analysis is carried out. We begin by specifying the dynamics of the stochastic processes underlying the filtering problem. We then introduce the source space of paths under consideration, followed by the target space of probability measures in which the optimal filter resides.

4.2.1 Regularity Conditions on the Partially-Observed System

We maintain the following regularity assumptions on the dynamics of the coupled system (4.2)-(4.3), which ensure well-posedness of the filtering problem and enable the subsequent approximation results.

Assumption 4.2.1 (Regularity Conditions: Dynamics)**.** *We assume the following uniform bounds:*

$$|a_1(t,y)_{i,j}| \le L \quad \textit{and} \quad |A_1(t,y)_{k,j}| \le L$$

for $y \in C([0:t],\mathbb{R}^{d_Y})$, $t \in [0:T]$, $i,j = 1,\dots,d_X$, *and* $k = 1,\dots,d_Y$.

Furthermore, the following integrability conditions hold:[2]

(i) $\int_0^T \mathbb{E}\Big[a_0(t,y_{[0:t]})_i^4 + b_1(t,y_{[0:t]})_{ij}^4 + b_2(t,y_{[0:t]})_{ij}^4\Big]\,dt < \infty$,

(ii) $\int_0^T \mathbb{E}\Big[b_2(t,y_{[0:t]})_{ij}^4\Big]\,dt < \infty$,

for $i,j = 1,\dots,d_X$ *and* $k,l = 1,\dots,d_Y$, *and all paths* $y_{[0:T]} \in C([0:T],\mathbb{R}^{d_Y})$.

We define

$$B \circ B \overset{\text{def.}}{=} B_1B_1^\top + B_2B_2^\top, \quad b \circ B \overset{\text{def.}}{=} b_1B_1^\top + b_2B_2^\top, \quad b \circ b \overset{\text{def.}}{=} b_1b_1^\top + b_2b_2^\top$$

and require that the matrix $B \circ B$ *is uniformly non-singular, i.e. its inverse is uniformly bounded. Moreover, there exist constants* $L_1, L_2 \in \mathbb{R}$ *and a non-decreasing right-continuous function* $K : [0:T] \to [0:1]$ *such that, for every* $x, y \in C([0:T],\mathbb{R}^{d_Y})$, *each* $k = 1,2$, *and* $i,j = 1,\dots,d$, *it holds for all* $0 \le t \le T$ *that*

(iii) $\Big|(B_k)_{i,j}(t,x) - (B_k)_{i,j}(t,y)\Big|^2 \le L_1 \int_0^t |x_s - y_s|^2\,dK(s) + L_2|x_t - y_t|^2$,

(iv) $(B_k)_{i,j}(t,x)^2 \le L_1 \int_0^t (1+|x_s|^2)\,dK(s) + L_2(1+|x_t|^2)$,

(v) $\int_0^T \mathbb{E}\Big[|A_1((t,Y_{[0:t]}))_{i,j}(X_t)_j|\Big]\,dt < \infty$,

(vi) $\mathbb{E}\Big[|(X_t)_j|\Big] < \infty, \quad \forall t \in [0:T]$,

(vii) $\mathbb{P}\Big(\int_0^T \Big(A_1(t,Y_{[0:t]})_{i,j}\mathbb{E}[(X_t)_j|\mathcal{Y}_t]\Big)^2 dt < \infty\Big) = 1$,

[2] In particular, they imply $\mathbb{E}\big[\sum_{j=1}^{d_X}(X_0)_j^4\big] < \infty$.

for indices $i = 1, \dots, d_Y$ *and* $j = 1, \dots, d_X$*, where* $|\cdot|$ *denotes the Euclidean norm on* $\mathbb{R}^{d_Y}$.

In addition, we impose the following further conditions on the dynamics (4.2)-(4.3)*:*

(viii) Local Lipschitz continuity in the path component uniformly in time,[3] *as well as global Lipschitz continuity in the time component of* a_0*,* a_1*,* $b \circ b$*,* $b \circ B$*,* A_0*,* A_1*,* $(b \circ B)$*, and* $(B \circ B)^{-1}$ *with respect to the* ℓ^2*-norm (or Frobenius norm when matrix-valued).*

(ix) Positive semi-definiteness of $(b \circ b)(t, y_{[0:t]}) - (b \circ B)(B \circ B)(b \circ B)^\top(t, y_{[0:t]})$ *for all* $t \in [0 : T]$ *and paths* $y_{[0:T]} \in C^1([0 : T], \mathbb{R}^{d_X})$*.*

(x) Let $G_t(y_{[0:t]})$ *solve* $\partial_t G_t(y_{[0:t]}) = \tilde{a}_1(t, y_{[0:t]}) G_t(y_{[0:t]})$ *with* $G_0(y_{[0:t]}) = I_{d_X}$*, where*

$$\tilde{a}_1^\top(t, y_{[0:t]}) \overset{\text{def.}}{=} a_1(t, y_{[0:t]}) - (b \circ B)(B \circ B) A_1(t, y_{[0:t]}).$$

There exist constants $K_1, K_2 > 0$ *such that* $K_1 \leq \operatorname{tr}(G_t(y_{[0:t]})) \leq K_2$ *uniformly, and* $K_3 > 0$ *such that* $\operatorname{tr}(A_1^\top (B \circ B)^{-1} A_1(t, y_{[0:t]})) \leq K_3$*.*

Finally, we require for the initial conditions that

(xi) the conditional distribution of X_0 *given* Y_0 *is Gaussian,*

(xii) the covariance matrix $\Sigma_0 := \mathbb{E}[(X_0 - \mu_0)(X_0 - \mu_0)^\intercal | Y_0]$ *is positive definite, where* $\mu_0 := \mathbb{E}[X_0 | Y_0]$*, and*

(xiii) μ_0 *and* Σ_0 *are Lipschitz continuous in* Y_0 *with respect to the* $\|\cdot\|_2$*-norm.*

4.2.2 Source Spaces: Regular Compact Subsets of Path Space

Fix a time horizon $T > 0$. For each $0 < t \leq T$, let $C([0 : t], \mathbb{R}^{d_Y})$ denote the set of continuous paths from $[0 : t]$ to $\mathbb{R}^{d_Y}$. We equip this set with the uniform norm

$$\|y\|_t \overset{\text{def.}}{=} \max_{0 \leq s \leq t} \|y_s\|_2,$$

making it a Banach space.

[3] That is, there exists a constant which is a local Lipschitz constant for the path at all times $t \in [0, T]$.

According to functional Itô calculus [15, Section 5.2.1], any path $y \in C([0:t], \mathbb{R}^{d_Y})$ can be embedded canonically into a path $\bar{y} \in C([0:T], \mathbb{R}^{d_Y})$ by extension after time t:

$$\bar{y}(s) \overset{\text{def.}}{=} \begin{cases} y(s), & 0 \leq s \leq t, \\ y(t), & t < s \leq T. \end{cases}$$

The map $\bar{\cdot} : C([0:t], \mathbb{R}^{d_Y}) \to C([0:T], \mathbb{R}^{d_Y})$ is a linear isometric embedding, since

$$\|\bar{y}\|_T = \max_{0 \leq s \leq T} \|y(s)\|_2 = \max_{0 \leq s \leq t} \|y(s)\|_2 = \|y\|_t.$$

Conversely, the restriction $\bar{y}_{[0:t]}$ of a path $\bar{y}$ to $[0:t]$ defines a non-expansive linear operator from $C([0:T], \mathbb{R}^{d_Y})$ to $C([0:t], \mathbb{R}^{d_Y})$.

Even in finite dimensions, certain regular functions cannot be approximated by standard deep neural networks with a number of trainable parameters polynomial in the reciprocal of the approximation error [188, Proposition A.3]. Thus, the curse of dimensionality remains unavoidable, and can only be more pronounced in infinite dimensions [117].

This curse can be mitigated either by restricting the function class (e.g. [2, 124, 211, 264]), or by restricting the regularity of the compact subsets of the input space on which uniform approximation is required (e.g. [180, 183, 204]). Recent work shows that such restrictions suffice to obtain universality for infinite-dimensional functions not taking values in linear spaces [182], although their necessity remains an open question. Explicit effects of dimension and diameter restrictions are studied in [1, Proposition 3.10].

Here, we adopt the second approach, since the function class is dictated by the stochastic filtering problem itself. We therefore restrict to compact subsets $K \subset C([0:T], \mathbb{R}^{d_Y})$ which are regular: either isometric to some compact Riemannian manifold, or consisting of piecewise linear paths with finitely many pieces. We exclude fractal-like subsets which cannot be compressed into finite dimensions without losing information (i.e. which admit no bi-Lipschitz embedding into a finite-dimensional normed space).

Assumption 4.2.2 (Domain Regularity)**.** *Fix a compact $K \subseteq C([0:T], \mathbb{R}^{d_Y})$ and a latent dimension $d_K \in \mathbb{N}_+$. Suppose one of the following:*

(i) Finite domains: *K is finite.*

(ii) Piecewise-linear domains: *There exist* $N_{\mathrm{p}} \in \mathbb{N}_+$, $0 = t_0 < \cdots < t_{N_{\mathrm{p}}} = T$, *and* $C_K > 0$ *such that* $y_\cdot \in K$ *iff*

(a) $y(t) = A^{(i)}t + b^{(i)}$ *for* $t \in (t_i, t_{i+1})$ *with some* $A^{(i)} \in \mathbb{R}^{d_Y \times d}$ *and* $b^{(i)} \in \mathbb{R}^{d_Y}$,

(b) $y_0 = 0$,

(c) $\max_{i=1,\ldots,N_{\mathrm{p}}} |y(t_i)| \leq C_K$.

(iii) Smooth domains: *There exists a compact, connected* d_K*-dimensional Riemannian manifold* $(\mathcal{M}, g)$ *such that* $(\mathcal{M}, d_g)$ *is isometric to* $(K, \|\cdot\|_T)$.

Finite domains correspond to any training set, and results here parallel interpolation results for deep learning [85, 181, 279], albeit with arbitrarily small error as in [231]. Smooth domains are also plentiful: by the Banach-Mazur theorem [16] and related embedding results [142, 252], $C([0:T], \mathbb{R}^{d_Y})$ contains isometric copies of any compact metric space, hence also of any compact connected Riemannian manifold.

Proposition 4.2.1 (Isometric Copies of Compact Riemannian Manifolds)**.** *Let* $\mathcal{X}$ *be a compact metric space. For every* $T > 0$ *and each* $d \in \mathbb{N}_+$*, there exists a compact* $K \subseteq C([0:T], \mathbb{R}^{d_Y})$ *and an isometry from* $\mathcal{X}$ *onto* K.

Proof. See Appendix C.1. □

Explicit estimates are available when $(K, d_\cdot)$ satisfies Assumption 4.2.2 (iii), provided the model manifold $\mathcal{M}$ is aspherical, i.e. its homotopy groups $\pi_i(\mathcal{M})$ vanish for all $i \geq 2$, as in the case of a torus (see [269, Chapter 7]).

Remark 4.2.1. Assumption 4.2.2 (ii) permits the compact domain K to consist of piecewise linear paths (with finitely many pieces). Although the observation dynamics considered in this chapter will almost certainly not generate such paths, this hypothesis is best understood as an *approximation device*.

Concretely, for any continuous path y of finite p-variation and any sequence of dissections D_n with mesh $|D_n| \to 0$, the piecewise linear interpolants y^{D_n} converge to y

uniformly; moreover, their p-variation is uniformly controlled, $\sup_n \|y^{D_n}\|_{p\text{-var};[0,T]} < \infty$; see [116, Thm. 5.25]. Since p-variation dominates the uniform topology, i.e.

$$\|y - \tilde{y}\|_T \leq \|y(0) - \tilde{y}(0)\| + \|y - \tilde{y}\|_{p\text{-var};[0,T]} \quad \text{for} \quad p \geq 1,$$

such piecewise linear approximations are arbitrarily accurate in the topology used throughout this chapter. See Remark 4.4.1 for more details. ◁

Domains satisfying Assumption 4.2.2 (ii) include piecewise linear interpolations of financial time-series data, or simulations thereof, since only finitely many inflection points are observable and prices are bounded.

Thus, Assumption 4.2.2 describes a rich family of compact subsets of path space with sufficient structure to be losslessly encoded into finitely many dimensions by the model's encoding layer, its *pathwise attention mechanism*.

4.2.3 Target Space: The 2-Wasserstein Space of Probability Measures

Let $1 \leq p \leq 2$ be fixed. The p-Wasserstein space $\mathcal{P}_p(\mathbb{R}^{d_X})$ consists of all probability measures on $\mathbb{R}^{d_X}$ with finite pth moment, i.e. $\mathbb{P} \in \mathcal{P}_p(\mathbb{R}^{d_X})$ if

$$\mathbb{E}_{X\sim\mathbb{P}}[\|X\|^p] < \infty.$$

The p-Wasserstein distance $\mathcal{W}_p$ between $\mathbb{P}$ and $\mathbb{Q}$ is defined via the Kantorovich problem:

$$\mathcal{W}_p(\mathbb{P},\mathbb{Q})^p \overset{\text{def.}}{=} \inf_{(X_1,X_2)\sim\pi;\, X_1\sim\mathbb{P},\, X_2\sim\mathbb{Q}} \mathbb{E}_\pi[\|X_1 - X_2\|^p],$$

where the infimum ranges over all couplings π of $\mathbb{P}$ and $\mathbb{Q}$.

In general, computing $\mathcal{W}_p$ is computationally demanding, with super-quadratic complexity. However, in certain cases the 2-Wasserstein distance admits closed-form expressions. For instance, when $\mathbb{P} = N(\mu^{(1)}, \Sigma^{(1)})$ and $\mathbb{Q} = N(\mu^{(2)}, \Sigma^{(2)})$ are Gaussian measures on $\mathbb{R}^{d_X}$ with invertible (hence positive-definite) covariance matrices, [96] showed

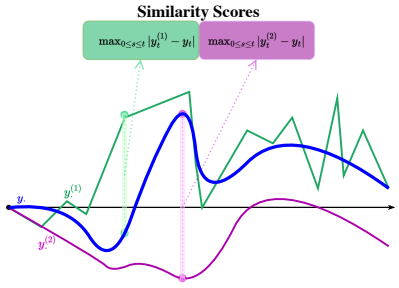
$$\mathcal{W}_2(\mathbb{P},\mathbb{Q})^2 = \|\mu^{(1)} - \mu^{(2)}\|_2^2 + \operatorname{tr}(\Sigma^{(1)}) + \operatorname{tr}(\Sigma^{(2)}) - 2\,\operatorname{tr}\left(\Sigma^{(2)^{1/2}}\Sigma^{(1)}\Sigma^{(2)^{1/2}}\right)^{1/2}. \tag{4.4}$$

4.3 The Filterformer: An Attention-Based Deep Filter

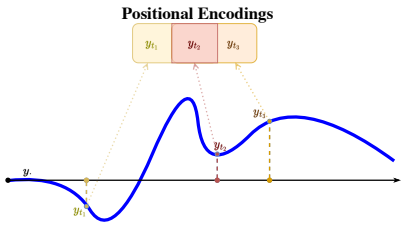
In this section we expand on the introduction given in Section 4.1.1, now with full technical detail.

4.3.1 Phase 1 — Encoding via Pathwise Attention

We introduce an attention-type feature encoder which adapts the classical *attention mechanism* of [12] and the graph attention mechanism of [278] to the context of continuous-time path inputs. Our encoder, termed *pathwise attention*, extracts salient features from a novel path $y_\cdot$. Analogous to attention mechanisms in natural language processing, it adaptively represents new inputs through similarity scores to previously observed reference paths.



(a) The *similarity score* component $\mathrm{sim}_T^{\theta_0}$ encodes a path into finite-dimensional vectorial data by comparing it to a dictionary of reference paths.



(b) The *positional encoding* component $\mathrm{post}_T^{\theta_0}$ enhances similarity scores by incorporating sampled positions from the input path.

Figure 4.1: Our *pathwise attention mechanism* attn_t^θ (see Definition 4.3.3) combines two methods for adaptive feature extraction. (a) Similarity scores rank the resemblance of a new path $x_\cdot$ to a library of reference paths. (b) Positional encodings sample points along $x_\cdot$, perturbing the similarity scores. Unlike basis-based methods, nearly all parameters are trainable, allowing the mechanism to adapt in a task-specific manner, just as deep MLPs outperform fixed basis-function regressions.

Defintion 4.3.1 (Similarity Score)**.** *Fix positive integers* N_{ref}, N_{sim}, d_Y, *and a time horizon* $T > 0$. *Given reference paths* $\{y^{(n)}\}_{n=1}^{N_{\mathrm{ref}}} \subset C([0,1], \mathbb{R}^{d_Y})$, *let* $B \in \mathbb{R}^{N_{\mathrm{sim}} \times N_{\mathrm{ref}}}$, $b \in \mathbb{R}^{N_{\mathrm{sim}}}$, $A \in \mathbb{R}^{N_{\mathrm{sim}} \times N_{\mathrm{sim}}}$, *and* $a \in \mathbb{R}^{N_{\mathrm{sim}}}$. *Set* $\theta_0 \overset{\text{def.}}{=} (A, B, a, b, \{y^{(n)}\}_{n=1}^{N_{\mathrm{ref}}})$.

The similarity score *is the map* $\mathrm{sim}_T^{\theta_0} : C([0,1], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N_{\mathrm{sim}}}$, *defined by*

$$\mathrm{sim}_T^{\theta_0}(y_\cdot) = \mathrm{Softmax}\left(A\,\mathrm{ReLU} \bullet \left(B\left(\|y_\cdot - y_\cdot^{(n)}\|_t\right)_{n=1}^{N_{\mathrm{ref}}} + b\right) + a\right). \tag{4.5}$$

As illustrated in Figure 4.1, similarity scores provide pathwise analogues of contextual keys and queries from NLP: they measure resemblance against either historical, stress-test, or synthetically generated reference paths.

Defintion 4.3.2 (Positional Encoding)**.** *With notation as in Definition 4.3.1, fix a positive integer* N_{pos}*, query times* $0 \leq t_1 < \cdots < t_{N_{\text{time}}} \leq T$*, a matrix* $V \in \mathbb{R}^{N_{\text{pos}} \times d_Y}$*, and a matrix* $U \in \mathbb{R}^{N_{\text{pos}} \times N_{\text{time}}}$*. Set* $\theta_1 \overset{\text{def.}}{=} (V, U, \{t_n\}_{n=1}^{N_{\text{time}}})$*.*

The positional encoding *is the map* $\text{post}_T^{\theta_1} : C([0,1], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N_{\text{pos}} \times d_Y}$ *given by*

$$\text{post}_T^{\theta_1}(y_\cdot) = U\left(\oplus_{j=1}^{N_{\text{time}}} y_{t_j}\right) + V. \tag{4.6}$$

The positional encoding augments similarity scores by incorporating snapshots of the path, ensuring that piecewise-linear trajectories (as in simulation studies) are sufficiently represented.

Defintion 4.3.3 (Pathwise Attention)**.** *In the notation of Definitions 4.3.1 and 4.3.2, let* $N_{\text{sim}} = N_{\text{pos}}$*. The* pathwise attention mechanism*, with parameter* $\theta \overset{\text{def.}}{=} (\theta_0, \theta_1, C)$*, is the map* $\text{attn}_T^{\theta} : [0,T] \times C([0,T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N+1}$ *defined by*

$$\text{attn}_T^{\theta}(t, y_\cdot) = \left(t\,,\, C\ \text{vec}\left(\text{sim}_T^{\theta_0}(y_\cdot) \odot \text{post}_T^{\theta_1}(y_\cdot)\right)\right), \tag{4.7}$$

for $y_\cdot \in C([0,1], \mathbb{R}^{d_Y})$*,* $0 \leq t \leq T$*, where* C *is an* $N \times (N_{\text{sim}} d_Y)$ *matrix.*

Proposition 4.3.1 (Pathwise Attention Encodes Regular Domains)**.** *Let* $K \subseteq C([0:T], \mathbb{R}^{d_Y})$ *satisfy Assumption 4.2.2. Then there exist* $N \in \mathbb{N}_+$ *and* θ *as in Definition 4.3.3 such that* attn_T^{θ} *restricts to a bi-Lipschitz embedding of* $(K, \|\cdot\|_T)$ *into* $(\mathbb{R}^{N+1}, \|\cdot\|_2)$*.*

Encoding dimension estimates are given in Table 4.1.

Proof. See Appendix C.3. □

Table 4.1: Complexity estimates for lossless pathwise attention encoding of regular compact domains.

Type of Compact Domain	Encoding Dimension (N)	Ass. 4.2.2
Finite	$\mathcal{O}(\log(\#K))$	(i)
Piecewise-linear (N_{p} pieces)	$\mathcal{O}(N_{\mathrm{p}}d)$	(ii)
Isometric compact Riemannian	Finite	(iii)
Aspherical compact Riemannian	$\mathcal{O}\big(\mathfrak{N}_{d_g}(\mathcal{M}, \mathrm{Vol}(\mathcal{M}, g)^{1/d_K})\big)$	(iii)

For $A \subset \mathcal{M}$ and $\delta > 0$, the covering number $\mathfrak{N}_{d_g}(A, \delta)$ is the smallest I such that A is covered by I geodesic balls of radius δ in $(\mathcal{M}, d_g)$.

4.3.2 Phase 2 — Transformation via Multi-Layer Perceptrons

We now recall the structure of standard MLPs, which provide the nonlinear transformations within the Filterformer.

Assumption 4.3.4 ([170]). *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous, non-affine, and differentiable at some $t \in \mathbb{R}$ with $\sigma'(t) \neq 0$.*

Examples include PReLU, $\tanh$, sine (SIREN), and Swish activations.

An MLP $\hat{f} : \mathbb{R}^N \to \mathbb{R}^{N'}$ with activation σ has iterative form

$$\begin{aligned} \hat{f}(x) &= A^{(J)}\, x^{(J)} + b^{(J)}, \\ x^{(j+1)} &= \sigma \bullet (A^{(j)} x^{(j)} + b^{(j)}), \quad j = 0, \dots, J-1, \\ x^{(0)} &= x, \end{aligned} \tag{4.8}$$

with $A^{(j)} \in \mathbb{R}^{d_{j+1} \times d_j}$, $b^{(j)} \in \mathbb{R}^{d_{j+1}}$, $d_0 = N$, $d_{J+1} = N'$.

4.3.3 Phase 3 — Decoding via Geometric Attention

The decoding stage employs a *geometric attention mechanism* mapping features into Gaussian measures, ensuring output distributions reside in $\mathcal{N}_{d_X}$. We denote by $P_{\Delta_{N'}}$ the projection of $\mathbb{R}^{N'}$ onto the probability simplex $\Delta_{N'}$; convexity ensures it is well-defined and 1-Lipschitz [26, Proposition 4.8].

Defintion 4.3.5 (Geometric Attention). *Let $N', d_X \in \mathbb{N}_+$. A geometric attention mechanism is a map $\text{g-attn}_{N'} : \mathbb{R}^{N'} \to \mathcal{N}_{d_X}$ of the form*

$$\text{g-attn}^{\vartheta}_{N'}(v) = \mathcal{N}_{d_X}\left(\sum_{n=1}^{N'} P_{\Delta_{N'}}(v)_n\, m^{(n)},\ \sum_{n=1}^{N'} P_{\Delta_{N'}}(v)_n\, (A^{(n)})^\top A^{(n)}\right),$$

with parameters $\vartheta = (m^{(n)}, A^{(n)})_{n=1}^{N'}$, *where* $m^{(n)} \in \mathbb{R}^{d_X}$ *and* $A^{(n)} \in \mathbb{R}^{d_X \times d_X}$.

4.3.4 The Filterformer Model

We can now formalise the Filterformer architecture, combining the three phases above.

Defintion 4.3.6 (Filterformer)**.** *Let* $d_X, d_Y \in \mathbb{N}_+$, $T > 0$, *and* $\sigma : \mathbb{R} \to \mathbb{R}$ *satisfy Assumption 4.3.4. A* Filterformer *is a function* $\hat{F} : C([0:T], \mathbb{R}^{d_Y}) \to \mathcal{N}_{d_X}$ *with representation*

$$\hat{F} = \text{g-attn}^{\vartheta}_{N'} \circ \hat{f} \circ \text{attn}^{\theta}_{T}, \tag{4.9}$$

where attn^{θ}_{T} *is pathwise attention (Definition 4.3.3),* $\hat{f}$ *is an MLP (4.8), and* $\text{g-attn}^{\vartheta}_{N'}$ *is geometric attention (Definition 4.3.5).*

4.4 Main Result

We are now ready to state our main theorem, which shows that the Filterformer (Definition 4.3.6) is capable of approximating the optimal filter for the coupled system (4.2)-(4.3). Our guarantees are non-asymptotic: they provide explicit error bounds in terms of the network's complexity.

Theorem 4.4.1 (The Filterformer Approximates the Optimal Filter)**.** *Let* $d_X, d_Y \in \mathbb{N}_+$ *and let* $K \subset C([0:T], \mathbb{R}^{d_Y})$ *satisfy Assumption 4.2.2. Suppose that the coupled system (4.2)-(4.3) satisfies Assumption 4.2.1. Then, for every* $T > 0$ *and every* $\varepsilon > 0$ *there exists a Filterformer* $\hat{F} : [0:T] \times C([0:T], \mathbb{R}^{d_Y}) \to \mathcal{N}_{d_X}$ *such that*

$$\max_{0 \le t \le T,\, y_\cdot \in K} \mathcal{W}_p\Big(\mathbb{P}(X_t \in \cdot \mid y_{[0:t]}), \hat{F}(t, y_\cdot)\Big) < \varepsilon, \tag{4.10}$$

for every $1 \le p \le 2$. *Complexity estimates for* $\hat{F}$ *are recorded in Table 4.2.*

Table 4.2: Complexity estimates for the Filterformer $\hat{F}$ in Theorem 4.4.1.

σ Regularity	Depth	Width	Encode (N)	Decode (N')
ReLU	$\mathcal{O}(\varepsilon^{-N'})$	$\mathcal{O}(\varepsilon^{-N'})$	$\mathcal{O}(1)$	$\mathcal{O}(\varepsilon^{-1})$
Smooth & Non-polynomial	$\mathcal{O}(\varepsilon^{-4N'-1})$	$N'+N+3$	$\mathcal{O}(1)$	$\mathcal{O}(\varepsilon^{-1})$
Polynomial & Non-affine	$\mathcal{O}(\varepsilon^{-8N'-6})$	$N'+N+4$	$\mathcal{O}(1)$	$\mathcal{O}(\varepsilon^{-1})$
Non-smooth & Non-polynomial	Finite	$N'+N+3$	$\mathcal{O}(1)$	$\mathcal{O}(\varepsilon^{-1})$

Remark 4.4.1 (Implications of Theorem 4.4.1 for piecewise linear domains). Piecewise linear functions (Assumption 4.2.2 (ii)) may be interpreted as a convenient *approximation domain*: Let $y \in C([0,T],\mathbb{R}^{d_Y})$ be an admissible observation path and choose piecewise linear interpolants $y^{(n)}$ along partitions of vanishing mesh, so that $\|y^{(n)} - y\|_T \to 0$ (cf. [116, Thm. 5.25]).

If the theorem is applied on a piecewise linear compact set K containing $\{y^{(n)} : n \in \mathbb{N}\}$, it yields the desired conclusion for each $y^{(n)}$. Assumption 4.2.2 (ii) then propagates this conclusion from $y^{(n)}$ to y with a transparent error term:

$$\mathcal{W}_2\Big(f_t(y^{(n)}), f_t(y)\Big) \;\le\; L_t\,\|y^{(n)} - y\|_T \;\longrightarrow\; 0.$$

Accordingly, the theorem may be read as holding for general observation paths up to an arbitrarily small approximation error, provided one is willing to work with a piecewise linear representation of the observed signal. This is for instance the case when data is only available at discrete sampling times. ◁

4.4.1 Proof Overview

We briefly outline the derivation of Theorem 4.4.1, the details are relegated to Appendix C.

Classical results [201, Theorem 12.6] show that under Assumption 4.2.1, the conditional law $\mathbb{P}(X_t \in \cdot \mid \mathcal{Y}_t)$ is Gaussian. More precisely, for each $0 \le t \le T$ and each path $y_{[0:t]} \in C([0:t],\mathbb{R}^{d_Y})$, the distribution admits the representation

$$\mathbb{P}(X_t \in \cdot \mid y_{[0:t]}) = f_t(y_{[0:t]}) = \mathcal{N}(\mu(y_{[0:t]}), \Sigma(y_{[0:t]})), \tag{4.11}$$

for suitable mean and covariance functions μ and Σ.

A first step towards approximability is to establish that f_t is locally Lipschitz in its path argument. This ensures that the conditional law depends in a controlled way on small perturbations of the observed trajectory.

Proposition 4.4.2 (Local Lipschitz Continuity of the Optimal Filter)**.** *Under Assumption 4.2.1, the map f_t in (1.4) is locally Lipschitz. In particular, for every $t \in [0:T]$, path $y^{(1)}_{\cdot} \in C^1([0:T],\mathbb{R}^{d_Y})$, and $\epsilon > 0$, there exists $C \ge 0$ such that for all $s \in [0:T]$ and $y^{(2)}_{\cdot} \in C^1([0:T],\mathbb{R}^{d_Y})$ with $|t-s| < \epsilon$ and $\|y^{(1)}_{[0:T]} - y^{(2)}_{[0:T]}\|_T < \epsilon$, one has*

$$\mathcal{W}_2\Big(\mathbb{P}(X_t \in \cdot \mid y^{(1)}_{[0:t]}),\, \mathbb{P}(X_s \in \cdot \mid y^{(2)}_{[0:s]})\Big) \le C\Big(\|y^{(1)}_{[0:T]} - y^{(2)}_{[0:T]}\|_T + |t-s|\Big).$$

Having established the regularity of the optimal filter, the next step is to verify that functions with the same domain, codomain, and regularity can be uniformly approximated by a Filterformer. This is formalised in the following approximation theorem.

Table 4.3: Complexity estimates for the Filterformer $\hat{F}$ in Proposition 4.4.3.

σ Regularity	Depth	Width	Encode (N)	Decode (N')
ReLU	$\mathcal{O}\big((LV(L))^{-N'}\varepsilon^{-N'}\big)$	$\mathcal{O}\big((LV(L))^{-N'}\varepsilon^{-N'}\big)$	$\mathcal{O}(1)$	$\mathcal{O}(L\varepsilon^{-1})$
Smooth & Non-polynomial	$\mathcal{O}(L^{4N'+1}\varepsilon^{-4N'-1})$	$N'+N+2$	$\mathcal{O}(1)$	$\mathcal{O}(L\varepsilon^{-1})$
Polynomial & Non-affine	$\mathcal{O}(L^{8N'+6}\varepsilon^{-8N'-6})$	$N'+N+3$	$\mathcal{O}(1)$	$\mathcal{O}(L\varepsilon^{-1})$
Non-smooth & Non-polynomial	Finite	$N'+N+2$	$\mathcal{O}(1)$	$\mathcal{O}(L\varepsilon^{-1})$

Here $V(t)$ denotes the inverse of $s \mapsto s^4 \log_3(t+2)$ on $[0,\infty)$, evaluated at $131t$.

Proposition 4.4.3 (Approximation Capacity of Filterformers)**.** *Let $d_X, d_Y \in \mathbb{N}_+$, $L > 0$, and $K \subset C([0:T],\mathbb{R}^{d_Y})$ satisfy Assumption 4.2.2. For every $0 < \varepsilon < 1/2$ and every L-Lipschitz function $f : ([0,T]\times K, |\cdot|\times\|\cdot\|_\infty) \to (\mathcal{N}_{d_X}, \mathcal{W}_2)$ there exists a Filterformer $\hat{F}$ such that*

$$\sup_{t,x\in[0:T]\times K} \mathcal{W}_p\Big(f(t,x.),\ \hat{F}(t,x.)\Big) < \varepsilon,$$

for all $1 \le p \le 2$. The corresponding depth, width, encoding dimension (N), and decoding dimension (N') are as in Table 4.3.

Proof of Theorem 4.4.1. Since the coupled system, defined in (4.2)-(4.3); satisfies Assumption 4.2.1, Proposition 4.4.2 implies that the optimal filter f_T in (1.4) is locally Lipschitz as a map from $([0:T]\times C([0:T],\mathbb{R}^{d_Y}), |\cdot|\times\|\cdot\|_T)$ to $(\mathcal{N}_{d_X},\mathcal{W}_2)$. Since K is compact, the restriction $f|_K$ is Lipschitz.

Applying Proposition 4.4.3, we conclude that for any $\varepsilon > 0$ there exists a Filterformer $\hat{F}$ with representation (4.9), and complexity as in Table 4.3, such that

$$\sup_{t,y\in[0:T]\times K} \mathcal{W}_p\Big(f(t,y),\ \hat{F}(t,y)\Big) < \varepsilon.$$

This proves Theorem 4.4.1. □

4.5 Conclusion

This chapter established a rigorous approximation theory for the *Filterformer*. Our main theorem (Theorem 4.4.1) showed that, under Assumption 4.2.1 and on regular compact subsets of path space (Assumption 4.2.2), the Filterformer can approximate the conditional law of the signal process uniformly in time, to arbitrary accuracy in $\mathcal{W}_p$ for $1 \leq p \leq 2$. The guarantees are non-asymptotic: the approximation error is controlled by the architectural complexity (depth, width, and encoding/decoding dimensions), with explicit rates summarised in Tables 4.2 and 4.3. The construction relies on three key components: a pathwise attention encoder embedding structured path domains into finite dimensions, an MLP transformation providing universal approximation power, and a geometric attention decoder mapping features into Gaussian measures while preserving Wasserstein control.

These theoretical guarantees complement the empirical successes of attention-based deep filters [184, 185], which is frequently trivially assumed under the misconception that "Kalman" implies linearity or finite dimensionality. By working directly with continuous-time paths and outputting distributional predictions in Wasserstein space, the Filterformer provides a principled architectural design that respects both temporal structure and probabilistic geometry.

The analysis draws on the conditional Gaussianity of the filter under Assumption 4.2.1 (cf. [201, Thm. 12.6]) and on regularity assumptions for the input domain. Although these conditions are natural in many partially observed diffusion settings, they also delineate the current scope. In particular, the necessity of restricting approximation to regular compact subsets of path space remains unresolved, reflecting broader insights in high-dimensional approximation where structural restrictions are generally unavoidable.

Several directions for further work arise naturally. First, a statistical learning theory for the Filterformer remains to be developed: generalisation bounds reflecting the attention encoder and geometric decoder could be obtained by adapting capacity measures or transport-theoretic stability arguments [23, 24, 153]. Second, in applications such as mathematical finance where only a single trajectory is observed, the challenge is to

train Filterformers with robust guarantees in this one-shot regime. In [199] we address this issue in a related context; but although the results may be extendable, they require substantial adaptation. Third, extending beyond Gaussian decoding to richer output families, studying robustness to model drift, and sharpening the complexity-accuracy trade-offs of the encoder all represent promising lines of enquiry.

Statement of Authorship

Title of Paper: *Transformers Can Solve Non-Linear and Non-Markovian Filtering Problems in Continuous Time For Conditionally Gaussian Signals*

Publication Status: Submitted for Publication at SIAM/ASA Journal on Uncertainty Quantification, July 2025

Publication Details: Blanka Horvath, Anastasis Kratsios, Yannick Limmer, and Xuwei Yang. *Transformers Can Solve Non-Linear and Non-Markovian Filtering Problems in Continuous Time For Conditionally Gaussian Signals*. 2025. arXiv: 2310.19603 [cs.LG]. URL: https://arxiv.org/abs/2310.19603

Contribution to the paper: This article is joint work with Blanka Horvath, Anastasis Kratsios, and Xuwei Yang. I contributed to all stages of the research, including the conceptual development, derivation of the Filterformer model class, formulation of the main theorem, and the construction and verification of its proof.

Blanka Horvath and Anastasis Kratsios acted in senior supervisory roles, providing high-level conceptual guidance throughout the project. Xuwei Yang supported the supervision, particularly during the technical development of the proof.

My main contributions to the proof included establishing the stability of the optimal filter and demonstrating the lossless encoding properties of the path-wise attention mechanism. The geometric aspects of the Filterformer's output were developed jointly with Anastasis Kratsios.

Xuwei Yang additionally supervised the proof of the regularity conditions of the optimal filter.

Blanka Horvath, in her capacity as my academic supervisor, provided critical conceptual feedback, guided the research direction, and played a central role in situating the work within the broader academic literature.

References

[1] Beatrice Acciaio, Anastasis Kratsios, and Gudmund Pammer. "Designing universal causal deep learning models: the geometric (hyper)transformer". In: *Math. Finance* 34.2 (2024), pp. 671–735. DOI: 10.1111/mafi.12389.

[2] Ben Adcock, Simone Brugiapaglia, Nick Dexter, and Sebastian Moraga. "Near-optimal learning of Banach-valued, high-dimensional functions via deep neural networks". In: *Neural Networks* 181 (Jan. 2025), p. 106761. DOI: 10.1016/j.neunet.2024.106761.

[3] R. Ahmad and R. M. Dudley. "Real Analysis and Probability." In: *The Statistician* 40.4 (1991), p. 457. DOI: 10.2307/2348742.

[4] Andrew L. Allan, Jost Pieper, and Josef Teichmann. *Rough SDEs and Robust Filtering for Jump-Diffusions*. 2025. arXiv: 2507.05930 [math.PR]. URL: https://arxiv.org/abs/2507.05930.

[5] Hervé Andrès, Alexandre Boumezoued, and Benjamin Jourdain. "Signature-based validation of real-world economic scenarios". In: *ASTIN Bulletin* 54.2 (Apr. 2024), pp. 410–440. DOI: 10.1017/asb.2024.12.

[6] Flavia Antonacci, Cristina Costantini, and Marco Papi. "Short-Term Interest Rate Estimation by Filtering in a Model Linking Inflation, the Central Bank and Short-Term Interest Rates". In: *Mathematics* 9.10 (2021). MDPI, p. 1152. DOI: 10.3390/math9101152.

[7] Reza Arabpour, John Armstrong, Luca Galimberti, Anastasis Kratsios, and Giulia Livieri. *Low-dimensional approximations of the conditional law of Volterra processes: a non-positive curvature approach*. 2024. arXiv: 2405.20094 [math.NA]. URL: https://arxiv.org/abs/2405.20094.

[8] John Armstrong, Damiano Brigo, and Bernard Hanzon. "Optimal projection filters with information geometry". In: *Information Geometry* 7.S1 (June 2023), pp. 525–540. DOI: 10.1007/s41884-023-00108-x.

[9] Imanol Perez Arribas, Cristopher Salvi, and Lukasz Szpruch. "Sig-SDEs model for quantitative finance". In: *Proceedings of the First ACM International Conference on AI in Finance*. ICAIF '20. ACM, Oct. 2020, pp. 1–8. DOI: 10.1145/3383455.3422553.

[10] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. "Coherent Measures of Risk". In: *Mathematical Finance* 9.3 (July 1999), pp. 203–228. DOI: 10.1111/1467-9965.00068.

[11] Simon H. Babbs and K. Ben Nowman. "Kalman Filtering of Generalized Vasicek Term Structure Models". In: *The Journal of Financial and Quantitative Analysis* 34.1 (Mar. 1999), p. 115. DOI: 10.2307/2676248.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL]. URL: https://arxiv.org/abs/1409.0473.

[13] David H. Bailey, Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu. "Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance". In: *Notices Amer. Math. Soc.* 61.5 (2014), pp. 458–471. DOI: 10.1090/noti1105.

[14] Yogesh Balaji, Mohammadmahdi Sajedi, Neha Mukund Kalibhat, Mucong Ding, et al. *Understanding Overparameterization in Generative Adversarial Networks*. 2021. arXiv: 2104.05605 [cs.LG]. URL: https://arxiv.org/abs/2104.05605.

[15] Vlad Bally, Lucia Caramellino, and Rama Cont. *Stochastic Integration by Parts and Functional Itô Calculus*. Springer International Publishing, 2016. ISBN: 9783319271286. DOI: 10.1007/978-3-319-27128-6.

[16] Stefan Banach. *Théorie des opérations linéaires*. Chelsea Publishing Co., New York, 1955, pp. vii+254. ISBN: 9780828401104.

[17] Tianzhe Bao, Syed Ali Raza Zaidi, Shengquan Xie, Pengfei Yang, Yihui Zhao, and Zhiqiang Zhang. "A deep Kalman filter network for hand kinematics estimation using sEMG". In: *Pattern Recognition Letters* 143 (2021), pp. 88–94. DOI: 10.1016/j.patrec.2021.01.001.

[18] Chris Bardgett, Elise Gourier, and Markus Leippold. "Inferring volatility dynamics and risk premia from the S&P 500 and VIX markets". In: *Journal of Financial Economics* 131.3 (2019), pp. 593–618. DOI: 10.1016/j.jfineco.2018.09.008.

[19] Ole E. Barndorff-Nielsen and Neil Shephard. "Econometric analysis of realized volatility and its use in estimating stochastic volatility models". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64.2 (2002), pp. 253–280. DOI: 10.1111/1467-9868.00336.

[20] Daniel Bartl. "Exponential utility maximization under model uncertainty for unbounded endowments". In: *Ann. Appl. Probab.* 29.1 (2019), pp. 577–612. DOI: 10.1214/18-AAP1428.

[21] Daniel Bartl, Samuel Drapeau, Jan Obłój, and Johannes Wiesel. "Sensitivity analysis of Wasserstein distributionally robust optimization problems". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 477.2256 (Dec. 2021). DOI: 10.1098/rspa.2021.0176.

[22] Daniel Bartl, Michael Kupper, and Ariel Neufeld. "Duality theory for robust utility maximisation". In: *Finance and Stochastics* 25.3 (2021), pp. 469–503. DOI: 10.1007/s00780-021-00455-6.

[23] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks". In: *J. Mach. Learn. Res.* 20 (2019), Paper No. 63, 17. ISSN: 1532-4435,1533-7928.

[24] Peter L. Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: risk bounds and structural results". In: *J. Mach. Learn. Res.* 3 (2002), pp. 463–482. DOI: 10.1162/153244303321897690.

[25] Giuliano Basso. "Fixed point theorems for metric spaces with a conical geodesic bicombing". In: *Ergodic Theory and Dynamical Systems* 38.5 (Feb. 2017), pp. 1642–1657. DOI: 10.1017/etds.2016.106.

[26] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Second. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. With a foreword by Hédy Attouch. Springer, Cham, 2017, pp. xix+619. ISBN: 978-3-319-48310-8. DOI: 10.1007/978-3-319-48311-5.

[27] Christian Beck, Sebastian Becker, Patrick Cheridito, Arnulf Jentzen, and Ariel Neufeld. "An efficient Monte Carlo scheme for Zakai equations". In: *Commun. Nonlinear Sci. Numer. Simul.* 126 (2023), Paper No. 107438. DOI: 10.1016/j.cnsns.2023.107438.

[28] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bert Melenberg, and Gijs Rennen. "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities". In: *Management Science* 59.2 (2013), pp. 341–357. DOI: 10.1287/mnsc.1120.1641.

[29] V. E. Beneš. "Existence of Optimal Strategies Based on Specified Information, for a Class of Stochastic Decision Problems". In: *SIAM Journal on Control* 8.2 (May 1970), pp. 179–188. DOI: 10.1137/0308012.

[30] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

[31] 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*. Institute of Mathematical Statistics, 2008, pp. 316–334. ISBN: 0940600749. DOI: 10.1214/193940307000000518.

[32] Ramaprasad Bhar. *Stochastic Filtering with Applications in Finance*. World Scientific Publishing Co. Pte. Ltd., 2010. ISBN: 9789814304863. DOI: 10.1142/9789814304863.

[33] Theerawat Bhudisaksang and Álvaro Cartea. "Adaptive Robust Control in Continuous Time". In: *SIAM Journal on Control and Optimization* 59.5 (2021), pp. 3912–3945. DOI: 10.1137/20M1336680.

[34] Theerawat Bhudisaksang, Álvaro Cartea, and Leandro Sánchez-Betancourt. "Adaptive-Robust Portfolio Optimisation". In: *SSRN* (2025). DOI: 10.2139/ssrn.5082206.

[35] Sara Biagini and Marco Frittelli. "On the super replication price of unbounded claims". In: *Ann. Appl. Probab.* 14.4 (2004), pp. 1970–1991. DOI: 10.1214/105051604000000459.

[36] Tomas Björk, Mark H. A. Davis, and Camilla Landén. "Optimal investment under partial information". In: *Math. Methods Oper. Res.* 71.2 (2010), pp. 371–399. DOI: 10.1007/s00186-010-0301-x.

[37] Fischer Black and Myron Scholes. "The pricing of options and corporate liabilities". In: *J. Polit. Econ.* 81.3 (1973), pp. 637–654. DOI: 10.1086/260062.

[38] Jose Blanchet and Karthyek Murthy. "Quantifying Distributional Model Risk via Optimal Transport". In: *Mathematics of Operations Research* 44.2 (2019), pp. 565–600. DOI: 10.1287/moor.2018.0936.

[39] H. Boedihardjo and X. Geng. “The uniqueness of signature problem in the non-Markov setting”. In: *Stochastic Process. Appl.* 125.12 (2015), pp. 4674–4701. DOI: 10.1016/j.spa.2015.07.012.

[40] Deivid Botina-Monsalve, Yannick Benezeth, and Johel Miteran. “Performance analysis of remote photoplethysmography deep filtering using long short-term memory neural network”. In: *Biomedical Engineering Online* 21 (2022), p. 69. DOI: 10.1186/s12938-022-01037-z.

[41] Léon Bottou. “Large-Scale Machine Learning with Stochastic Gradient Descent”. In: *Proceedings of COMPSTAT’2010*. Physica-Verlag HD, 2010, pp. 177–186. ISBN: 9783790826043. DOI: 10.1007/978-3-7908-2604-3_16.

[42] Léon Bottou. “Stochastic Gradient Descent Tricks”. In: *Neural Networks: Tricks of the Trade*. Springer Berlin Heidelberg, 2012, pp. 421–436. ISBN: 9783642352898. DOI: 10.1007/978-3-642-35289-8_25.

[43] Jean-Philippe Bouchaud and Marc Potters. *On a Generalisation of the Marcenko-Pastur Problem*. 2020. DOI: https://doi.org/10.1103/PhysRevE.102.062117.

[44] Leo Breiman. “Bagging predictors”. In: *Machine Learning* 24.2 (1996), pp. 123–140. DOI: 10.1007/BF00058655.

[45] D. Brigo, B. Hanzon, and F. Le Gland. “A differential geometric approach to nonlinear filtering: the projection filter”. In: *Proceedings of 1995 34th IEEE Conference on Decision and Control*. Vol. 4. CDC-95. IEEE, 1998, pp. 4006–4011. DOI: 10.1109/cdc.1995.479232.

[46] Damiano Brigo, Bernard Hanzon, François Le Gland, and Francois Le Gland. “Approximate Nonlinear Filtering by Projection on Exponential Manifolds of Densities”. In: *Bernoulli* 5.3 (June 1999), p. 495. DOI: 10.2307/3318714.

[47] Andrew Brock, Jeff Donahue, and Karen Simonyan. “Large scale GAN training for high fidelity natural image synthesis”. In: *International Conference on Learning Representations (ICLR)*. 2019. URL: https://openreview.net/forum?id=B1xsqj09Fm.

[48] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. “GRU-ODE-Bayes: continuous modeling of sporadically-observed time series”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[49] Bernard Bru, Henri Heinich, and Jean-Claude Lootgieter. “Distances de Lévy et extensions des théoremès de la limite centrale et de Glivenko-Cantelli”. In: *Publ. Inst. Statist. Univ. Paris* 37.3-4 (1993), pp. 29–42.

[50] H. Buehler, L. Gonon, J. Teichmann, and B. Wood. “Deep hedging”. In: *Quant. Finance* 19.8 (2019), pp. 1271–1291. DOI: 10.1080/14697688.2019.1571683.

[51] Hans Buehler, Blanka Horvath, Yannick Limmer, and Thorsten Schmidt. *Uncertainty-Aware Strategies: A Model-Agnostic Framework for Robust Financial Optimization through Subsampling*. 2025. arXiv: 2506.07299 [q-fin.CP]. URL: https://arxiv.org/abs/2506.07299.

[52] Hans Buehler, Phillip Murray, Mikko S. Pakkanen, and Ben Wood. *Deep Hedging: Learning to Remove the Drift under Trading Frictions with Minimal Equivalent Near-Martingale Measures*. 2022. arXiv: 2111.07844 [q-fin.CP]. URL: https://arxiv.org/abs/2111.07844.

[53] Hans Buehler, Phillip Murray, and Ben Wood. *Deep Bellman Hedging*. 2024. arXiv: 2207.00932 [q-fin.CP]. URL: https://arxiv.org/abs/2207.00932.

[54] Hans Bühler, Blanka Horvath, Terry Lyons, Imanol Perez Arribas, and Ben Wood. *A Data-driven Market Simulator for Small Data Environments*. 2020. arXiv: 2006.14498 [q-fin.ST]. URL: https://arxiv.org/abs/2006.14498.

[55] F. O. Bunnin, Y. Guo, and Y. Ren. "Option pricing under model and parameter uncertainty using predictive densities". In: *Statistics and Computing* 12.1 (Jan. 2002), pp. 37–44. DOI: 10.1023/a:1013116204872.

[56] Russel E. Caflisch. "Monte Carlo and quasi-Monte Carlo methods". In: *Acta Numerica* 7 (1998), pp. 1–49. DOI: 10.1017/S0962492900002804.

[57] John Y. Campbell, Andrew W. Lo, and A. Craig MacKinlay. *The Econometrics of Financial Markets*. Princeton, NJ: Princeton University Press, 1997. ISBN: 978-0-691-04301-2.

[58] Álvaro Cartea, Ryan Donnelly, and Sebastian Jaimungal. "Algorithmic Trading with Model Uncertainty". In: *SIAM Journal on Financial Mathematics* 8.1 (Jan. 2017), pp. 635–671. DOI: 10.1137/16m106282x.

[59] Álvaro Cartea, Sebastian Jaimungal, and Zhen Qin. "Model Uncertainty in Commodity Markets". In: *SIAM Journal on Financial Mathematics* 7.1 (2016), pp. 1–33. DOI: 10.1137/15M1027243.

[60] Umut Çetin. "Financial equilibrium with asymmetric information and random horizon". In: *Finance and Stochastics* 22.1 (Nov. 2017), pp. 97–126. DOI: 10.1007/s00780-017-0348-0.

[61] Ren-Raw Chen and Louis Scott. "Maximum Likelihood Estimation for a Multifactor Equilibrium Model of the Term Structure of Interest Rates". In: *The Journal of Fixed Income* 3.3 (Dec. 1993), pp. 14–31. DOI: 10.3905/jfi.1993.408090.

[62] Tao Chen and Michael Ludkovski. "A Machine Learning Approach to Adaptive Robust Utility Maximization and Hedging". In: *SIAM Journal on Financial Mathematics* 12.3 (2021), pp. 1226–1256. DOI: 10.1137/20M1336023.

[63] Patrick Cheridito and Balint Gersey. "Computation of conditional expectations with guarantees". In: *J. Sci. Comput.* 95.1 (2023), Paper No. 12, 30. DOI: 10.1007/s10915-023-02130-8.

[64] Patrick Cheridito, Arnulf Jentzen, and Florian Rossmannek. "Efficient Approximation of High-Dimensional Functions With Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.7 (July 2022), pp. 3079–3093. DOI: 10.1109/tnnls.2021.3049719.

[65] Ilya Chevyrev and Andrey Kormilitzin. *A Primer on the Signature Method in Machine Learning*. 2025. arXiv: 1603.03788 [stat.ML]. URL: https://arxiv.org/abs/1603.03788.

[66] Ilya Chevyrev and Terry Lyons. "Characteristic functions of measures on geometric rough paths". In: *Ann. Probab.* 44.6 (2016), pp. 4049–4082. DOI: 10.1214/15-AOP1068.

[67] Ilya Chevyrev and Harald Oberhauser. "Signature Moments to Characterize Laws of Stochastic Processes". In: *Journal of Machine Learning Research* 23.176 (2022), pp. 1–42. URL: https://www.jmlr.org/papers/v23/20-1200.html.

[68] Yinlam Chow and Mohammad Ghavamzadeh. "Algorithms for CVaR optimization in MDPs". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 3509–3517.

[69] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. "Risk-sensitive and robust decision-making: a CVaR optimization approach". In: *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, 2015, pp. 1522–1530.

[70] Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. "Patch-level Routing in Mixture-of-Experts is Provably Sample-efficient for Convolutional Neural Networks". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 6074–6114. URL: https://proceedings.mlr.press/v202/chowdhury23a.html.

[71] Charles K. Chui and Guanrong Chen. *Kalman Filtering*. Springer International Publishing, 2017. ISBN: 9783319476124. DOI: 10.1007/978-3-319-47612-4.

[72] J. M. C. Clark. "The Design of Robust Approximations to the Stochastic Differential Equations of Nonlinear Filtering". In: *Communication Systems and Random Process Theory*. Springer Netherlands, 1978, pp. 721–734. ISBN: 9789401175777. DOI: 10.1007/978-94-011-7577-7_41.

[73] J. M. C. Clark and D. Crisan. "On a robust version of the integral representation formula of nonlinear filtering". In: *Probab. Theory Related Fields* 133.1 (2005), pp. 43–56. DOI: 10.1007/s00440-004-0412-5.

[74] Lucas Clarté, Adrien Vandenbroucque, Guillaume Dalle, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. "Analysis of bootstrap and subsampling in high-dimensional regularized regression". In: *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*. UAI '24. Barcelona, Spain: JMLR.org, 2024.

[75] Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. "Optimal stable nonlinear approximation". In: *Found. Comput. Math.* 22.3 (2022), pp. 607–648. DOI: 10.1007/s10208-021-09494-z.

[76] Samuel N. Cohen, Christoph Reisinger, and Sheng Wang. "Detecting and Repairing Arbitrage in Traded Option Prices". In: *Applied Mathematical Finance* 27.5 (Sept. 2020), pp. 345–373. DOI: 10.1080/1350486x.2020.1846573.

[77] Samuel N. Cohen, Christoph Reisinger, and Sheng Wang. "Arbitrage-Free Neural-SDE Market Models". In: *Applied Mathematical Finance* 30.4 (2023), pp. 1–46. DOI: 10.1080/1350486X.2023.2257217.

[78] Samuel N. Cohen and Martin Tegnér. "European Option Pricing with Stochastic Volatility Models Under Parameter Uncertainty". In: *Frontiers in Stochastic Analysis–BSDEs, SPDEs and their Applications*. Springer International Publishing, 2019, pp. 123–167. ISBN: 9783030222857. DOI: 10.1007/978-3-030-22285-7_5.

[79] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer London, 2001. ISBN: 9781447136750. DOI: 10.1007/978-1-4471-3675-0.

[80] Rama Cont. "Empirical properties of asset returns: stylized facts and statistical issues". In: *Quantitative Finance* 1.2 (2001), pp. 223–236. DOI: 10.1080/713665670.

[81] Rama Cont. "Model uncertainty and its impact on the pricing of derivative instruments". In: *Math. Finance* 16.3 (2006), pp. 519–547. DOI: 10.1111/j.1467-9965.2006.00281.x.

[82] D. Crisan, J. Diehl, P. K. Friz, and H. Oberhauser. "Robust filtering: correlated noise and multidimensional observation". In: *Ann. Appl. Probab.* 23.5 (2013), pp. 2139–2160. DOI: 10.1214/12-AAP896.

[83] Christa Cuchiero, Guido Gazzani, and Irene Klein. *Risk measures under model uncertainty: a Bayesian viewpoint*. 2022. arXiv: 2204.07115 [q-fin.RM]. URL: https://arxiv.org/abs/2204.07115.

[84] Christa Cuchiero, Wahid Khosrawi, and Josef Teichmann. "A Generative Adversarial Network Approach to Calibration of Local Stochastic Volatility Models". In: *Risks* 8.4 (2020). DOI: 10.3390/risks8040101.

[85] Christa Cuchiero, Martin Larsson, and Josef Teichmann. "Deep neural networks, generic universal interpolation, and controlled ODEs". In: *SIAM J. Math. Data Sci.* 2.3 (2020), pp. 901–919. DOI: 10.1137/19M1284117.

[86] Theo Darsinos and Stephen Satchell. "Bayesian forecasting of options prices". In: *Forecasting Expected Returns in the Financial Markets*. Elsevier, 2007, pp. 151–175. ISBN: 9780750683210. DOI: 10.1016/b978-075068321-0.50008-x.

[87] P. Date and K. Ponomareva. "Linear and non-linear filtering in mathematical finance: a review". In: *IMA Journal of Management Mathematics* 22.3 (June 2010), pp. 195–211. DOI: 10.1093/imaman/dpq008.

[88] M. H. A. Davis. "On a multiplicative functional transformation arising in nonlinear filtering theory". In: *Z. Wahrsch. Verw. Gebiete* 54.2 (1980), pp. 125–139. DOI: 10.1007/BF00531444.

[89] M. H. A. Davis and M. P. Spathopoulos. "Pathwise nonlinear filtering for nondegenerate diffusions with noise correlation". In: *SIAM J. Control Optim.* 25.2 (1987), pp. 260–278. DOI: 10.1137/0325016.

[90] DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, et al. *DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model*. 2024. arXiv: 2405.04434 [cs.CL]. URL: https://arxiv.org/abs/2405.04434.

[91] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, et al. *DeepSeek-V3 Technical Report*. 2025. arXiv: 2412.19437 [cs.CL]. URL: https://arxiv.org/abs/2412.19437.

[92] Pierre Del Moral. "Nonlinear filtering: Interacting particle resolution". In: *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 325.6 (Sept. 1997), pp. 653–658. DOI: 10.1016/s0764-4442(97)84778-7.

[93] Pierre Del Moral. *Mean field simulation for Monte Carlo integration*. Vol. 126. Monographs on Statistics and Applied Probability. CRC Press, Boca Raton, FL, 2013, pp. xlvii+578. ISBN: 978-1-4665-0405-9.

[94] P.M. Djuric, J.H. Kotecha, Jianqui Zhang, Yufei Huang, et al. "Particle filtering". In: *IEEE Signal Processing Magazine* 20.5 (2003), pp. 19–38. DOI: 10.1109/MSP.2003.1236770.

[95] Yan Dolinsky and H. Mete Soner. “Martingale optimal transport and robust hedging in continuous time”. In: *Probab. Theory Related Fields* 160.1-2 (2014), pp. 391–427. DOI: 10.1007/s00440-013-0531-y.

[96] D. C. Dowson and B. V. Landau. “The Fréchet distance between multivariate normal distributions”. In: *J. Multivariate Anal.* 12.3 (1982), pp. 450–455. DOI: 10.1016/0047-259X(82)90077-X.

[97] Fayçal Drissi. “Solvability of Differential Riccati Equations and Applications to Algorithmic Trading with Signals”. In: *Applied Mathematical Finance* 29.6 (2022), pp. 457–493. DOI: 10.1080/1350486X.2023.2241130.

[98] Jin-Chuan Duan and Jean-Guy Simonato. “Estimating and Testing Exponential-Affine Term Structure Models by Kalman Filter”. In: *Review of Quantitative Finance and Accounting* 13.2 (Sept. 1999), pp. 111–135. DOI: 10.1023/a:1008304625054.

[99] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge, 2009, pp. xvi+196. ISBN: 978-0-521-88427-3. DOI: 10.1017/CBO9780511581274.

[100] Moritz Duembgen and L. C. G. Rogers. “Estimate nothing”. In: *Quantitative Finance* 14.12 (Nov. 2014), pp. 2065–2072. DOI: 10.1080/14697688.2014.951678.

[101] Paul Dupuis, Matthew R. James, and Ian Petersen. “Robust Properties of Risk-Sensitive Control”. In: *Mathematics of Control, Signals, and Systems* 13.4 (Dec. 2000), pp. 318–332. DOI: 10.1007/pl00009872.

[102] Bradley Efron. “Bootstrap Methods: Another Look at the Jackknife”. In: *Annals of Statistics* 7.1 (1979), pp. 1–26. DOI: 10.1214/aos/1176344552.

[103] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993. DOI: 10.1201/9780429246593.

[104] Nicole El Karoui, Monique Jeanblanc-Picqué, and Steven E. Shreve. “Robustness of the Black and Scholes formula”. In: *Math. Finance* 8.2 (1998), pp. 93–126. DOI: 10.1111/1467-9965.00047.

[105] Robert J. Elliott and Cody. B. Hyndman. “Parameter estimation in commodity markets: A filtering approach”. In: *Journal of Economic Dynamics and Control* 31.7 (July 2007), pp. 2350–2373. DOI: 10.1016/j.jedc.2006.07.005.

[106] William Fedus, Barret Zoph, and Noam Shazeer. “Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity”. In: *Journal of Machine Learning Research* 23.120 (2022), pp. 1–39. URL: http://jmlr.org/papers/v23/21-0998.html.

[107] Robert Fildes, Andrew C. Harvey, Mike West, and Jeff Harrison. “Forecasting, Structural Time Series Models and the Kalman Filter”. In: *The Journal of the Operational Research Society* 42.11 (Nov. 1991), p. 1031. DOI: 10.2307/2583225.

[108] Hans Föllmer and Alexander Schied. “Convex Measures of Risk and Trading Constraints”. In: *Finance and Stochastics* 6.4 (2002), pp. 429–447. DOI: 10.1007/s007800200072.

[109] Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. Berlin, Boston: De Gruyter, 2004. ISBN: 9783110212075. DOI: doi:10.1515/9783110212075.

[110] Hans Föllmer and Alexander Schied. "Probabilistic aspects of finance". In: *Bernoulli* 19.4 (2013), pp. 1306–1326. DOI: 10.3150/12-BEJSP05.

[111] Hans Föllmer and Alexander Schied. *Stochastic finance*. De Gruyter Graduate. An introduction in discrete time, Fourth revised and extended edition of [MR1925197]. De Gruyter, Berlin, 2016, pp. xii+596. ISBN: 978-3-11-046345-3. DOI: 10.1515/9783110463453.

[112] Philip Hans Franses and Dick van Dijk. *Nonlinear Time Series Models in Empirical Finance*. Cambridge, UK: Cambridge University Press, 2000. ISBN: 0-521-77041-6.

[113] Rüdiger Frey and Wolfgang Runggaldier. "Pricing credit derivatives under incomplete information: a nonlinear-filtering approach". In: *Finance and Stochastics* 14.4 (July 2010), pp. 495–526. DOI: 10.1007/s00780-010-0129-5.

[114] Rüdiger Frey and Thorsten Schmidt. "Pricing and hedging of credit derivatives via the innovations approach to nonlinear filtering". In: *Finance Stoch.* 16.1 (2012), pp. 105–133. DOI: 10.1007/s00780-011-0153-0.

[115] Rüdiger Frey and Carlos A. Sin. "Bounds on European option prices under stochastic volatility". In: *Math. Finance* 9.2 (1999), pp. 97–116. DOI: 10.1111/1467-9965.00064.

[116] Peter Friz and Nicolas Victoir. *Master course notes on Rough Paths*. See Theorem 5.25 and related discussion. TU Berlin, 2009. URL: https://page.math.tu-berlin.de/~friz/master4_May6th.pdf.

[117] Luca Galimberti, Anastasis Kratsios, and Giulia Livieri. *Designing Universal Causal Deep Learning Models: The Case of Infinite-Dimensional Dynamical Systems from Stochastic Analysis*. 2025. arXiv: 2210.13300 [math.DS]. URL: https://arxiv.org/abs/2210.13300.

[118] Alireza Ghahtarani, Ahmed Saif, and Alireza Ghasemi. "Robust Portfolio Selection Problems: A Comprehensive Review". In: *Operational Research* 22.4 (2022), pp. 3203–3264. DOI: 10.1007/s12351-022-00690-5.

[119] Patryk Gierjatowicz, Marc Sabate-Vidales, David Šiška, Łukasz Szpruch, and Žan Žurič. "Robust Pricing and Hedging via Neural Stochastic Differential Equations". In: *Journal of Computational Finance* 26.3 (2022), pp. 1–32. DOI: 10.21314/JCF.2022.025.

[120] Itzhak Gilboa and David Schmeidler. "Maxmin expected utility with nonunique prior". In: *J. Math. Econom.* 18.2 (1989), pp. 141–153. DOI: 10.1016/0304-4068(89)90018-9.

[121] Clark R. Givens and Rae Michael Shortt. "A class of Wasserstein metrics for probability distributions". In: *Michigan Math. J.* 31.2 (1984), pp. 231–240. DOI: 10.1307/mmj/1029003026.

[122] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer New York, 2003. ISBN: 9780387216171. DOI: 10.1007/978-0-387-21617-1.

[123] Paul Glasserman and Xunyu Xu. "Robust Portfolio Control With Stochastic Factor Dynamics". In: *Operations Research* 62.4 (2014), pp. 812–832. DOI: 10.1287/opre.2014.1280.

[124] Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. "Approximation bounds for random neural networks and reservoir systems". In: *The Annals of Applied Probability* 33.1 (Feb. 2023). DOI: 10.1214/22-aap1806.

[125] Lukas Gonon and Josef Teichmann. “Linearized filtering of affine processes using stochastic Riccati equations”. In: *Stochastic Process. Appl.* 130.1 (2020), pp. 394–430. DOI: 10.1016/j.spa.2019.03.016.

[126] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[127] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, et al. “Generative adversarial nets”. In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[128] Google. *Gemini*. Google. URL: https://gemini.google.com/.

[129] Mikhael Gromov. “Filling Riemannian manifolds”. In: *Journal of Differential Geometry* 18.1 (Jan. 1983). DOI: 10.4310/jdg/1214509283.

[130] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, et al. *DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence*. 2024. arXiv: 2401.14196 [cs.SE]. URL: https://arxiv.org/abs/2401.14196.

[131] Ivan Guo, Nicolas Langrené, Gregoire Loeper, and Wei Ning. “Robust Utility Maximization Under Model Uncertainty via a Penalization Approach”. In: *SSRN Electronic Journal* (2020). DOI: 10.2139/ssrn.3612503.

[132] A. Gupta. *A Bayesian approach to financial model calibration, uncertainty measures and optimal hedging*. University of Oxford, 2010. URL: https://books.google.de/books?id=MZBD0AEACAAJ.

[133] Nikhil Gupta and Jason Yip. “DBRX: Creating an LLM from Scratch Using Databricks”. In: *Databricks Data Intelligence Platform*. Apress, 2024, pp. 311–330. ISBN: 9798868804441. DOI: 10.1007/979-8-8688-0444-1_13.

[134] Patrick Hagan, Deep Kumar, Andrew Lesniewski, and Diana Woodward. “Managing Smile Risk”. In: *Wilmott Magazine* 1 (Jan. 2002), pp. 84–108. URL: http://www.wilmott.com/pdfs/021118-smile.eps.

[135] Ben Hambly and Terry Lyons. “Uniqueness for the signature of a path of bounded variation and the reduced path group”. In: *Ann. of Math. (2)* 171.1 (2010), pp. 109–167. DOI: 10.4007/annals.2010.171.109.

[136] James D. Hamilton. “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle”. In: *Econometrica* 57.2 (Mar. 1989), p. 357. DOI: 10.2307/1912559.

[137] Lars Peter Hansen and Thomas J. Sargent. “Robustness and ambiguity in continuous time”. In: *Journal of Economic Theory* 146.3 (May 2011), pp. 1195–1223. DOI: 10.1016/j.jet.2011.01.004.

[138] LarsPeter Hansen and Thomas J. Sargent. “Robust Control and Model Uncertainty”. In: *American Economic Review* 91.2 (May 2001), pp. 60–66. DOI: 10.1257/aer.91.2.60.

[139] Jian He, Asma Khedher, and Peter Spreij. *A Kalman particle filter for online parameter estimation with applications to affine models*. 2019. arXiv: 1905.08552 [stat.CO]. URL: https://arxiv.org/abs/1905.08552.

[140] Juha Heinonen. *Lectures on analysis on metric spaces*. Universitext. Springer-Verlag, New York, 2001, pp. x+140. ISBN: 0-387-95104-0. DOI: 10.1007/978-1-4613-0131-8.

[141] Jakob Heiss, Florian Krach, Thorsten Schmidt, and Félix B. Tambe-Ndonfack. *Nonparametric Filtering, Estimation and Classification using Neural Jump ODEs*. 2025. arXiv: 2412.03271 [stat.ML]. URL: https://arxiv.org/abs/2412.03271.

[142] Stanislav Hencl. "Isometrical embeddings of separable Banach spaces into the set of nowhere approximatively differentiable and nowhere Hölder functions". In: *Proc. Amer. Math. Soc.* 128.12 (2000), pp. 3505–3511. DOI: 10.1090/S0002-9939-00-05595-7.

[143] Calypso Herrera, Florian Krach, Anastasis Kratsios, Pierre Ruyssen, and Josef Teichmann. "Denise: Deep Robust Principal Component Analysis for Positive Semidefinite Matrices". In: *Transactions on Machine Learning Research* (2023). ISSN: 2835-8856. URL: https://openreview.net/forum?id=D45gGvUZp2.

[144] Calypso Herrera, Florian Krach, and Josef Teichmann. "Neural Jump Ordinary Differential Equations: Consistent Continuous-Time Prediction and Filtering". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=JFKR3WqwyXR.

[145] Sebastian Herrmann, Johannes Muhle-Karbe, and Frank Thomas Seifried. "Hedging with small uncertainty aversion". In: *Finance Stoch.* 21.1 (2017), pp. 1–64. DOI: 10.1007/s00780-016-0309-z.

[146] Steven L. Heston. "A closed-form solution for options with stochastic volatility with applications to bond and currency options". In: *Rev. Financ. Stud.* 6.2 (1993), pp. 327–343. DOI: 10.1093/rfs/6.2.327.

[147] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. "Bayesian model averaging: a tutorial". In: *Statist. Sci.* 14.4 (1999). With comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors, pp. 382–417. DOI: 10.1214/ss/1009212519.

[148] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. 2nd ed. Cambridge University Press, 2012. DOI: 10.1017/CBO9781139020411.

[149] Blanka Horvath, Zacharia Issa, and Aitor Muguruza. *Clustering Market Regimes using the Wasserstein Distance*. 2021. arXiv: 2110.11848 [q-fin.CP]. URL: https://arxiv.org/abs/2110.11848.

[150] Blanka Horvath, Anastasis Kratsios, Yannick Limmer, and Xuwei Yang. *Transformers Can Solve Non-Linear and Non-Markovian Filtering Problems in Continuous Time For Conditionally Gaussian Signals*. 2025. arXiv: 2310.19603 [cs.LG]. URL: https://arxiv.org/abs/2310.19603.

[151] Blanka Horvath, Aitor Muguruza, and Mehdi Tomas. "Deep learning volatility: a deep neural network perspective on pricing and calibration in (rough) volatility models". In: *Quantitative Finance* 21.1 (2021), pp. 11–27. DOI: 10.1080/14697688.2020.1817974.

[152] Siavash Hosseinyalamdary. "Deep Kalman Filter: Simultaneous Multi-Sensor Integration and Modelling; A GNSS/IMU Case Study". In: *Sensors* 18.5 (2018). DOI: 10.3390/s18051316.

[153] Songyan Hou, Parnian Kassraie, Anastasis Kratsios, Andreas Krause, and Jonas Rothfuss. "Instance-dependent generalization bounds via optimal transport". In: *J. Mach. Learn. Res.* 24 (2023), Paper No. [349], 51. ISSN: 1532-4435,1533-7928.

[154] Aytaç İlhan, Mattias Jonsson, and Ronnie Sircar. "Optimal static-dynamic hedges for exotic options under convex risk measures". In: *Stochastic Process. Appl.* 119.10 (2009), pp. 3608–3632. DOI: 10.1016/j.spa.2009.06.009.

[155] Zacharia Issa, Blanka Horvath, Maud Lemercier, and Cristopher Salvi. "Non-adversarial training of Neural SDEs with signature kernel scores". In: *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*. 2023. URL: https://papers.nips.cc/paper_files/paper/2023/file/2460396f2d0d421885997dd1612ac56b-Paper-Conference.pdf.

[156] Garud Iyengar, Henry Lam, and Tianyu Wang. "Hedging against Complexity: Distributionally Robust Optimization with Parametric Approximation". In: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 206. Proceedings of Machine Learning Research. PMLR, Apr. 2023, pp. 9976–10011. URL: https://proceedings.mlr.press/v206/iyengar23a.html.

[157] Garud N. Iyengar. "Robust Dynamic Programming". In: *Mathematics of Operations Research* 30.2 (May 2005), pp. 257–280. DOI: 10.1287/moor.1040.0129.

[158] Eric Jacquier and Robert Jarrow. "Bayesian analysis of contingent claim model error". In: *Journal of Econometrics* 94.1–2 (Jan. 2000), pp. 145–180. DOI: 10.1016/s0304-4076(99)00020-2.

[159] Sebastian Jaimungal. "Reinforcement learning and stochastic optimisation". In: *Finance and Stochastics* 26.1 (Dec. 2021), pp. 103–129. DOI: 10.1007/s00780-021-00467-2.

[160] Sebastian Jaimungal, Silvana M. Pesenti, Ye Sheng Wang, and Hariom Tatsat. "Robust Risk-Aware Reinforcement Learning". In: *SIAM Journal on Financial Mathematics* 13.1 (2022), pp. 213–226. DOI: 10.1137/21M144640X.

[161] Alireza Javaheri, Delphine Lautier, and Alain Galli. "Filtering in Finance". In: *Wilmott* 2003.3 (May 2003), pp. 67–83. DOI: 10.1002/wilm.42820030315.

[162] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, et al. *Mixtral of Experts*. 2024. arXiv: 2401.04088 [cs.LG]. URL: https://arxiv.org/abs/2401.04088.

[163] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. *A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem*. 2017. arXiv: 1706.10059 [q-fin.CP]. URL: https://arxiv.org/abs/1706.10059.

[164] Michael Johannes and Nicholas Polson. "MCMC Methods for Continuous-Time Financial Econometrics (Chapter 13)". In: *Handbook of Financial Econometrics: Applications*. Vol. 2. Handbooks in Finance. San Diego: Elsevier, 2010, pp. 1–72. DOI: https://doi.org/10.1016/B978-0-444-53548-1.50003-9.

[165] William B. Johnson and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space". In: *Conference in modern analysis and probability (New Haven, Conn., 1982)*. Vol. 26. Contemp. Math. Amer. Math. Soc., Providence, RI, 1984, pp. 189–206. DOI: 10.1090/conm/026/737400.

[166] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems". In: *Journal of Basic Engineering* 82.1 (Mar. 1960), pp. 35–45. DOI: 10.1115/1.3662552.

[167] R. E. Kalman and R. S. Bucy. "New Results in Linear Filtering and Prediction Theory". In: *Journal of Basic Engineering* 83.1 (Mar. 1961), pp. 95–108. DOI: 10.1115/1.3658902.

[168] Karin Usadi Katz and Mikhail G. Katz. "Bi-Lipschitz approximation by finite-dimensional imbeddings". In: *Geometriae Dedicata* 150.1 (Apr. 2010), pp. 131–136. DOI: 10.1007/s10711-010-9497-4.

[169] Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. "Deep Signature Transforms". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/d2cdf047a6674cef251d56544a3cf029-Paper.pdf.

[170] Patrick Kidger and Terry Lyons. "Universal Approximation with Deep Narrow Networks". In: *Proceedings of Thirty Third Conference on Learning Theory*. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 2306–2327. URL: https://proceedings.mlr.press/v125/kidger20a.html.

[171] Patrick Kidger and Terry Lyons. *Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU*. 2021. arXiv: 2001.00706 [cs.LG]. URL: https://arxiv.org/abs/2001.00706.

[172] Peter Klibanoff, Massimo Marinacci, and Sujoy Mukerji. "A Smooth Model of Decision Making under Ambiguity". In: *Econometrica* 73.6 (2005), pp. 1849–1892. URL: https://EconPapers.repec.org/RePEc:ecm:emetrp:v:73:y:2005:i:6:p:1849-1892.

[173] Susanne Klöppel and Martin Schweizer. "Dynamic indifference valuation via convex risk measures". In: *Math. Finance* 17.4 (2007), pp. 599–627. DOI: 10.1111/j.1467-9965.2007.00317.x.

[174] Frank Hyneman Knight. *Risk, uncertainty and profit*. Vol. 31. Hart, Schaffner & Marx Price Essays. Boston and New York, Houghton Mifflin Company, 1921, p. 381.

[175] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. "On Convergence and Stability of GANs". In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://openreview.net/forum?id=ryepFJbA-.

[176] Alexei Kondratyev and Christian Schwarz. "The Market Generator". In: *Journal of Financial Data Science* 2.3 (2020), pp. 78–87. DOI: 10.3905/jfds.2020.1.049.

[177] Florian Krach, Marc Nübel, and Josef Teichmann. *Optimal Estimation of Generic Dynamics by Path-Dependent Neural Jump ODEs*. 2024. arXiv: 2206.14284 [stat.ML]. URL: https://arxiv.org/abs/2206.14284.

[178] Florian Krach, Josef Teichmann, and Hanna Wutte. *Robust Utility Optimization via a GAN Approach*. 2024. arXiv: 2403.15243 [q-fin.CP]. URL: https://arxiv.org/abs/2403.15243.

[179] Dmitry Kramkov and Mihai Sîrbu. "Asymptotic analysis of utility-based hedging strategies for small number of contingent claims". In: *Stochastic Processes and their Applications* 117.8 (2007), pp. 1606–1620. DOI: 10.1016/j.spa.2007.04.014.

[180] Anastasis Kratsios. "Universal Regular Conditional Distributions via Probabilistic Transformers". In: *Constructive Approximation* 57.3 (Mar. 2023), pp. 1145–1212. DOI: 10.1007/s00365-023-09635-3.

[181] Anastasis Kratsios, Valentin Debarnot, and Ivan Dokmanić. "Small transformers compute universal metric embeddings". In: *J. Mach. Learn. Res.* 24 (2023), Paper No. [170], 48. ISSN: 1532-4435,1533-7928.

[182] Anastasis Kratsios, Chong Liu, Matti Lassas, Maarten V. de Hoop, and Ivan Dokmanić. *An Approximation Theory for Metric Space-Valued Functions With A View Towards Deep Learning*. 2023. arXiv: 2304.12231 [cs.LG]. URL: https://arxiv.org/abs/2304.12231.

[183] Anastasis Kratsios and Leonie Papon. "Universal Approximation Theorems for Differentiable Geometric Deep Learning". In: *Journal of Machine Learning Research* 23.196 (2022), pp. 1–73. URL: http://jmlr.org/papers/v23/21-0716.html.

[184] Rahul Krishnan, Uri Shalit, and David Sontag. "Structured Inference Networks for Nonlinear State Space Models". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (Feb. 2017). DOI: 10.1609/aaai.v31i1.10779.

[185] Rahul G. Krishnan, Uri Shalit, and David Sontag. *Deep Kalman Filters*. 2015. arXiv: 1511.05121 [stat.ML]. URL: https://arxiv.org/abs/1511.05121.

[186] Harold J. Kushner. "A robust discrete state approximation to the optimal nonlinear filter for a diffusion". In: *Stochastics* 3.2 (1979), pp. 75–83. ISSN: 0090-9491.

[187] Peter Lakner. "Optimal trading strategy for an investor: the case of partial information". In: *Stochastic Processes and their Applications* 76.1 (Aug. 1998), pp. 77–97. DOI: 10.1016/s0304-4149(98)00032-5.

[188] Samuel Lanthaler and Andrew M. Stuart. *The Parametric Complexity of Operator Learning*. 2025. arXiv: 2306.15924 [cs.LG]. URL: https://arxiv.org/abs/2306.15924.

[189] Delphine Lautier * and Alain Galli. "Simple and extended Kalman filters: an application to term structures of commodity prices". In: *Applied Financial Economics* 14.13 (Sept. 2004), pp. 963–973. DOI: 10.1080/0960310042000233629.

[190] Peter D. Lax. *Linear algebra and its applications*. second. Pure and Applied Mathematics (Hoboken). Wiley-Interscience, Hoboken, NJ, 2007, pp. xvi+376. ISBN: 978-0-471-75156-4.

[191] Yunong Lee and Andrew J. Majda. "State estimation and prediction using clustered particle filters". In: *Proceedings of the National Academy of Sciences* 113.51 (2016), pp. 14609–14614. DOI: 10.1073/pnas.1617398113.

[192] Daniel Levin, Terry Lyons, and Hao Ni. *Learning from the past, predicting the statistics for the future, learning an evolving system*. 2016. arXiv: 1309.0260 [q-fin.ST]. URL: https://arxiv.org/abs/1309.0260.

[193] Pingzhi Li, Zhenyu Zhang, Prateek Yadav, Yi-Lin Sung, et al. "Merge, Then Compress: Demystify Efficient SMoE with Hints from Its Routing Policy". In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=eFWG9Cy3WK.

[194] Shujian Liao, Terry Lyons, Weixin Yang, and Hao Ni. *Learning stochastic differential equations using RNN with log signature features*. 2019. arXiv: 1908.08286 [cs.LG]. URL: https://arxiv.org/abs/1908.08286.

[195] Shujian Liao, Hao Ni, Lukasz Szpruch, Magnus Wiese, Marc Sabate-Vidales, and Baoren Xiao. *Conditional Sig-Wasserstein GANs for Time Series Generation*. 2023. DOI: https://doi.org/10.1111/mafi.12423.

[196] Yuan Liao, Xinjie Ma, Andreas Neuhierl, and Linda Schilling. *The Uncertainty of Machine Learning Predictions in Asset Pricing*. 2025. arXiv: 2503.00549 [econ.EM]. URL: https://arxiv.org/abs/2503.00549.

[197] Gabriele Lillacci and Mustafa Khammash. "Parameter Estimation and Model Selection in Computational Biology". In: *PLoS Computational Biology* 6.3 (Mar. 2010), e1000696. DOI: 10.1371/journal.pcbi.1000696.

[198] Yannick Limmer and Blanka Horvath and. "Robust Hedging GANs: Towards Automated Robustification of Hedging Strategies". In: *Applied Mathematical Finance* 31.3 (2024), pp. 164–201. DOI: 10.1080/1350486X.2024.2440661.

[199] Yannick Limmer, Anastasis Kratsios, Xuwei Yang, Raeid Saqur, and Blanka Horvath. *Higher-Order Transformer Derivative Estimates for Explicit Pathwise Learning Guarantees*. 2025. arXiv: 2405.16563 [cs.LG]. URL: https://arxiv.org/abs/2405.16563.

[200] Erik Lindström, Jonas Ströjby, Mats Brodén, Magnus Wiktorsson, and Jan Holst. "Sequential calibration of options". In: *Computational Statistics & Data Analysis* 52.6 (Feb. 2008), pp. 2877–2891. DOI: 10.1016/j.csda.2007.08.009.

[201] Robert S. Liptser and Albert N. Shiryaev. *Statistics of random processes. II*. expanded. Vol. 6. Applications of Mathematics (New York). Applications, Translated from the 1974 Russian original by A. B. Aries, Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2001, pp. xvi+402. ISBN: 3-540-63928-4.

[202] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Zhiyong Gao, and Ming-Ting Sun. "Deep Kalman Filtering Network for Video Compression Artifact Reduction". In: *Computer Vision – ECCV 2018*. Springer International Publishing, 2018, pp. 591–608. ISBN: 9783030012649. DOI: 10.1007/978-3-030-01264-9_35.

[203] Guo Lu, Xiaoyun Zhang, Wanli Ouyang, Dong Xu, Li Chen, and Zhiyong Gao. "Deep Non-Local Kalman Network for Video Compression Artifact Reduction". In: *IEEE Transactions on Image Processing* 29 (2019), pp. 1725–1737. DOI: 10.1109/TIP.2019.2943214.

[204] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. "Deep network approximation for smooth functions". In: *SIAM J. Math. Anal.* 53.5 (2021), pp. 5465–5506. DOI: 10.1137/20M134695X.

[205] Eva Lütkebohmert, Thorsten Schmidt, and Julian Sester. "Robust deep hedging". In: *Quant. Finance* 22.8 (2022), pp. 1465–1480. DOI: 10.1080/14697688.2022.2056073.

[206] Terry Lyons. "Rough paths, signatures and the modelling of functions on streams". In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. IV*. Kyung Moon Sa, Seoul, 2014, pp. 163–184.

[207] Terry Lyons and Hao Ni. "Expected signature of Brownian motion up to the first exit time from a bounded domain". In: *Ann. Probab.* 43.5 (2015), pp. 2729–2762. DOI: 10.1214/14-AOP949.

[208] Fabio Maccheroni, Massimo Marinacci, and Aldo Rustichini. "Ambiguity aversion, robustness, and the variational representation of preferences". In: *Econometrica* 74.6 (2006), pp. 1447–1498. DOI: 10.1111/j.1468-0262.2006.00716.x.

[209] Pascal J. Maenhout. "Robust Portfolio Rules and Asset Pricing". In: *Review of Financial Studies* 17.4 (2004), pp. 951–983. DOI: 10.1093/rfs/hhh003.

[210] Mihaela Manoliu and Stathis Tompaidis. "Energy futures prices: term structure models with Kalman filter estimation". In: *Applied Mathematical Finance* 9.1 (Mar. 2002), pp. 21–43. DOI: 10.1080/13504860210126227.

[211] Carlo Marcati, Joost A. A. Opschoor, Philipp C. Petersen, and Christoph Schwab. "Exponential ReLU Neural Network Approximation Rates for Point and Edge Singularities". In: *Foundations of Computational Mathematics* 23.3 (June 2022), pp. 1043–1127. DOI: 10.1007/s10208-022-09565-9.

[212] T. Magri Materossi. "Il problema dei fili di Lecher". In: *Il Nuovo Cimento* 1.3 (June 1943), pp. 190–204. DOI: 10.1007/bf02958691.

[213] Alexander J. McNeil and Rüdiger Frey. "Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach". In: *Journal of Empirical Finance* 7.3–4 (Nov. 2000), pp. 271–300. DOI: 10.1016/s0927-5398(00)00012-8.

[214] Nicolai Meinshausen. "Causality from a Distributional Robustness Point of View". In: *Proceedings of the IEEE Data Science Workshop (DSW)*. 2018, pp. 6–10. DOI: 10.1109/DSW.2018.8439889.

[215] Robert C. Merton. "Lifetime Portfolio Selection under Uncertainty: The Continuous-Time Case". In: *The Review of Economics and Statistics* 51.3 (Aug. 1969), p. 247. DOI: 10.2307/1926560.

[216] Robert C. Merton. "Theory of Rational Option Pricing". In: *The Bell Journal of Economics and Management Science* 4.1 (1973), pp. 141–183. DOI: 10.2307/3003143.

[217] Richard O. Michaud. "The Markowitz Optimization Enigma: is "Optimized" Optimal?" In: *SSRN Electronic Journal* (2014). DOI: 10.2139/ssrn.2387669.

[218] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. 2014. arXiv: 1411.1784 [cs.LG]. URL: https://arxiv.org/abs/1411.1784.

[219] Peyman Mohajerin Esfahani and Daniel Kuhn. "Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations". In: *Mathematical Programming* 171.1–2 (July 2017), pp. 115–166. DOI: 10.1007/s10107-017-1172-1.

[220] James Morrill, Adeline Fermanian, Patrick Kidger, and Terry Lyons. *A Generalised Signature Method for Multivariate Time Series Feature Extraction*. 2021. arXiv: 2006.00873 [cs.LG]. URL: https://arxiv.org/abs/2006.00873.

[221] James R. Munkres. *Topology*. Second edition of [MR0464128]. Prentice Hall, Inc., Upper Saddle River, NJ, 2000, pp. xvi+537. ISBN: 0-13-181629-2.

[222] Phillip Murray, Ben Wood, Hans Buehler, Magnus Wiese, and Mikko S. Pakkanen. "Deep Hedging: Continuous Reinforcement Learning for Hedging of General Portfolios across Multiple Risk Aversions". In: *Proceedings of the 3rd ACM International Conference on AI in Finance (ICAIF 2022)*. 2022. DOI: 10.1145/3533271.3561731.

[223] Hongseok Namkoong and John C Duchi. "Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/4588e674d3f0faf985047d4c3f13ed0d-Paper.pdf.

[224] Ariel Neufeld. "Buy-and-hold property for fully incomplete markets when super-replicating Markovian claims". In: *Int. J. Theor. Appl. Finance* 21.8 (2018), pp. 1850051, 12. DOI: 10.1142/S0219024918500516.

[225] Ariel Neufeld and Marcel Nutz. "Superreplication under volatility uncertainty for measurable claims". In: *Electron. J. Probab.* 18 (2013), no. 48, 14. DOI: 10.1214/EJP.v18-2358.

[226] A. Martina Neuman, Yuying Xie, and Qiang Sun. "Restricted Riemannian geometry for positive semidefinite matrices". In: *Linear Algebra and its Applications* 665 (May 2023), pp. 153–195. DOI: 10.1016/j.laa.2023.01.026.

[227] John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. third. Princeton University Press, Princeton, N.J., 1980, pp. xx+641. ISBN: 0-691-00362-9.

[228] Hao Ni, Lukasz Szpruch, Marc Sabate-Vidales, Baoren Xiao, Magnus Wiese, and Shujian Liao. "Sig-wasserstein GANs for time series generation". In: *Proceedings of the Second ACM International Conference on AI in Finance*. ICAIF'21. ACM, Nov. 2021, pp. 1–8. DOI: 10.1145/3490354.3494393.

[229] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. "Deep Exploration via Bootstrapped DQN". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf.

[230] M. P. Owen. "Utility-based optimal hedging in incomplete markets". In: *Annals of Applied Probability* 12.2 (2002), pp. 605–644. DOI: 10.1214/aoap/1026915621.

[231] Sejun Park, Jaeho Lee, Chulhee Yun, and Jinwoo Shin. "Provable Memorization via Deep Neural Networks using Sub-linear Parameters". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Vol. 134. Proceedings of Machine Learning Research. PMLR, Aug. 2021, pp. 3627–3661. URL: https://proceedings.mlr.press/v134/park21a.html.

[232] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al. "PyTorch: an imperative style, high-performance deep learning library". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019. DOI: 10.1063/5.0063300.

[233] Gareth W. Peters, Mark Briers, Pavel V. Shevchenko, and Arnaud Doucet. "Calibration and filtering for multi factor commodity models with seasonality: Incorporating panel data from futures contracts". In: *Methodology and Computing in Applied Probability* 15.4 (Dec. 2013), pp. 841–874. DOI: 10.1007/s11009-012-9286-7.

[234] I.R. Petersen, M.R. James, and P. Dupuis. “Minimax optimal control of stochastic uncertain systems with relative entropy constraints”. In: *IEEE Transactions on Automatic Control* 45.3 (2000), pp. 398–412. DOI: 10.1109/9.847720.

[235] Guergana Petrova and Przemysław Wojtaszczyk. “Lipschitz widths”. In: *Constr. Approx.* 57.2 (2023), pp. 759–805. DOI: 10.1007/s00365-022-09576-3.

[236] Iosif Pinelis. *Local Lipschitzness of parameterization of Gaussians in Wasserstein space*. MathOverflow. URL:https://mathoverflow.net/q/449287 (version: 2023-06-22). 2020. URL: https://mathoverflow.net/q/449287.

[237] Jamshid Pirgazi and Ali Reza Khanteymoori. “A robust gene regulatory network inference method base on Kalman filter and linear regression”. In: *PLoS ONE* 13.7 (2018), e0200094. DOI: 10.1371/journal.pone.0200094.

[238] Marcin Pitera and Thorsten Schmidt. “Estimating and backtesting risk under heavy tails”. In: *Insurance: Mathematics and Economics* 104 (May 2022), pp. 1–14. DOI: 10.1016/j.insmatheco.2022.01.006.

[239] Marcin Pitera, Thorsten Schmidt, and Łukasz Stettner. *A novel scaling approach for unbiased adjustment of risk estimators*. 2023. arXiv: 2312.05655 [q-fin.RM]. URL: https://arxiv.org/abs/2312.05655.

[240] Dimitris N. Politis. “Scalable subsampling: computation, aggregation and inference”. In: *Biometrika* 111.1 (2024), pp. 347–354. DOI: 10.1093/biomet/asad021.

[241] Dimitris N. Politis and Joseph P. Romano. “Large sample confidence regions based on subsamples under minimal assumptions”. In: *Annals of Statistics* 22.4 (1994), pp. 2031–2050. DOI: 10.1214/aos/1176325770.

[242] Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer Series in Statistics. Springer, 1999. DOI: 10.1007/978-1-4612-1554-7.

[243] Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. “From Sparse to Soft Mixtures of Experts”. In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=jxpsAj7ltE.

[244] Alec Radford, Luke Metz, and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. arXiv: 1511.06434 [cs.LG]. URL: https://arxiv.org/abs/1511.06434.

[245] Hamed Rahimian and Sanjay Mehrotra. “Distributionally Robust Optimization: A Review”. In: *SIAM Review* 61.3 (2019), pp. 513–544. DOI: 10.1137/16M1082468.

[246] Hamed Rahimian and Sanjay Mehrotra. “Frameworks and Results in Distributionally Robust Optimization”. In: *Open Journal of Mathematical Optimization* 3 (July 2022), pp. 1–85. DOI: 10.5802/ojmo.15.

[247] Bruno Remillard. *Statistical Methods for Financial Engineering*. Chapman and Hall/CRC, Apr. 2016. ISBN: 9781439856956. DOI: 10.1201/b14285.

[248] Gerasimos G. Rigatos. *State-Space Approaches for Modelling and Control in Financial Engineering*. Springer International Publishing, 2017. ISBN: 9783319528663. DOI: 10.1007/978-3-319-52866-3.

[249] Herbert Robbins and Sutton Monro. “A Stochastic Approximation Method”. In: *The Annals of Mathematical Statistics* 22.3 (Sept. 1951), pp. 400–407. DOI: 10.1214/aoms/1177729586.

[250] James C. Robinson. *Dimensions, embeddings, and attractors*. Vol. 186. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 2011, pp. xii+205. ISBN: 978-0-521-89805-8.

[251] R Tyrrell Rockafellar, Stanislav Uryasev, et al. "Optimization of conditional value-at-risk". In: *Journal of risk* 2 (2000), pp. 21–42. DOI: 10.21314/JOR.2000.038.

[252] L. Rodríguez-Piazza. "Every separable Banach space is isometric to a space of continuous nowhere differentiable functions". In: *Proc. Amer. Math. Soc.* 123.12 (1995), pp. 3649–3654. DOI: 10.2307/2161889.

[253] Tom Ryder, Andrew Golightly, A. Stephen McGough, and Dennis Prangle. "Black-Box Variational Inference for Stochastic Differential Equations". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 4423–4432. URL: https://proceedings.mlr.press/v80/ryder18a.html.

[254] El Mehdi Saad, Nicolas Verzelen, and Alexandra Carpentier. "Active Ranking of Experts Based on their Performances in Many Tasks". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 29490–29513. URL: https://proceedings.mlr.press/v202/saad23b.html.

[255] Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. "The signature kernel is the solution of a Goursat PDE". In: *SIAM J. Math. Data Sci.* 3.3 (2021), pp. 873–899. DOI: 10.1137/20M1366794.

[256] R. Sameni, M.B. Shamsollahi, C. Jutten, and M. Babaie-Zade. "Filtering noisy ECG signals using the extended Kalman filter based on a modified dynamic ECG model". In: *Computers in Cardiology, 2005*. 2005, pp. 1017–1020. DOI: 10.1109/CIC.2005.1588283.

[257] Raeid Saqur, Anastasis Kratsios, Florian Krach, Yannick Limmer, et al. "Filtered not Mixed: Stochastic Filtering-Based Online Gating for Mixture of Large Language Models". In: *International Conference on Learning Representations (ICLR)*. 2025. URL: https://openreview.net/forum?id=ecIvumCyAj.

[258] Leonard J. Savage. "The foundations of statistics". In: *Naval Research Logistics Quarterly* 1.3 (1954), pp. 236–236. DOI: https://doi.org/10.1002/nav.3800010316.

[259] Eduardo Schwartz and James E. Smith. "Short-Term Variations and Long-Term Dynamics in Commodity Prices". In: *Management Science* 46.7 (July 2000), pp. 893–911. DOI: 10.1287/mnsc.46.7.893.12034.

[260] Eduardo S. Schwartz. "The Stochastic Behavior of Commodity Prices: Implications for Valuation and Hedging". In: *The Journal of Finance* 52.3 (July 1997), pp. 923–973. DOI: 10.1111/j.1540-6261.1997.tb02721.x.

[261] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[262] Steven E. Shreve. *Stochastic calculus for finance. I*. Springer Finance. The binomial asset pricing model. Springer-Verlag, New York, 2004, pp. xvi+187. ISBN: 0-387-40100-8.

[263] Steven E. Shreve. *Stochastic calculus for finance. II*. Springer Finance. Continuous-time models. Springer-Verlag, New York, 2004, pp. xx+550. ISBN: 0-387-40101-6.

[264] Jonathan W. Siegel and Jinchao Xu. "Characterization of the Variation Spaces Corresponding to Shallow Neural Networks". In: *Constructive Approximation* 57.3 (Feb. 2023), pp. 1109–1132. DOI: 10.1007/s00365-023-09626-4.

[265] Aman Sinha, Hongseok Namkoong, and John Duchi. "Certifiable Distributional Robustness with Principled Adversarial Training". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=Hk6kPgZA-.

[266] Stavros J. Sioutis. *Calibration and Filtering of Exponential Lévy Option Pricing Models*. 2017. arXiv: 1705.04780 [q-fin.PR]. URL: https://arxiv.org/abs/1705.04780.

[267] A. N. Sirjaev. "Some explicit formulae in a problem on "disorder"". In: *Teor. Verojatnost. i Primenen.* 10 (1965), pp. 380–385. ISSN: 0040-361x.

[268] A. N. Sirjaev. *Optimal stopping rules*. Vol. Vol. 8. Applications of Mathematics. Translated from the Russian by A. B. Aries. Springer-Verlag, New York-Heidelberg, 1978, pp. x+217. ISBN: 0-387-90256-2.

[269] Edwin H. Spanier. *Algebraic Topology*. Springer New York, 1981. ISBN: 9781468493221. DOI: 10.1007/978-1-4684-9322-1.

[270] Ruslan Leontyevich Stratonovich. "Application of the Markov processes theory to optimal filtering". In: *Radio Engineering and Electronic Physics* 5 (1960), pp. 1–19.

[271] Ruslan Leontyevich Stratonovich. "Optimum nonlinear systems which bring about a separation of a signal with constant parameters from noise". In: *Non-Linear Transformations of Stochastic Processes*. Elsevier, 1965, pp. 454–466. ISBN: 9781483232300. DOI: 10.1016/b978-1-4832-3230-0.50042-0.

[272] Brad Suchoski, Steve Stage, Heidi Gurung, and Prasith Baccam. "GPU Accelerated Parallel Processing for Large-Scale Monte Carlo Analysis: COVID-19 Parameter Estimation and New Case Forecasting". In: *Frontiers in Applied Mathematics and Statistics* 8 (2022). DOI: 10.3389/fams.2022.818016.

[273] Xian-He Sun and Lionel M. Ni. "Another View on Parallel Speedup". In: *Proceedings of the 1990 ACM/IEEE Conference on Supercomputing*. 1990, pp. 324–333. DOI: 10.1109/SUPERC.1990.130037.

[274] Mahdi Tabassian, Somayeh Akbari, Sandro Queirós, and Jan D'hooge. *Deep Spatiotemporal Clutter Filtering of Transthoracic Echocardiographic Images: Leveraging Contextual Attention and Residual Learning*. 2025. arXiv: 2401.13147 [eess.IV]. URL: https://arxiv.org/abs/2401.13147.

[275] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. *Policy Gradient for Coherent Risk Measures*. 2015. arXiv: 1502.03919 [cs.AI]. URL: https://arxiv.org/abs/1502.03919.

[276] Allan G. Timmermann and Massimo Guidolin. "Option Prices under Bayesian Learning: Implied Volatility Dynamics and Predictive Densities". In: *SSRN Electronic Journal* (2001). DOI: 10.2139/ssrn.264655.

[277] L. N. Vasershtein. "Markov processes over denumerable products of spaces describing large system of automata". In: *Problems Inform. Transmission* 5.3 (1969), pp. 47–52. DOI: 10.1016/s0016-0032(33)90010-1.

[278] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. *Graph Attention Networks*. 2018. arXiv: 1710.10903 [stat.ML]. URL: https://arxiv.org/abs/1710.10903.

[279] Roman Vershynin. "Memory Capacity of Neural Networks with Threshold and Rectified Linear Unit Activations". In: *SIAM Journal on Mathematics of Data Science* 2.4 (Jan. 2020), pp. 1004–1033. DOI: 10.1137/20m1314884.

[280] Cédric Villani. *Optimal transport*. Vol. 338. Grundlehren der mathematischen Wissenschaften. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973. ISBN: 978-3-540-71049-3. DOI: 10.1007/978-3-540-71050-9.

[281] Harrison Waldon, Fayçal Drissi, Yannick Limmer, Uljad Berdica, Jakob Nicolaus Foerster, and Alvaro Cartea. "DARE: The Deep Adaptive Regulator for Control of Uncertain Continuous-Time Systems". In: *ICML 2024 Workshop: Foundations of Reinforcement Learning and Control – Connections and Perspectives*. 2024. URL: https://openreview.net/forum?id=vTtldsicU9.

[282] Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994. ISBN: 9780429170591. DOI: https://doi.org/10.1201/b14876.

[283] Curt Wells. *The Kalman Filter in Finance*. Springer Netherlands, 1996. ISBN: 9789401586115. DOI: 10.1007/978-94-015-8611-5.

[284] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. "Distributionally Robust Convex Optimization". In: *Operations Research* 62.6 (2014), pp. 1358–1376. DOI: 10.1287/opre.2014.1314.

[285] Magnus Wiktorsson and Erik Lindström. "Fast Simultaneous Calibration and Quadratic Hedging under Parameter Uncertainty". In: *8th World Congress of the Bachelier Finance Society*. 2014. URL: https://bachelliercongress.com/2014/abstracts/all/1084.html.

[286] W. M. Wonham. "Some Applications of Stochastic Differential Equations to Optimal Nonlinear Filtering". In: *Journal of the Society for Industrial and Applied Mathematics Series A Control* 2.3 (Jan. 1964), pp. 347–369. DOI: 10.1137/0302028.

[287] David Wu and Sebastian Jaimungal. "Robust Risk-Aware Option Hedging". In: *Applied Mathematical Finance* 30.3 (2023), pp. 153–174. DOI: 10.1080/1350486X.2023.2301354.

[288] Mingxin Xu. "Risk measure pricing and hedging in incomplete markets". In: *Annals of Finance* 2.1 (Oct. 2005), pp. 51–71. DOI: 10.1007/s10436-005-0023-x.

[289] Dmitry Yarotsky and Anton Zhevnerchuk. "The phase diagram of approximation rates for deep neural networks". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc., 2020. ISBN: 9781713829546.

[290] Moshe Zakai. "On the optimal filtering of diffusion processes". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 11.3 (1969), pp. 230–243. DOI: 10.1007/bf00536382.

[291] Fengshan Zhang, Yongkui Zou, Shimin Chai, and Yanzhao Cao. “Numerical analysis of a time discretized method for nonlinear filtering problem with Lévy process observations”. In: *Advances in Computational Mathematics* 50.4 (July 2024). DOI: 10.1007/s10444-024-10169-w.

Appendices

As the statistician George E. P. Box wrote, "All models are wrong, but some models are useful." What he meant by that is that all models are simplifications of the universe, as they must necessarily be. As another mathematician said, "The best model of a cat is a cat." [...] The key is in remembering that a model is a tool to help us understand the complexities of the universe, and never a substitute for the universe itself."

— Nate Silver

A Appendix of Chapter 2

Contents

A.1 Proofs of Section 2.2.1 125
A.2 Rough path theory and distances of stochastic processes 127
A.2.1 The signature of a path 128
A.2.2 The expected signature of a stochastic process 129
A.2.3 Signature Wasserstein-1 metric (SigW_1) and Signature-MMD (SigMMD) 130
A.3 Setting of Herrmann, Muhle-Karbe, and Seifried 131
A.4 Out-of-sample Testing 132
A.5 Inverse Calibration in the Black-Scholes model 135

A.1 Proofs of Section 2.2.1

Proof of Theorem 2.2.1. Let $z \in [0,1]$, $\bar{z} := 1 - z$ and $\varphi, \bar{\varphi} \in \Phi$. Then,

$$\begin{aligned}\sup_{\mathbb{P}\in\mathcal{P}} \mathcal{L}(z\varphi + \bar{z}\bar{\varphi}, \mathbb{P}) &= \sup_{\mathbb{P}\in\mathcal{P}} \Big(U^{\mathbb{P}}\Big(z((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}}\Big) + \bar{z}((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}})\Big) - \alpha(\mathbb{P})\Big) \\ &\leq \sup_{\mathbb{P}\in\mathcal{P}} \Big(zU^{\mathbb{P}}\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}}\Big) + \bar{z}U^{\mathbb{P}}\Big((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}}\Big) - \alpha(\mathbb{P})\Big)\,, \end{aligned} \tag{A.1}$$

where we applied the convexity of $U^{\mathbb{P}}$ in the last step. Now, we observe that for functions $f, g : \mathcal{P} \to \mathbb{R}$ it holds that $\sup_{\mathbb{P}\in\mathcal{P}} f(\mathbb{P}) + g(\mathbb{P}) \leq \sup_{\mathbb{P}\in\mathcal{P}} f(\mathbb{P}) + \sup_{\mathbb{P}\in\mathcal{P}} g(\mathbb{P})$, hence,

$$(\text{A.1}) \leq z \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{L}(\varphi, \mathbb{P}) + \bar{z} \sup_{\mathbb{P}\in\mathcal{P}} \mathcal{L}(\bar{\varphi}, \mathbb{P}),$$

what proves the claim. □

Proof of Theorem 2.2.2. Assume that $\mathbf{X}_1, \mathbf{X}_2 \in \mathcal{X}$ and let $z \in [0,1]$ with $\bar{z} := 1 - z$. Then,

$$\pi(zX_1 + \bar{z}X_2) \tag{A.2}$$
$$= \inf_{\varphi \in \Phi} \left(\sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T + (z\mathbf{X}_1^{\mathbb{P}} + \bar{z}\mathbf{X}_2^{\mathbb{P}}) \right) - \alpha(\mathbb{P}) \right) \right)$$
$$= \inf_{\varphi, \bar{\varphi} \in \Phi} \left(\sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left(z(\varphi \cdot \mathbf{S}^{\mathbb{P}})_T + \bar{z}(\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T + (z\mathbf{X}_1^{\mathbb{P}} + \bar{z}\mathbf{X}_2^{\mathbb{P}}) \right) - \alpha(\mathbb{P}) \right) \right) \tag{A.3}$$

due to convexity of Φ. We then apply the convexity of $U^{\mathbb{P}}$ and see that

$$\text{(A.3)} \tag{A.4}$$
$$\leq \inf_{\varphi, \bar{\varphi} \in \Phi} \left(\sup_{\mathbb{P} \in \mathcal{P}} \left(zU^{\mathbb{P}} \left((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T + \mathbf{X}_1^{\mathbb{P}} \right) + \bar{z}U^{\mathbb{P}} \left((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T + \mathbf{X}_2^{\mathbb{P}} \right) - \alpha(\mathbb{P}) \right) \right)$$
$$= \inf_{\varphi, \bar{\varphi} \in \Phi} \left(\sup_{\mathbb{P} \in \mathcal{P}} \left(zU^{\mathbb{P}} \left((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T + \mathbf{X}_1^{\mathbb{P}} \right) - z\alpha(\mathbb{P}) + \bar{z}U^{\mathbb{P}} \left((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T + \mathbf{X}_2^{\mathbb{P}} \right) - \bar{z}\alpha(\mathbb{P}) \right) \right) \tag{A.5}$$

As before, we use the fact that for functions $f, g : \mathcal{P} \to \mathbb{R}$ it holds that $\sup_{\mathbb{P} \in \mathcal{P}} f(\mathbb{P}) + g(\mathbb{P}) \leq \sup_{\mathbb{P} \in \mathcal{P}} f(\mathbb{P}) + \sup_{\mathbb{P} \in \mathcal{P}} g(\mathbb{P})$, hence,

$$\begin{aligned} \text{(A.5)} &\leq \inf_{\varphi, \bar{\varphi} \in \Phi} \Bigg[z \sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T + \mathbf{X}_1^{\mathbb{P}} \right) - \alpha(\mathbb{P}) \right) \\ &\qquad + \bar{z} \sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T + \mathbf{X}_2^{\mathbb{P}} \right) - \alpha(\mathbb{P}) \right) \Bigg] \\ &= z\pi(\mathbf{X}_1) + \bar{z}\pi(\mathbf{X}_2). \end{aligned}$$

With that, we have shown that π is convex. Furthermore, cash-invariance and monotonicity follows directly from properties of $U^{\mathbb{P}}$. □

Proof of Lemma 2.2.3. Under the assumption of the Lemma, observe that the

$$\begin{aligned} & p(\mathbf{C}_T) \\ &= \inf_{\bar{\varphi} \in \Phi} \sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}} \right) - \alpha(\mathbb{P}) \right) - \inf_{\bar{\varphi} \in \Phi} \sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left((\bar{\varphi} \cdot \mathbf{S}_T^{\mathbb{P}}) - \alpha(\mathbb{P}) \right) \right. \\ &= \inf_{\bar{\varphi} \in \Phi} \sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left(((\bar{\varphi} - \varphi) \cdot \mathbf{S}^{\mathbb{P}})_T - p_0 \right) - \alpha(\mathbb{P}) \right) - \inf_{\bar{\varphi} \in \Phi} \sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T - \alpha(\mathbb{P}) \right) \right. \\ &= p_0 + \inf_{\bar{\varphi} \in \Phi} \sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T - \alpha(\mathbb{P}) \right) - \inf_{\bar{\varphi} \in \Phi} \sup_{\mathbb{P} \in \mathcal{P}} \left(U^{\mathbb{P}} \left((\bar{\varphi} \cdot \mathbf{S}^{\mathbb{P}})_T - \alpha(\mathbb{P}) \right) \right), \end{aligned}$$

where we used in the second step that $\bar{\varphi} - \varphi \in \Phi$. The above implies $p(\mathbf{C}_T) = p_0$ and therefore concludes the proof. □

Proof of Theorem 2.2.4. Let $\varphi \in \Phi$ and $z \in [0,1]$ with $\bar{z} := 1 - z$. Furthermore, fix $\mathbb{P}, \bar{\mathbb{P}} \in \mathcal{P}$. We can rewrite the loss function as

$$\mathcal{L}(\mathbb{P}, \varphi) = \inf_{w \in \mathbb{R}} w + \mathbb{E}_{\mathbb{P}}\left[u\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}} - w\Big)\right] - \alpha(\mathbb{P})$$

and conclude

$$\begin{aligned}\mathcal{L}(z\mathbb{P} + \bar{z}\bar{\mathbb{P}}, \varphi) = \inf_{w \in \mathbb{R}} w + z\mathbb{E}_{\mathbb{P}}\left[u\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}} - w\Big)\right] \\ + \bar{z}\mathbb{E}_{\bar{\mathbb{P}}}\left[u\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}} - w\Big)\right] - \alpha(z\mathbb{P} + \bar{z}\bar{\mathbb{P}}). \quad \text{(A.6)}\end{aligned}$$

Next, we observe that for functions $f, g : \mathbb{R} \to \mathbb{R}$ it holds that

$$\inf_{w \in \mathbb{R}} f(w) + g(w) \geq \inf_{w \in \mathbb{R}} f(w) + \inf_{w \in \mathbb{R}} g(w),$$

so,

$$\begin{aligned}\text{(A.6)} &\geq zU^{\mathbb{P}}\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}}\Big) + \bar{z}U^{\bar{\mathbb{P}}}\Big((\varphi \cdot \mathbf{S}^{\mathbb{P}})_T - \mathbf{C}_T^{\mathbb{P}}\Big) - \alpha(z\mathbb{P} + \bar{z}\bar{\mathbb{P}}) \\ &\geq z\mathcal{L}(\varphi, \mathbb{P}) + \bar{z}\mathcal{L}(\varphi, \bar{\mathbb{P}}),\end{aligned}$$

where we employed convexity of α in the last step. This proves the assertion. □

A.2 Rough path theory and distances of stochastic processes

The objective of this section is to revisit the concepts of path and expected signatures of a stochastic process, which characterize it and its path law respectively, to establish a distance on stochastic processes. While our framework is capable of employing various notions of distances, the following concepts first established in rough path theory deserve particular mention, since they are model-independent and have proven to be effective in a range of applications, see for instance [9, 169, 228].

In this work, we restrict ourselves to the class of continuous paths on a time interval $J \subseteq \mathbb{R}$ with values in $\mathbb{R}^d$ and finite p-variation, i.e. $\mathcal{C}_0^p(J, \mathbb{R}^d)$. For a detailed discussion of the signature feature set for parametrized paths, refer to [65, 192, 206].

A.2.1 The signature of a path

Fix by $T((\mathbb{R}^d)) := \oplus_{k=0}^{\infty}(\mathbb{R}^d)^{\otimes k}$ a tensor algebra, the space the signature of a $\mathbb{R}^d$-valued path takes values in. Then, the signature of a path is defined as follows.[1]

Defintion A.2.1 (Signature)**.** *Let $X \in \mathcal{C}_0^1(J, \mathbb{R}^d)$ such that the integrals in* (A.7) *are well-defined. The* signature of the path X *is defined as* $\mathrm{Sig}(X_J) := (1, X_J^1, \dots, X_J^k, \dots) \in T((\mathbb{R}^d))$*, where*

$$X_J^k = \int\limits_{(t_1,\dots,t_k)\in J^k,\, t_1<\dots<t_k} \cdots \int dX_{t_1} \otimes \cdots \otimes dX_{t_k}, \quad k \in \mathbb{N}. \tag{A.7}$$

Moreover, let us denote the truncated signature of the path X *at degree $M \in \mathbb{N}$ by writing* $\mathrm{Sig}_M(X_J) := (1, X_J^1, \dots, X_J^M)$.

The first feature of the signature map that we want to highlight is *uniqueness*. In particular, a signature can determine the path up to tree-like equivalence, see [39, 135]. However, since the path signature is invariant with respect to time re-parametrizations and its starting value, it is necessary to augment the observed path before applying the feature map. We will denote the tree-reduced set of $\mathcal{C}_0^1(J, \mathbb{R}^d)$ by $\tilde{\mathcal{C}}_0^1(J, \mathbb{R}^d)$

Remark A.2.1 (Path augmentations). Common path augmentations that have been suggested in [220] include (a) the time augmentation, (b) the visibility transform, and (c) the lead-lag transform. These augmentations encode the time parametrization, the starting point, and the lagged process into the path by adding further dimensions. In detail, a path of dimension d results into a path taking values in $\mathbb{R}^{2d+1}$ after applying augmentations (a), (b), and (c). ◁

Another feature of the signature map is *universality*, that means that every continuous functional on the path space $\tilde{\mathcal{C}}_0^1(J, \mathbb{R}^d)$ can be approximated arbitrarily well by linear functionals of truncated signatures when the degree of the signature is high enough. This is summarized in the subsequent theorem.

[1] For more background on this, refer to [65] or the appendix of [228] and the references mentioned therein.

Theorem A.2.1 (Universality of Signatures [192]). *For any continuous functional $f : K \to \mathbb{R}$, where $K \subseteq \tilde{\mathcal{C}}_0^1(J, \mathbb{R}^d)$ is a compact set, and any $\epsilon > 0$, there is a functional $L \in T((\mathbb{R}^d))^*$ such that*

$$\sup_{X \in K} |f(X) - L(\mathrm{Sig}(X))| < \epsilon.$$

Remark A.2.2 (Implementation). Since data of stochastic processes is usually provided in a discretized form, we use linear interpolation between the values to elevate the data to continuous paths. To compute the signatures, the open source library signatory [171] which operates in PyTorch [232] is used. ◁

A.2.2 The expected signature of a stochastic process

Let $\mathbf{X}$ be a stochastic process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with paths in $\tilde{\mathcal{C}}_0^p(J, \mathbb{R}^d)$, such that (A.7) is well-defined almost surely. Then, by path-wise applying the signature map, we obtain the expected signature $\mathbb{E}_{\mathbb{P}}[\mathrm{Sig}(\mathbf{X})]$. Of course, the expected signature is only well-defined if $\mathrm{Sig}(\mathbf{X})$ has finite expectation w.r.t. $\mathbb{P}$. The following result states that the expected signature characterizes the law of the stochastic process.

Theorem A.2.2 ([66]). *If $\mathbf{X}$ and $\mathbf{Y}$ are two $\tilde{\mathcal{C}}_0^1(J, \mathbb{R}^d)$-valued random variables,*

$$\mathbb{E}_{\mathbb{P}}[\mathrm{Sig}(\mathbf{X})] = \mathbb{E}_{\mathbb{P}}[\mathrm{Sig}(\mathbf{Y})]$$

, and $\mathbb{E}_{\mathbb{P}}[\mathrm{Sig}(\mathbf{X})]$ has a finite radius of convergence, then $\mathbf{X} \overset{d}{=} \mathbf{Y}$.

To put this another way, under the regularity condition, the distribution of some process $\mathbf{X}$ under $\mathbb{P}$ on the path space is characterized by $\mathbb{E}_{\mathbf{X} \sim \mathbb{P}}[\mathrm{Sig}(\mathbf{X})]$. This means, the expected signature of a random process can be viewed as an analogy of the moment generating function of a d-dimensional random variable. In [207] it is for instance shown that the expected Stratonovich signature of a Brownian motion determines its law. For a more detailed discussion on the characterization of stochastic processes by moments of signatures, see [67].

A.2.3 Signature Wasserstein-1 metric (SigW_1) and Signature-MMD (SigMMD)

As a next step, we will briefly discuss the metrics which are used to construct the penalty term α in (2.1) that are based on the signature framework. We begin by defining the Sig-Wasserstein distance. Note that in certain circumstances the following distances, introduced in recent literature are equivalent to one-another, however their numerical approximation and computational efficiency may have some discrepancies in applications. We therefore include a reminder and analysis with both metrics here and postpone the discussion of consolidating our findings to the identification of the ideal metric to future work.

Defintion A.2.2 (The Signature Wasserstein-1 metric)**.** *Assume $\mathbb{P}, \mathbb{Q}$ are two measures on the path space $\tilde{\mathcal{C}}_0^1(J, \mathbb{R}^d)$ and $\mathbf{X}$ is a stochastic process taking values in this space, then the* signature Wasserstein-1 metric on path space *(SigW_1), is defined as*

$$\mathrm{SigW}_1(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}} \left[f(\mathrm{Sig}(\mathbf{X})) \right] - \mathbb{E}_{\mathbf{X} \sim \mathbb{Q}} \left[f(\mathrm{Sig}(\mathbf{X})) \right],$$

where the supremum is over all the 1-Lipschitz functions $f : \tilde{\mathcal{C}}_0^1(J, \mathbb{R}^d) \to \mathbb{R}$ with Lipschitz norm smaller than 1, i.e. for all $x_1, x_2 \in \tilde{\mathcal{C}}_0^1(J, \mathbb{R}^d)$ it holds that $|f(x_1) - f(x_2)| \leq |x_1 - x_2|$.

This distance is based on the fact that signatures are representing stochastic processes in a random variable-like manner and the fact that the Wasserstein-metric provides a way to measure the distance between vector valued random variables. This metric goes back to [277]. While our framework is capable of employing a penalty that is defined via a maximization problem (see Section 2.3), we follow the steps presented in [228] to find and approximation of this metric and incorporate the analytic solution thereof. We begin with an application of Theorem A.2.1,

$$\mathrm{SigW}_1(\mathbb{P}, \mathbb{Q}) = \sup_{\substack{\|L\| \leq 1 \\ L \text{ is linear}}} \mathbb{E}_{\mathbf{X} \sim \mathbb{P}} \left[L(\mathrm{Sig}(\mathbf{X})) \right] - \mathbb{E}_{\mathbf{X} \sim \mathbb{Q}} \left[L(\mathrm{Sig}(\mathbf{X})) \right], \tag{A.8}$$

where the norm on L is chosen as the l_2-norm of the linear coefficients in L. Then we approximate $\mathrm{SigW}_1(\mathbb{P}, \mathbb{Q})$ by $\mathrm{SigW}_1^{(M)}(\mathbb{P}, \mathbb{Q})$, where we truncate the signatures in (A.8)

at level $M \in \mathbb{N}$. Since functionals on finite dimensional spaces can be represented by a vector multiplication, the optimization problem admits the analytic solution

$$\mathrm{SigW}_1^{(M)}(\mathbb{P}, \mathbb{Q}) = \left| \mathbb{E}_{\mathbf{X}\sim\mathbb{P}} \left[\mathrm{Sig}_M(\mathbf{X}) \right] - \mathbb{E}_{\mathbf{X}\sim\mathbb{Q}} \left[\mathrm{Sig}_M(\mathbf{X}) \right] \right|. \tag{A.9}$$

where $|\cdot|$ is the l_2 norm on the truncated tensor algebra $T^M((\mathbb{R}^d))$ and $M \in \mathbb{N}$ is the truncation level. For more details on this derivation, see [14, 228].

Eventually, note that the above derivation leads to the square of (A.9) for the *signature maximum mean discrepancy* (SigMMD) as defined in [67] if one chooses the signature with truncation level M as feature map, see [228].

Remark A.2.3 (Kernel trick and PDE representation). The approximation (A.9) relies on the accuracy of Theorem A.2.1 and in practice, where low values of M are used, this may lead to larger error terms. Moreover, to obtain sufficient precision for Monte-Carlo estimates of the expectations in (A.9), high sample sizes may be necessary to avoid significant bias.[2]

There are efforts to circumvent this approximation whilst maintaining a computable metric: As it is presented in [67], the SigMMD has a representation using the signature kernel, for which Salvi, Cass, Foster, Lyons, and Yang in [255] show that it is possible to compute as the solution of a Gorsuat-PDE. This leads to potential speed-ups, as it is indicated by the benchmarking performed in [155]. ◁

A.3 Setting of Herrmann, Muhle-Karbe, and Seifried

The authors Herrmann, Muhle-Karbe, and Seifried prove that for small uncertainty-aversions $1/\gamma$ the optimal strategy is given by

$$\begin{array}{rccc} \delta^\gamma : & [0,T] \times \mathbb{R}_+ \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ & (t,s,y) & \longmapsto & \delta^{\mathrm{BS}}(t,s) + \gamma \delta^{\mathrm{U}}(t,s,y) \end{array}$$

where $\delta^{\mathrm{BS}}(\cdot,\cdot)$ is the Black-Scholes delta and the uncertainty correction $\delta^{\mathrm{BS}}(\cdot,\cdot)$ is defined as

$$\begin{array}{rccc} \delta^{\mathrm{U}} : & [0,T] \times \mathbb{R}_+ \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ & (t,s,y) & \longmapsto & \frac{\partial}{\partial s} w(t,s,y), \end{array}$$

[2] Our findings suggest that low sample sizes tend to under-estimate the volatility of the underlying process.

where the source function $w : [0,T] \times \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ which is given by the following PDE

$$\frac{\partial}{\partial t} w(t,s) + \frac{1}{2}\sigma^2 s^2 \frac{\partial}{\partial s^2} w(t,s) + \frac{T\Big(\sigma^{\mathbb{Q}} s^2 \gamma^{\mathrm{BS}}(t,s)\Big)^2}{4} = 0, \quad \forall t, s \in [0,T] \times \mathbb{R}_+,$$
$$w(T,s) = 0, \quad \forall s \in \mathbb{R}_+.$$

Hereby, the term $\gamma^{\mathrm{BS}}(\cdot,\cdot)$ refers to the Black-Scholes gamma. The term δ^{U} is visualized for $t = T/2$ in the right-hand side of Figure 2.2.

A.4 Out-of-sample Testing

In this section, we investigate an out-of-sample test to evaluate a hedging strategy in an uncertainty setting. A straightforward approach is to fix some true underlying dynamics and perform the hedge-algorithm, as an agent would, on small data environments thereof and evaluate the strategies on the true dynamics.

To make this precise, stipulate that $\mathcal{G}_{\xi^*}$ is a generator for the true underlying dynamics for a fixed parameter $\xi^* \in \Xi$. This generator is not available to the actors, i.e. its data is not used to train components of the deep hedge. Moreover, we let dist be an arbitrary *calibration metric*, that is a function

$$\mathrm{dist} : \begin{array}{ccc} \mathbb{R}^{|I| \times N+1 \times d'} \times \Xi & \longrightarrow & \mathbb{R} \\ \Big((S^I_{t_n})_{n\in\{0,\dots,N\}}, \xi\Big) & \longmapsto & \mathrm{dist}\Big((S^I_{t_n})_{n\in\{0,\dots,N\}}, \xi\Big), \end{array}$$

that determines the distance of some set of paths $(S^I_{t_n})_{n\in\{0,\dots,N\}}$ indexed by some index set $I \subseteq \mathbb{N}$ to the generator's parameter ξ. The reader may for instance think of a Black-Scholes type generator (see Example 2.4.1) and dist the absolute distance of the (average) realized volatility on $(S^I_{t_n})_{n\in\{0,\dots,N\}}$ to the generator's volatility parameter $\sigma = \xi$.

The test can then be described with the following three steps:

1. Scenario generation.

   (a) We generate $M \in \mathbb{N}$ batches of Brownian increments, $(\Delta W^{I,m}_{t_n})_{n\in\{1,\dots,N\}}$ where $m \in \{1,\dots,M\}$ refers to the batch which is then indexed by some index set $I \subseteq \mathbb{N}$. We will refer to each of the M batches as a *scenario*.

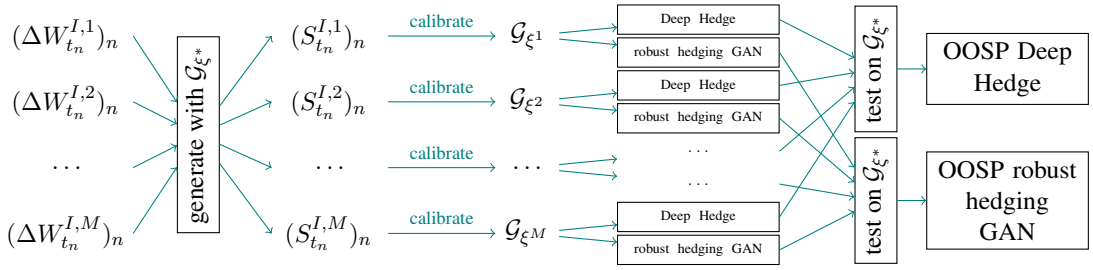


Figure A.1: Structure of the out-of-sample performance test on scenarios $\{1, \dots, M\}$.

(b) Next, the generator is applied to the batches of increments to obtain batches of paths

$$(S^{I,m}_{t_n})_{n\in\{1,\dots,N\}} := \left\{\mathcal{G}_{\xi^*}\left((\Delta W^{m,i}_{t_n})_{n\in\{1,\dots,N\}}\right)\right\}_{i\in I}$$

for every $m \in \{1, \dots, M\}$. We refer to a batch as a scenario.

(c) For every scenario $m \in \{1, \dots, M\}$, we fit a generator $\mathcal{G}_\cdot$ to the corresponding batch of paths. This means to find for every a parameter $\xi_m \in \Xi$ s.t.

$$\xi_m := \arg\min_{\xi\in\Xi} \operatorname{dist}\left((S^{I,m}_{t_n})_{n\in\{0,\dots,N\}}, \xi\right).$$

2. Training.

(a) For each scenario $m \in \{1, \dots, M\}$, we train a deep hedge on the respective generator $\mathcal{G}_{\xi_m}$.

(b) Likewise, we use the generator $\mathcal{G}_{\xi^m}$ for each scenario $m \in \{1, \dots, M\}$ to construct the robust hedging GAN and train it adversarially.

3. Evaluation. We can now evaluate the out-of-sample performance (OOSP) (implemented as loss) of both approaches. For this, let $\{\varphi^{(m)}\}_{m\in\{1,\dots,M\}}$ correspond to the strategies arising from the M batches. If $\mathbb{P}^*$ is the measure that matches the dynamics implied by the reference generator $\mathcal{G}_\xi$, the out-of-sample performance corresponds to

$$\left\{-U^{\mathbb{P}^*}\left(\sum_{n=1}^{N} {\varphi^{(m)}_{t_{n-1}}}^\intercal \Delta \mathbf{S}^{\mathbb{P}^*}_{t_n} - \mathbf{C}_T\right)\right\}_{m\in\{1,\dots,M\}}. \tag{A.10}$$

The test structure is visualized in Figure A.1. It is clear, that the set size $|I|$ dictates the level of model uncertainty that the scenarios inherit, while M indicates the precision when statistically evaluating (A.10).

Backward test The above introduced test comes with a high computational demand, since it is necessary to retrain deep hedge and robust hedging GAN for each scenario. Instead, we suggest calibrating different generators to one model, where in every case the former may be the measure indicating the true underlying for the latter.

To make this precise, assume that we estimated a measure $\mathbb{Q}$ under which we know the dynamics of $(\mathbf{S}^{\mathbb{Q}}_{t_n})_{n\in\{0,\dots,N\}}$. These are expressed by the generator $\mathcal{G}_{\xi^*}$ for some parameter $\xi^* \in \Xi$. With that, the test reads as follows:

1. Scenario generation.

   (a) As before, we generate $M \in \mathbb{N}$ batches of Brownian increments, denoted by $(\Delta W^{I,m}_{t_n})_{n\in\{1,\dots,N\}}$. Hereby, $m \in \{1, ..., M\}$ refers to the batch which is then indexed by some index set $I \subseteq \mathbb{N}$. We will refer to each of the M batches as a *scenario*.

   (b) For every $m \in \{1, \dots, M\}$ we now find parameters ξ^m such that $\mathcal{G}_{\xi_m}$ represents the true underlying distribution in this scenario

$$\xi_m := \underset{\xi\in\Xi}{\arg\min}\, \mathrm{dist}\left(\left\{\mathcal{G}_\xi\left((\Delta W^{m,i}_{t_n})_{n\in\{1,\dots,N\}}\right)\right\}_{i\in I}, \xi^*\right).$$

2. Training.

   (a) We train a deep hedge on the generator representing the estimated model $\mathcal{G}_{\xi^*}$.

   (b) Likewise, we use the estimated generator $\mathcal{G}_{\xi^*}$ to construct the robust hedging GAN and train it adversarially.

3. Evaluation. We can now evaluate the out-of-sample performance (OOSP) (implemented as loss) of both approaches. For this, let $\{\varphi\}_{m\in\{1,\dots,M\}}$ correspond to a strategy from one of the both procedures. If for every $m \in \{1, \dots, M\}$, $\mathbb{P}^m$ is the measure that matches the dynamics implied by the generator $\mathcal{G}_{\xi_m}$, the out-of-sample performance corresponds to

$$\left\{ - U^{\mathbb{P}^m}\Big(\sum_{n=1}^{N} {\varphi_{t_{n-1}}}^{\intercal} \Delta \mathbf{S}^{\mathbb{P}^m}_{t_n} - \mathbf{C}_T \Big) \right\}_{m\in\{1,\dots,M\}}. \tag{A.11}$$

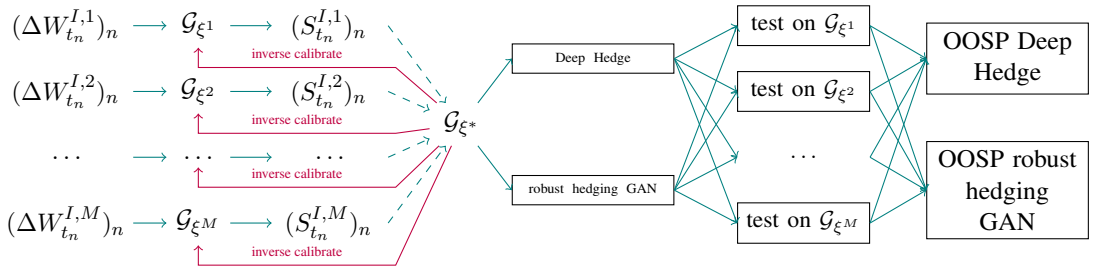


Figure A.2: Structure of the inverse out-of-sample performance test on scenarios $\{1, \dots, M\}$.

The test structure is visualized in Figure A.2. Usually, the size of J is chosen reasonably high to avoid estimation errors on this side, while the size of I dictates the level of uncertainty the scenarios are generated with. Again, M indicates the precision with which (A.11) is evaluated statistically.

A.5 Inverse Calibration in the Black-Scholes model

Let $(\mathbf{W}_{t_n})_{n\in\{0,\dots,N\}}$ be a 1-dimensional Brownian motion under a measure $\mathbb{P}$ and denote by $(\Delta\mathbf{W}_{t_n})_{n\in\{1,\dots,N\}}$ its increments, i.e. $\Delta\mathbf{W}_{t_n} = \mathbf{W}_{t_n} - \mathbf{W}_{t_{n-1}}$ for every $n \in \{1, \dots, N\}$. Then the corresponding Black-Scholes path for parameters $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$ is given by

$$\mathbf{S}_n = S_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right) t_n + \sigma \mathbf{W}_n\right), \quad \forall n \in \{1, \dots, N\},$$

what, given the assumption that $\Delta t := t_n - t_{n-1}$ for all $n \in \{1, \dots, N\}$, implies

$$\log\left(\frac{\mathbf{S}_n}{\mathbf{S}_{n-1}}\right) = \left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\Delta\mathbf{W}_n, \quad \forall n \in \{1, \dots, N\}. \tag{A.12}$$

From this, we see that

$$\sigma = \sqrt{\frac{1}{\Delta t}\mathbb{V}_{\mathbb{P}}\left[\log\left(\frac{\mathbf{S}_n}{\mathbf{S}_{n-1}}\right)\right]}, \quad \forall n \in \{1, \dots, N\},$$

with $\mathbb{V}_{\mathbb{P}}$ being the variance operator under $\mathbb{P}$. This derives our estimation formula of σ, i.e. given a path $(S_n)_{n\in\{0,\dots,N\}} \in \mathbb{R}^{N+1}$, we estimate

$$\hat{\sigma} = \frac{1}{\sqrt{\Delta t}}\widehat{\mathrm{Std}}\left[\left\{\log\left(\frac{S_n}{S_{n-1}}\right)\right\}_{n\in\{1,\dots,N\}}\right],$$

where $\widehat{\mathrm{Std}}[\{\cdot\}_{n\in\{1,\dots,N\}}]$ is an estimator for the standard deviation. Plugging in (A.12), we obtain

$$\begin{aligned}\hat{\sigma} &= \frac{1}{\sqrt{\Delta t}}\widehat{\mathrm{Std}}\left[\left\{\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\Delta W_{t_n}\right\}_{n\in\{1,\dots,N\}}\right], \\ &= \sigma\widehat{\mathrm{Std}}\left[\{Z_n\}_{n\in\{1,\dots,N\}}\right], \end{aligned} \tag{A.13}$$

with $(\Delta W_{t_n})_{n\in\{1,\dots,N\}}$ constituting the sampled Brownian increments that generated the instance of the path $(S_n)_{n\in\{0,\dots,N\}} \in \mathbb{R}^{N+1}$ and $Z_n := \Delta W_{t_n}/\sqrt{\Delta t}$ for all $n \in \{1,\dots,N\}$.

This means, if we want to inverse calibrate σ to a fixed $\hat{\sigma}$ for the random noise $(Z_n)_{n\in\{1,\dots,N\}}$, we obtain

$$\sigma = \hat{\sigma}\left(\widehat{\mathrm{Std}}\left[\{Z_n\}_{n\in\{1,\dots,N\}}\right]\right)^{-1}.$$

Genius ideas are often simple.

— Proverb

B Appendix of Chapter 3

Contents

B.1 Basics on Risk Measures 137
B.2 Results for Section 3.3 138
B.3 Results for Section Section 3.5 146
B.3.1 Details of Theorem 3.5.1 146
B.3.2 Results for Section 3.5.1 147

B.1 Basics on Risk Measures

Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $X \in L^0(\Omega, \mathcal{F}, \mathbb{P})$. Then, the entropy [109] for risk aversion $\lambda > 0$ is defined as follows

$$\mathrm{Entr}_\lambda(X) := -\frac{1}{\lambda} \log\left(\mathbb{E}_\mathbb{P}\left[\exp\left(-\lambda X\right)\right]\right). \tag{B.1}$$

For normal random variables, it is useful to exchange the entropic risk-measure with the mean-variance operator [212], since it offers higher numerical stability. The mean-variance for risk-aversion $\lambda > 0$ is defined by

$$\mathrm{MeanVar}_\lambda(X) := \mathbb{E}_\mathbb{P}\left[X\right] - \frac{1}{2}\lambda \mathbb{V}_\mathbb{P}\left[X\right]. \tag{B.2}$$

If X is normally distributed, the Laplace transform equals

$$\mathbb{E}_\mathbb{P}\left[\exp\left(-\lambda X\right)\right] = \exp\left(-\lambda \mathbb{E}_\mathbb{P}\left[X\right] + {}^{\lambda}\!/_{2}\mathbb{V}_\mathbb{P}\left[X\right]\right), \tag{B.3}$$

such that mean-variance and negative entropy coincide.

In the above setting, the conditional value at risk (CVaR) [109] is defined for a risk-aversion parameter $\alpha \in [0,1]$ as

$$-\mathrm{CVaR}_\alpha(X) := \mathbb{E}_\mathbb{P}\left[X \middle| X \leq \mathrm{VaR}_\alpha(X)\right], \tag{B.4}$$

where for $\gamma \in [0,1]$

$$\mathrm{VaR}_\gamma(X) = -\inf\left\{x \in \mathbb{R} : F_X(x) > \gamma\right\}.$$

Remark B.1.1. Note we can rewrite

$$\mathrm{CVaR}(X) = \frac{1}{\alpha}\left(\mathbb{E}[X1_{\{X \leq x_\alpha\}}] + x_\alpha(\alpha - \mathbb{P}(X < x_\alpha))\right),$$

where x_α is the lower α-quantile. This particularly simplifies for continuous distribution functions, since the term $x_\alpha(\alpha - \mathbb{P}(X < x_\alpha)) = 0$ vanishes. ◁

B.2 Results for Section 3.3

Lemma B.2.1. *For $X \sim \mathcal{N}(\nu, \tau^2)$ with $\tau, \nu, \lambda \in \mathbb{R}$,*

$$\arg\max_{a \in \mathbb{R}} \mathrm{Entr}_\lambda(aX) = \frac{\nu}{\lambda\tau^2}$$

Proof. As remarked above, under normality, the entropy coincides with mean-variance, such that, the left-hand side reduces to $a\nu - \lambda a^2\tau^2/2$ which is optimized by the right-hand side. □

Lemma B.2.2. *Assume that $\Delta S_1 \sim \mathcal{N}(\mu, \sigma^2)$ and $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2 N^{-1})$. Then,*

$$\begin{aligned}\mathrm{Entr}_\lambda\left(x\hat{\mu}\Delta S_1\right) &= \mathrm{Entr}_\lambda\left(x\mu\Delta S_1\right) - \frac{\lambda}{2}\mathbb{V}_\mathbb{P}\left[x(\hat{\mu}-\mu)\Delta S_1\right] \\ &= x\mu^2\left(1 - \frac{\lambda}{2}x\sigma^2\right) - \frac{\lambda}{2N}x^2\sigma^2(\mu^2+\sigma^2).\end{aligned}$$

Proof. We start with the first equation: due to equivalence of Entr and MeanVar for normal random variables,

$$\mathrm{Entr}_\lambda\left(x\hat{\mu}\Delta S_1\right) = \mathrm{MeanVar}_\lambda\left(x\hat{\mu}\Delta S_1\right)$$

$$= \mathbb{E}_{\mathbb{P}}\left[x\hat{\mu}\Delta S_1\right] - \frac{\lambda}{2}\mathbb{V}_{\mathbb{P}}\left[x\hat{\mu}\Delta S_1\right]$$
$$= x\mathbb{E}_{\mathbb{P}}\left[\mu\Delta S_1\right] - \frac{x^2\lambda}{2}\mathbb{V}_{\mathbb{P}}\left[\mu\Delta S_1 + (\hat{\mu}-\mu)\Delta S_1\right].$$

Next, observe that

$$\begin{aligned}
&\mathbb{V}_{\mathbb{P}}\left[\mu\Delta S_1 + (\hat{\mu}-\mu)\Delta S_1\right] \\
&\quad = \mathbb{E}_{\mathbb{P}}\left[(\mu\Delta S_1)^2 + 2\mu(\hat{\mu}-\mu)\Delta S_1^2 + (\hat{\mu}-\mu)^2\Delta S_1^2\right] - \mathbb{E}_{\mathbb{P}}\left[\mu\Delta S_1 + (\hat{\mu}-\mu)\Delta S_1\right]^2 \\
&\quad = \mathbb{E}_{\mathbb{P}}\left[(\mu\Delta S_1)^2\right] + \mathbb{E}_{\mathbb{P}}\left[(\hat{\mu}-\mu)^2\Delta S_1^2\right] - \mathbb{E}_{\mathbb{P}}\left[\mu\Delta S_1\right]^2 - \mathbb{E}_{\mathbb{P}}\left[(\hat{\mu}-\mu)\Delta S_1\right]^2 \\
&\quad = \mathbb{V}_{\mathbb{P}}\left[\mu\Delta S_1\right] + \mathbb{V}_{\mathbb{P}}\left[(\hat{\mu}-\mu)\Delta S_1\right]. \qquad \text{(B.5)}
\end{aligned}$$

The second equation follows by observing that by iterated conditional expectations,

$$\mu\mathbb{E}_{\mathbb{P}}\left[\Delta S_1\right] = \mathbb{E}_{\mathbb{P}}\left[\Delta S_1\right] = \mu^2, \qquad \text{(B.6)}$$

together with

$$\mathbb{V}_{\mathbb{P}}\left[(\hat{\mu}-\mu)\Delta S_1\right] = \mathbb{E}_{\mathbb{P}}\left[(\hat{\mu}-\mu)^2\right]\mathbb{E}_{\mathbb{P}}\left[\Delta S_1^2\right] = \frac{\sigma^2}{N}(\mu^2+\sigma^2).$$

□

Lemma B.2.3. *The "out-of-sample" performance of* $a_{\text{plug-in}}$ *as in* (3.1)*, under the assumptions of Lemma B.2.2 is*

$$\text{MeanVar}_\lambda^{\mathbb{P}}(a_{\text{plug-in}}\Delta S_1) = \frac{\mu^2}{2\lambda\sigma^2} - \frac{\mu^2+\sigma^2}{2N\lambda\sigma^2} \qquad \text{(B.7)}$$

Proof. The explicit form of the strategy is given by Lemma B.2.1 and the statement follows from Lemma B.2.2 with $x = 1/(\lambda\sigma^2)$. □

Lemma B.2.4. *The "out-of-sample" performance of* a_{mix} *as in* (3.10)*, under the assumptions of Lemma B.2.2, is*

$$\text{Entr}_\lambda(a_{\text{mix}}\Delta S_1) = \frac{\mu^2 N}{\lambda\sigma^2(N+1)}\left(1 - \frac{\lambda}{2}\sigma^2\right) - \frac{N(\mu^2+\sigma^2)}{2\lambda\sigma^2(N+1)^2}$$

Proof. This follows directly from Lemma B.2.1 and Lemma B.2.2. □

Lemma 3.3.1. *The entropic uncertainty-aware policy* (3.3) *admits the representation*

$$a_{\mathrm{u-a}} \;=\; \frac{\hat{\mu}}{\lambda\hat{\sigma^2}\left(1+\frac{\lambda'}{\lambda N}\right)} \;=\; \frac{\lambda N}{\lambda N+\lambda'}\, a_{\text{plug-in}}, \tag{3.7}$$

and, under $\hat{\sigma^2} \equiv \sigma^2$*, its out-of-sample performance is*

$$\begin{aligned} \mathrm{Entr}_\lambda(a_{\mathrm{u-a}}\,\Delta S_1) &\equiv \mathrm{MeanVar}^{\mathbb{P}}_\lambda(a_{\mathrm{u-a}}\Delta S_1) \\ &= \frac{x_N}{\lambda\sigma^2}\left(\mu^2\left(1-\frac{x_N}{2}\right)-\frac{x_N(\mu^2+\sigma^2)}{2N}\right), \\ x_N &:= \frac{\lambda N}{\lambda N+\lambda'}. \end{aligned} \tag{3.8}$$

Proof. By the same arguments as in Lemma 3.3.1 and $a \in \mathcal{A}$,

$$\begin{aligned} &\mathrm{MeanVar}^{\mathfrak{P}}_{\lambda'}\left(\hat{\mathrm{MeanVar}}_\lambda\left(\{a(\Delta S_t+\epsilon)\}_{t\in\{-N+1,\dots,0\}}\right)\right) \\ &\qquad = \mathrm{MeanVar}^{\mathfrak{P}}_{\lambda'}\left(a(\hat{\mu}+\epsilon)-\frac{\lambda}{2}a^2\hat{\sigma^2}\right) \end{aligned}$$

and thus

$$\begin{aligned} &\mathrm{MeanVar}^{\mathfrak{P}}_{\lambda'}\left(a(\hat{\mu}+\epsilon)-\frac{\lambda}{2}a^2\hat{\sigma^2}\right) \\ &\qquad = \mathbb{E}_{\mathfrak{P}}\left[a(\hat{\mu}+\epsilon)-\frac{\lambda}{2}a^2\hat{\sigma^2}\right]-\frac{\lambda'}{2}\mathbb{V}_{\mathfrak{P}}\left[a(\hat{\mu}+\epsilon)-\frac{\lambda}{2}a^2\hat{\sigma^2}\right] \\ &\qquad = a\hat{\mu}-\frac{\lambda}{2}a^2\hat{\sigma^2}-\frac{\lambda'}{2}a^2\mathbb{V}_{\mathfrak{P}}\left[\epsilon\right] \\ &\qquad = a\hat{\mu}-\frac{\lambda}{2}a^2\hat{\sigma^2}\left(1+\frac{\lambda'}{\lambda N}\right). \end{aligned}$$

Hence, the strategy $a_{\mathrm{u-a}}$ maximizing the objective can be expressed as

$$a_{\mathrm{u-a}} := \frac{\hat{\mu}}{\lambda\hat{\sigma^2}(1+\frac{\lambda'}{\lambda N})}.$$

A straightforward computation shows that

$$a_{\mathrm{u-a}} = \frac{\lambda N}{\lambda N+\lambda'}a_{\text{plug-in}}.$$

By employing results (B.6) and (B.5), we receive for $x \in \mathbb{R}$

$$\begin{aligned} \mathrm{MeanVar}^{\mathbb{P}}_\lambda\left(x\, a_{\text{plug-in}}\Delta S_1\right) &= \mathrm{MeanVar}^{\mathbb{P}}_\lambda\left(x\frac{\hat{\mu}}{\lambda\sigma^2}\Delta S_1\right) \\ &= x\mathbb{E}_{\mathbb{P}}\left[\frac{\hat{\mu}}{\lambda\sigma^2}\Delta S_1\right]-\frac{\lambda}{2}x^2\mathbb{V}_{\mathbb{P}}\left[\frac{\hat{\mu}}{\lambda\sigma^2}\Delta S_1\right] \end{aligned}$$

$$
\begin{aligned}
&= x\mathbb{E}_{\mathbb{P}}\left[\frac{\mu}{\lambda\sigma^2}\Delta S_1\right] - \frac{\lambda}{2}x^2\left(\mathbb{V}_{\mathbb{P}}\left[\frac{\mu}{\lambda\sigma^2}\Delta S_1\right] + \frac{\mu^2+\sigma^2}{N\lambda^2\sigma^2}\right)\\
&= x\frac{\mu^2}{\lambda\sigma^2} - x^2\frac{2N\mu^2+\mu^2+\sigma^2}{2N\lambda\sigma^2}\\
&= \frac{x}{\lambda\sigma^2}\left(\mu^2\left(1-\frac{x}{2}\right) - \frac{x(\mu^2+\sigma^2)}{2N}\right)
\end{aligned}
$$

and hence for $x_N := \lambda N/(\lambda N + \lambda')$,

$$
\mathrm{MeanVar}^{\mathbb{P}}_{\lambda}\left(a_{\mathrm{u-a}}\Delta S_1\right) = \frac{x_N}{\lambda\sigma^2}\left(\mu^2\left(1-\frac{x_N}{2}\right) - \frac{x_N(\mu^2+\sigma^2)}{2N}\right).
$$

□

Lemma B.2.5. *The uncertainty-aware out-performs the plug-in policy if and only if*

$$
\mu^2\frac{(N-1)\lambda' - 2N\lambda}{2N\lambda+\lambda'} < \sigma^2,
$$

implying that the robustification is necessarily successful, whenever

$$
\lambda' < \frac{2N}{N-1}\lambda.
$$

Proof. Set

$$
x_N = \frac{\lambda N}{\lambda N + \lambda'}.
$$

The out-of-sample value of the uncertainty-aware policy is

$$
\frac{x_N}{\lambda\sigma^2}\left[\mu^2\left(1-\tfrac{x_N}{2}\right) - \frac{x_N(\mu^2+\sigma^2)}{2N}\right],
$$

while the plug-in policy yields

$$
\frac{\mu^2}{2\lambda\sigma^2} - \frac{\mu^2+\sigma^2}{2N\lambda\sigma^2}.
$$

Thus the former exceeds the latter exactly when

$$
\frac{x_N}{\lambda\sigma^2}\left[\mu^2(1-\tfrac{x_N}{2}) - \tfrac{x_N(\mu^2+\sigma^2)}{2N}\right] > \frac{\mu^2}{2\lambda\sigma^2} - \frac{\mu^2+\sigma^2}{2N\lambda\sigma^2}.
$$

Multiply both sides by $2N\lambda\sigma^2 > 0$:

$$
2Nx_N\mu^2\left(1-\tfrac{x_N}{2}\right) - x_N^2(\mu^2+\sigma^2) > N\mu^2 - (\mu^2+\sigma^2).
$$

Expand and collect terms:

$$2Nx_N\mu^2 - Nx_N^2\mu^2 - x_N^2\mu^2 - x_N^2\sigma^2 > N\mu^2 - \mu^2 - \sigma^2,$$

$$\mu^2\Big(2Nx_N - Nx_N^2 - x_N^2 - N + 1\Big) > -\sigma^2(1 - x_N^2).$$

Note that

$$\begin{aligned} 2Nx_N - Nx_N^2 - x_N^2 - N + 1 &= -N(x_N^2 - 2x_N + 1) - (x_N^2 - 1) \\ &= -(1 - x_N)\Big(N(1 - x_N) - (1 + x_N)\Big). \end{aligned}$$

Hence

$$-\mu^2(1 - x_N)\Big(N(1 - x_N) - (1 + x_N)\Big) > -\sigma^2(1 - x_N^2),$$

$$\mu^2(1 - x_N)\Big(N(1 - x_N) - (1 + x_N)\Big) < \sigma^2(1 - x_N^2).$$

Since $0 < x_N < 1$, divide by $(1 - x_N)(1 + x_N)$ to get

$$\mu^2\frac{N(1 - x_N) - (1 + x_N)}{1 + x_N} < \sigma^2.$$

Finally, substitute

$$1 - x_N = \frac{\lambda'}{\lambda N + \lambda'}, \quad 1 + x_N = \frac{2\lambda N + \lambda'}{\lambda N + \lambda'}.$$

Then

$$\begin{aligned} N(1 - x_N) - (1 + x_N) &= \frac{N\lambda' - (2\lambda N + \lambda')}{\lambda N + \lambda'} = \frac{(N - 1)\lambda' - 2N\lambda}{\lambda N + \lambda'}, \\ 1 + x_N &= \frac{2\lambda N + \lambda'}{\lambda N + \lambda'}. \end{aligned}$$

Thus

$$\frac{N(1 - x_N) - (1 + x_N)}{1 + x_N} = \frac{(N - 1)\lambda' - 2N\lambda}{2N\lambda + \lambda'},$$

and the inequality becomes

$$\mu^2\frac{(N - 1)\lambda' - 2N\lambda}{2N\lambda + \lambda'} < \sigma^2,$$

as claimed. Finally, since $2N\lambda + \lambda' > 0$, this holds in particular whenever $(N - 1)\lambda' - 2N\lambda < 0$, i.e.

$$\lambda' < \frac{2N}{N - 1}\lambda.$$

□

Lemma 3.3.2. *For $U = \mathrm{CVaR}_\alpha$, the uncertainty-aware policy (3.4) is*

$$a_{\mathrm{u-a'}} \;=\; \mathrm{sign}(\hat{\mu})\,\frac{\left(|\hat{\mu}| - A\sqrt{\hat{\sigma^2}/N}\right)_+}{\lambda\hat{\sigma^2}} \;=\; \left(a_{\text{plug-in}} - \mathrm{sign}(\hat{\mu})\frac{A\sqrt{\hat{\sigma^2}/N}}{\lambda\hat{\sigma^2}}\right)\mathbb{1}_{\{\hat{\mu}^2 \geq A^2\,\frac{\hat{\sigma^2}}{N}\}},$$

where $A := \varphi(\Phi^{-1}(\alpha))/\alpha$, with φ and Φ the standard normal density and cdf. Under $\hat{\sigma^2} \equiv \sigma^2$, the OOSP is explicit: $\mathrm{Entr}_\lambda(a_{\mathrm{u-a'}}\Delta S_1) = f(\mu, \sigma, \lambda, \alpha, N)$ for an analytically computable function f.

Proof. Since we have chosen $\epsilon \sim \mathcal{N}(0, \hat{\sigma^2}/N)$, we can compute for every $a \in \mathcal{A}$

$$\begin{aligned}\mathrm{CVaR}_\alpha^{\mathfrak{P}}\left(\hat{\mathrm{MeanVar}}_\lambda\left(\{a(\Delta S_t + \epsilon)\}_{t\in\{-N+1,\dots,0\}}\right)\right) &= \mathrm{CVaR}_\alpha^{\mathfrak{P}}\left(a(\hat{\mu}+\epsilon) - \frac{\lambda}{2}a^2\hat{\sigma^2}\right)\\ &= \left(a\hat{\mu} - \frac{\lambda}{2}a^2\hat{\sigma^2}\right) - A|a|\sqrt{\frac{\hat{\sigma^2}}{N}}\end{aligned}$$

where $A := \varphi(\Phi^{-1}(\alpha))/\alpha$. Here, $\varphi(\cdot)$ denotes the probability density function and $\Phi^{-1}(\cdot)$ the quantile function of the standard normal distribution. Note that we have to use the absolute value of a to apply that formula, since negative factorized random variables have their worst scenarios at the other tail.

Point-wise maximization of the above expression results in

$$a_{\mathrm{u-a'}} = \mathrm{sign}(\hat{\mu})\frac{\left(|\hat{\mu}| - A\sqrt{\frac{\hat{\sigma^2}}{N}}\right)_+}{\lambda\hat{\sigma^2}} \tag{B.8}$$

where $\mathrm{sign}(\cdot)$ is the sign operator and $(\cdot)_+$ is the restriction to the positive orthant.

Eventually, we calculate again the out-of-sample performance

$$\mathrm{MeanVar}_\lambda^{\mathbb{P}}\left(a_{\mathrm{u-a'}}\Delta S_1\right) = \mathrm{MeanVar}_\lambda^{\mathbb{P}}\left(\mathrm{sign}(\hat{\mu})\frac{\left(|\hat{\mu}| - A\sqrt{\frac{\hat{\sigma^2}}{N}}\right)_+}{\lambda\hat{\sigma^2}}\Delta S_1\right). \tag{B.9}$$

Using the assumption $\hat{\sigma^2} \equiv \sigma^2$, we see that

$$\mathrm{MeanVar}_\lambda^{\mathbb{P}}\left(\mathrm{sign}(\hat{\mu})\frac{\left(|\hat{\mu}| - A\sqrt{\frac{\hat{\sigma^2}}{N}}\right)_+}{\lambda\hat{\sigma^2}}\Delta S_1\right)$$

$$= \mu\mathbb{E}_{\mathbb{P}}\left[\operatorname{sign}(\hat{\mu})\frac{\left(|\hat{\mu}| - A\sqrt{\frac{\hat{\sigma^2}}{N}}\right)_+}{\lambda\hat{\sigma^2}}\right] - \frac{\lambda\sigma^2}{2}\mathbb{E}_{\mathbb{P}}\left[\left(\frac{\left(|\hat{\mu}| - A\sqrt{\frac{\hat{\sigma^2}}{N}}\right)_+}{\lambda\hat{\sigma^2}}\right)^2\right]$$
$$- \frac{\lambda\mu^2}{2}\mathbb{E}_{\mathbb{P}}\left[\operatorname{sign}(\hat{\mu})\frac{\left(|\hat{\mu}| - A\sqrt{\frac{\hat{\sigma^2}}{N}}\right)_+}{\lambda\hat{\sigma^2}}\right]$$

We compute

$$2\mathbb{E}_{\mathbb{P}}\left[\operatorname{sign}(\hat{\mu})\left(|\hat{\mu}| - A\sqrt{\frac{\hat{\sigma^2}}{N}}\right)_+\right]$$
$$= \frac{\mu^2\sqrt{N}\mathrm{Erf}\left(\frac{\sqrt{\left(A\sigma+\mu\sqrt{N}\right)^2}}{\sqrt{2}\sigma}\right)}{\sqrt{\left(A\sigma+\mu\sqrt{N}\right)^2}} + \mu\mathrm{Erf}\left(\frac{\mu\sqrt{N}-A\sigma}{\sqrt{2}\sigma}\right)$$
$$- \frac{A\mu\sqrt{N}\sigma\mathrm{Erf}\left(\frac{\sqrt{\left(A\sigma+\mu\sqrt{N}\right)^2}}{\sqrt{2}\sigma}\right)}{\sqrt{N\left(A\sigma+\mu\sqrt{N}\right)^2}} - \frac{A\sigma\mathrm{Erf}\left(\frac{\mu\sqrt{N}-A\sigma}{\sqrt{2}\sigma}\right)}{\sqrt{N}}$$
$$+ \frac{A\sigma\mathrm{Erfc}\left(\frac{A\sigma+\mu\sqrt{N}}{\sqrt{2}\sigma}\right)}{\sqrt{N}} - \frac{\sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\left(A\sigma+\mu\sqrt{N}\right)^2}{2\sigma^2}}}{\sqrt{N}} + \frac{\sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\left(\mu\sqrt{N}-A\sigma\right)^2}{2\sigma^2}}}{\sqrt{N}}$$
$$- \frac{A\sigma}{\sqrt{N}} + \frac{\sqrt{\mu^2 N}\mathrm{Erf}\left(\frac{\sqrt{\mu^2 N}}{\sqrt{2}\sigma}\right)}{\sqrt{N}} - \mu\mathrm{Erf}\left(\frac{\mu\sqrt{N}}{\sqrt{2}\sigma}\right) + 2\mu,$$

where $\mathrm{Erf}(\cdot) := 2/\sqrt{\pi}\int_0^{\cdot}\exp(-x^2)dx$ and $\mathrm{Erfc}(\cdot) = 1 - \mathrm{Erf}(\cdot)$. Furthermore,

$$2N\mathbb{E}_{\mathbb{P}}\left[\left(\left(|\hat{\mu}| - A\sqrt{\frac{\hat{\sigma^2}}{N}}\right)_+\right)^2\right]$$
$$= \left(\left(A^2+1\right)\sigma^2 - 2A\mu\sqrt{N}\sigma + \mu^2 N\right)\mathrm{Erfc}\left(\frac{A\sigma-\mu\sqrt{N}}{\sqrt{2}\sigma}\right)$$
$$+ \left(\left(A^2+1\right)\sigma^2 + 2A\mu\sqrt{N}\sigma + \mu^2 N\right)\mathrm{Erfc}\left(\frac{A\sigma+\mu\sqrt{N}}{\sqrt{2}\sigma}\right)$$
$$- \sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\left(A\sigma+\mu\sqrt{N}\right)^2}{2\sigma^2}}\left(A\sigma\left(e^{\frac{2A\mu\sqrt{N}}{\sigma}}+1\right) - \mu\sqrt{N}\left(e^{\frac{2A\mu\sqrt{N}}{\sigma}}-1\right)\right).$$

Thus, we obtain a closed form expression of (B.9). □

Corollary 3.3.3. *For the policy* (3.3)*,*

$$\lim_{\lambda' \to 0} a_{\mathrm{u-a}} = a_{\text{plug-in}}, \qquad \lim_{\lambda' \to \infty} a_{\mathrm{u-a}} \equiv 0, \qquad \lim_{N \to \infty} a_{\mathrm{u-a}} = a_{\text{oracle}}.$$

Analogously, for (3.4)*,*

$$\lim_{\alpha \to 1} a_{\mathrm{u-a'}} = a_{\text{plug-in}}, \qquad \lim_{\alpha \to 0} a_{\mathrm{u-a'}} \equiv 0, \qquad \lim_{N \to \infty} a_{\mathrm{u-a'}} = a_{\text{oracle}}.$$

Proof of Corollary 3.3.3. We begin with the entropic uncertainty-aware strategy from Lemma 3.3.1, which has the analytic form

$$a_{\mathrm{u-a}} = \frac{\lambda N}{\lambda N + \lambda'} a_{\text{plug-in}} =: x_N a_{\text{plug-in}},$$

where $x_N := \frac{\lambda N}{\lambda N + \lambda'}$.

- As $\lambda' \to 0$, we have $x_N \to 1$, hence $a_{\mathrm{u-a}} \to a_{\text{plug-in}}$.
- As $\lambda' \to \infty$, we have $x_N \to 0$, hence $a_{\mathrm{u-a}} \to 0$.
- As $N \to \infty$, we have $x_N \to 1$, so again $a_{\mathrm{u-a}} \to a_{\text{plug-in}}$.

  But since $a_{\text{plug-in}} = \hat{\mu}/(\lambda \hat{\sigma^2})$ and $\hat{\mu} \to \mu$, $\hat{\sigma^2} \to \sigma^2$ as $N \to \infty$, we conclude $a_{\text{plug-in}} \to a_{\text{oracle}} := \mu/(\lambda \sigma^2)$. Therefore, $a_{\mathrm{u-a}} \to a_{\text{oracle}}$ as $N \to \infty$.

Next, for the CVaR-based uncertainty-aware strategy from Lemma 3.3.2, we have

$$a_{\mathrm{u-a'}} = \left(a_{\text{plug-in}} - \operatorname{sign}(\hat{\mu}) \frac{A \sqrt{\frac{\hat{\sigma^2}}{N}}}{\lambda \hat{\sigma^2}} \right) \mathbb{1}_{\{\hat{\mu}^2 \geq A^2 \frac{\hat{\sigma^2}}{N}\}},$$

where $A := \varphi(\Phi^{-1}(\alpha))/\alpha$.

- As $\alpha \to 1$, we have $A \to 0$, so the indicator condition $\hat{\mu}^2 \geq A^2 \hat{\sigma^2}/N$ is always satisfied and the correction term vanishes. Therefore, $a_{\mathrm{u-a'}} \to a_{\text{plug-in}}$.
- As $\alpha \to 0$, we have $A \to \infty$, so the indicator becomes zero with probability one, hence $a_{\mathrm{u-a'}} \equiv 0$.
- As $N \to \infty$, the term $A\sqrt{\hat{\sigma^2}/N} \to 0$ and $\hat{\mu} \to \mu$, $\hat{\sigma^2} \to \sigma^2$, thus the indicator becomes 1 for large N and the correction vanishes. Therefore, $a_{\mathrm{u-a'}} \to a_{\text{oracle}} := \mu/(\lambda \sigma^2)$.

□

B.3 Results for Section Section 3.5

B.3.1 Details of Theorem 3.5.1

Theorem 3.5.1. *[CVaR-Stochastic Gradient Descent] Let* $\Phi = \{\varphi \in \mathbb{R}^d : \|\varphi\| \leq B\}$, *and suppose* $\varphi \mapsto J(X(a_\varphi), \mathbb{P})$ *is convex and* ρ*-Lipschitz for all* $\mathbb{P} \sim \mathfrak{P}$, *i.e.*

$$\|\nabla_\varphi J(X(a_\varphi), \mathbb{P})\| \leq \rho \quad \mathfrak{P}\text{-a.s.}.$$

For $\alpha \in (0, 1)$, *define*

$$\mathcal{L}(\varphi) := \mathrm{CVaR}_\alpha \left(J(X(a_\varphi), \mathbb{P}) \,|\, \mathbb{P} \sim \mathfrak{P} \right).$$

Let $(\varphi_t)_{t\geq 0}$ *be the sequence generated by Algorithm 1 with step size* $\eta_t = B/(\rho\sqrt{t})$, *and set* $\bar{\varphi}_t := t^{-1} \sum_{s=1}^{t} \varphi_s$. *Then, for all* $t \geq 1$,

$$\mathbb{E}[\mathcal{L}(\bar{\varphi}_t)] - \min_{\varphi\in\Phi} \mathcal{L}(\varphi) \leq \frac{B\rho}{\sqrt{t}}.$$

Proof. **Convexity and Lipschitz continuity.** Using the Rockafellar–Uryasev representation of CVaR [251],

$$\mathrm{CVaR}_\alpha(Z) = \min_{\zeta\in\mathbb{R}} \left\{ \zeta + \frac{1}{\alpha}\mathbb{E}\left[(Z-\zeta)_+\right] \right\},$$

we write

$$\mathcal{L}(\varphi) = \min_{\zeta\in\mathbb{R}} \left\{ \zeta + \frac{1}{\alpha}\mathbb{E}_{\mathbb{P}\sim\mathfrak{P}}\left[(J(X(a_\varphi), \mathbb{P}) - \zeta)_+\right] \right\}.$$

Because $J(X(a_\varphi), \mathbb{P})$ is convex in φ and the hinge function $(\cdot - \zeta)_+$ is jointly convex, the full objective is convex in φ. The bound on $\|\nabla_\varphi J\|$ implies that $\mathcal{L}$ is ρ-Lipschitz.

Unbiased sub-gradient via random k**.** Let $J_\varphi(\mathbb{P}) := J(X(a_\varphi), \mathbb{P})$ and let $\zeta_\varphi := \mathrm{VaR}_\alpha(J_\varphi)$. Then,

$$\nabla_\varphi \mathcal{L}(\varphi) = \frac{1}{\alpha}\mathbb{E}_{\mathbb{P}\sim\mathfrak{P}}\left[\nabla_\varphi J(X(a_\varphi), \mathbb{P}) \cdot \mathbf{1}_{J_\varphi(\mathbb{P})\leq\zeta_\varphi}\right].$$

We now draw m i.i.d. samples $\mathbb{P}_1, \ldots, \mathbb{P}_m \sim \mathfrak{P}$, compute $J_i := J(X(a_\varphi), \mathbb{P}_i)$ and $g_i := \nabla_\varphi J(X(a_\varphi), \mathbb{P}_i)$, and recall that k is a random integer with $\mathbb{E}[k] = \alpha m$. Let $I(\varphi)$ be the indices of the k smallest J_i. Define the stochastic gradient estimator

$$\widehat{g}(\varphi) := \frac{1}{k} \sum_{i\in I(\varphi)} g_i.$$

By symmetry and independence, the distribution of the k smallest elements ensures

$$\mathbb{E}[\widehat{g}(\varphi)] = \frac{1}{\alpha}\mathbb{E}_{\mathbb{P}\sim\mathfrak{P}}\left[\nabla_\varphi J(X(a_\varphi),\mathbb{P})\cdot \mathbf{1}_{J_\varphi(\mathbb{P})\leq\zeta_\varphi}\right] = \nabla_\varphi \mathcal{L}(\varphi).$$

Hence, $\widehat{g}(\varphi)$ is an unbiased stochastic sub-gradient. Its norm is bounded by ρ by assumption.

Apply SGD convergence. All conditions of [261, Theorem 14.8] are satisfied: convex domain, unbiased sub-gradient with norm $\leq \rho$, Lipschitz objective, and learning rate $\eta_t = B/(\rho\sqrt{t})$. The theorem then yields the expected bound

$$\mathbb{E}[\mathcal{L}(\bar{\varphi}_t)] - \min_{\varphi\in\Phi}\mathcal{L}(\varphi) \leq \frac{B\rho}{\sqrt{t}}.$$

□

B.3.2 Results for Section 3.5.1

We compute the objective of example of Section 3.5.1

Lemma B.3.1. *Assume that* $\Delta S_t \sim \mathcal{N}_d(\hat{\mu},\Sigma)$. *Then, the entropy is given by*

$$J(X(a)) \;=\; a^\top\hat{\mu} \;-\; \frac{\lambda}{2}\,a^\top\Sigma\,a. \tag{B.10}$$

Proof. Since $X(a)$ is normally distributed, the entropy coincides with the mean-variance objective, as already remarked. Note that $X(a) = a^\top\Delta S_t \sim \mathcal{N}(a^\top\hat{\mu}, a^\top\Sigma\,a)$. The result then follows by Equation (B.2). □

Lemma B.3.2. *The plug-in portfolio as defined in Equation (3.1) for the above setting is*

$$a_{\text{plug-in}} = \frac{1}{\lambda}\,\Sigma^{-1}\,\hat{\mu}.$$

Proof. This result is obtained by maximizing Equation (B.10) in a. This equation is quadratic in a, such that the maximum is given by the first order condition

$$\hat{\mu} - \lambda\Sigma a = 0.$$

□

Lemma 3.5.2. *The* CVaR*-uncertainty-aware strategy*

$$\arg\max_{a\in\mathbb{R}^d} \mathrm{CVaR}_\alpha\Big(\mathrm{Entr}_\lambda(a^\top \Delta S_1)\,\Big|\,\mu\sim\mathcal{N}(\hat{\mu},\Sigma/N)\Big)$$

is

$$a_{\mathrm{u-a'}} = \frac{1}{\lambda}\left(1-\frac{\kappa_\alpha}{\sqrt{N}\,\|\Sigma^{-1/2}\hat{\mu}\|}\right)_+ \Sigma^{-1}\hat{\mu},$$

where $\kappa_\alpha = \varphi(\Phi^{-1}(\alpha))/\alpha$. *The zero-investment region* $a_{\mathrm{u-a'}} = 0$ *occurs whenever* $\|\Sigma^{-1/2}\hat{\mu}\| \le \kappa_\alpha/\sqrt{N}$.

Proof. We compute

$$\begin{aligned}
a_{\mathrm{u-a'}} &= \arg\max_{a\in\mathbb{R}} \mathrm{CVaR}_\alpha\Big(\mathrm{Entr}_\lambda\Big(a\Delta S_1 \,\Big|\, S_1\sim\mathcal{N}(\mu,\sigma)\Big)\,\Big|\,\mu\sim\mathcal{N}(\hat{\mu},\Sigma/N)\Big)\\
&= \arg\max_{a\in\mathbb{R}} \mathrm{CVaR}_\alpha\Big(a^\intercal\mu \;-\; \frac{\lambda}{2}\,a^\intercal\Sigma\,a \,\Big|\, \mu\sim\mathcal{N}(\hat{\mu},\Sigma/N)\Big)\\
&= \arg\max_{a\in\mathbb{R}} \mathrm{CVaR}_\alpha\Big(a^\intercal\mu \,\Big|\, \mu\sim\mathcal{N}(\hat{\mu},\Sigma/N)\Big) \;-\; \frac{\lambda}{2}\,a^\intercal\Sigma\,a\\
&= \arg\max_{a\in\mathbb{R}} \underbrace{a^\top\hat{\mu} - \kappa_\alpha\sqrt{\tfrac{a^\top\Sigma a}{N}} - \frac{\lambda}{2}\,a^\top\Sigma\,a}_{:=\widetilde{J}(a)},
\end{aligned}$$

where the last line follows from CVaR_α on normal random variables.

(i) Zero-investment region at $a=0$. The subgradient of $\widetilde{J}$ at 0 is

$$\partial\widetilde{J}(0) = \left\{\hat{\mu} - \kappa_\alpha\tfrac{\Sigma^{1/2}u}{\sqrt{N}} \;\middle|\; \|u\|=1\right\}.$$

Hence 0 is optimal iff $\max_{\|u\|=1}\langle\hat{\mu},u\rangle \le \kappa_\alpha/\sqrt{N}\max_{\|u\|=1}\|\Sigma^{1/2}u\|$, i.e. $\|\Sigma^{-1/2}\hat{\mu}\| \le \kappa_\alpha/\sqrt{N}$.

(ii) Outside the zero-investment region. For $\|\Sigma^{-1/2}\hat{\mu}\| > \kappa_\alpha/\sqrt{N}$, symmetry implies the maximizer lies on the ray $a = c\,\Sigma^{-1}\hat{\mu},\ c>0$. Set $m = \|\Sigma^{-1/2}\hat{\mu}\|$, so

$$a^\top\hat{\mu} = c\,m^2, \qquad \sqrt{a^\top\Sigma a} = c\,m.$$

Then

$$\widetilde{J}(c) = c\,m^2 - \kappa_\alpha\frac{c\,m}{\sqrt{N}} - \frac{\lambda}{2}\,c^2\,m^2.$$

Differentiating w.r.t. c and setting to zero yields

$$m^2 - \kappa_\alpha\frac{m}{\sqrt{N}} - \lambda\,c\,m^2 = 0 \quad\Longrightarrow\quad c^* = \frac{1}{\lambda}\Big(1-\frac{\kappa_\alpha}{\sqrt{N}\,m}\Big).$$

Since $c^* > 0$ precisely when $m > \kappa_\alpha/\sqrt{N}$, we conclude

$$a_{\mathrm{u-a'}} = c^* \, \Sigma^{-1}\hat{\mu} = \frac{1}{\lambda}\Big(1 - \frac{\kappa_\alpha}{\sqrt{N}\,\|\Sigma^{-1/2}\hat{\mu}\|}\Big)_+ \Sigma^{-1}\hat{\mu},$$

as claimed. □

If we knew what we were doing, it would not be called research, would it?

— Albert Einstein

C Proofs of Chapter 4

Contents

C.1 Isometric Copies of Every Compact Riemannian Manifold in the Path Space — Proof of Proposition 4.2.1 151
C.2 Stability of the Optimal Filter — Proof of Proposition 4.4.2 152
C.2.1 Lipschitz-Continuity in the Path-Component of the Optimal Filter 153
C.2.2 Lipschitz-Continuity in the Time-Component of the Optimal Filter 161
C.2.3 Lower Loewner order bound 164
C.3 Stable and Lossless Encoding — Proof of Proposition 4.3.1 165
C.4 Approximability of Locally Lipschitz Maps By Filterformers — Proof of Proposition 4.4.3 172
C.4.1 Polish QAS Space Structure on the space of Positive Semi-Definite Matrices 172
C.4.2 Proof of the Main Approximation Lemma 178
C.5 An Auxiliary Lemma — I. Pinelis 180

C.1 Isometric Copies of Every Compact Riemannian Manifold in the Path Space — Proof of Proposition 4.2.1

Proof of Proposition 4.2.1. Since $(\mathcal{X}, d_{\mathcal{X}})$ is compact then the Kuratowksi embedding $\varphi_1 : x \mapsto d_{\mathcal{X}}(\cdot, x)$ is an isometric embedding of $\mathcal{X}$ into the Banach space $C(\mathcal{X})$ with

its uniform norm (since $\mathcal{X}$ is compact); where $d_{\mathcal{X}}$ denotes the metric on $\mathcal{X}$. By the Banach-Mazur theorem, there exists an isometric embedding of $\varphi_2 : C(\mathcal{X}) \to C([0,1])$. Since the map $\psi_3 : C([0:1]) \to C([0:T], \mathbb{R}^{d_Y})$, given by $f \mapsto (f(\cdot/T), 0, \ldots, 0)$, is an isometric embedding; then $\varphi \overset{\text{def.}}{=} \varphi_3 \circ \varphi_2 \circ \varphi_1$ is an isometric embedding of $\mathcal{X}$ into $C([0:T], \mathbb{R}^{d_Y})$. □

C.2 Stability of the Optimal Filter — Proof of Proposition 4.4.2

We now prove the local Lipschitz stability of the optimal filter (1.4), for the coupled system (4.2)-(4.3), as a function of its observed/input path. The proof builds on a series of Lemmata, which demonstrate the local Lipschitz continuity of the parameters defining the optimal filter.

We first observe that, Assumption 4.2.1 grants us access to [201, Theorem 12.7], which implies that $\mu_t \overset{\text{def.}}{=} \mu(y_{[0:t]})$ and $\Sigma_t \overset{\text{def.}}{=} \Sigma(y_{[0:t]})$ are

$$\begin{aligned} d\mu_t = [a_0(t, y_{[0:t]}) + a_1(t, y_{[0:t]})\mu_t]dt + [(b \circ B)(t, y_{[0:t]}) + \Sigma_t A_1^\top(t, y_{[0:t]})] \\ \times (B \circ B)^{-1}(t, y_{[0:t]})[dy_t - (A_0(t, y_{[0:t]}) + A_1(t, y_{[0:t]}))\mu_t dt] \end{aligned} \tag{C.1}$$

$$\begin{aligned} \partial_t \Sigma_t = a_1(t, y_{[0:t]})\Sigma_t + \Sigma_t a_1^\top(t, y_{[0:t]}) + (b \circ b)(t, y_{[0:t]}) \\ -[(b \circ B)(t, y_{[0:t]}) + \Sigma_t A_1^\top(t, y_{[0:t]})](B \circ B)^{-1}(t, y_{[0:t]}) \\ \times [(b \circ B)(t, y_{[0:t]}) + \Sigma_t A_1^\top(t, y_{[0:t]})]^\top \end{aligned} \tag{C.2}$$

with initial conditions $\mu_0 \overset{\text{def.}}{=} \mathbb{E}[X_0|Y_0]$ and $\Sigma_0 \overset{\text{def.}}{=} \mathbb{E}[(X_0 - \mu_0)(X_0 - \mu_0)^\top|Y_0]$; where

$$B \circ B \overset{\text{def.}}{=} B_1 B_1^\top + B_2 B_2^\top, \quad b \circ B \overset{\text{def.}}{=} b_1 B_1^\top + b_2 B_2^\top, \quad b \circ b \overset{\text{def.}}{=} b_1 b_1^\top + b_2 b_2^\top.$$

Furthermore, [201, Theorem 12.7] implies that Σ_t is positive definite for all $t \in [0:T]$ since we have assume that Σ_0 in Assumption 4.2.1.

To clarify the argument's direction, we first record the proof of Proposition 4.4.2. We subsequently derive the lemmata on which it relies.

Proof of Proposition 4.4.2. Fix $t \in [0:T]$, a path $y_\cdot^{(1)} \in C^1([0:T], \mathbb{R}^{d_Y})$ and $\epsilon > 0$. By Lemma C.2.7 we can fix a $K_1 > 0$ such that for every time $0 < s < T$ holds

$$\mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y_{[0:t]}^{(1)}), \mathbb{P}(X_s \in \cdot | y_{[0:s]}^{(1)})) \leq K_1 |t-s|_t.$$

By Lemma C.2.1 we can fix a $K_2 > 0$ such that for every path $y_\cdot^{(2)} \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y_{[0:T]}^{(1)} - y_{[0:T]}^{(2)}\|_T < \epsilon$ holds

$$\mathcal{W}_2(\mathbb{P}(X_u \in \cdot | y_{[0:u]}^{(1)}), \mathbb{P}(X_u \in \cdot | y_{[0:u]}^{(2)})) \leq K_2 \|y_{[0:T]}^{(1)} - y_{[0:T]}^{(2)}\|_T.$$

We see by the triangular inequality of the $\mathcal{W}_2$-distance (that holds as we handle normal distributions, see (4.11)),

$$\begin{aligned}
&\mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y_{[0:t]}^{(1)}), \mathbb{P}(X_s \in \cdot | y_{[0:s]}^{(2)})) \\
&\quad \leq \mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y_{[0:t]}^{(1)}), \mathbb{P}(X_s \in \cdot | y_{[0:s]}^{(1)})) + \mathcal{W}_2(\mathbb{P}(X_s \in \cdot | y_{[0:s]}^{(1)}), \mathbb{P}(X_s \in \cdot | y_{[0:s]}^{(2)})) \\
&\quad \leq \max\{K_1, K_2\} \big(\|y_{[0:T]}^{(1)} - y_{[0:T]}^{(2)}\|_T + |t-s|. \big)
\end{aligned}$$

□

C.2.1 Lipschitz-Continuity in the Path-Component of the Optimal Filter

Lemma C.2.1 (Local Lipschitz - the Path-Component of the Optimal Filter)**.** *Under Assumption 4.2.1, f from (1.4) is locally Lipschitz-continuous uniformly in time. In particular, for every path $y_\cdot^{(1)} \in C^1([0:T], \mathbb{R}^{d_Y})$ and $\epsilon > 0$ there exists constant $C \geq 0$ such that for all paths $y_\cdot^{(2)} \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y_{[0:t]}^{(1)} - y_{[0:t]}^{(2)}\|_t < \epsilon$ holds*

$$\mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y_{[0:t]}^{(1)}), \mathbb{P}(X_t \in \cdot | y_{[0:t]}^{(2)})) \leq C \|y_{[0:T]}^{(1)} - y_{[0:T]}^{(2)}\|_T, \quad \forall t \in [0:T].$$

Proof of Lemma C.2.1. By (4.11), Lemma C.2.10, and Lemma C.5.1 there exists a non-negative constant $\bar{K} < \infty$ satisfying the following for every pair of paths $y_\cdot^{(1)}, y_\cdot^{(2)} \in C([0:T], \|\cdot\|_2)$ and all time $t \in [0:T]$

$$\begin{aligned}
&\mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y_{[0:t]}^{(1)}), \mathbb{P}(X_t \in \cdot | y_{[0:t]}^{(2)})) \\
&\qquad \leq \bar{K} \sqrt{\|\mu(y_{[0:t]}^{(1)}) - \mu(y_{[0:t]}^{(2)})\|_2^2 + \|\Sigma(y_{[0:t]}^{(1)}) - \Sigma(y_{[0:t]}^{(2)})\|_2^2}.
\end{aligned}$$

We see from Lemma C.2.2 and Lemma C.2.5, there exists a non-negative constant $K < \infty$, satisfying

$$\begin{aligned} \|\Sigma(y^{(1)}_{[0:t]}) - \Sigma(y^{(2)}_{[0:t]})\|_2^2 &\le K^2 \|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t^2, \\ \|\mu(y^{(1)}_{[0:t]}) - \mu(y^{(2)}_{[0:t]})\|_2^2 &\le K^2 \|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t^2, \end{aligned} \qquad \forall t \in [0:T]$$

Therefore, for each $y^{(1)}_\cdot, y^{(2)}_\cdot \in C([0:T], \mathbb{R}^{d_Y})$ we have

$$\mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y^{(1)}_{[0:t]}), \mathbb{P}(X_t \in \cdot | y^{(2)}_{[0:t]})) \le \sqrt{2} K \bar{K} \|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t, \qquad \forall t \in [0:T]$$

and the statement follows. □

Lemma C.2.2 (Local Lipschitz-continuity of μ in the path)**.** *Under Assumption 4.2.1 (viii) $\mu_\cdot$ from (C.1) is locally Lipschitz-continuous uniformly in time. This means that for every path $y^{(1)}_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$ there exists constant $K \in \mathbb{R}$, $\epsilon > 0$ such that for all paths $y^{(2)}_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t < \epsilon$ holds*

$$\left\|\mu^{(1)}_t - \mu^{(2)}_t\right\|_2 \le K \left\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\right\|_t, \qquad \forall t \in [0:T].$$

Proof of Lemma C.2.2. Define

$$\begin{aligned} \Phi_1(s, y_{[0:s]}) \overset{\text{def.}}{=} a_1(s, y_{[0:s]}) - [(b \circ B)(s, y_{[0:s]}) + \Sigma_t A_1^\top(s, y_{[0:s]})] \\ \times (B \circ B)^{-1}(s, y_{[0:s]})[A_0(s, y_{[0:s]}) + A_1(s, y_{[0:s]})], \end{aligned} \tag{C.3}$$

$$\Phi_2(s, y_{[0:s]}) \overset{\text{def.}}{=} a_0(s, y_{[0:s]}) + [(b \circ B)(s, y_{[0:s]}) + \Sigma_t A_1^\top(s, y_{[0:s]})](B \circ B)^{-1}(s, y_{[0:s]}). \tag{C.4}$$

Note that conditions of Lemma C.2.3 and Lemma C.2.4 are satisfied, therefore, Φ_1 and Φ_2 are globally Lipschitz w.r.t. time and locally Lipschitz w.r.t. the path.

By (C.1) and for $t \in [0, T]$, μ_t can be written as

$$\mu_t = \mu_0 + \int_0^t a_0(s, y_{[0:s]})ds + \int_0^t \Phi_2(s, y_{[0:s]})dy_{[0:s]} + \int_0^t \Phi_1(s, y_{[0:s]})\mu_s ds.$$

We further have that

$$\begin{aligned} \mu^{(1)}_t - \mu^{(2)}_t = \mu^{(1)}_0 - \mu^{(2)}_0 + \int_0^t a_0(s, y^{(1)}_{[0:s]}) - a_0(s, y^{(2)}_{[0:s]})ds \\ + \int_0^t \Phi_2(s, y^{(1)}_{[0:s]})dy^{(1)}_{[0:s]} - \int_0^t \Phi_2(s, y^{(2)}_{[0:s]})dy^{(2)}_{[0:s]} \end{aligned}$$

$$+\int_0^t \Phi_1(s, y^{(1)}_{[0:s]}) - \Phi_1(s, y^{(2)}_{[0:s]})\mu^{(2)}_s ds + \int_0^t \Phi_1(s, y^{(1)}_{[0:s]})\,(\mu^{(1)}_s - \mu^{(2)}_s)ds.$$

By the triangle inequality, Jensen's inequality, and Cauchy-Schwarz inequality, we have

$$\left\|\mu^{(1)}_t - \mu^{(2)}_t\right\|_2 \leq \gamma(t) + \int_0^t \left\|\Phi_1(s, y^{(1)}_{[0:s]})\right\|_2 \cdot \left\|\mu^{(1)}_s - \mu^{(2)}_s\right\|_2 ds,$$

where

$$\begin{aligned}\gamma(t) \overset{\text{def.}}{=} \left\|\mu^{(1)}_0 - \mu^{(2)}_0\right\|_2 &+ \underbrace{\left\|\int_0^t a_0(s, y^{(1)}_{[0:s]}) - a_0(s, y^{(2)}_{[0:s]})ds\right\|_2}_{\text{(I)}} \\ &+ \underbrace{\left\|\int_0^t \Phi_2(s, y^{(1)}_{[0:s]})dy^{(1)}_{[0:s]} - \int_0^t \Phi_2(s, y^{(2)}_{[0:s]})dy^{(2)}_{[0:s]}\right\|_2}_{\text{(II)}} \\ &+ \underbrace{\left\|\int_0^t \Phi_1(s, y^{(1)}_{[0:s]}) - \Phi_1(s, y^{(2)}_{[0:s]})\mu^{(2)}_s ds\right\|_2}_{\text{(III)}}.\end{aligned}$$

For an application of Grönwall's inequality, it is sufficient to show $\|\gamma(t)\|_2 < \infty$. We have, due to a_0 being Lipschitz with constant $K_1 < \infty$,

$$\begin{aligned}\text{(I)} &\leq \int_0^t \|a_0(s, y^{(1)}_{[0:s]}) - a_0(s, y^{(2)}_{[0:s]})\|_2 ds \\ &\leq \int_0^t K_1 \|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\|_s ds \\ &\leq tK_1 \|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t.\end{aligned}$$

Further, we see that

$$\text{(II)} \leq \underbrace{\left\|\int_0^t \Phi_2(s, y^{(2)}_{[0:s]}) - \Phi_2(s, y^{(1)}_{[0:s]})dy^{(2)}_{[0:s]}\right\|_2}_{\text{(IV)}} + \underbrace{\left\|\int_0^t \Phi_2(s, y^{(1)}_{[0:s]})d(y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]})\right\|_2}_{\text{(V)}}.$$

Using Itô's isometry, we have

$$\begin{aligned}\text{(IV)}^2 &= \int_0^t \left\|\Phi_2(s, y^{(2)}_{[0:s]}) - \Phi_2(s, y^{(1)}_{[0:s]})\right\|_2^2 d\,(B \circ B)(s, y^{(2)}_{[0:s]}) \\ &\leq K_2^2 \|y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]}\|_t^2 \int_0^t d\,(B \circ B)(s, y^{(2)}_{[0:s]}),\end{aligned}$$

where the last inequality holds due to (path-)Lipschitz continuity of Φ_2 uniformly in time with constant $K_2 < \infty$, given by Lemma C.2.4. Next, let $\{[s_{i-1} : s_i] | i \in \{1, \ldots, I\}\}$,

$I \in \mathbb{N}$ be a fine enough partition of $[0:t]$, we see due to the (time-) Lipschitz continuity of $B \circ B$ together with the definition of the Riemann-Stieltjes integral that

$$\begin{aligned}(\text{IV})^2 &\leq \epsilon_I + K_2^2 \|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t^2 \cdot \sum_{i=1}^{I} \|(B\circ B)(y^{(2)}_{[0:s_i]}) - (B\circ B)(y^{(2)}_{[0:s_{i-1}]})\|_2 \\ &\leq \epsilon_I + K_2^2 T \|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t^2.\end{aligned}$$

As $\epsilon_I \to 0$ for $I \to \infty$, we obtain $(\text{IV}) \leq K_2\sqrt{T}\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t$.

Integrating by parts, we find that

$$\begin{aligned}(\text{V}) &= \Big\| \Phi_2(t, y^{(1)}_{[0:t]})(y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]}) - \Phi_2(0, y^{(1)}_0)(y^{(2)}_0 - y^{(1)}_0) \\ &\qquad - \int_0^t (y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]})\, d\Phi_2(s, y^{(1)}_{[0:s]}) \Big\|_2 \\ &\leq \Big\| \Phi_2(t, y^{(1)}_{[0:t]})(y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]}) - \Phi_2(0, y^{(1)}_0)(y^{(2)}_0 - y^{(1)}_0) \Big\|_2 \\ &\qquad + \Big\| \int_0^t (y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]})\, d\Phi_2(s, y^{(1)}_{[0:s]}) \Big\|_2 \\ &\leq 2 \max_{0\leq s\leq T} \Big\| \Phi_2(s, y^{(1)}_{[0:s]}) \Big\|_2 \cdot \Big\| y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]} \Big\|_t + \underbrace{\Big\| \int_0^t (y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]})\, d\Phi_2(s, y^{(1)}_{[0:s]}) \Big\|_2}_{(\text{VI})}.\end{aligned}$$

Lemma C.2.4 yields (time-)Lipschitz continuity of Φ_2, implying

$$\max_{s\in[0:T]} \|\Phi_2(s, y^{(1)}_{[0:s]})\|_2 =: K_3 < \infty,$$

together with (path-) Lipschitz continuity of Φ_2 uniformly in time with constant $K_4 < \infty$ yields

$$\begin{aligned}(\text{VI}) &\leq \epsilon_I + \sum_{i=1}^{I} \Big\| y^{(2)}_{[0:s_i]} - y^{(1)}_{[0:s_i]} \Big\|_2 \Big\| \Phi_2(s_i, y^{(1)}_{[0:s_i]}) - \Phi_2(s_{i-1}, y^{(1)}_{[0:s_{i-1}]}) \Big\|_2 \\ &\leq \epsilon_I + K_4 t \|y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]}\|_t,\end{aligned}$$

which holds for $\epsilon_I \to 0$ for $I \to \infty$; we conclude that

$$(\text{II}) \leq \left(K_2\sqrt{T} + K_3 + K_4 T\right) \|y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]}\|_t.$$

Next, we consider

$$(\text{III}) \leq \int_0^t \Big\| \Phi_1(s, y^{(1)}_{[0:s]}) - \Phi_1(s, y^{(2)}_{[0:s]}) \Big\|_2 \Big\| \mu^{(2)}_s \Big\|_2 ds \leq K_5 t \|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t \max_{s\in[0:t]} \Big\| \mu^{(2)}_s \Big\|_2.$$

due to (path-)Lipschitz continuity of Φ_1 uniformly in time with constant K_5 which is obtained by Lemma C.2.3. We conclude, since $\max_{s\in[0:T]} \|\mu_s^{(2)}\|_2 := K_6 < \infty$ due to Lemma C.2.8, that $\gamma(t) \leq K\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t$ with $K := K_1 + K_2\sqrt{T} + K_3 + K_4 T + K_5 K_6 T$.

With that, we are able to apply Grönwall's inequality and obtain

$$\begin{aligned}\left\|\mu_t^{(1)} - \mu_t^{(2)}\right\|_2 &\leq \gamma(t) + \int_0^t \gamma(s)\left\|\Phi_1(s, y^{(1)}_{[0:s]})\right\|_2 \exp\left(\int_s^t \left\|\Phi_1(r, y^{(1)}_{[0:r]})\right\|_2 dr\right) ds \\ &\leq K\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t \\ &\qquad \left(1 + \int_0^t \left\|\Phi_1(s, y^{(1)}_{[0:s]})\right\|_2 \exp\left(\int_s^t \left\|\Phi_1(r, y^{(1)}_{[0:r]})\right\|_2 dr\right) ds\right).\end{aligned}$$

Note that Φ_1 is (time-)Lipschitz continuous, as shown in Lemma C.2.3, therefore implying that $\max_{s\in[0:T]} \|\Phi_1(s, y^{(1)}_{[0:s]})\|_2 < \infty$. As constants K_1, K_2, K_3, K_3, K_4, K_5, and K_6 were chosen independently of $t \in [0:T]$ and Assumption 4.2.1 (xiii) holds, we conclude the proof. □

Lemma C.2.3. *Let Assumption 4.2.1 (viii) hold. Then,*

1. *for any path $y \in C([0:T]), \mathbb{R}^d)$ there is a constant $K \in \mathbb{R}$ s.t. for all $t > s \in [0:T]$ holds*

$$\|\Phi_1(t, y_{[0:t]}) - \Phi_1(s, y_{[0:s]})\|_2 \leq K|t - s|,$$

   where Φ_1 is defined as in (C.3).

2. *Also, for every path $y^{(1)}_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$ there exists a constant $C \geq 0$ such that for all $\epsilon > 0$, paths $y^{(2)}_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y^{(1)}_{[0:T]} - y^{(2)}_{[0:T]}\|_T < \epsilon$, and times $t \in [0:T]$ holds*

$$\left\|\Phi_1(t, y^{(1)}_{[0:t]}) - \Phi_1(t, y^{(2)}_{[0:t]})\right\|_2 \leq C\left\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\right\|_t.$$

Proof of Lemma C.2.3. 1. The statement holds if a_1, $(b\circ B)(B\circ B)^{-1}A_0$, $(b\circ B)(B\circ B)^{-1}A_1$, $\Sigma_\cdot A_1(B\circ B)^{-1}A_0$, $\Sigma_\cdot A_1(B\circ B)^{-1}A_1$ are globally Lipschitz continuous w.r.t. the time component. As a_1, $(b\circ B)$, $\Sigma_\cdot$, $(B\circ B)^{-1}$, A_0, A_1 satisfy this already (see Lemma C.2.9) and with this are also bounded on the interval $[0:T]$, their product is globally Lipschitz continuous w.r.t. time as well.

2. The assertion holds if a_1, $(b \circ B)(B \circ B)^{-1}A_0$, $(b \circ B)(B \circ B)^{-1}A_1$, $\Sigma_\cdot A_1(B \circ B)^{-1}A_0$, $\Sigma_\cdot A_1(B \circ B)^{-1}A_1$ are locally Lipschitz continuous w.r.t. the path component uniformly in time. As a_1, $(b \circ B)$, $\Sigma_\cdot$, $(B \circ B)^{-1}$, A_0, A_1 are locally Lipschitz continuous w.r.t. the path component (see Lemma C.2.5) uniformly in time, this also follows for their product.

□

Lemma C.2.4. *Let Assumption 4.2.1 (viii) be fulfilled. Then,*

1. *for any path $y \in C([0:T]), \mathbb{R}^d)$ there is a constant $C \geq 0$ s.t. for all $t > s \in [0:T]$ holds*

$$\|\Phi_2(t, y_{[0:t]}) - \Phi_2(s, y_{[0:s]})\|_2 \leq C|t - s|,$$

where Φ_2 is defined as in (C.4).

2. *Additionally, for every path $y_\cdot^{(1)} \in C^1([0:T], \mathbb{R}^{d_Y})$ there exists constant $C \geq 0$, such that for all $\epsilon > 0$, paths $y_\cdot^{(2)} \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t < \epsilon$, and times $t \in [0:T]$ holds*

$$\left\|\Phi_2(t, y^{(1)}_{[0:t]}) - \Phi_2(t, y^{(2)}_{[0:t]})\right\|_2 \leq C\left\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\right\|_t.$$

Proof of Lemma C.2.4. 1. The assertion holds if a_0, $(b \circ B)(B \circ B)^{-1}$, $\Sigma_\cdot A_1(B \circ B)^{-1}$ are globally Lipschitz continuous w.r.t. the time component. As a_0, $(b \circ B)$, $\Sigma_\cdot$, $(B \circ B)^{-1}$, A_1 satisfy this already (see Lemma C.2.9) and with this are also bounded on the interval $[0:T]$, their product is globally Lipschitz continuous w.r.t. time as well.

2. The statement holds if a_0, $(b \circ B)(B \circ B)^{-1}$, $\Sigma_\cdot A_1(B \circ B)^{-1}$ are locally Lipschitz continuous w.r.t. the path component. As a_0, $(b \circ B)$, $\Sigma_\cdot$, $(B \circ B)^{-1}$, A_1 are locally Lipschitz continuous w.r.t. the path component (see Lemma C.2.5) this also follows for their product.

□

Lemma C.2.5 (Local Lipschitz-continuity of Σ)**.** *Under Assumption 4.2.1 (viii), $\Sigma_\cdot$ from (C.2) is locally Lipschitz-continuous uniformly in time. This means, for every path $y^{(1)}_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$ there exists constant $C \in \mathbb{R}$ such that for all $\epsilon > 0$, paths $y^{(2)}_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t < \epsilon$, and times $t \in [0:T]$ holds*

$$\left\|\Sigma^{(1)}_t - \Sigma^{(2)}_t\right\|_2 \leq C \left\|y^1_{[0:t]} - y^2_{[0:t]}\right\|_t,$$

where $\|\cdot\|_2$ on the left-hand side refers to the Frobenius norm.

Proof of Lemma C.2.5. From (C.2) and for $t \in [0:T]$, we obtain the integral representation of Σ_t

$$\begin{aligned}\Sigma_t =& \Sigma_0 + \int_0^t a_1(t, y_{[0:s]})\Sigma_s + \Sigma_s a_1^\top(s, y_{[0:s]}) + (b \circ b)(s, y_{[0:s]}) \\ &-[(b \circ B)(s, y_{[0:s]}) + \Sigma_s A_1^\top(s, y_{[0:s]})](B \circ B)^{-1}(s, y_{[0:s]}) \\ &\times[(b \circ B)(s, y_{[0:s]}) + \Sigma_s A_1^\top(s, y_{[0:s]})]^\top ds.\end{aligned}$$

Let $\Sigma^{(1)}_t$, $\Sigma^{(2)}_t$ be the covariance matrices corresponding to paths $y^{(1)}_\cdot, y^{(2)}_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$ where $y^{(1)}_\cdot$ is arbitrary and $y^{(2)}_\cdot$ s.t. $\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t < \epsilon$. The difference $\Sigma^{(1)}_t - \Sigma^{(2)}_t$ satisfies the integral representation

$$\begin{aligned}\Sigma^{(1)}_t - \Sigma^{(2)}_t = \Sigma^{(1)}_0 - \Sigma^{(2)}_0 &+ \int_0^t \Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s) ds \\ &+ \int_0^t (\Sigma^{(1)}_s - \Sigma^{(2)}_s)\Xi_2(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s) \\ &+ \Xi_2^\top(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)(\Sigma^{(1)}_s - \Sigma^{(2)}_s) ds,\end{aligned}$$

where we denote

$$\begin{aligned}&\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s) \\ &\quad \stackrel{\text{def.}}{=} (a_1(s, y^{(1)}_{[0:s]}) - a_1(s, y^{(2)}_{[0:s]}))\Sigma^{(1)}_s + \Sigma^{(1)}_s(a_1^\top(s, y^{(1)}_{[0:s]}) - a_1^\top(s, y^{(2)}_{[0:s]})) \\ &\qquad + (b \circ b)(s, y^{(1)}_{[0:s]}) - (b \circ b)(s, y^{(2)}_{[0:s]}) \\ &\qquad - \Sigma^{(2)}_s[A_1^\top(B \circ B)^{-1}(b \circ B)(s, y^{(1)}_{[0:s]}) - A_1^\top(B \circ B)^{-1}(b \circ B)(s, y^{(2)}_{[0:s]})] \\ &\qquad - [(b \circ B)(B \circ B)^{-1}A_1(s, y^{(1)}_{[0:s]}) - (b \circ B)(B \circ B)^{-1}A_1(s, y^{(2)}_{[0:s]})]\Sigma^{(2)}_s \\ &\qquad - \Sigma^{(2)}_s[A_1^\top(B \circ B)^{-1}A_1(s, y^{(1)}_{[0:s]}) - A_1^\top(B \circ B)^{-1}A_1(s, y^{(2)}_{[0:s]})]\Sigma^{(1)}_s\end{aligned}$$

$$- [(b \circ B)(B \circ B)^{-1}(b \circ B)^\top(s, y^{(1)}_{[0:s]}) - (b \circ B)(B \circ B)^{-1}(b \circ B)^\top(s, y^{(2)}_{[0:s]})] \tag{C.5}$$

and

$$\Xi_2(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s) \overset{\text{def.}}{=} a_1^\top(s, y^{(2)}_s) + A_1^\top(s, y^{(1)}_{[0:s]})(B \circ B)^{-1}(s, y^{(1)}_{[0:s]}) \times (b \circ B)(s, y^{(1)}_{[0:s]}) + A_1^\top(s, y^{(1)}_{[0:s]})(B \circ B)^{-1}(s, y^{(1)}_{[0:s]})A_1(s, y^{(1)}_{[0:s]})\Sigma^{(1)}_s. \tag{C.6}$$

By the triangular inequality, Jensen's inequality, and Cauchy-Schwarz inequality, we have

$$\begin{aligned}\left\|\Sigma^{(1)}_t - \Sigma^{(2)}_t\right\|_2 \leq& \left\|\Sigma^{(1)}_0 - \Sigma^{(2)}_0\right\|_2 + \int_0^t \left\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)\right\|_2 ds \\ &+ 2\int_0^t \left\|\Sigma^{(1)}_s - \Sigma^{(2)}_s\right\|_2 \cdot \left\|\Xi_2(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)\right\|_2 ds.\end{aligned}$$

By Grönwall's inequality, it holds that

$$\begin{aligned}\left\|\Sigma^{(1)}_t - \Sigma^{(2)}_t\right\|_2 \leq& \left\|\Sigma^{(1)}_0 - \Sigma^{(2)}_0\right\|_2 + \int_0^t \left\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)\right\|_2 ds \\ &+ \int_0^t \left(\left\|\Sigma^{(1)}_0 - \Sigma^{(2)}_0\right\|_2 + \int_0^s \left\|\Xi_1(r, y^{(1)}_{[0:r]}, y^{(2)}_{[0:r]}, \Sigma^{(1)}_r, \Sigma^{(2)}_r)\right\|_2 dr\right) \cdot \\ &\cdot 2\left\|\Xi_2(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)\right\|_2 \cdot \\ &\cdot \exp\left(2\int_s^t \left\|\Xi_2(r, y^{(1)}_{[0:r]}, y^{(2)}_{[0:r]}, \Sigma^{(1)}_r, \Sigma^{(2)}_r)\right\|_2 dr\right) ds.\end{aligned} \tag{C.7}$$

From Lemma C.2.6, we know that there exists a constant $K \in \mathbb{R}$ such that

$$\left\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)\right\|_2 \leq K\left\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\right\|_s. \tag{C.8}$$

Combining (C.7) and (C.8), we obtain

$$\begin{aligned}\left\|\Sigma^{(1)}_t - \Sigma^{(2)}_t\right\|_2 \leq& \left(\left\|\Sigma^{(1)}_0 - \Sigma^{(2)}_0\right\|_2 + KT\left\|y^{(1)} - y^{(2)}\right\|_T\right) \\ &\cdot \left[1 + \int_0^t 2\left\|\Xi_2(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)\right\|_2 \right. \\ &\left. \cdot \exp\left(2\int_s^t \left\|\Xi_2(r, y^{(1)}_{[0:r]}, y^{(2)}_{[0:r]}, \Sigma^{(1)}_r, \Sigma^{(2)}_r)\right\|_2 dr\right) ds\right].\end{aligned} \tag{C.9}$$

Since a_1, A_1, $(B \circ B)^{-1}$ and $b \circ B$ are globally (time-) Lipschitz continuous, Ξ_2 is uniformly bounded, according to (C.6). Therefore, from (C.9) we obtain a constant $C \in \mathbb{R}$ such that

$$\left\|\Sigma^{(1)}_t - \Sigma^{(2)}_t\right\|_2 \leq C(\left\|\Sigma^{(1)}_0 - \Sigma^{(2)}_0\right\|_2 + \left\|y^{(1)} - y^{(2)}\right\|_T).$$

We conclude the proof due to Assumption 4.2.1 (xiii). □

Lemma C.2.6. *Let Ξ_1 be defined as in (C.5), and let Assumption 4.2.1 (viii) hold. Then, for every path $y_\cdot^{(1)} \in C^1([0:T], \mathbb{R}^{d_Y})$ there exists constant $K \in \mathbb{R}$, such that for all $\epsilon > 0$, paths $y_\cdot^{(2)} \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t < \epsilon$, and times $s \in [0:T]$ holds*

$$\left\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)\right\|_2 \leq K \left\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\right\|_s.$$

Proof of Lemma C.2.6. Since a_1, $b \circ b$, $b \circ B$, A_1, $(B \circ B)^{-1}$ are locally Lipschitz continuous in the path component uniformly over all times, we conclude that a_1, $b \circ b$, $A_1^\top (B \circ B)^{-1}(b \circ B)$, $(b \circ B)(B \circ B)^{-1} A_1$, $A_1^\top (B \circ B)^{-1} A_1$, and $(b \circ B)(B \circ B)^{-1}(b \circ B)^\top$ have the same property. By this, together with the triangular inequality, there exists a constant $K_1 \in \mathbb{R}$

$$\begin{aligned}
&\left\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma^{(1)}_s, \Sigma^{(2)}_s)\right\| \\
&\quad \leq 2K_1 \left\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\right\|_s \left(\left\|\Sigma^{(1)}_s\right\|_2 + \left\|\Sigma^{(2)}_s\right\|_2\right) + K_1 \left\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\right\|_s \left\|\Sigma^{(1)}_s\right\|_2 \left\|\Sigma^{(2)}_s\right\|_2 \\
&\qquad\qquad + K_1 \left\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\right\|_s \\
&\quad = K_1 \left\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\right\|_s \left(1 + 2\|\Sigma^{(1)}_s\|_2 + 2\|\Sigma^{(2)}_s\|_2 + \left\|\Sigma^{(1)}_s\right\|_2 \left\|\Sigma^{(2)}_s\right\|_2\right).
\end{aligned}$$

With that, it is left to show that $\|\Sigma^{(1)}_s\|_2 < \infty$ and $\|\Sigma^{(2)}_s\|_2 < K_2$ for some $K_2 \in \mathbb{R}$. The former is trivially fulfilled due to the continuity of $a_1, b \circ b, b \circ B, A_1, (B \circ B)^{-1}$. Since Σ_t is continuous in the path-component and the closure of the set

$$\{y \in C^1([0:T]) \,:\, \|y_{[0:t]} - y^{(1)}_{[0:t]}\|_2 < \epsilon\}$$

is compact, we know $\|\Sigma^{(2)}_s\|_2$ is bounded. □

C.2.2 Lipschitz-Continuity in the Time-Component of the Optimal Filter

Lemma C.2.7 (Global Lipschitz-Cont. in the Time-Comp. of the Optimal Filter)**.** *Under Assumption 4.2.1, f_t from (1.4) is locally Lipschitz-continuous in the time-component. In particular, for every path $y_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$, $t \in [0:T]$, and $\epsilon > 0$ there exists constant $C \geq 0$ such that for all times $s \in [0:T]$ with $\|t - s\|_t < \epsilon$ holds*

$$\mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y_{[0:t]}), \mathbb{P}(X_s \in \cdot | y_{[0:s]})) \leq C|t - s|.$$

Proof of Lemma C.2.7. Analogously to the proof of Lemma C.2.1, we argue that by (4.11), Lemma C.2.10, and Lemma C.5.1 there exists a non-negative constant $\bar{K} < \infty$ satisfying the following for every path $y_\cdot \in C([0:T], \|\cdot\|_2)$ and all times $t, s \in [0:T]$

$$\mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y_{[0:t]}), \mathbb{P}(X_s \in \cdot | y_{[0:s]})) \\ \leq \bar{K} \sqrt{\|\mu(y_{[0:t]}) - \mu(y_{[0:s]})\|_2^2 + \|\Sigma(y_{[0:t]}) - \Sigma(y_{[0:s]}\|_2^2}.$$

We see from Lemma C.2.8 and Lemma C.2.9 that there exists a non-negative Lipschitz-constant $K < \infty$ for both μ and Σ such that

$$\mathcal{W}_2(\mathbb{P}(X_t \in \cdot | y_{[0:t]}), \mathbb{P}(X_s \in \cdot | y_{[0:s]})) \leq \sqrt{2}\bar{K}K|t-s|,$$

and the statement follows. □

Lemma C.2.8 (Global Lipschitz-continuity of μ in time)**.** *Let the Assumption 4.2.1 (viii) be fulfilled. Then, for any path $y \in C([0:T]), \mathbb{R}^d)$ there is a constant $K \in \mathbb{R}$ s.t. for all $t > s \in [0:T]$ holds*

$$\|\mu_t - \mu_s\|_2 \leq K|t-s|$$

where μ is defined as in (C.1).

Proof of Lemma C.2.8. With the notation of Lemma C.2.2, we have

$$\|\mu_t - \mu_s\|_2 \\ \leq \underbrace{\left\| \int_s^t a_0(u, y_{[0:u]}) du \right\|_2}_{\text{(I)}} + \underbrace{\left\| \int_s^t \Phi_2(u, y_{[0:u]}) dy_{[0:u]} \right\|_2}_{\text{(II)}} + \underbrace{\left\| \int_s^t \Phi_1(u, y_{[0:u]}) \mu_u du \right\|_2}_{\text{(III)}}.$$

We see that (I) $\leq K|t-s|$ as $\max_{u \in [0:T]} \|a_0(u, y_{[0:u]})\|_2 < \infty$ due to (time-)Lipschitz-continuity of a_0.

By the use of integration by parts,

$$\text{(II)} = \left\| \Phi_2(t, y_{[0:t]}) y_{[0:t]} - \Phi_2(s, y_{[0:s]}) y_{[0:s]} - \int_0^t y_{[0:u]} \, d\Phi_2(u, y_{[0:u]}) \right\|_2 \\ \leq \underbrace{\left\| \Phi_2(t, y_{[0:t]}) y_{[0:t]} - \Phi_2(0, y_0) y_0 \right\|_2}_{\text{(IV)}} + \underbrace{\left\| \int_0^t y_{[0:u]} \, d\Phi_2(u, y_{[0:u]}) \right\|_2}_{\text{(V)}},$$

we see that (IV) $\leq K\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t$ due to (path-)Lipschitz continuity of Φ_2. Moreover, let $\{[u_{i-1} : u_i] : i \in \{1, \dots, I\}\}, I \in \mathbb{N}$ be a fine enough partition of $[s : t]$, then we see due to global (time-)Lipschitz continuity of Φ_2 and the definition of the Riemann-Stieltjes integral that

$$\begin{aligned}(\text{V}) &\leq \epsilon_I + \sum_{i=1}^{I} \|y_{[0:u_i]}\|_2 \left\|\Phi_2(u_i, y_{[0:u_i]}) - \Phi_2(u_{i-1}, y_{[0:u_{i-1}]})\right\|_2 \\ &\leq \epsilon_I + \|y_{[0:t]}\|_t \sum_{i=1}^{I} K|u_i - u_{i-1}|\end{aligned}$$

and, since $\epsilon_I \to 0$ for $I \to \infty$, it follows (II) $\leq K|t - s|$.

Due to (I) $< \infty$ and (II) $< \infty$, we can apply Grönwall's inequality to

$$\begin{aligned}\|\mu_t\|_2 \leq \overbrace{\left\| \int_0^t a_0(u, y_{[0:u]})du \right\|_2 + \left\| \int_0^t \Phi_2(u, y_{[0:u]})dy_{[0:u]} \right\|_2}^{\stackrel{\text{def.}}{=} \alpha(t)} \\ + \int_0^t \left\|\Phi_1(u, y_{[0:u]})\right\|_2 \left\|\mu_u\right\|_2 du,\end{aligned}$$

which yields

$$\|\mu_t\|_2 \leq \alpha(t) + \int_0^t \alpha(s) \left\|\Phi_1(s, y_{[0:s]})\right\|_2 \exp\left(\int_s^t \left\|\Phi_1(u, y_{[0:u]})\right\|_2 du \right) ds \leq Kt.$$

where the last equality holds due to $\alpha(s) \leq Ks$ for all $s \in [0 : T]$ and (time-)Lipschitz continuity of Φ_1, implying $\max_{u \in [0:t]} \|\Phi_1(u, y_{[0:u]})\|_2 < \infty$.

With that, we have

$$(\text{III}) \leq \int_s^t \left\|\Phi_1(u, y_{[0:u]})\right\|_2 \left\|\mu_u\right\|_2 du \leq \max_{u \in [0:t]} \|\Phi_1(u, y_{[0:u]})\|_2 KT|t - s|,$$

which proves the claim. □

Lemma C.2.9 (Global Lipschitz-continuity of Σ in time)**.** *Let Assumption 4.2.1 (viii) hold. Then, for any path $y \in C([0 : T]), \mathbb{R}^d)$ there is a constant $K \in \mathbb{R}$ s.t. for all $t > s \in [0 : T]$ holds*

$$\|\Sigma_t - \Sigma_s\|_2 \leq K|t - s|.$$

where Σ is defined as in (C.2).

Proof of Lemma C.2.9. Note that $\Sigma_\cdot$ is continuous in time. From (C.2), we obtain

$$\begin{aligned}\|\Sigma_t - \Sigma_s\|_2 = \int_s^t &\|a_1(t, y_{[0:u]})\Sigma_u\|_2 + \|\Sigma_u a_1^\top(u, y_{[0:u]})\|_2 + \|(b \circ b)(u, y_{[0:u]})\|_2 \\ &+ \|(b \circ B)(u, y_{[0:u]})(B \circ B)^{-1}(u, y_{[0:u]})(b \circ B)(u, y_{[0:u]})^\top\|_2 \\ &+ \|(b \circ B)(u, y_{[0:u]})(B \circ B)^{-1}(u, y_{[0:u]})\Sigma_u A_1^\top(u, y_{[0:u]})^\top\|_2 \\ &+ \|\Sigma_u A_1^\top(u, y_{[0:u]})(B \circ B)^{-1}(u, y_{[0:u]})(b \circ B)(u, y_{[0:u]})^\top\|_2 \\ &+ \|\Sigma_u A_1^\top(u, y_{[0:u]})(B \circ B)^{-1}(u, y_{[0:u]})\Sigma_u A_1^\top(u, y_{[0:u]})^\top\|_2 du.\end{aligned}$$

As all components are continuous, we can bound them by their maximal value on $[0:T]$. With this, we conclude $\|\Sigma_t - \Sigma_s\|_2 \leq K|t-s|$. □

C.2.3 Lower Loewner order bound

Lemma C.2.10 (Lower Loewner order bound)**.** *Let Σ be defined as in Equation (C.2) and Assumption 4.2.1 (ix), (x), hold. Then there exists an $r > 0$, s.t. for all times $t \in [0:T]$ and paths $y_\cdot \in C^1([0:T], \mathbb{R}^{d_Y})$ holds that $rI_d \preccurlyeq \Sigma_t$ (where $\preccurlyeq$ is the Loewner bound).*

Proof of Lemma C.2.10. By the Courant-Fisher Theorem, see [148], the statement is equivalent to finding a lower bound on the eigenvalues of Σ, which we will show by finding an upper bound on the Eigenvalues of Σ^{-1}, which exists as it is positive definite.

Following the proof of [201, Theorem 12.7], we note that

$$\begin{aligned}\partial_t \Sigma_t^{-1} = &- \tilde{a}_1^\top(t, y_{[0:t]})\Sigma_t^{-1} - \Sigma_t^{-1}\tilde{a}_1(t, y_{[0:t]}) + A_1^\top(B \circ B)^{-1}A_1(t, y_{[0:t]}) \\ &- \Sigma_t^{-1}\left[(b \circ b)(t, y_{[0:t]}) - (b \circ B)(B \circ B)(b \circ B)^\top(t, y_{[0:t]})\right]\Sigma_t^{-1}\end{aligned}$$

with

$$\tilde{a}_1^\top(t, y_{[0:t]}) \overset{\text{def.}}{=} a_1(t, y_{[0:t]}) - (b \circ B)(B \circ B)A_1(t, y_{[0:t]}).$$

Now, let $G_t(y_{[0:t]})$ be a solution of $\partial_t G_t(y_{[0:t]}) = \tilde{a}_1(t, y_{[0:t]})G_t(y_{[0:t]})$ with $G_0(y_{[0:t]}) = I_{d_X}$. Then, as pointed out in [201, Theorem 12.7], we arrive due to $(b \circ b)(t, y_{[0:t]}) - (b \circ B)(B \circ B)(b \circ B)^\top(t, y_{[0:t]})$ being positive semi-definite at

$$\operatorname{tr}(G_t(y_{[0:t]})\Sigma_t^{-1}G_t(y_{[0:t]})^\top)$$

$$\leq \operatorname{tr}(\Sigma_0^{-1}) + \int_0^T \operatorname{tr}\left(G_s(y_{[0:s]})A_1^\top (B \circ B)^{-1} A_1(s, y_{[0:s]}) G_s(y_{[0:s]})^\top\right) ds.$$

As $G_t(y_{[0:t]})$ is positive definite, we conclude by submultiplicativity of positive semi-definite matrices that

$$\begin{aligned}\operatorname{tr}(\Sigma_t^{-1}) &\leq \operatorname{tr}(G_t(y_{[0:t]})^{-1} G_t(y_{[0:t]}) \Sigma_t^{-1} G_t(y_{[0:t]})^\top (G_t(y_{[0:t]})^{-1})^\top) \\ &\leq \operatorname{tr}(G_t(y_{[0:t]})^{-1})^2 \operatorname{tr}(G_t(y_{[0:t]}) \Sigma_t^{-1} G_t(y_{[0:t]})^\top)\end{aligned}$$

and

$$\begin{aligned}&\operatorname{tr}\left(G_s(y_{[0:s]})A_1^\top (B \circ B)^{-1} A_1(s, y_{[0:s]}) G_s(y_{[0:s]})^\top\right) \\ &\qquad \leq \operatorname{tr}\left(G_s(y_{[0:s]})\right)^2 \operatorname{tr}\left(A_1^\top (B \circ B)^{-1} A_1(s, y_{[0:s]})\right).\end{aligned}$$

As we assumed that there exist constants $K_1, K_2 > 0$ s.t. uniformly $K_1 \leq \operatorname{tr}(G_t(y_{[0:t]})) \leq K_2$ as well as $K_3 > 0$ s.t. $\operatorname{tr}(A_1^\top (B \circ B)^{-1} A_1(t, y_{[0:t]})) \leq K_3$ we conclude $\operatorname{tr}(\Sigma_t^{-1}) < K$ uniformly for some constant $K > 0$. As the trace is the sum of all Eigenvalues, this proves the claim. □

C.3 Stable and Lossless Encoding — Proof of Proposition 4.3.1

In the case where the "latent geometry" of K has additional structure, we may guarantee that the attention mechanism attn_T is linearly-stable with linearly-stable inverse, how to generate the reference paths $y^{(1)}, \ldots, y^{(N_{\mathrm{ref}})}$ in $C([0:T], \mathbb{R}^{d_Y})$, and quantitative estimates on how many paths must be generated.

Lemma C.3.1 (Stable Lossless Feature Maps – Riemannian Case)**.** *Under Assumption 4.2.2 (iii), there exists a constant $C_{d_K} > 0$, depending only on d_K, and a $\delta > 0$ depending only on (M, g) such that any δ-packing $\{y^{(n)}\}_{n=1}^{d_K}$ of K, there are matrices A, B, b, V, U, C and vectors a, b for which the parameter θ, as in Definition 4.3.3, defining a pathwise attention mechanism $\mathrm{attn}_T^\theta : C([0:T], \mathbb{R}^{d_Y}) \to \mathbb{R}^N$ which restricts to a bi-Lipschitz embedding of $(K, \|\cdot\|_T)$ into $(\mathbb{R}^{N+1}, \|\cdot\|_2)$. Furthermore, $\|\theta\|_0 \in \mathcal{O}(N^2)$. Moreover, if $(\mathcal{M}, g)$ is aspherical then:*

(i) $\delta \in \mathcal{O}(\mathrm{Vol}(\mathcal{M}, g)^{1/d_K})$,

(ii) $\|\theta\|_0 \in \mathcal{O}(N^2)$,

(iii) $C_{d_K} \in \mathcal{O}({d_K}^{3(d_K+1)/2})$

In particular, the attention mechanism attn_T *satisfies [182, Setting 3.6 (i)] for any Borel probability measure* $\mathbb{P}_{\mathrm{in}}$ *on* K.

Proof of Lemma C.3.1. *Step 1:* N_{ref}*-estimate on the Riemannian manifold* $(\mathcal{M}, g)$. Since $(K, \|\cdot\|_T)$ be isometric to $(\mathcal{M}, d_g)$ where d_g is the geodesic distance on an aspherical compact d_K-dimensional Riemannian manifold $(\mathcal{M}, g)$, then Gromov's systolic inequality [129, Theorem 0.1.A] implies that the systole of $\mathcal{M}$ denoted by $\mathrm{sys}(\mathcal{M})$ satisfies

$$\mathrm{sys}(\mathcal{M}) \leq \tilde{C}_{d_K} \ \mathrm{Vol}(\mathcal{M}, g)^{1/\dim(\mathcal{M})}, \tag{C.10}$$

where $\mathrm{Vol}(\mathcal{M}, g)$ denotes the Riemannian volume of $(\mathcal{M}, g)$ and $\tilde{C}_{d_K} > 0$ is a universal constant only depending on d_K satisfying

$$0 < \tilde{C}_{d_K} < 6(d_K + 1){d_K}^{d_K} \sqrt{(d_K + 1)!}\,.$$

By Stirling's approximation, $\sqrt{(d_K + 1)!} \in \mathcal{O}({d_K}^{(d_K+1)/2})$; whence it follows that $\tilde{C}_{d_K} \in \mathcal{O}({d_K}^{3(d_K+1)/2})$.

In the proof of [168, Theorem 1] (circa [168, Equation (1.1)]) we see that if $\delta = \mathrm{Sys}(\mathcal{M})/10$ then given any δ-net $\widetilde{\mathbb{X}}$ in $(\mathcal{M}, g)$, the map $\varphi_2 : (\mathcal{M}, d_g) \mapsto (\mathbb{R}^{\#\widetilde{\mathbb{X}}}, \|\cdot\|_2)$ given for any $p \in M$ by

$$\varphi_2 : p \mapsto \Big(d_g(p, u)\Big)_{u \in \widetilde{\mathbb{X}}}$$

is a bi-Lipschitz embedding. Set $C_{d_K} \overset{\text{def.}}{=} \tilde{C}_{d_K}/10$ and $N_{\mathrm{ref}} \overset{\text{def.}}{=} \#\widetilde{\mathbb{X}}$. Therefore,

$$\delta \leq C_{d_K} \ \mathrm{Vol}(\mathcal{M}, g)^{1/d_K}.$$

Enumerate $\widetilde{\mathbb{X}} \overset{\text{def.}}{=} \{u_n\}_{n=1}^{N_{\mathrm{ref}}}$. We note that if K is not aspherical but if it is only isometric to a closed Riemannian manifold, then the conclusion still holds, however, without this explicit upper bound on δ.

Step 2: Building the Feature Map with Well-posed Inverse. Let $\varphi_1 : (\mathcal{M}, d_g) \to (K, \|\cdot\|_T)$ be any isometry, which we have postulated to exist. Set $\mathbb{X} \overset{\text{def.}}{=} \{\varphi(u) : u \in \widetilde{\mathbb{X}}\} = \{y^{(n)}\}_{n=1}^{N_{\text{ref}}}$ where, for $n = 1, \dots, N_{\text{ref}}$ we define $y^{(n)} \overset{\text{def.}}{=} \varphi_1(u_n)$. Define

$$\varphi : C([0:T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N_{\text{ref}}}, \quad y_\cdot \mapsto (\|y_\cdot - y_\cdot^{(n)}\|_T)_{n=1}^{N_{\text{ref}}}.$$

Observe that, for every path $y \in K$, the following holds

$$\begin{aligned}
\varphi(y_\cdot) &\overset{\text{def.}}{=} \left(\|y_\cdot - y_\cdot^{(n)}\|_T\right)_{n=1}^{N_{\text{ref}}} \\
&= \left(\left\|\varphi_1 \circ \varphi_1|^{-1}_{\varphi_1(K)}(y_\cdot) - \varphi_1 \circ \varphi_1|^{-1}_{\varphi_1(K)}(y_\cdot^{(n)})\right\|_T\right)_{n=1}^{N_{\text{ref}}} \\
&= \left(d_g\left(\varphi_1|^{-1}_{\varphi_1(K)}(y_\cdot), \varphi_1|^{-1}_{\varphi_1(K)}(y_\cdot^{(n)})\right)\right)_{n=1}^{N_{\text{ref}}} \qquad \text{(C.11)} \\
&= \left(d_g\left(\varphi_1|^{-1}_{\varphi_1(K)}(y_\cdot), u^n\right)\right)_{n=1}^{N_{\text{ref}}} \qquad \text{(C.12)} \\
&= \varphi_2 \circ \varphi_1|^{-1}_{\varphi_1(K)}(y_\cdot),
\end{aligned}$$

where (C.11) holds since φ_2 is an isometry and (C.12) holds unambiguously since φ_1 is a bijection from $\mathcal{M}$ to K. Since every isometry is a bi-Lipschitz map, the compositions of bi-Lipschitz maps is again bi-Lipschitz, and since we have just shown that $\varphi|_K = \varphi_2 \circ \varphi_1|^{-1}_{\varphi(K)}$ then, $\varphi|_K$ is a bi-Lipschitz embedding of K into the N_{ref}-dimensional Euclidean space.

Step 3: Aligning to a Hyperplane in $\mathbb{R}^{N_{\text{sim}}}$ with a Shallow ReLU Neural Network. Set $N_{\text{sim}} \overset{\text{def.}}{=} N_{\text{ref}} + 1$. Let b be the zero vector in $\mathbb{R}^{2N_{\text{ref}}}$. We now consider a variation of the example on [64, page 3], the respective $N_{\text{ref}} \times 2N_{\text{ref}}$ and $2N_{\text{ref}} \times N_{\text{ref}}$ block-matrices A_1 and B

$$A_1 = \begin{pmatrix} I_{N_{\text{ref}}} & -I_{N_{\text{ref}}} \end{pmatrix} \text{ and } B = \begin{pmatrix} I_{N_{\text{ref}}} \\ -I_{N_{\text{ref}}} \end{pmatrix}$$

are such that the ReLU neural network $\tilde{\psi} : \mathbb{R}^{N_{\text{ref}}} \to \mathbb{R}^{N_{\text{ref}}}$, with $\tilde{\psi}(u) \overset{\text{def.}}{=} A_1 \operatorname{ReLU} \bullet (Bu + b)$ satisfies $\tilde{\psi}(u) = u$, for each $u \in \mathbb{R}^{N_{\text{ref}}}$. Consider the $N_{\text{sim}} \times N_{\text{ref}}$ block-matrix A_2 and the vector $a \in \mathbb{R}^{N_{\text{sim}}}$ given by

$$A_2 = \begin{pmatrix} I_{N_{\text{ref}}} \\ 0 \end{pmatrix} \text{ and } a_i = \begin{cases} 0 & \text{if } i = 1, \dots, N_{\text{ref}}, \\ 1 & \text{if } i = N_{\text{sim}}. \end{cases}$$

Set $A \overset{\text{def.}}{=} A_2 A_1$, $\psi \overset{\text{def.}}{=} A \operatorname{ReLU} \bullet (Bu + b) + a$, and note that ψ is a ReLU neural network. A direct computation shows that, $\|A\|_0 = 2N_{\text{ref}}$, $\|B\|_0 = 2N_{\text{ref}}$, $\|a\|_0 = 1$, and $\|b\|_0 = 0$.

Step 4: Injectivity of Softmax Function on Hyperplane. Note that, for each $u \in \mathbb{R}^{N_{\text{ref}}}$

$$\psi(u) = (u_1, \dots, u_N, 1)^\top,$$

therefore bijectively ψ maps $\mathbb{R}^{N_{\text{ref}}}$ onto the N_{ref}-dimensional hyperplane $H \overset{\text{def.}}{=} \{(u,1) : u \in \mathbb{R}^{N_{\text{ref}}}\}$ in $\mathbb{R}^{N_{\text{sim}}}$. Observe that the Softmax function, given for any $u \in \mathbb{R}^{N_{\text{sim}}}$ by $\mathrm{Softmax} : u \mapsto (e^{u_j} / \sum_{k=1}^{N_{\text{ref}}} e^{u_k})_{j=1}^{N_{\text{sim}}}$, maps H surjectively and continuously onto the image set

$$\mathrm{Softmax}(\mathbb{R}^{N_{\text{sim}}}) \subseteq \mathrm{int}(\Delta_{N_{\text{sim}}}) \overset{\text{def.}}{=} \left\{ v \in [0,1]^{N_{\text{sim}}} : \sum_{i=1}^{N_{\text{sim}}} v_i = 1 \right\};$$

i.e. of the interior of the N_{sim}-simplex. Since

$$v \mapsto \Big(\ln(v_i) - \ln(v_{N_{\text{sim}}}) + 1 \Big)_{i=1}^{N_{\text{sim}}} \tag{C.13}$$

is a continuous right-inverse of $\mathrm{Softmax}$ on the interior of the image set $\mathrm{Softmax}(\mathbb{R}^{N_{\text{sim}}})$, the softmax function defines a continuous bijection from H onto $\mathrm{Softmax}(\mathbb{R}^{N_{\text{sim}}})$.

Therefore, the map $\Psi \overset{\text{def.}}{=} \mathrm{Softmax} \circ \psi \circ \varphi$ is a continuous and injective when restricted to K. Since K is compact and Ψ is continuous then $\Psi(K)$ is a compact subset of the interior $\mathrm{int}(\Delta_{N_{\text{sim}}})$ of the N_{sim}-simplex, then,

$$C^\star \overset{\text{def.}}{=} \max_{u \in \Phi(K)} \min_{v \in \Delta_{N_{\text{sim}}} \setminus \mathrm{int}(\Delta_{N_{\text{sim}}})} \|u - v\| > 0.$$

Therefore, $\Psi(K)$ is contained in the set

$$\Delta^\star_{N_{\text{sim}}} \overset{\text{def.}}{=} \{ t^\star (u - \bar{\Delta}_{N_{\text{sim}}}) + \bar{\Delta}_{N_{\text{sim}}} \},$$

where $\bar{\Delta}_{N_{\text{sim}}} \overset{\text{def.}}{=} (1/N_{\text{sim}}, \dots, 1/N_{\text{sim}})$ is the barycenter of the N_{sim}-simplex $\Delta_{N_{\text{sim}}}$ for some $t^\star \in [0,1)$ such that $\Psi(K) \subseteq \Delta^\star_{N_{\text{sim}}}$ (which is possible since $C^\star > 0$). Since the map (C.13) is locally Lipschitz on $\mathrm{int}(\Delta_{N_{\text{sim}}})$, it is Lipschitz on the compact set $\Delta^\star_{N_{\text{sim}}}$ then Ψ has a Lipschitz inverse on its image since it is the composition of Lipschitz functions with Lipschitz inverses.

Step 5: Representation as an Attention Mechanism. Set $\theta_0 = (A, B, a, b, \{y_\cdot^{(n)}\}_{n=1}^{N_{\text{ref}}})$ and observe that the map Ψ is of the form (4.5). We continue by choosing a map of

the form (4.6). For this, let $N_{\text{time}} \in \mathbb{N}$, set $N_{\text{pos}} \overset{\text{def.}}{=} N_{\text{sim}}$ and let U be the $N_{\text{pos}} \times N_{\text{time}}$-dimensional zero matrix, and let V be the $N_{\text{pos}} \times d_Y$-dimensional matrix whose entries are all equal to 1. Set $\theta_1 \overset{\text{def.}}{=} (U, V, \{n/N_{\text{pos}}\}_{n=1}^{N_{\text{time}}})$. Define the encoding dimension $N \overset{\text{def.}}{=} N_{\text{sim}}$ and let C be the $N \times N \cdot d_Y$-dimensional matrix given by $Cx = (x_{n \cdot d_Y})_{n=1}^{N}$ for every $x \in \mathbb{R}^{N d_Y}$. Set $\theta \overset{\text{def.}}{=} (\theta_0, \theta_1, C)$ and observe that

$$\operatorname{attn}_T^{\theta}(t, y_\cdot) \overset{\text{def.}}{=} \Big(t, C \operatorname{vec}\Big(\operatorname{sim}_T^{\theta_0}(y_\cdot) \odot \operatorname{post}_T^{\theta_1}(y_\cdot)\Big)\Big) = \Big(1_{[0:T]} \times \Psi\Big)(t, y_\cdot)$$

for each $y_\cdot \in K_\cdot$ and every $y \in [0 : T]$. Consequentially, $\operatorname{attn}_T^{\theta_\cdot}$ is a bi-Lipschitz embedding of $[0 : T] \times K$ to $\mathbb{R}^{1+N}$. Note that $\|C\|_0 = N_{\text{pos}}$, $\|U\|_0 = 0$, and $\|V\|_0 = N_{\text{pos}} d_Y$. Thus, $\|\theta\|_0 \in \mathcal{O}(N_{\text{pos}} \cdot d_Y + N_{\text{ref}})$.

This concludes the proof. □

Lemma C.3.2 (Stable Lossless Feature Maps – Finite Case)**.** *Under Assumption 4.2.2 (i) and let* $N_{\text{ref}} \overset{\text{def.}}{=} \#K$. *Set* $N = 192\lceil \log(N_{\text{ref}}) \rceil$ *and* $\{y_i\}_{i=1}^{k}$ *enumerate* K. *There is a parameter* θ *as in Definition 4.3.3 defining a pathwise attention mechanism* $\operatorname{attn}_T^{\theta} : C([0 : T], \mathbb{R}^{d_Y}) \to \mathbb{R}^N$ *which restricts to a* $(2^{-1/2}, (3N_{\text{ref}}/6)^{1/2})$*-bi-Lipschitz embedding of* $(K, \|\cdot\|_T)$ *into* $(\mathbb{R}^{1+N}, \|\cdot\|_2)$.

In particular, attn_T *satisfies [182, Setting 3.6 (i)].*

Proof of Lemma C.3.2. We argue similarly to the proof of Lemma C.3.1, with only Steps 1 and 2 of its proof being replaced with the following argument.

Steps 1-2 (Modified): Building the Feature Map.

Enumerate $K = \{x_n\}_{n=1}^{N_{\text{ref}}}$. The map $\varphi_1 : (K, \|\cdot\|_T) \to (\mathbb{R}^{N_{\text{ref}}}, \|\cdot\|_\infty)$ given for every $x \in K$ by

$$x \mapsto (\|x - x_n\|_t)_{n=1}^{N_{\text{ref}}}$$

is an isometric embedding, called the *Kuratowski embedding* (see [140, page 99]). The optimal constants for the equivalence of the Euclidean $\|\cdot\|_2$ and max $\|\cdot\|_\infty$ norms are given by

$$\|u\|_\infty \leq \|u\|_2 \quad \text{and} \quad \|u\|_2 \leq {N_{\text{ref}}}^{1/2} \|u\|_\infty \tag{C.14}$$

for each $u \in \mathbb{R}^{N_{\text{ref}}}$. Therefore, the "set theoretic identity map" $\varphi_2 : (\mathbb{R}^{N_{\text{ref}}}, \|\cdot\|_\infty) \to (\mathbb{R}^{N_{\text{ref}}}, \|\cdot\|_2)$ is bi-Lipschitz with optimal (shrinking and expansion) constants given

by (C.14). Define $N_{\text{sim}} \overset{\text{def.}}{=} \lceil 48 \ln(N_{\text{ref}}) \rceil$. By the Johnson-Lindenstrauss Lemma, with "small" constant given in the derivation of [99, Theorem 2.1][1], exists a linear map $\varphi_3 : (\mathbb{R}^{N_{\text{ref}}}, \|\cdot\|_2) \to (\mathbb{R}^{N_{\text{sim}}}, \|\cdot\|_2)$, i.e. $\varphi_3(u) = A_0\, u$ for some $N_{\text{sim}} \times N_{\text{ref}}$-matrix A_0, satisfying: For every $u, v \in \mathbb{R}^{N_{\text{ref}}}$

$$2^{-1/2} \|u - v\|_2 \leq \|A_0\, u - A_0\, v\|_2 \leq (3/2)^{1/2} \|u - v\|_2.$$

Consequentially, the map $\varphi : (K, \|\cdot\|_T) \to (\mathbb{R}^{N_{\text{sim}}}, \|\cdot\|_2)$ given by $\varphi \overset{\text{def.}}{=} \varphi_3 \circ \varphi_2 \circ \varphi_1$ is $(2^{-1/2}, (3N_{\text{ref}}/2)^{1/2})$-bi-Lipschitz, since the Kuratowksi embedding φ_1 is an isometry and φ_2 satisfies (C.14); this is because for each $x, \tilde{x} \in K$ we have

$$\begin{aligned}
2^{-1/2} \|x - \tilde{x}\|_T =& 2^{-1/2} \|\varphi_1(x) - \varphi_1(\tilde{x})\|_\infty \\
\leq& 2^{-1/2} \|\varphi_2 \circ \varphi_1(x) - \varphi_2 \circ \varphi_1(\tilde{x})\|_2 \\
\leq& \|\varphi_3 \circ \varphi_2 \circ \varphi_1(x) - \varphi_3 \circ \varphi_2 \circ \varphi_1(\tilde{x})\|_2 \\
\overset{\text{def.}}{=}& \|\varphi(x) - \varphi(\tilde{x})\|_2 \\
\overset{\text{def.}}{=}& \|\varphi_3 \circ \varphi_2 \circ \varphi_1(x) - \varphi_3 \circ \varphi_2 \circ \varphi_1(\tilde{x})\|_2 \\
\leq& (3/2)^{1/2} \|\varphi_2 \circ \varphi_1(x) - \varphi_2 \circ \varphi_1(\tilde{x})\|_2 \\
\leq& (3/2)^{1/2}\, {N_{\text{ref}}}^{1/2} \|\varphi_1(x) - \varphi_1(\tilde{x})\|_\infty \\
=& (3N_{\text{ref}}/2)^{1/2} \|x - \tilde{x}\|_T.
\end{aligned}$$

Remaining Steps. The rest of the proof is identical to Steps 3 to 5 of the proof of Lemma C.3.1 but with B defined instead as $B \overset{\text{def.}}{=} B_1 A_0$, where $B_1 = (I_{N_{\text{sim}}}\ I_{N_{\text{sim}}})^\top$.[2]

This concludes the proof. □

Lemma C.3.3 (Stable Lossless Feature Maps – Linear Case)**.** *Fix $N_{\text{p}} \in \mathbb{N}_+$, $0 = t_0 < \cdots < t_{N_{\text{p}}} = T$, and a constant $C_K > 0$. Let $K \subset C([0:T], \mathbb{R}^{d_Y})$ satisfy Assumption 4.2.2 (ii). Then, there is a parameter θ, as in Definition 4.3.3, defining a pathwise attention mechanism $\text{attn}^\theta_T : C([0:T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N+1}$, with $N \overset{\text{def.}}{=} N_{\text{p}} \cdot d_Y$ which*

[1] We use this formulation with the constant in $4 * 48$ as opposed to $4 * 200$ in the standard derivation; found for example in [165].

[2] Here N_{sim} corresponds to $N_{\text{sim}} - 1$ in Step 3.

restricts to a bi-Lipschitz embedding of $(K, \|\cdot\|_T)$ *into* $(\mathbb{R}^{N+1}, \|\cdot\|_2)$. *Furthermore,* $\|\theta\|_0 \in \mathcal{O}(N_{\mathrm{p}} \cdot d_Y)$.

In particular, attn_T *satisfies [182, Setting 3.6 (i)].*

Proof of Lemma C.3.3. Let $N_{\mathrm{ref}} \in \mathbb{N}_+$, set $N_{\mathrm{sim}} \overset{\text{def.}}{=} N_{\mathrm{p}}$, and define A, B as zero-matrices as well as a, b as zero-vectors according to the dimensions in Definition 4.3.1. By fixing $\theta_0 = (A, B, a, b)$, we observe that

$$\mathrm{sim}_T^{\theta_0}(y_\cdot) = 1 \in \mathbb{R}^{N_{\mathrm{p}}} \tag{C.15}$$

for all $y \in C([0:T], \mathbb{R}^{d_Y})$.

Set $N_{\mathrm{time}} \overset{\text{def.}}{=} N_{\mathrm{pos}} \overset{\text{def.}}{=} N_{\mathrm{p}}$. Then, let V be the zero matrix and U the identity matrix according to the dimensions in Definition 4.3.2[3]. Set $\theta_1 = (U, V, \{t_n\}_{n=1}^{N_{\mathrm{p}}})$ and with that,

$$\mathrm{pos}_T^{\theta_1}(y_\cdot) = (y_1, \ldots, y_{N_{\mathrm{p}}}). \tag{C.16}$$

In particular, the map $\mathrm{pos}_T^{\theta_1}$ in (C.16) defines a linear bijection between K and the set $B_{C_K}(0)^{N_{\mathrm{p}}}$, where $B_{C_K}(0) := \{z \in \mathbb{R}^{d_Y} : |z| \leq C_K\}$. Since $B_{C_K}(0)^{N_{\mathrm{p}}}$ is a subset of the finite-dimensional Banach space $(\mathbb{R}^{N_{\mathrm{p}} \times d_Y}, \|\cdot\|_2)$ and all norms on a finite-dimensional normed are equivalent, then, $\mathrm{pos}_T^{\theta_1}$ is a bi-Lipschitz embedding of K into $(\mathbb{R}^{N_{\mathrm{p}} \times d_Y}, \|\cdot\|_2)$.

Fix $N \overset{\text{def.}}{=} N_{\mathrm{p}} \cdot d_Y$ and let C be the $N \times N$-dimensional matrix with $C_{i,j} = 1$ for $i, j = 1, \ldots, N$. Set $\theta = (\theta_0, \theta_1, C)$. By construction $\|\theta\|_0 \in \mathcal{O}(N_{\mathrm{p}} \cdot d_Y)$. Then,

$$\mathrm{attn}_T^{\theta}(t, y_\cdot) \overset{\text{def.}}{=} \Big(t, C\ \mathrm{vec}\Big(\mathrm{sim}_T^{\theta_0}(y_\cdot) \odot \mathrm{post}_T^{\theta_1}(y_\cdot)\Big)\Big) = \Big(t, (y_1, \ldots, y_{N_{\mathrm{p}}})\Big). \tag{C.17}$$

Together, (C.15) and (C.16) imply that attn_T^{θ} defines a bi-Lipschitz embedding of K into the $N+1$-dimensional Euclidean space. □

Proof of Proposition 4.3.1. Follows directly from Lemmas C.3.1 to C.3.3. □

[3]The point t_0 is not sampled since every path $y_\cdot$ in K satisfies $y_0 = 0$ and, thus, there is no need to sample it at time 0.

C.4 Approximability of Locally Lipschitz Maps By Filterformers — Proof of Proposition 4.4.3

We now prove our main universal approximation theorem, Proposition 4.4.3. We show that the target space/codomain of any considered Filterformer is geometrically regular, in the sense of [1, 182]. Using this fact, we then combine and apply the results of [180, 182] to deduce the result.

C.4.1 Polish QAS Space Structure on the space of Positive Semi-Definite Matrices

Let $d_X \in \mathbb{N}_+$, and Sym_{0,d_X} denote the set of $d_X \times d_X$ symmetric positive semi-definite matrices and let $\|\cdot\|_F$ denote the Frobenius norm on the set of $d_X \times d_X$ matrices.[4] Consider the 2-product metric on $\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ given for any $(m^{(1)}, A), (m^{(2)}, B)$ by

$$\mathrm{d}_{2,F}\Big((m^{(1)}, A), (m^{(2)}, B)\Big)^2 \overset{\text{def.}}{=} \|m^{(1)} - m^{(2)}\|^2 + \|A - B\|_F^2. \tag{C.18}$$

We show that $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathrm{d}_{2,F})$ satisfies the conditions of [182, Theorem 3.7]; namely [182, Setting 3.6 (iii)]. This requires defining a few maps first. For $q \in \mathbb{N}$ define the map $\mathcal{Q}_q : \mathbb{R}^{d_X + d_X^2} \to \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ by sending any $(m, A) \in \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ to

$$\mathcal{Q}_q((m, A)) \overset{\text{def.}}{=} (m, A^\top A).$$

The family $\mathcal{Q} \overset{\text{def.}}{=} (\mathcal{Q}_q)_{q \in \mathbb{N}_+}$ quantizes $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathrm{d}_{2,F})$, in the sense of [1, Definition 3.2].

Next, we consider the so-called *mixing function*

$$\eta : \cup_{N \in \mathbb{N}_+} \Delta_N \times (\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X})^N \to \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X},$$

where $N \in \mathbb{N}_+$ and Δ_N is the N-simplex, defined for any $N \in \mathbb{N}_+$, $w \in \Delta_N$, and $(m^{(1)}, A^{(1)}), \ldots, (m^{(N)}, A^{(N)}) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_X^2}$ by

$$\eta\Big(w, \{(m^{(n)}, A^{(n)}\}_{n=1}^N\Big) \overset{\text{def.}}{=} \sum_{n=1}^N w_n \cdot (m^{(n)}, A^{(n)}). \tag{C.19}$$

[4] The 0 emphasizes that there is no rank-restriction on the matrices in Sym_{0,d_X} unlike, for example, in [143, 226].

Note that by the convexity of $\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ and the fact that each $w \in \Delta_N$, for some $N \in \mathbb{N}_+$, then η does indeed take values in $\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$. The mixing function η will be used to inscribe "abstract geodesic simplices" in $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathrm{d}_{2,F})$ thereby endowing it with the structure of an approximately simplicial space, in the sense of [1, Definition 3.1].

Lemma C.4.1. *$(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathcal{Q}, \eta)$ is a barycentric QAS space, in the sense of [1, Definition 3.4]. In particular, it satisfies [182, Setting 3.6 (iii)].*

Proof of Lemma C.4.1. We begin by showing that $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathcal{Q}, \eta)$ is a QAS space, as defined in [1, Definition 3.4]. We also show that it is barycentric, meaning that it admits a 1-barycenter map as defined, for example, in [25, Section 3.2].

QAS Space Structure. Since Sym_{0,d_X} is a closed convex subset of space $(\mathbb{R}^{d\times d}, \|\cdot\|_F)$. Since the Cartesian product of closed subsets is a closed subset of $(\mathbb{R}^{d^2+d}, d_{2,F})$ and since the product of convex sets is again convex by [26, Proposition 3.6] then, $\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ is a closed and convex subset of the normed linear space $(\mathbb{R}^{d+d^2}, d_{2,F})$.

For $q \in \mathbb{N}$ define the map $\mathcal{Q}_q : \mathbb{R}^{d+d^2} \to \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ as sending any $(m, A) \in \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ to

$$\mathcal{Q}_q((m, A)) \overset{\text{def.}}{=} (m, A^\top A).$$

The map $\mathcal{Q}_q$ is a surjection since every symmetric matrix is the square of some $d \times d$ matrix. Therefore, $\mathcal{Q}_\cdot = (\mathcal{Q}_q)_{q\in\mathbb{N}_+}$ trivially satisfies [1, Definition 3.2] with modulus of quantizability, defined on [1, page 12], given by $\mathcal{Q}_K(\varepsilon) = D_1 \overset{\text{def.}}{=} d(d+1)$ for every compact subset K of $\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ and for every $\varepsilon > 0$.

Observe that, for any $N \in \mathbb{N}_+$, $w \in \Delta_N$, and $(m^{(1)}, A^{(1)}), \ldots, (m^{(N)}, A^{(N)}) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d\times d}$ we have

$$\begin{aligned}
&d_{2,F}\Big(\eta\Big(w, \{(m^{(n)}, A^{(n)}\}_{n=1}^N\Big), y_i\Big) \\
&\quad= \Big(\Big\|\Big(\sum_{n=1}^N w_n\, m^{(n)}\Big) - m^{(i)}\Big\|^2 + \Big\|\Big(\sum_{n=1}^N w_n\, A^{(n)}\Big) - A^{(i)}\Big\|_F^2\Big)^{1/2} \\
&\quad= \Big(\Big\|\sum_{n=1}^N w_n\, m^{(n)} - \sum_{n=1}^N w_n\, m^{(i)}\Big\|^2 + \Big\|\sum_{n=1}^N w_n\, A^{(n)} - \sum_{n=1}^N w_n\, A^{(i)}\Big\|_F^2\Big)^{1/2} \qquad \text{(C.20)} \\
&\quad\le \Big(\sum_{n=1}^N w_n\Big)^2 \Big\|m^{(n)} - m^{(i)}\Big\|^2 + \Big(\sum_{n=1}^N w_n\Big)^2 \Big\|A^{(n)}) - A^{(i)})\Big\|_F^2\Big)^{1/2}
\end{aligned}$$

$$\leq \sum_{n=1}^{N} w_n \left(\left\| m^{(n)} - m^{(i)} \right\|^2 + \left\| A^{(n)}) - A^{(i)}) \right\|_F^2 \right)^{1/2}$$
$$= 1 \cdot \sum_{n=1}^{N} w_n \, d_{2,F}\Big((m^{(n)}, A^{(n)}), (m^{(i)}, A^{(i)}) \Big)^1,$$

for $i = 1, \ldots, N$, where (C.20) holds since $\sum_{n=1}^{N} w_n = 1$ since $w \in \Delta_N$. Thus, η is a mixing function and therefore $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F})$ is approximately simplicial, as defined in [1, Definition 3.1]. Consequentially, $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F}, \mathcal{Q}, \eta)$ is a QAS space; as defined in [1, Definition 3.4].

Barycentricity. Since $(\mathbb{R}^{d_X} \times \mathrm{Sym}_d, d_{2,F})$ is a normed linear space then [49] shows that the only contracting barycenter map is given by $\mathcal{P}_1(\mathbb{R}^{d_X} \times \mathrm{Sym}_d, d_{2,F}) \ni \mathbb{P} \to \mathbb{E}_{X\sim\mathbb{P}}[X] \in \mathbb{R}^{d_X} \times \mathrm{Sym}_d$. Since Sym_{0,d_X} is a closed convex subset of the normed linear space Sym_d then Jensen's inequality, as formulated in [3, Theorem 10.2.6], implies that for each $\mathbb{P} \in \mathcal{P}_1(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F})$ we have $\mathbb{E}_{X\sim\mathbb{P}}[X] \in \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$; where $\mathbb{E}_{X\sim\mathbb{P}}[X]$ denote the Bochner integral of a random variable with law $\mathbb{P}$. Consequentially, $\mathcal{P}_1(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F}) \ni \mathbb{P} \to \mathbb{E}_{X\sim\mathbb{P}}[X] \in \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ is a contracting barycenter map. Thus, $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F})$ is barycentric metric space.

Since is a barycentric QAS space then, [182, Setting 3.6 (iii)] is satisfied. □

We are now ready to prove the following result, as we have established the barycentricity and the QAS space structure of $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F}, \eta, \mathcal{Q}_\cdot)$,

Lemma C.4.2. *Fix an activation function σ satisfying Assumption 4.3.4. For every $K \subset C([0:T], \mathbb{R}^{d_X})$ satisfying Assumption 4.2.2, each $0 < \delta, \alpha \leq 1$, $0 \leq L$, and every (L,α)-Hölder[5] function $f : [0:T] \times K \to \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ there is a map $\hat{g} : [0:T] \times C([0:T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ satisfying the uniform estimate*

$$\max_{(t,y_\cdot)\in[0:T]\times K} d_{2,F}(\hat{g}(t,y_\cdot), f(t,y_\cdot)) < \delta \tag{C.21}$$

with representation

$$\hat{g}(t,y_\cdot) = \sum_{i=1}^{N'} P_{\Delta_{N'}}(\hat{f} \circ \mathrm{attn}^\theta(t,y_\cdot))_i \cdot (m^{(i)}, (A^{(i)})^\top A^{(i)}), \tag{C.22}$$

[5] That is, f is α Hölder with optimal Hölder coefficient L.

where $N' \in \mathbb{N}_+$, $m^{(1)}, \dots, m^{(N')} \in \mathbb{R}^{d_X}$, $A^{(1)}, \dots, A^{(N')} \in \mathbb{R}^{d \times d}$, $P_{\Delta_{N'}} : \mathbb{R}^{N'} \to \Delta_{N'}$ *is the Euclidean (orthogonal) projection onto the* N'*-simplex, and an MLP* $\hat{f} : \mathbb{R}^N \to \mathbb{R}^{N'}$ *with activation function* σ*. The depth, width, encoding dimension* (N)*, and decoding dimension* (N') *are recorded in Table C.1.*

Table C.1: Complexity Estimates for transformer-type model $\hat{g}$ in Lemma C.4.2.

σ Regularity	Depth	Width	Encode (N)	Decode (N')
ReLU	$\mathcal{O}\big((LV(L))^{-N-1}\,\varepsilon^{-N-1}\big)$	$\mathcal{O}\big((LV(L))^{-N-1}\,\varepsilon^{-N-1}\big)$	$\mathcal{O}(1)$	$\mathcal{O}\big(L\varepsilon^{-1}\big)$
Smooth & Non-poly.	$\mathcal{O}\left(L^{4N+5}\,\varepsilon^{-4N-5}\right)$	$\mathcal{O}(L\epsilon^{-1}+N+3)$	$\mathcal{O}(1)$	$\mathcal{O}\big(L\varepsilon^{-1}\big)$
Poly. & Non-affine	$\mathcal{O}\big(L^{8N+14}\,\varepsilon^{-8m-14}\big)$	$\mathcal{O}(L\epsilon^{-1}+N+4)$	$\mathcal{O}(1)$	$\mathcal{O}\big(L\varepsilon^{-1}\big)$
$C(\mathbb{R})$ & Non-poly.	Finite	$\mathcal{O}(L\epsilon^{-1}+N+3)$	$\mathcal{O}(1)$	$\mathcal{O}\big(L\varepsilon^{-1}\big)$

Where $V(t)$ is the inverse of $s \mapsto s^4 \log_3(t+2)$ on $[0,\infty)$ evaluated at $131t$.

Proof of Lemma C.4.2. We work in the notation of [182, Theorem 3.7], or rather, its quantitative version [182, Lemma 5.10]. Our objective is to apply [182, Theorem 3.7] by verifying each of the conditions of [182, Setting 3.6].

Step 1: Implementing a Bi-Lipschitz Feature Map with the Attention Layer.

Remark C.4.1. We first show that the conditions of [182, Setting 3.6 (i)] are met, by verifying that the parameters of the attention layer (4.7) can be chosen such that attn^θ is a suitable feature map. ◁

When convenient, let attn^θ be as in either of Lemmas C.3.1 to C.3.3 depending on which assumption of Assumptions 4.2.2 (i), (ii), or (iii) holds. For convenience, we denote the map attn by φ. These lemmata show that the map φ is a bi-Lipschitz embedding of $([0:T]\times K, \|\cdot\|\times\|\cdot\|_T)$ into a Euclidean space $(\mathbb{R}^N, \|\cdot\|_2)$.

We also observe that every bi-Lipschitz map is a quasi-symmetric map[6], as defined on [140, page 78]. Thus, [140, Theorem 12.1] implies that $(\mathcal{M}, d_g)$ is a doubling metric space, as defined on [140, page 81]. Since $(\mathcal{M}, d_g)$ and $(K, \|\cdot\|_T)$ since both are isometric then $(K, \|\cdot\|_T)$ is also a doubling metric space (see [250, Lemma 9.6 (v)]). Thus, $([0:T]\times K, \|\cdot\|\times\|\cdot\|_T)$ is a doubling metric space. We have thus verified [182, Setting 3.6 (i)].

[6] See [140, page 78].

Step 2: Feature Space Geometry. Since the codomain of φ is simply a Euclidean space, then, the constant sequence of identity maps $\{T^i \overset{\text{def.}}{=} 1_{\mathbb{R}^{N+1}}\}_{i=1}^{\infty}$ are trivially finite-rank linear operators realizing the bounded approximation property on any compact subset of $\mathbb{R}^{N+1}$. That is, for each non-empty compact $A \subseteq \mathbb{R}^{N+1}$,

$$\lim_{i\mapsto\infty} \max_{u\in A} \|u - T^i(u)\|_2 = \lim_{i\mapsto\infty} \max_{u\in A} \|u - 1_{\mathbb{R}^N}(u)\|_2 = 0$$

and the operator norm $\|1_{\mathbb{R}^{N+1}}\|_{\text{op}} = 1$. Thus, $(T^i)_{i\in\mathbb{N}}$ implements the 1-BAP (1-bounded approximation property) on $\mathbb{R}^{N+1}$ for every i. Therefore, [182, Setting 3.6 (ii)] holds.

Step 3: Geometry of $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathcal{Q}, \eta)$. Lemma C.4.1 shows that $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathcal{Q}, \eta)$ is barycentric and it is a QAS space with quantized mixing function, see [182, page 7], given for any $N' \in \mathbb{N}_+$, $u \in \mathbb{R}^{N'}$, and $(m^{(1)}, A^{(1)}), \ldots, (m^{(N)}, A^{(N')}) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_X^2}$ by

$$\hat{\eta}\Big(w, \{(m^{(n)}, A^{(n)}\}_{n=1}^{N'}\Big) \overset{\text{def.}}{=} \sum_{n=1}^{N'} P_{\Delta_{N'}}(w_n) \cdot (m^{(n)}, (A^{(n)})^\top A^{(n)}). \tag{C.23}$$

This verifies [182, Setting 3.6 (iii)].

Step 4: Determining The Euclidean Universal Approximator.

Remark C.4.2. We now verify that the class of all MLPs with activation function σ satisfying Assumption 4.3.4, thus assumption [182, Setting 3.6 (iv)] holds. The case where $\sigma = \mathrm{ReLU}$ and $\sigma \neq \mathrm{ReLU}$ are treated separately. ◁

First, consider the case where $\sigma \neq \mathrm{ReLU}$. Then, for each $N, N', c \in \mathbb{N}_+$ let $\mathcal{F}_{N,N',c}$ denote the family of maps $f : \mathbb{R}^N \to \mathbb{R}^{N'}$ with representation (4.8) and satisfying

$$J \leq c \text{ and } \max_{j=0,\ldots,J} d_j \leq N + N' + 3. \tag{C.24}$$

Since σ was assumed to satisfy Assumption 4.3.4 then by [183, Theorem 9], as formulated in [183, Proposition 53], implies that $\mathcal{F}_\cdot$ is a universal approximator in the sense of [182, Definition 2.11]. Moreover, its rate function is recorded in [183, Proposition 53]. Therefore, $\mathcal{F}_\cdot$, as defined in (C.24), verifies [182, Setting 3.6 (iv)].

Next, suppose that $\sigma = \mathrm{ReLU}$. Then, for each $N, N', c \in \mathbb{N}_+$ let $\mathcal{F}_{N,N',c}$ denote the family of maps $f : \mathbb{R}^N \to \mathbb{R}^{N'}$ with representation (4.8) and satisfying

$$J \leq c \text{ and } \max_{j=0,\dots,J} d_j \leq c. \tag{C.25}$$

By [170, Theorem 1.1], $\mathcal{F}_\cdot$ is a universal approximator in the sense of [182, Definition 2.11]. Moreover, its rate function is given in [117, Theorem 1], as recorded in [117, Table 1]. In either case, $\mathcal{F}_\cdot$, as defined in (C.25), verifies [182, Setting 3.6 (iv)].

Step 5: Applying [182, Theorem 3.7]. Steps 1 though 5 verify that the conditions of [182, Theorem 3.7] are indeed met. Furthermore, we have just shown that we are in the special case where the feature decomposition $\{([0,T] \times K, \varphi)\}$ of $([0,T] \times K, \|\cdot\|_T)$, as defined in [182, Definition 3.4], is a singleton. Therefore, by [182, Lemma 5.10] for every $\varepsilon > 0$ there is a map $\hat{F} : [0,T] \times K \to \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ with representation[7]

$$\begin{aligned}
\hat{F}(y_\cdot) =& \beta_{\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}} \left(\delta_{\hat{\eta}(P_{\Delta_{N'}} \circ \hat{f}_n \circ \varphi(y_\cdot), (Z_n)_{n=1}^N)} \right) && \text{(C.26)} \\
=& \hat{\eta}(P_{\Delta_{N'}} \circ \hat{f}_n \circ \varphi(y_\cdot), (Z_n)_{n=1}^N) && \text{(C.27)} \\
=& \sum_{n=1}^{N'} \left(P_{\Delta_{N'}} \circ \hat{f}_n \circ \varphi(y_\cdot) \right) \cdot (m^{(n)}, (A^{(n)})^\top A^{(n)}) && \text{(C.28)}
\end{aligned}$$

for each $y_\cdot \in K$ satisfying

$$\sup_{y_\cdot \in K} d_{2,F}\left(\hat{F}(y_\cdot), f(y_\cdot) \right) < \varepsilon_A + \varepsilon_Q + \varepsilon_E = \varepsilon,$$

where $\varepsilon_E = 0$ and $\varepsilon_A \overset{\text{def.}}{=} \varepsilon_Q \overset{\text{def.}}{=} \varepsilon/2$ and where $\hat{f} \in \mathcal{F}_{d_n,N_n,c_n}$, $Z_n \in \mathbb{R}^{N_n \times D_n}$, for some positive integers d_n, c_n, D_n, and $N_1, \dots, N_N \in \mathbb{N}_+$ recorded in [182, Table 3], and $\beta_{\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}}$ is a 1-Lipschitz barycenter map on $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F})$ (which exists by Lemma C.4.1). We observe that that (C.27) follows from the fact that the barycenter map $\beta_{\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}}$ is a right-inverse of the map $\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X} \ni (m, B) \mapsto \delta_{(m,B)} \in \mathcal{P}_1(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X})$, where $\mathcal{P}_1(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X})$ denotes the 1-Wasserstein space on $(\mathcal{P}_1(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}), d_{2,F})$ and (C.28) follows from the expression for $\hat{\eta}$ given in (C.23). Consequentially, (C.26) to (C.28) reduces to (C.22).

[7] Since the barycentric decomposition $\{(K, \varphi)\}$ of $(K, \|\cdot\|_T, \mu)$ has exactly one part then the partition of unity [182, Setting 3.7] is trivial and $\psi_1(x_\cdot) = 1$ for each $x_\cdot \in K$.

Step 6: Tallying Parameters. Since the quantitative version of [182, Theorem 3.7] held, namely [182, Lemma 5.10], then we obtain the following parameter estimates

(a) $N = \mathcal{N}_{\text{pack}}\Big(K, C_{d_K}\ \text{Vol}(\mathcal{M}, g)^{1/d_K}\Big) \in \mathcal{O}(1)$,

(b) c is recorded in Table C.1 as the depth of the network $\hat{f}$,

(c) The expression of N' is recorded, in detail, atop [182, page 46] and is

$$N' \leq \left(C_{K,1}^{\lceil \frac{1}{4\alpha} \rceil}\right)^{\log_2(\text{diam}(K)) - \frac{1}{\alpha}\log_2(\epsilon_A/(2LC_{K:2}))} \in \mathcal{O}_K\Big(L/\varepsilon\Big),$$

for constants $C_{K,1}, C_{K,2}, C_K > 0$ depending only on the compact set K and on the mixing function η, $\alpha = 1$ as f is 1-Hölder, and where $\mathcal{O}$ suppresses a constant depending only on K and on the mixing function η in (C.19).

This concludes the proof. □

C.4.2 Proof of the Main Approximation Lemma

In what follows, for any $d \in \mathbb{N}_+$, let $\overline{\mathcal{N}_d}$ denote the set of Gaussian measure on $\mathbb{R}^d$ equipped with the 2-Wasserstein metric $\mathcal{W}_2$.

The Lemmata in the previous sections together with the main results of [180] and [182] to deduce our main approximation theoretic tool, namely, Proposition 4.4.3.

Proof of Proposition 4.4.3. *Step 1: Bounded The Local Lipschitz Stability of ρ on $f(K)$.* By Lemma C.5.1, the map $\varrho : (\mathcal{N}_{d_X}, \mathcal{W}_2) \to (\mathbb{R}^{d_X} \times \text{Sym}_{0,d_X}, d_{2,F})$ is continuous. By definition of the product topology, see [221, page 114], the projection map $\pi : (\mathbb{R}^{d_X}, \text{Sym}_{0,d_X}, d_{2,F}) \ni (\mu, \Sigma) \to \Sigma \in (\text{Sym}_{0,d_X}, \|\cdot\|_F)$ is continuous. Since the composition of continuous functions is again continuous, then the map

$$g : ([0,T] \times K, |\cdot| \times \|\cdot\|_T) \ni (t, y.) \mapsto \pi \circ \varrho \circ f(t, y.) \in (\mathbb{R}, |\cdot|)$$

is continuous. By [221, Theorem 26.5], $\tilde{K} \overset{\text{def.}}{=} g(K)$ is a compact subset of $(\mathbb{R}, |\cdot|)$.

By [190, Theorem 9.2.6 - page 130] the map $\lambda_{\text{min}} : (\text{Sym}_{0,d_X}, \|\cdot\|_F) \to (\mathbb{R}, |\cdot|)$ which sends any $d_X \times d_X$ square matrix Σ to its minimal eigenvalue $\lambda_{\text{min}}(\Sigma)$ is continuous. Since

every continuous function with compact domain achieves its minimum on its domain then, there exists some $\Sigma_0 \in \tilde{K}$ minimizing $\lambda_{\min}$; by which we mean that

$$\lambda_{\min}(\Sigma_0) = \min_{\Sigma \in g(K)} \lambda_{\min}(\Sigma) < \infty. \tag{C.29}$$

Since f takes values in $\mathcal{N}_{d_X}$ then $\pi \circ \varrho(f(x))$ is positive definite, for each $x \in K$. In particular, $\lambda_{\min}(\Sigma_0) > 0$. Consequentially, (C.29) implies that

$$0 < r \overset{\text{def.}}{=} \lambda_{\min}(\Sigma_0) = \min_{\Sigma \in g(K)} \lambda_{\min}(\Sigma) < \infty. \tag{C.30}$$

By Lemma C.2.10, we have that there is an $r > 0$ such that $r = \lambda_{\min}(\Sigma_t)$ for all $0 \leq t \leq T$. Consequentially, Lemma C.5.1 implies that the map $\varrho : (\mathcal{N}_{d_X}, \mathcal{W}_2) \to (\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F})$ is Lipschitz on $\tilde{K}$ with Lipschitz constant bounded-above by $\mathrm{Lip}\left(\varrho | f(K)\right)$. Consequentially, we have that

$$\mathrm{Lip}\left(\varrho \circ f | K\right) \leq \mathrm{Lip}\left(\varrho | f(K)\right) \mathrm{Lip}\left(f | K\right) \leq \max\left\{1, \frac{\sqrt{d}}{2\sqrt{r}}\right\} L. \tag{C.31}$$

Step 2: Approximating $\varrho \circ f$ on K. Fix $\varepsilon > 0$ and fix the "perturbed approximation error"

$$\delta \overset{\text{def.}}{=} \min\left\{1, \frac{\varepsilon}{\sqrt{d} 2\sqrt{L\sqrt{T^2 + \mathrm{diam}(K)^2} + 1}}\right\}. \tag{C.32}$$

We apply Lemma C.4.2 to deduce that there exists a map $\hat{F} : [0:T] \times K \to \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ with representation (C.22) satisfying the uniform estimate

$$\max_{(t,y.) \in [0:T] \times K} d_{2,F}(\hat{F}(t,y.), \varrho \circ f(t,y.)) < \delta. \tag{C.33}$$

Since Lemma C.5.1 showed that ϱ has a locally bi-Lipschitz homeomorphism then, in particular, ϱ^{-1} exists and it is Lipschitz continuous. Consider the 1-thickening of $\varrho \circ f(K)$ defined by

$$\bar{B}_1 \overset{\text{def.}}{=} \{(m,\Sigma) \in \mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X} : (\exists (t,y.) \in [0:T] \times K)\, d_{2,F}(f(t,y.),(m,\Sigma)) \leq 1\}.$$

Since δ was defined, in (C.32), to be at-most 1, then (C.32) implies that

$$\hat{F}(K) \subseteq \bar{B}_1. \tag{C.34}$$

From (C.33) and Lemma C.5.1 we deduce the uniform estimate

$$\begin{aligned}
&\max_{(t,y.)\in[0:T]\times K} \mathcal{W}_2\Big(\varrho^{-1}\circ F(t,y.), f(t,y.)\Big) && \text{(C.35)}\\
&= \max_{(t,y.)\in[0:T]\times K} \mathcal{W}_2\Big(\varrho^{-1}\circ \hat{F}(t,y.), \varrho^{-1}\circ\varrho\circ f(t,y.)\Big)\\
&\leq \mathrm{Lip}\left(\varrho^{-1}|\bar{B}_1\right) \max_{(t,y.)\in[0:T]\times K} d_{2,F}\Big(\hat{F}(t,y.), \varrho\circ f(t,y.)\Big)\\
&\leq \mathrm{Lip}\left(\varrho^{-1}|\bar{B}_1\right)\delta\\
&\leq \mathrm{Lip}\left(\varrho^{-1}|\bar{B}_1\right)\frac{\varepsilon}{\sqrt{d}2\sqrt{L\sqrt{T^2+\mathrm{diam}(K)^2}+1}} && \text{(C.36)}\\
&\leq \varepsilon, && \text{(C.37)}
\end{aligned}$$

where (C.36) followed from Lemma C.5.1 and the definition of $\bar{B}_1$ and bound (C.37) followed from the estimate (C.31) for $\mathrm{Lip}(\varrho^{-1}|\bar{B}_1)$. Since $\mathcal{W}_p \leq \mathcal{W}_2$ for all $1\leq p\leq 2$, see [280, Remark 6.6] then, the estimates in (C.35)-(C.37) imply that

$$\max_{(t,y.)\in[0:T]\times K} \mathcal{W}_p\Big(\hat{F}(t,y), f(y)\Big) \leq \varepsilon, \qquad \text{(C.38)}$$

for all $1\leq p\leq 2$; as claimed.

Step 3: Counting Parameters. Using δ, as defined in (C.33), in the place of ε in Table C.1 and noting that $\delta^{-1}\in\mathcal{O}(\varepsilon^{-1})$ yields the conclusion.

□

C.5 An Auxiliary Lemma — I. Pinelis

This section records an auxiliary lemma due to Iosif Pinelis, [236]. We include the result and its proof here to keep our manuscript self-contained.

Lemma C.5.1 ([236]). *Fix $d\in\mathbb{N}_+$ and $R,r>0$. For every $m^{(1)},m^{(2)}\in\mathbb{R}^d$ and each $d\times d$ symmetric positive semi-definite matrix A,B satisfying: $\|A\|_F,\|B\|_F\leq R$ and $A-r\cdot I_d$ and $B-r\cdot I_d$ are positive semi-definite*[8] *then the following lower-bound holds*

$$\frac{1}{\min\{1,\sqrt{d}(2\sqrt{R})\}}\sqrt{\|m^{(1)}-m^{(2)}\|^2+\|A-B\|_F^2}\leq \mathcal{W}_2\Big(\mathcal{N}(m^{(1)},A),\mathcal{N}(m^{(2)},B)\Big).$$

[8] I.e. $A,B\geq r\cdot I_d$ where $\geq$ is the partial ordering on the set of $d\times d$-dimensional symmetric positive-definite matrices given by $A\geq B$ if and only if $A-B$ is positive semi-definite.

Moreover, the following upper-bound also holds

$$\mathcal{W}_2\Big(\mathcal{N}(m^{(1)},A),\mathcal{N}(m^{(2)},B)\Big)\leq \max\left\{1,\frac{\sqrt{d}}{2\sqrt{r}}\right\}\sqrt{\|m^{(1)}-m^{(2)}\|^2+\|A-B\|_F^2}.$$

In particular, the map $\varrho:(\overline{\mathcal{N}}_d,\mathcal{W}_2)\to(\mathbb{R}^d\times \mathrm{Sym}_{0,d},\|\cdot\|\times\|\cdot\|_F)$ *is locally-Lipschitz; where* $\|\cdot\|\times\|\cdot\|_F$ *denotes the product of the Euclidean norm on* $\mathbb{R}^d$ *and the Fröbenius norm on the space of* $d\times d$*-dimensional symmetric positive semi-definite matrices* $\mathrm{Sym}_{0,d}$.

Proof of Lemma C.5.1. By [121, Proposition 7] the 2-Wasserstein distance between $\mathcal{N}(m^{(1)},A)$ and $\mathcal{N}(m^{(2)},B)$ satisfies

$$\mathcal{W}_2(\mathcal{N}(m^{(1)},A),\mathcal{N}(m^{(2)},B))=\sqrt{\|m^{(1)}-m^{(2)}\|^2+W_2(\mathcal{N}(0,A),\mathcal{N}(0,B))^2}.$$

Thus, $\|m^{(1)}-m^{(2)}\|\leq\mathcal{W}_2(\mathcal{N}(m^{(1)},A),\mathcal{N}(m^{(2)},B))$. Moreover, for any unit vector $u\in\mathbb{R}^n=\mathbb{R}^{n\times 1}$ we have

$$\begin{aligned}\mathbb{E}_{X\sim\mathcal{N}(0,A),Y\sim\mathcal{N}(0,B)}\|X-Y\|^2\geq&\mathbb{E}_{X\sim\mathcal{N}(0,A),Y\sim\mathcal{N}(0,B)}(u^\top X-u^\top Y)^2\\ \geq&(\sqrt{u^\top Au}-\sqrt{u^\top Bu})^2,\end{aligned}$$

where the last inequality holds since $u^\top X$ and $u^\top Y$ have Gaussian law with zero-mean random with respective variances $u^\top Au$ and $u^\top Bu$. Again using the inequality $\|m^{(1)}-m^{(2)}\|\leq\mathcal{W}_2(\mathcal{N}(m^{(1)},A),\mathcal{N}(m^{(2)},B))$, for any unit vector $u\in\mathbb{R}^n=\mathbb{R}^{n\times 1}$, we have

$$\begin{aligned}\mathcal{W}_2(\mathcal{N}(a,A),\mathcal{N}(b,B))&\geq\mathcal{W}_2(\mathcal{N}(0,A),\mathcal{N}(0,B))\\ &\geq|\sqrt{u^\top Au}-\sqrt{u^\top Bu}|\\ &=\frac{|u^\top Au-u^\top Bu|}{\sqrt{u^\top Au}+\sqrt{u^\top Bu}}\\ &\geq\frac{|u^\top(A-B)u|}{\sqrt{\|A\|}+\sqrt{\|B\|}}\\ &=\frac{\|A-B\|}{\sqrt{\|A\|}+\sqrt{\|B\|}}\end{aligned}$$

for some unit vector $u\in\mathbb{R}^n=\mathbb{R}^{n\times 1}$, where $\|M\|$ is the spectral norm of a matrix M. So,

$$\|A-B\|\leq(\sqrt{\|A\|}+\sqrt{\|B\|})\,\mathcal{W}_2(\mathcal{N}(m^{(1)},A),\mathcal{N}(m^{(2)},B)).$$

The conclusion follows upon combining $\|m^{(1)} - m^{(2)}\| \leq \mathcal{W}_2(N(m^{(1)}, A), N(m^{(2)}, B))$ and $\|A - B\| \leq (\sqrt{\|A\|} + \sqrt{\|B\|})\mathcal{W}_2(\mathcal{N}(m^{(1)}, A), \mathcal{N}(m^{(2)}, B))$ together with the observation that $\|\cdot\|_{\lambda:2} \leq \|\cdot\|_F \leq \sqrt{d}\|\cdot\|_{\lambda:2}$; where

$$\|A\|_{\lambda:2} \stackrel{\text{def.}}{=} \sqrt{\max_{i=1,\ldots,d} \lambda_{\max}(C^\top C)}$$

denotes the spectral norm on the set of $d \times d$ matrices and where $\lambda_{\max}(C^\top C)$ denotes the largest eigenvalue $C^\top C$ for a given $d \times d$ matrix C. □