Transformers Can Solve Non-Linear and Non-Markovian Filtering Problems in Continuous Time For Conditionally Gaussian Signals

Blanka Horvath[*] [†] Anastasis Kratsios[‡] [§] Yannick Limmer[¶] [‖] Xuwei Yang[‡] [**]

Abstract

The use of attention-based deep learning models in stochastic filtering, e.g. transformers and deep Kalman filters, has recently come into focus; however, the potential for these models to solve stochastic filtering problems remains largely unknown. The paper provides an affirmative answer to this open problem in the theoretical foundations of machine learning by showing that a class of continuous-time transformer models, called *filterformers*, can approximately implement the conditional law of a broad class of non-Markovian and conditionally Gaussian signal processes given noisy continuous-time (possibly non-Gaussian) measurements. Our approximation guarantees hold uniformly over sufficiently regular compact subsets of continuous-time paths, where the worst-case 2-Wasserstein distance between the true optimal filter and our deep learning model quantifies the approximation error. Our construction relies on two new customizations of the standard attention mechanism: The first can losslessly adapt to the characteristics of a broad range of paths since we show that the attention mechanism implements bi-Lipschitz embeddings of sufficiently regular sets of paths into low-dimensional Euclidean spaces; thus, it incurs no "dimension reduction error". The latter attention mechanism is tailored to the geometry of Gaussian measures in the 2-Wasserstein space. Our analysis relies on new stability estimates of robust optimal filters in the conditionally Gaussian setting.

Keywords: Bayesian Filtering, Approximation, Neural Networks, Non-Markovian.

1 Introduction

In a wide variety of scientific domains, from mathematical finance [67, 119] to medicine [104] and evolutionary biology [87], one is often interested in estimating an unobservable *signal* process $X_\cdot$ based on noisy *observations* $Y_\cdot$. This leads to the classical problem of *optimal (stochastic) filtering*, which seeks the best reconstruction of the signal process given the observed data. For instance, in mathematical finance, market sentiment can act as a signal, and the impact of the sentiment on market prices can act as observations. Although this problem is well studied and admits a unique solution in the form of an infinite-dimensional recursion, e.g. [110, 111, 116, 124, 7], the resulting non-linear filters are nearly always computationally intractable and so are traditional approximation schemes; e.g. particle fitlers [39, 41, 43] or linearized surrogates [57]. This challenge has sparked the exploration of deep learning approaches to stochastic filtering, motivated by the success of neural networks in most challenging computational problems, which have shown promising empirical performance [81, 24, 5]. Nevertheless, the fundamental question: *"Can neural networks solve the stochastic filtering problem?"* remains open.

This paper provides an *affirmative* answer to this question for a broad range of non-Markovian signal processes evolving according to non-linear dynamics, given general (possibly non-Gaussian) observations, in

[*]University of Oxford, UK
[†]horvath@maths.ox.ac.uk
[‡]1) McMaster University, Canada and 2) Vector Institute, Canada
[§]kratsioa@mcmaster.ca
[¶]University of Oxford, UK
[‖]yannick.limmer@maths.ox.ac.uk
[**]yangx212@mcmaster.ca

continuous time up to an arbitrarily small approximation error. However, only the latter of these two can be directly measured. More precisely, here both of the processes are often assumed to have continuous paths and are defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[0:T]}, \mathbb{P})$. The objective of the *robust stochastic filtering problem*, studied by [27, 82, 35, 36, 26, 30], is to identify a continuous function $f_t : C([0:t], \mathbb{R}^{d_Y}) \to \mathcal{P}(\mathbb{R}^{d_X})$, with $d_X, d_Y \in \mathbb{N}$ for $t > 0$, satisfying

$$f_t(Y_{[0:t]}) = \mathbb{P}(X_t \in \cdot | \mathcal{F}_t^Y), \tag{1}$$

where $\mathbb{P}(X_t \in \cdot | \mathcal{F}_t^Y)$ is the conditional law of the d_X-dimensional signal process X_t given the σ-algebra generated by $(Y_s)_{s\in[0:t]}$. The key innovation in (1) is the continuity, and the uniqueness, of f_t. In contrast, a Borel f_t satisfying (1) exists by elementary measure-theoretic [70, Theorem 6.3]. The continuity of f_t is qualified by equipping the set of Borel probability measures on $\mathbb{R}^{d_X}$, denoted by $\mathcal{P}(\mathbb{R}^{d_X})$, with the weak topology and by equipping the set of continuous paths from $[0 : t]$ to $\mathbb{R}^{d_Y}$, denoted by $C([0 : t], \mathbb{R}^{d_Y})$, with the uniform norm.

When one has access to a *robust representation* (1), then they can reliably predict the conditional law of X_t even subject to imperfections on the observed historical data in $y_{[0:t]} \in C([0 : t], \mathbb{R}^{d_Y})$. These robust representations are particularly invaluable in mathematical finance, where continuous streams of data are often noisy. More broadly, stochastic filters are indispensable in situations where latent parameters influence or obscure market factors. Applications include the computation of optimal investment under partial information [83, 18], the estimation of volatility from observed intra-day stock prices [10], estimation of interest rates [23, 46, 4, 68], estimation of spot price estimation for commodity futures [106, 105, 92, 85, 50], hedging of credit derivatives [52, 53], estimation of equilibria under asymmetric information [22], and calibration of option pricing models [88, 120]. Pathwise, or so-called robust, formulations were later derived in [26, 31]. 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 [60, 118, 17, 33, 101].

Although the stochastic filtering problem is a well-understood mathematical problem. Many of the fundamental questions in stochastic filtering, such as existence and dynamics for the evolution of the conditional law of signal process X_t given the realizations of the measurement process $Y_\cdot$, were solved in a series of classical [110, 111, 124, 116, 117] and contemporary [125] papers. Nevertheless, the infinite-dimensionality of general stochastic filters, being measure-valued path-dependent processes, makes the problem computationally intractable.

The exception to this rule is so-called finite-dimensional filters. These encapsulate rare situations in which the dynamics of $X_\cdot$ and $Y_\cdot$ give way to optimal filters that are finitely parameterized. Examples include the *Kalman-Bucy filters* [71, 72] where closed-forms are derived under the assumption of Markovian Gaussian noise and affine OU-type dynamics of $X_\cdot$ and $Y_\cdot$, the [121] filter where all involved quantitative are finite-state Markov processes, and the Beňes filter [16] which relies a particular set of one-dimensional dynamics.

The computational intractability of the general filtering problem leads to the use of approximately optimal filters. These approaches include particle filters which aim at dynamically approximating optimal filters using an evolving interacting particle system [40, 44, 42], linear relaxations of the optimal filtering functional for affine processes which can be numerically computed by solving specific stochastic Riccati equations [58], or by projection of the optimal filter onto finite-dimensional manifolds of exponential families [20, 19, 3] where it can be tracked using finitely parameterized representations. The latter two approaches work well if the coupled system $(X_\cdot, Y_\cdot)$ follows the postulated affine dynamics, and the former particle-filtering approach is well-suited to low-dimensional settings. Nevertheless, one would like to have access to (approximate) optimal finite-dimensional filters for a broader range of situations.

One possibility could be via deep learning, since deep neural networks have demonstrated that they can efficiently solve a variety of previously intractable numerical problems. The expressivity of neural networks on classical learning problems has inspired the use of deep learning approaches to the numerical stochastic filtering problems; e.g. [24, 15, 103, 37, 64, 75, 62], and the development of measure-valued deep learning models; e.g. [1]. At the forefront of these methods, are the class of deep Kalman filters (DKFs) [81], these are the primary object of study in this paper. Broadly speaking, these models, and our FF, are deep neural networks which map historical data to Gaussian measures. Nevertheless, though DKF are inspired by stochastic filters, there is still little connection between them and the stochastic filtering problem (1).

This paper constructs a FF model which can uniformly approximate the map f_t in the robust representation (1), to arbitrary precision, for any pair of non-Markovian $\mathbb{F} \stackrel{\text{def.}}{=} (\mathcal{F}_s)_{s\in[0:T]}$-adapted *conditionally Gaussian* stochastic processes studied by [89].

Informal Theorem 1 (Universal Conditionally-Gaussian Filtering)**.**
Let $(X_\cdot, Y_\cdot)$ be the partially-observed system (2)-(3) below, subject to mild regularity conditions. For any suitable compact subset $K \subseteq C([0:T], \mathbb{R}^{d_Y})$, every $T > 0$, and any approximation error $\varepsilon > 0$ there exists a Filterformer $\hat{F} : [0:T] \times C([0:T] \times \mathbb{R}^{d_Y}) \to \mathcal{N}_{d_X}$ satisfying the uniform estimate

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

for every $1 \le p \le 2$.

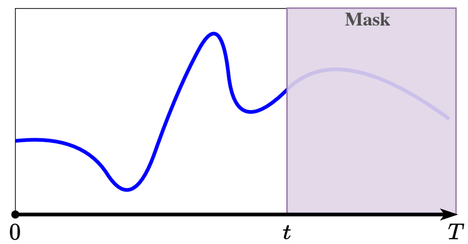


Figure 1: The *Filterformer* model in Informal Theorem 1 has two inputs: a time t and a continuous path $y_\cdot$ defined up to some time $T \ge t$. The first parameter t acts as a mask hiding the future evolution of the observation (input) path $y_\cdot$ beyond the current time t. We note that there is no loss of generality in assuming that the path $y_\cdot$ is defined up to time T, since any path can be trivially extended beyond the current time t by setting $y_s = y_t$ for all $s \in [t, T]$; e.g. as in the Functional Itô Calculus [49, 51, 29].
The mask, i.e. the parameter t, fills the same analogous role to recursions in classical stochastic filters. That is, by varying the time parameter t, the prediction of the FFs evolves into the future without having to retrain the FF model.

Our uniform estimates hold for coupled stochastic differential equations governed by

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

$$dY_t = [A_0(t, Y_{[0:t]}) + A_1(t, Y_{[0:t]})X_t]dt + \sum_{i=1}^{2} B_i(t, Y_{[0:t]})dW_t^{(i)} \tag{3}$$

where a_0 and A_0 respectively take values in $\mathbb{R}^{d_X}$ 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 $d_X \times d_X$, $d_Y \times d_X$, $d_X \times d_X$, $d_X \times d_Y$, $d_Y \times d_X$, 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 $\mathcal{B}_t^d \stackrel{\text{def.}}{=} \sigma(C([0:t], \mathbb{R}^d))$ denotes the σ-algebra generated by continuous paths on $[0:t]$ to $\mathbb{R}^d$. We will always assume that the filtration $\mathbb{F}$ is right-continuous and that the probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ supports independent $\mathbb{F}$-adapted Brownian motions, $W_\cdot^{(1)} \stackrel{\text{def.}}{=} (W_t^{(1)})_{0\le t\le T}$ and $W_\cdot^{(2)} \stackrel{\text{def.}}{=} (W_t^{(2)})_{0\le t\le T}$, of respective dimensions d_X and d_Y, for positive integers d_X and d_Y. The requirements for the dynamics (2) and (3) are detailed in Assumption 1 below.

1.1 The Filterformer Model

The proposed Filterformer (FF), illustrated in Figure 2, generates predictions on the finite-dimensional metric space $(\mathcal{N}_{d_X}, \mathcal{W}_2)$ whose points are non-singular Gaussian measures, and whose distance between points is quantified by the 2-Wasserstein distance. These predictions are generated from the continuous input paths via the following three-phase process.

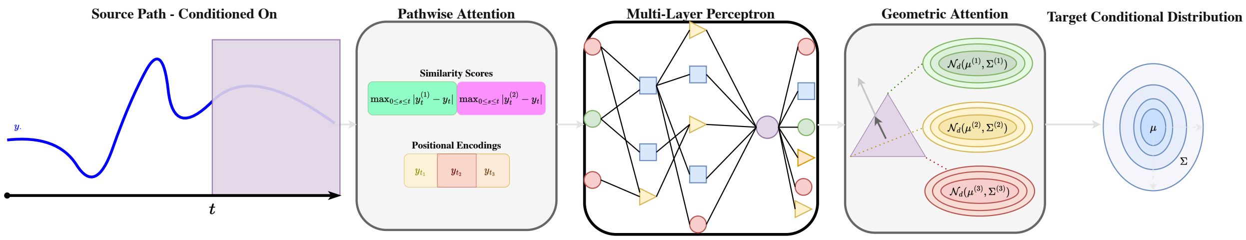


Figure 2: Architecture of a Deep Filter which can approximate the conditional law of the signal process $X.$ given source paths, possibly (but not necessarily) taken by the observation process.

Phase 1. The observed infinite-dimensional continuous path is encoded into a finite-length real vector. Surprisingly, our proposed encoding layer, coined the *pathwise attention mechanism*, adaptively implements a stable and lossless encoding for a broad class of compact subsets of K paths in $C([0:T], \mathbb{R}^{d_Y})$. By *losslessness*, we mean that it is injective for a broad class of compact subsets of K include classes of piecewise linear paths, any finite (training) set, any K which is isometric to a closed Riemannian manifold. By stability, we mean it in the sense of constructive approximation theory, e.g. [28, 98], namely that the map encoding layer is Lipschitz and so is its inverse. *Stability* is desirable since it implies that minor numerical errors, e.g. rounding, do not lead to drastically different downstream predictions. Finally, by *adaptivity* we mean that the encoding layer is is devoid of any projection onto a finite-dimensional (Schauder) basis, and its parameters can be chosen to suit the specific geometry of the compact set of paths K on which the approximation of the optimal filter is to be performed.

Phase 2. In the next phase, the FF processes the vectors encoding the observed paths and generates "deep features" which are then passed along to the output layer for prediction generation. This phase leverages the efficient approximation capacity of MLPs, e.g. [122, 74, 76, 90], to flexibly process the encoded features in a task-specific manner, by adaptive to specific dynamics of the coupled system (2)-(3).

Phase 3. In its final processing phase, the FF decodes the "deep features" generated by the MLP to $\mathcal{N}_{d_X}$-valued predictions, which are then used to approximately implement the predictions of the optimal filter (1). This decoder is a modified instance of the *geometric attention mechanism* of [1]. Intuitively, this layer translates Euclidean data to a point in a generalized geodesic convex hull of a finite number containing the image of K under the optimal filter.

1.2 Organization of Paper

Section 2 formalizes the basic required regularity conditions on the coupled system (2)-(3) and it overviews any background material. Section 3 rigorously introduces the relevant FF model. Section 4 contains the paper'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 FF model and the local Lipschitz regularity of the robust filtering map (1) are discussed. Detailed proofs of our results are relegated to Section 6.

Notation

This section serves as a reference, which records the notation used throughout our manuscript. 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:* $f \bullet x \stackrel{\text{def.}}{=} (f(x_n))_{n=1}^N$ for any $N \in \mathbb{N}_+$.
2. *Rectified Linear Unit (ReLU):* $\operatorname{ReLU} : \mathbb{R} \to \mathbb{R}$ given by $\operatorname{ReLU}(t) \stackrel{\text{def.}}{=} \max\{0, t\}$.
3. *Rowwise Product:* $v \odot X \stackrel{\text{def.}}{=} (v_i\, X_{i,j})_{i=1,\dots,N,\, j=1,\dots,d}$.

4. *Softmax Function:* $\operatorname{Softmax}(x) \stackrel{\text{def.}}{=} (e^{x_n}/\sum_{i=1}^N e^{x_i})_{n=1}^N$.
5. *Sparsity:* $\|(A_{i,j})_{i,j=1}^{I,J}\|_0 \stackrel{\text{def.}}{=} \#\{A_{i,j} \neq 0 : \, i,j = 1,\dots,I\}$.
6. *Vector Concatenation:* $\oplus_{j=1}^N x_j \stackrel{\text{def.}}{=} (x_1,\dots,x_J)^\top$ is the $J \times N$-matrix whose j^{th} row is x_j.
7. *Vectorization:* $\operatorname{vec}((A_{i,j})_{i,j=1}^{I,J}) \stackrel{\text{def.}}{=} (A_{1,1},\dots,A_{I,1},\dots,A_{1,J},\dots,A_{I,J})$.
8. *Euclidean Norm:* $\|A\| = (\sum_{i,j=1}^{I,J} |A_{i,j}|^2)^{1/2}$.
9. *Supremum Norm:* $\|A\|_\infty = \sup\{A_{i,j} : \, i = 1,...,I, j = 1,...,J\}$.

2 The Setting

This section formalizes the setting in which our analysis takes place. We begin by specifying the dynamics of the stochastic processes in the filtering problem. Next, the source space of paths considered is introduced. Subsequently, we formulate the target space of probability measures wherein the optimal filter lies.

2.1 Regularity Conditions on the Partially-Observed System

We maintain the following basic regularity assumptions on the dynamics of the coupled system (2)-(3).

Assumption 1 (Regularity Conditions: Dynamics)**.** *We will assume the following uniform bounds:*

$$|a_1(t,y)_{i,j}| \leq L \text{ and } |A_1(t,y)_{k,j}| \leq L$$

for $y \in C([0:t],\mathbb{R}^{d_Y})$, $t \in [0:T]$, $i,j = 1,\dots,d_X$, and $k = 1,\dots,d_Y$. Further, we require the following integrability conditions:[1]

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

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

hold 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

$$B \circ B \stackrel{\text{def.}}{=} B_1 B_1^\top + B_2 B_2^\top, \quad b \circ B \stackrel{\text{def.}}{=} b_1 B_1^\top + b_2 B_2^\top, \quad b \circ b \stackrel{\text{def.}}{=} b_1 b_1^\top + b_2 b_2^\top.$$

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

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

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

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

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

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

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

We further impose the following assumptions on the dynamics (2), (3):

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

(viii) *Local Lipschitz continuity in the path component uniformly in time*[2]*, 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, $(b \circ B)$, $(B \circ B)^{-1}$ *with respect to the* l^2*-norm, or Frobenius norm if 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 times* $t \in [0:T]$ *and paths* $y_{[0:T]} \in C^1([0:T], \mathbb{R}^{d_X})$.

(x) *Let* $G_t(y_{[0:t]})$ *be a solution of* $\partial_t G_t(y_{[0:t]}) = \tilde{a}_1(t, y_{[0:t]}) G_t(y_{[0:t]})$ *with* $G_0(y_{[0:t]}) = I_{d_X}$ *for any path* $y_{[0:T]} \in C^1([0:T], \mathbb{R}^{d_X})$*, 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 $K_1, K_2 > 0$ *s.t. uniformly* $K_1 \leq \operatorname{tr}(G_t(y_{[0:t]})) \leq K_2$ *as well as* $K_3 > 0$ *s.t.* $\operatorname{tr}(A_1^\top (B \circ B)^{-1} A_1(t, y_{[0:t]})) \leq K_3$.

Eventually we require for the initial conditions that

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

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

(xiii) *Lipschitz continuity of* μ_0, Σ_0 *in* Y_0 *w.r.t. the* $\| \cdot \|_2$*-norm holds.*

2.2 Source Spaces: Regular Compact Subsets of Path Space

Fix a time-horizon $T > 0$. For every $0 < t \leq T$, let $C([0:t], \mathbb{R}^{d_Y})$ denote the set of continuous "paths", i.e. functions, from $[0:t]$ to the d-dimensional Euclidean space $\mathbb{R}^{d_Y}$. We equip this set with the uniform norm, making it a Banach space, where the norm of a path $y_\cdot \in C([0:t], \mathbb{R}^{d_Y})$ is defined by

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

where $\| \cdot \|_2$ is the Euclidean norm on $\mathbb{R}^{d_Y}$.

For every $0 < t \leq T$, similarly to the Horizontal Extension of a path in the Functional Itô calculus [8, Section 5.2.1], we may canonically embed any path $y \in C([0:t], \mathbb{R}^{d_Y})$ into a path $\bar{y} \in C([0:T], \mathbb{R}^{d_Y})$ via by extending the "frozen version after time t" defined by

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

Moreover, note that the map $\bar{\cdot} : C([0:t], \mathbb{R}^{d_Y}) \to C([0:T], \mathbb{R}^{d_Y})$ is a linear isometric embedding of the Banach spaces, since

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

since $y(s) = y(t)$ for all $t < s \leq T$. Conversely, the restriction $\bar{y}_{[0:t]}$ of a path $\bar{y} \in C([0:T], \mathbb{R}^{d_Y})$ to any shorter time-interval $[0:t]$ for $0 \leq t \leq T$ defines a non-expansive linear operator from $C([0:T], \mathbb{R}^{d_Y})$ to $C([0:t], \mathbb{R}^{d_Y})$ since

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

Therefore, in what follows, we will always consider the domain of our path space to be the Banach space $C([0:T], \mathbb{R}^{d_Y})$. As show in [84, Proposition A.3], even in finite dimensions, there are rather regular functions which cannot be approximated by standard deep neural networks parameterized determined by a number of

[2]By *local Lipschitz continuity uniformly in time* we mean that there exists a constant that is a local Lipschitz-constant for the path at all times $t \in [0, T]$.

trainable parameters which is a polynomial in the reciprocal approximation error. In other words, the curse of dimensionality is generally unavoidable when approximating rather regular functions between finite-dimensional Banach spaces, and thus the problem can only be exacerbated in finite-dimensions; see [54].

The curse of dimensionality cannot be broken, but can typically be avoided either by restricting the class of functions being approximated, e.g. in [2, 93, 107, 56], or the regularity of the compact subsets of the input space on which the uniform approximation is to hold; e.g. in [76, 91, 79]. Recent advances in infinite-dimensional approximation of functions not taking values in a linear space – as is the case in our results – show that such restrictions are a sufficient requirement for obtaining universality, see [78]. However, the necessity of such restrictions on compact subsets is still unknown. The explicit effects of restricting the (fractal) dimension and diameter of compact sets of which a deep neural network approximation is to hold is explicitly studied in [1, Proposition 3.10].

We adopt the second approach, since we cannot restrict the function class, which is determined by the stochastic filtering problem. We therefore restrict to compact subsets K of the input space which are regular, in that they are either isometric to some compact Riemannian manifold or they are comprised of piecewise linear paths with finitely many pieces. In particular, we exclude fractal-like subsets of the path space $C([0:T],\mathbb{R}^{d_Y})$ which need not be compressible into finitely many dimensions without "loosing information" (i.e. for which there is no bi-Lipschitz embedding into a finite-dimensional normed space). We denote the Riemannian volume of a connected (C^2) Riemannian manifold $(\mathcal{M},g)$ by $\operatorname{Vol}(\mathcal{M},g)$, its geodesic distance function is d_g.

Assumption 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 either that:*

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

(ii) Piecewise Linear Domains: *There are $N_{\mathrm{p}} \in \mathbb{N}_+$, $0 = t_0 < \cdots < t_{N_{\mathrm{p}}} = T$, and $C_K > 0$ such that $y_\cdot \in K$ if and only if: $y_0 = 0$, $\max_{i=1,\dots,N_{\mathrm{p}}} |y(t_i)| \leq C_K$, and for $i = 0,\dots,N_{\mathrm{p}}-1$ there is a $d \times d$ matrix $A^{(i)}$ and a $b^{(i)} \in \mathbb{R}^{d_Y}$ satisfying $y(t) = A^{(i)}t + b^{(i)}$ forallt $\in (t_i, t_{i+1})$.*

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

Evidently, their domains K satisfying Assumption 2 (i) are easily exhibited, and they correspond to any *training set*. One can interpret results in this case as paralleling interpolation results for deep learning, e.g. [32, 114, 77], but allowing for some arbitrarily small error as in [97]. Compact subsets of the path-space $C([0:T],\mathbb{R}^{d_Y})$ satisfying Assumption 2 (iii) are also plentiful. Indeed, using Banach-Mazur theorem [9] one can show that $C([0:T],\mathbb{R}^{d_Y})$ contains isometric copies of any compact metric space, and in particular, every compact and connected Riemannian manifold (see Proposition 4 in the supplement).

We obtain explicit estimates for the case where $(K,d_\cdot)$ satisfies Assumption 2 (iii), if the manifold $\mathcal{M}$ is topologically regular, in that sense that it is *aspherical*. This means that their homotopy groups $\pi_i(K)$ all vanish for all indices $i \geq 2$, e.g. if $\mathcal{M}$ is a torus. See e.g. [109, Chapter 7] for definitions and details on homotopy groups. Domains K satisfying Assumption 2 (ii) include any piecewise linear interpolation of real-world financial time-series data, or simulated data. This is because only a finite number of inflection points can be observed, and prices are not unbounded. In this way, Assumption 2 summarizes a rich family of compact subsets of the path-space $C([0:T],\mathbb{R}^{d_Y})$ which possess a sufficient degree of structure to be losslessly encoded into finitely many dimensions by our model's encoding layer, called its *pathwise attention mechanism.*

2.3 Target Space - The 2-Wasserstein Space of Probability Measures

Fix $1 \leq p \leq 2$. The p-Wasserstein space $\mathcal{P}_p(\mathbb{R}^{d_X})$ consists of all probability measures on $\mathbb{R}^{d_X}$ with finite second moment; i.e. $\mathbb{P} \in \mathcal{P}_p(\mathbb{R}^{d_X})$ if $\mathbb{E}_{X\sim\mathbb{P}}[\|X\|^p] < \infty$.The 2-Wasserstein metric $\mathcal{W}_p$ on $\mathbb{P}_p(\mathbb{R}^{d_X})$ is given by minimizing the optimal cost of transporting mass between any two measures $\mathbb{P}$ and $\mathbb{Q}$ via randomized transport plans. This can be formalized by the Kanotovich problem

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

Generally, the p-Wasserstein distance between measures can be computationally taxing, requiring a super-quadratic complexity to compute [112]. However, the 2-Wasserstein distance is not always computationally intractable as for instance in the case where $\mathbb{P} = N(\mu^{(1)}, \Sigma^{(1)})$ and $\mathbb{Q} = N(\mu^{(2)}, \Sigma^{(2)})$ are Gaussian measures on $\mathbb{R}^{d_X}$, [45] showed that it admits the following closed-form expression

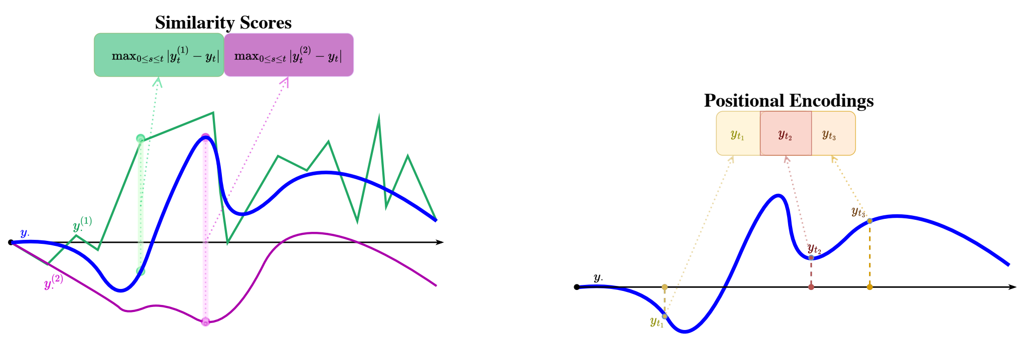
$$\mathcal{W}_2(\mathbb{P}, \mathbb{Q})^2 = \|\mu^{(1)} - \mu^{(2)}\|_2^2 + \mathrm{tr}(\Sigma^{(1)}) + \mathrm{tr}(\Sigma^{(2)}) - 2\ \mathrm{tr}\left({\Sigma^{(2)}}^{1/2} \Sigma^{(1)} {\Sigma^{(2)}}^{1/2}\right)^{1/2}, \tag{4}$$

provided that $\Sigma^{(1)}$ and $\Sigma^{(2)}$ are invertible; in particular, they are positive-definite.

3 The Model: An Attention-Based Deep Filter

3.1 Phase 1 - Encoding via Pathwise Attention

We consider an attention-type feature encoding which modifies the *attention mechanism* of [6] and the graph attention mechanism of [113] to the context of continuous-time path sources/inputs. Our attention mechanism, called *pathwise attention*, encodes key information about a novel path $y_\cdot$ and, like those attention mechanisms, allows us to adaptively encode any newly observed path in terms of its similarity scores to “known/previously generated paths”.



(a) The *similarity score* component $\mathrm{sim}_T^{\theta_0}$ encodes a path into finite-dimensional vectorial data, by relating its similarity to a dictionary of previously observed “contextual paths”.

(b) The *positional encoding* component $\mathrm{post}_T^{\theta_0}$ of the attention mechanism allows the similarity scores to further adapt to any novel path by extracting sample-positions on the novel paths.

Figure 3: Our *pathwise attention mechanism* attn_t^{θ}, in Definition 3, relies on two methods for adaptive feature extraction. The first component of the mechanism, in 3a, in is a similarity score which ranks the similarity of any novel path $x_\cdot$ to a dictionary of “recognized/saved” reference paths. The second component of the mechanism, in 3b, samples points on the novel path $x_\cdot$ and perturbs the similarity scores by them. The adeptness of the pathwise attention mechanism, unlike basis-based methods, is rooted in the fact that nearly all of its parameters are trainable, which allows it to be tasks-specific (similarly to the advantage of deep MLPs have over regression via basis-functions).

Definition 1 (Similarity Score)**.** *Fix positive integers N_{ref}, N_{sim}, d_Y, and a “time horizon” $T > 0$. Given a set of N_{ref} distinct paths $\{y^{(n)}\}_{n=1}^{N_{\mathrm{ref}}}$ in $C([0,1], \mathbb{R}^{d_Y})$, an $N_{\mathrm{sim}} \times N_{\mathrm{ref}}$-matrix B, $b \in \mathbb{R}^{N_{\mathrm{sim}}}$, an $N_{\mathrm{sim}} \times N_{\mathrm{sim}}$-matrix A, and $a \in \mathbb{R}^{N_{\mathrm{sim}}}$. Denote $\theta_0 \stackrel{\text{def.}}{=} (A, B, a, b, \{y^{(n)}\}_{n=1}^{N_{\mathrm{ref}}})$.*

The similarity score, is a map $\mathrm{sim}_T^{\theta_0} : C([0,1], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N_{\mathrm{sim}}}$, mapping any $y_\cdot \in C([0,1], \mathbb{R}^{d_Y})$ to

$$\mathrm{sim}_T^{\theta_0} : y_\cdot \mapsto \mathrm{Softmax}\bigg(A\,\mathrm{ReLU} \bullet \Big(B\big(\|y_\cdot - y_\cdot^{(n)}\|_t\big)_{n=1}^{N_{\mathrm{ref}}} + b\Big) + a\bigg). \tag{5}$$

Illustrated in Figure 3, our pathwise attention mechanism extracts features by reinterpreting two of the key features of attention mechanisms from natural language processing (NLP). First, it utilizes *similarity scores* which rank the likeness of any novel path against a dictionary of references paths. These reference paths can either arise from real/stress historical (market) scenarios, they can be generated synthetically, or a combination of either. These similarity scores are illustrated in Figure 3a, and they serve as pathwise analogues of contextual keys and queries in attention mechanisms in NLP.

Definition 2 (Positional Encoding). *In the notation of Definition 1, fix a positive integer N_{pos}, and set of query times $0 \leq t_1 < \cdots < t_{N_{\mathrm{time}}} \leq T$, an $N_{\mathrm{pos}} \times d_Y$ matrix V, and an $N_{\mathrm{pos}} \times N_{\mathrm{time}}$ matrix U. Set $\theta_1 \stackrel{\text{def.}}{=} (V, U, \{t_n\}_{n=1}^{N_{\mathrm{time}}})$. The positional encoding, is a map $\mathrm{post}_T^{\theta_0} : C([0,1], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N_{\mathrm{pos}} \times d_Y}$, defined by*

$$\mathrm{post}_T^{\theta_1}(y_\cdot) \stackrel{\text{def.}}{=} U\left(\oplus_{j=1}^{N_{\mathrm{time}}} y_{t_j}\right) + V, \tag{6}$$

for any $y_\cdot \in C([0,1], \mathbb{R}^{d_Y})$.

Next, the *positional encoding* of any novel path, illustrated in Figure 3b, encodes snapshots of this path at various times. These indicate changes in any sample path, and they fully determined piecewise linear paths with finitely many pieces (as are generated in simulation studies). This data then combines with the weights of the similarity scores to produce the features encodings of our pathwise attention mechanism.

Definition 3 (Pathwise Attention). *In the notation of Definitions 1 and 2, set $N_{\mathrm{sim}} = N_{\mathrm{pos}}$. A pathwise attention, with parameter $\theta \stackrel{\text{def.}}{=} (\theta_0, \theta_1, C)$, is the map $\mathrm{attn}_T^{\theta} : [0,T] \times C([0,T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N+1}$ with representation*

$$\mathrm{attn}_T^{\theta}(t, y_\cdot) \stackrel{\text{def.}}{=} \Big(t, C\ \mathrm{vec}\left(\mathrm{sim}_T^{\theta_0}(y_\cdot) \odot \mathrm{post}_T^{\theta_1}(y_\cdot)\right)\Big) \tag{7}$$

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

Proposition 1 (Pathwise Attention Losslessly Encodes Regular Domains). *Let $K \subseteq C([0:T], \mathbb{R}^{d_Y})$ satisfy Assumption 2. Then, there exists an $N \in \mathbb{N}_+$ and a parameter θ as in Definition 3 such that, attn_T^{θ} restricts to a bi-Lipschitz embedding of $(K, \|\cdot\|_T)$ into $(\mathbb{R}^{N+1}, \|\cdot\|_2)$.*

Estimates for the encoding dimension N are recorded in Table 1 on a case-by-base basis.

Table 1: Complexity Estimates for Lossless Pathwise Attention Encoding of Regular Compact Domains.

Type of Compact Domain	Encoding Dimension (N)	Ass. 2
Finite	$\mathcal{O}(\log(\#K))$	(i)
P.W. Lin. (N_{p} Pieces)	$\mathcal{O}(N_{\mathrm{p}} d)$	(ii)
Iso. Comp. Riemann.	Finite	(iii)
Iso. Comp. Riemann & Aspherical	$\mathcal{O}\big(\mathfrak{N}_{d_g}(\mathcal{M}, \mathrm{Vol}(\mathcal{M}, g)^{1/d_K})\big)$	(iii)

Fix an $A \subset \mathcal{M}$ and $\delta > 0$. The quantity $\mathfrak{N}_{d_g}(A, \delta)$ denotes the minimum number, I, of points $\{x_i\}_{i=1}^I \subseteq A$ for which every $x \in A$ is contained in a geodesic ball of radius δ about some x_i, for $i \in \{1, \ldots, I\}$.

3.2 Phase 2 - Multi-Layered Perceptrons (MLPs) Transformation

Let us briefly recall the structure of a standard MLP, before formalizing our FF. In what follows, we fix a real-valued *activation function* σ defined on $\mathbb{R}$ satisfying:

Assumption 3 ([74] Condition). *The activation function σ is continuous, non-affine, and there exists an $t \in \mathbb{R}$ such that σ is differentiable at t and $\sigma'(t) \neq 0$.*

Examples of activation functions satisfying this condition are the PReLU function $t \mapsto \max\{0, t\} + a \min\{0, t\}$, for a hyperparameter $a \in \mathbb{R}$, the tanh function used in the numerical PDE literature [38], the sine function used in SIRENs [108], and the Swish map $t \mapsto \frac{t}{1+e^{-\beta t}}$ of [100], where $0 \leq \beta$ is a hyperparameter.

Fix an encoding dimension N and a target dimension N'; both of which are positive integers. A *multi-layer perceptron* (MLP), also called feedforward neural network, from $\mathbb{R}^N$ to $\mathbb{R}^{N'}$ with activation function σ is a map $\hat{f}:\mathbb{R}^N \to \mathbb{R}^{N'}$ with iterative representation: for each $x \in \mathbb{R}^N$

$$\begin{aligned}\hat{f}(x) &\overset{\text{def.}}{=} A^{(J)}\,x^{(J)} + b^{(J)},\\ x^{(j+1)} &\overset{\text{def.}}{=} \sigma \bullet \big(A^{(j)}\,x^{(j)} + b^{(j)}\big) \qquad \text{for } j = 0,\dots,J-1,\\ x^{(0)} &\overset{\text{def.}}{=} x.\end{aligned} \tag{8}$$

where for $j = 0,\dots,J-1$, each $A^{(j)}$ is a $d_{j+1} \times d_j$ matrix, $b^{(j)} \in \mathbb{R}^{d_{j+1}}$, $d_0 = N$ and $d_{J+1} = N'$.

3.3 Phase 3 - Decoding via Geometric Attention Mechanism

Fix $N' \in \mathbb{N}_+$. In what follows, we denote the orthogonal projection of $\mathbb{R}^{N'}$ onto the N'-simplex $\Delta_{N'} \overset{\text{def.}}{=} \{w \in [0,1]^{N'} : \sum_{n=1}^{N'} w_n = 1\}$by $P_{\Delta_{N'}}$. The closedness and convexity of $\Delta_{N'}$ implies that $P_{\Delta_{N'}}$ is a well-defined 1-Lipschitz map, see e.g. [14, Proposition 4.8].

Definition 4 (Geometric Attention Mechanism)**.** *Let $N', d_X \in \mathbb{N}_+$. A geometric attention mechanism is a map* $\text{g-attn}_{N'}:\mathbb{R}^{N'} \to \mathcal{N}_{d_X}$ *with representation*

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

where $m^{(1)},\dots,m^{(N')} \in \mathbb{R}^{d_X}$, $A^{(1)},\dots,A^{(N')} \in \mathbb{R}^{d_X \times d_X}$; and $\vartheta \overset{\text{def.}}{=} (m^{(n)}, A^{(n)})_{n=1}^{N'}$.

3.4 The FF Model

We may now formalize the Filterformer model of Figure 2.

Definition 5 (Filterformer (FF))**.** *Let $d_X, d_Y \in \mathbb{N}_+$, $T > 0$, and an activation function $\sigma:\mathbb{R}\to\mathbb{R}$. A function $\hat{F}: C([0:T],\mathbb{R}^{d_Y}) \to \mathcal{N}_{d_X}$ is called a Filterformer if it admits the representation*

$$\hat{F} = \text{g-attn}_{N'}^{\vartheta} \circ \hat{f} \circ \text{attn}_T^{\theta} \tag{9}$$

where $\text{g-attn}_{N'}^{\vartheta}$ *is as in Definition 4,* attn_T^{θ} *is as in Definition 3, and $\hat{f}$ is an MLP as in (8); such that the composition (9) is well-defined.*

4 Main Result

We are now ready to state our main theorem, which shows that the FF (9) is indeed capable of asymptotically optimally filtering the coupled system (2)-(3). Our guarantees are of a non-asymptotic form; in that they depend on the complexity of the network.

Theorem 1 (The FF Can Approximate the Optimal Filter)**.** *Let $d_X, d_Y \in \mathbb{N}_+$ and $K \subset C([0:T],\mathbb{R}^{d_Y})$ satisfy Assumption 2. Suppose that the coupled system (2)-(3) satisfies Assumption 1. For every $T > 0$ there exists a FF $\hat{F}: [0:T] \times C([0:T]\times\mathbb{R}^{d_Y}) \to \mathcal{N}_{d_X}$ satisfying the uniform estimate*

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

for every $1 \le p \le 2$. Furthermore, Table 2 records the complexity estimates of $\hat{F}$.

Table 2: Complexity Estimates for the FF model $\hat{F}$ in Theorem 1.

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

4.1 Proof Overview

This section overviews the derivation of our main result, namely, Theorem 1.

Results in [89] to show that $\mathbb{P}(X_t \in \cdot|\mathcal{Y}_t)$ is conditionally Gaussian, they denote an explicit formulation of the resulting distribution which can then analyze directly. In particular, Assumption 1 allows the application of [89, Theorem 12.6] and we obtain that for each $0 \le t \le T$ and each path $y_{[0:t]} \in C([0:t],\mathbb{R}^{d_Y})$ the probability measure $f_t(y_{[0:t]})$ is a d_X-dimensional Gaussian measure. We denote the mean and covariance of this measure by $\mu(y_{[0:t]})$ and $\Sigma(y_{[0:t]})$. Thus,

$$\mathcal{N}(\mu(y_{[0:t]}),\Sigma(y_{[0:t]})) \sim \mathbb{P}(X_t \in \cdot|y_{[0:t]}) \overset{\text{def.}}{=} f_t(y_{[0:t]}), \tag{10}$$

for each $0 \le t \le T$ and each path $y_{[0:t]} \in C([0:t],\mathbb{R}^{d_Y})$.

A function is uniformly approximable by a deep learning model if it is continuous. If there is no clear favourable structure in the function being approximated, e.g. smoothness [123, 56] or neural network-like structure [25, 94], then the best available approximation rates are those corresponding to those where the target function is Lipschitz; see [34]. Thus, the first step to obtaining approximability guarantees for the optimal filter by deep neural networks, which depend on relatively few trainable parameters, is to show that the optimal filter is a locally-Lipschitz function of observed paths.

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

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

Having established the regularity of the optimal filter for the coupled system (2)-(3), we need to verify that maps sharing the same domain (input space), codomain (output space), and regularity as the optimal filter are approximable by the proposed FF model. This is the content of the next proposition, which acts as its standalone approximation theorem for the FF.

Table 3: Complexity Estimates for transformer-type model $\hat{F}$ in Proposition 3.

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

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

Proposition 3 (Approximation Capacity of FFs)**.** *Let $d_X, d_Y \in \mathbb{N}_+$, $L > 0$, and $K \subset C([0:T], \mathbb{R}^{d_Y})$ satisfying Assumption 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 FF $\hat{F}$ satisfying the uniform estimate*

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

for every $1 \leq p \leq 2$. Furthermore, the depth, width, encoding dimension (N) and Decoding dimension (N') are all recorded in Table 3 depending on the activation function σ.

We now have at our disposal, the tools to prove our main result.

Proof. **of Theorem 1**
Since the coupled system (2)-(3) satisfies Assumption 1, Proposition 2 implies that the optimal filter f_T in (1) is a locally-Lipschitz map from $([0:T], C([0:T], \mathbb{R}^{d_Y}), |\cdot| \times \|\cdot\|_T)$ to $(\mathcal{N}_{d_X}, \mathcal{W}_2)$. Since $K \subseteq C([0:T], \mathbb{R}^{d_Y})$ is compact then, $f|_K$ is Lipschitz.

Since $K \subset C([0:T], \mathbb{R}^{d_Y})$ satisfies Assumption 2 and $0 < \varepsilon < 1/2$ then, Proposition 3 implies that there is a FF $\hat{F} : C([0:T], \mathbb{R}^{d_Y}) \to \mathcal{N}_{d_X}$ with representation (9) and complexity recorded in Table 3 satisfying

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

This completes the proof of Theorem 1. □

5 Discussion and Future Work

Our main result, namely Theorem 1, showed that there are Filterformers which can solve the approximately stochastic filtering problem in continuous time for the coupled system (2)-(3), to arbitrary accuracy. This complements the experimental work of [81, 80, 24, 5], which shows that deep learning models seemingly can approximately filter and that such models can be trained to offer state-of-the-art empirical performance.

From a theoretical perspective, it would be interesting to analyze the statistical learning theoretical properties of FFs. One would expect that elementary generalization bounds could be derived by using the bounds on VC-dimension derived in [11] and a version of the results of [12] or by directly modifying the transport theoretic arguments of [66], to incorporate the effects of the pathwise and geometric attention layers in (9). Nevertheless, for problems in mathematical finance where one only observes a single training path (one-shot learning), the most interesting question in that direction is: *How to train the FFs in a way such that they have robust statistical guarantees when only a single training path is observed.*

6 Proofs

This section contains the derivations of our paper's results.

6.1 Stability of the Optimal Filter - Proof of Proposition 2

We now prove the local Lipschitz stability of the optimal filter (1), for the coupled system (2)-(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 1 grants us access to [89, Theorem 12.7], which implies that $\mu_t \stackrel{\text{def.}}{=} \mu(y_{[0:t]})$ and $\Sigma_t \stackrel{\text{def.}}{=} \Sigma(y_{[0:t]})$ are

$$\begin{aligned} d\mu_t = {} & [a_0(t, y_{[0:t]}) + a_1(t, y_{[0:t]})\mu_t]dt + [(b \circ B)(t, y_{[0:t]}) + \Sigma_t A_1^\top(t, y_{[0:t]})] \\ & \times (B \circ B)^{-1}(t, y_{[0:t]})[dy_t - (A_0(t, y_{[0:t]}) + A_1(t, y_{[0:t]}))\mu_t dt] \end{aligned} \tag{11}$$

$$\partial_t \Sigma_t = a_1(t, y_{[0:t]})\Sigma_t + \Sigma_t a_1^\top(t, y_{[0:t]}) + (b \circ b)(t, y_{[0:t]}) \tag{12}$$

$$-[(b\circ B)(t,y_{[0:t]})+\Sigma_t A_1^\top(t,y_{[0:t]})](B\circ B)^{-1}(t,y_{[0:t]})$$
$$\times[(b\circ B)(t,y_{[0:t]})+\Sigma_t A_1^\top(t,y_{[0:t]})]^\top$$

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

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

Furthermore, [89, 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 1.

To clarify the argument's direction, we first record the proof of Proposition 2. We subsequently derive the lemmata on which it relies.

Proof. **of Proposition 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 7 we can fix a $K_1>0$ such that for every time $0<s<T$ holds

$$\mathcal{W}_2(\mathbb{P}(X_t\in\cdot|y^{(1)}_{[0:t]}),\mathbb{P}(X_s\in\cdot|y^{(1)}_{[0:s]}))\le K_1|t-s|_t.$$

By Lemma 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 $\|y^{(1)}_{[0:T]}-y^{(2)}_{[0:T]}\|_T<\epsilon$ holds

$$\mathcal{W}_2(\mathbb{P}(X_u\in\cdot|y^{(1)}_{[0:u]}),\mathbb{P}(X_u\in\cdot|y^{(2)}_{[0:u]}))\le K_2\|y^{(1)}_{[0:T]}-y^{(2)}_{[0:T]}\|_T.$$

We see by the triangular inequality of the $\mathcal{W}_2$-distance (that holds as we handle normal distributions, see (10)),

$$\begin{aligned}&\mathcal{W}_2(\mathbb{P}(X_t\in\cdot|y^{(1)}_{[0:t]}),\mathbb{P}(X_s\in\cdot|y^{(2)}_{[0:s]}))\\ &\quad\le \mathcal{W}_2(\mathbb{P}(X_t\in\cdot|y^{(1)}_{[0:t]}),\mathbb{P}(X_s\in\cdot|y^{(1)}_{[0:s]}))+\mathcal{W}_2(\mathbb{P}(X_s\in\cdot|y^{(1)}_{[0:s]}),\mathbb{P}(X_s\in\cdot|y^{(2)}_{[0:s]}))\\ &\quad\le \max\{K_1,K_2\}\big(\|y^{(1)}_{[0:T]}-y^{(2)}_{[0:T]}\|_T+|t-s|\big)\end{aligned}$$

what concludes the proof. □

6.1.1 Lipschitz-Continuity in the Path-Component of the Optimal Filter

Lemma 1 (Local Lipschitz - the Path-Component of the Optimal Filter)**.** *Under Assumption 1, f from (1) 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 $C\ge 0$ such that for all paths $y_\cdot^{(2)}\in C^1([0:T],\mathbb{R}^{d_Y})$ with $\|y^{(1)}_{[0:t]}-y^{(2)}_{[0:t]}\|_t<\epsilon$ holds*

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

Proof. **of Lemma 1** By (10), Lemma 10, and Lemma 16 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 $t\in[0:T]$

$$\mathcal{W}_2(\mathbb{P}(X_t\in\cdot|y^{(1)}_{[0:t]}),\mathbb{P}(X_t\in\cdot|y^{(2)}_{[0:t]}))\le\bar{K}\sqrt{\|\mu(y^{(1)}_{[0:t]})-\mu(y^{(2)}_{[0:t]})\|_2^2+\|\Sigma(y^{(1)}_{[0:t]})-\Sigma(y^{(2)}_{[0:t]})\|_2^2}.$$

We see from Lemma 2 and Lemma 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 $y_\cdot^{(1)},y_\cdot^{(2)}\in C([0:T],\mathbb{R}^{d_Y})$ 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 2 (Local Lipschitz-continuity of μ in the path). *Under Assumption 1 (viii) $\mu_.$ from (11) is locally Lipschitz-continuous uniformly in time. This means that for every path $y_.^{(1)} \in C^1([0:T], \mathbb{R}^{d_Y})$ there exists constant $K \in \mathbb{R}$, $\epsilon > 0$ such that for all paths $y_.^{(2)} \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y_{[0:t]}^{(1)} - y_{[0:t]}^{(2)}\|_t < \epsilon$ holds*

$$\big\|\mu_t^{(1)} - \mu_t^{(2)}\big\|_2 \le K\big\|y_{[0:t]}^{(1)} - y_{[0:t]}^{(2)}\big\|_t, \qquad \forall t \in [0:T].$$

Proof. **of Lemma 2** Define

$$\begin{aligned}\Phi_1(s, y_{[0:s]}) &\stackrel{\text{def.}}{=} a_1(s, y_{[0:s]}) - [(b \circ B)(s, y_{[0:s]}) + \Sigma_t A_1^\top(s, y_{[0:s]})] \\ &\qquad\qquad \times (B \circ B)^{-1}(s, y_{[0:s]})[A_0(s, y_{[0:s]}) + A_1(s, y_{[0:s]})], \end{aligned} \tag{13}$$

$$\Phi_2(s, y_{[0:s]}) \stackrel{\text{def.}}{=} a_0(s, y_{[0:s]}) + [(b \circ B)(s, y_{[0:s]}) + \Sigma_t A_1^\top(s, y_{[0:s]})](B \circ B)^{-1}(s, y_{[0:s]}). \tag{14}$$

Note that conditions of Lemma 3 and Lemma 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 (11) 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

$$\begin{aligned}\mu_t^{(1)} - \mu_t^{(2)} = \mu_0^{(1)} - \mu_0^{(2)} &+ \int_0^t a_0(s, y_{[0:s]}^{(1)}) - a_0(s, y_{[0:s]}^{(2)})ds \\ &+ \int_0^t \Phi_2(s, y_{[0:s]}^{(1)})dy_{[0:s]}^{(1)} - \int_0^t \Phi_2(s, y_{[0:s]}^{(2)})dy_{[0:s]}^{(2)} \\ &+ \int_0^t \Phi_1(s, y_{[0:s]}^{(1)}) - \Phi_1(s, y_{[0:s]}^{(2)})\mu_s^{(2)} ds + \int_0^t \Phi_1(s, y_{[0:s]}^{(1)})\,(\mu_s^{(1)} - \mu_s^{(2)})ds.\end{aligned}$$

By the triangle inequality, Jensen's inequality, and Cauchy-Schwarz inequality, we have

$$\big\|\mu_t^{(1)} - \mu_t^{(2)}\big\|_2 \le \gamma(t) + \int_0^t \big\|\Phi_1(s, y_{[0:s]}^{(1)})\big\|_2 \cdot \big\|\mu_s^{(1)} - \mu_s^{(2)}\big\|_2 ds,$$

where

$$\begin{aligned}\gamma(t) \stackrel{\text{def.}}{=} \big\|\mu_0^{(1)} - \mu_0^{(2)}\big\|_2 &+ \underbrace{\Big\|\int_0^t a_0(s, y_{[0:s]}^{(1)}) - a_0(s, y_{[0:s]}^{(2)})ds\Big\|_2}_{\text{(I)}} \\ &+ \underbrace{\Big\|\int_0^t \Phi_2(s, y_{[0:s]}^{(1)})dy_{[0:s]}^{(1)} - \int_0^t \Phi_2(s, y_{[0:s]}^{(2)})dy_{[0:s]}^{(2)}\Big\|_2}_{\text{(II)}} \\ &+ \underbrace{\Big\|\int_0^t \Phi_1(s, y_{[0:s]}^{(1)}) - \Phi_1(s, y_{[0:s]}^{(2)})\mu_s^{(2)} ds\Big\|_2}_{\text{(III)}}.\end{aligned}$$

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 $K_1 < \infty$,

$$\text{(I)} \le \int_0^t \|a_0(s, y_{[0:s]}^{(1)}) - a_0(s, y_{[0:s]}^{(2)})\|_2 ds \le \int_0^t K_1\|y_{[0:s]}^{(1)} - y_{[0:s]}^{(2)}\|_s ds \le tK_1\|y_{[0:t]}^{(1)} - y_{[0:t]}^{(2)}\|_t.$$

Further, we see that

$$(\text{II}) \le \underbrace{\left\| \int_0^t \Phi_2(s, y^{(2)}_{[0:s]}) - \Phi_2(s, y^{(1)}_{[0:s]}) dy^{(2)}_{[0:s]} \right\|_2}_{(\text{IV})} + \underbrace{\left\| \int_0^t \Phi_2(s, y^{(1)}_{[0:s]}) d(y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]}) \right\|_2}_{(\text{V})}.$$

Using Itô's isometry, we have

$$\begin{aligned}(\text{IV})^2 &= \int_0^t \left\| \Phi_2(s, y^{(2)}_{[0:s]}) - \Phi_2(s, y^{(1)}_{[0:s]}) \right\|_2^2 d\,(B \circ B)(s, y^{(2)}_{[0:s]}) \\ &\le K_2^2 \| y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]} \|_t^2 \int_0^t d\,(B \circ B)(s, y^{(2)}_{[0:s]})\end{aligned}$$

where the last inequality holds due to (path-)Lipschitz continuity of Φ_2 uniformly in time with constant $K_2 < \infty$, given by Lemma 4. Next, let $\{[s_{i-1} : s_i] | i \in \{1, \dots, I\}\}$, $I \in \mathbb{N}$ 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

$$\begin{aligned}(\text{IV})^2 &\le \epsilon_I + K_2^2 \| y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]} \|_t^2 \cdot \sum_{i=1}^{I} \| (B \circ B)(y^{(2)}_{[0:s_i]}) - (B \circ B)(y^{(2)}_{[0:s_{i-1}]}) \|_2 \\ &\le \epsilon_I + K_2^2 T \| y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]} \|_t^2.\end{aligned}$$

As $\epsilon_I \to 0$ for $I \to \infty$, we obtain $(\text{IV}) \le K_2 \sqrt{T} \| y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]} \|_t$.

Integrating by parts, we find that

$$\begin{aligned}(\text{V}) &= \left\| \Phi_2(t, y^{(1)}_{[0:t]})(y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]}) - \Phi_2(0, y^{(1)}_0)(y^{(2)}_0 - y^{(1)}_0) - \int_0^t (y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]})\, d\Phi_2(s, y^{(1)}_{[0:s]}) \right\|_2 \\ &\le \left\| \Phi_2(t, y^{(1)}_{[0:t]})(y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]}) - \Phi_2(0, y^{(1)}_0)(y^{(2)}_0 - y^{(1)}_0) \right\|_2 + \left\| \int_0^t (y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]})\, d\Phi_2(s, y^{(1)}_{[0:s]}) \right\|_2 \\ &\le 2 \max_{0 \le s \le T} \left\| \Phi_2(s, y^{(1)}_{[0:s]}) \right\|_2 \cdot \left\| y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]} \right\|_t + \underbrace{\left\| \int_0^t (y^{(2)}_{[0:s]} - y^{(1)}_{[0:s]})\, d\Phi_2(s, y^{(1)}_{[0:s]}) \right\|_2}_{(\text{VI})}.\end{aligned}$$

Lemma 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

$$(\text{VI}) \le \epsilon_I + \sum_{i=1}^{I} \left\| y^{(2)}_{[0:s_i]} - y^{(1)}_{[0:s_i]} \right\|_2 \left\| \Phi_2(s_i, y^{(1)}_{[0:s_i]}) - \Phi_2(s_{i-1}, y^{(1)}_{[0:s_{i-1}]}) \right\|_2 \le \epsilon_I + K_4 t \| y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]} \|_t,$$

which holds for $\epsilon_I \to 0$ for $I \to \infty$; we conclude that

$$(\text{II}) \le \left(K_2 \sqrt{T} + K_3 + K_4 T \right) \| y^{(2)}_{[0:t]} - y^{(1)}_{[0:t]} \|_t.$$

Next, we consider

$$(\text{III}) \le \int_0^t \left\| \Phi_1(s, y^{(1)}_{[0:s]}) - \Phi_1(s, y^{(2)}_{[0:s]}) \right\|_2 \left\| \mu^{(2)}_s \right\|_2 ds \le K_5 t \| y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]} \|_t \max_{s \in [0:t]} \left\| \mu^{(2)}_s \right\|_2$$

what holds due to (path-)Lipschitz continuity of Φ_1 uniformly in time with constant K_5 which is obtained by Lemma 3. We conclude, since $\max_{s \in [0:T]} \| \mu^{(2)}_s \|_2 := K_6 < \infty$ due to Lemma 8, that $\gamma(t) \le K \| y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]} \|_t$ 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

$$
\begin{aligned}
\big\|\mu_t^{(1)} - \mu_t^{(2)}\big\|_2 &\leq \gamma(t) + \int_0^t \gamma(s) \big\|\Phi_1(s, y_{[0:s]}^{(1)})\big\|_2 \exp\Big(\int_s^t \big\|\Phi_1(r, y_{[0:r]}^{(1)})\big\|_2 dr\Big) ds \\
&\leq K \|y_{[0:t]}^{(1)} - y_{[0:t]}^{(2)}\|_t \Big(1 + \int_0^t \big\|\Phi_1(s, y_{[0:s]}^{(1)})\big\|_2 \exp\Big(\int_s^t \big\|\Phi_1(r, y_{[0:r]}^{(1)})\big\|_2 dr\Big) ds\Big)
\end{aligned}
$$

Note that Φ_1 is (time-)Lipschitz continuous, as shown in Lemma 3, therefore implying that $\max_{s\in[0:T]} \|\Phi_1(s, y_{[0:s]}^{(1)})\|_2 < \infty$. 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 1 (xiii) holds, we conclude the proof. □

Lemma 3. *Let Assumption 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 (13).

2. *Also, for every path $y_\cdot^{(1)} \in C^1([0:T], \mathbb{R}^{d_Y})$ there exists a constant $C \geq 0$ such that for all $\epsilon > 0$, paths $y_\cdot^{(2)} \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y_{[0:T]}^{(1)} - y_{[0:T]}^{(2)}\|_T < \epsilon$, and times $t \in [0:T]$ holds*

$$
\big\|\Phi_1(t, y_{[0:t]}^{(1)}) - \Phi_1(t, y_{[0:t]}^{(2)})\big\|_2 \leq C \big\|y_{[0:t]}^{(1)} - y_{[0:t]}^{(2)}\big\|_t.
$$

Proof. **of Lemma 3**

1. The statement holds if a_1, $(b \circ B)(B \circ B)^{-1} A_0$, $(b \circ B)(B \circ B)^{-1} A_1$, $\Sigma.A_1(B \circ B)^{-1} A_0$, $\Sigma.A_1(B \circ B)^{-1} A_1$ are globally Lipschitz continuous w.r.t. the time component. As a_1, $(b \circ B)$, $\Sigma.$, $(B \circ B)^{-1}$, A_0, A_1 satisfy this already (see Lemma 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, $(b \circ B)(B \circ B)^{-1} A_0$, $(b \circ B)(B \circ B)^{-1} A_1$, $\Sigma.A_1(B \circ B)^{-1} A_0$, $\Sigma.A_1(B \circ B)^{-1} A_1$ are locally Lipschitz continuous w.r.t. the path component uniformly in time. As a_1, $(b \circ B)$, $\Sigma.$, $(B \circ B)^{-1}$, A_0, A_1 are locally Lipschitz continuous w.r.t. the path component (see Lemma 5) uniformly in time, this also follows for their product.

□

Lemma 4. *Let Assumption 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 (14).

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 $\|y_{[0:t]}^{(1)} - y_{[0:t]}^{(2)}\|_t < \epsilon$, and times $t \in [0:T]$ holds*

$$
\big\|\Phi_2(t, y_{[0:t]}^{(1)}) - \Phi_2(t, y_{[0:t]}^{(2)})\big\|_2 \leq C \big\|y_{[0:t]}^{(1)} - y_{[0:t]}^{(2)}\big\|_t.
$$

Proof. **of Lemma 4**

1. The assertion holds if a_0, $(b \circ B)(B \circ B)^{-1}$, $\Sigma.A_1(B \circ B)^{-1}$ are globally Lipschitz continuous w.r.t. the time component. As a_0, $(b \circ B)$, $\Sigma.$, $(B \circ B)^{-1}$, A_1 satisfy this already (see Lemma 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 5) this also follows for their product.

$\square$

Lemma 5 (Local Lipschitz-continuity of Σ)**.** *Under Assumption 1 (viii), $\Sigma_{\cdot}$ from (12) is locally Lipschitz-continuous uniformly in time. This means, for every path $y_{\cdot}^{(1)}\in C^1([0:T],\mathbb{R}^{d_Y})$ there exists constant $C\in\mathbb{R}$ such that for all $\epsilon>0$, paths $y_{\cdot}^{(2)}\in C^1([0:T],\mathbb{R}^{d_Y})$ with $\|y_{[0:t]}^{(1)}-y_{[0:t]}^{(2)}\|_t<\epsilon$, and times $t\in[0:T]$ holds*

$$\big\|\Sigma_t^{(1)}-\Sigma_t^{(2)}\big\|_2\leq C\big\|y_{[0:t]}^1-y_{[0:t]}^2\big\|_t,$$

where $\|\cdot\|_2$ on the left-hand side refers to the Frobenius norm.

Proof. **of Lemma 5** From (12) and for $t\in[0:T]$, we obtain the integral representation of Σ_t

$$\begin{aligned}\Sigma_t=&\Sigma_0+\int_0^t a_1(t,y_{[0:s]})\Sigma_s+\Sigma_s a_1^\top(s,y_{[0:s]})+(b\circ b)(s,y_{[0:s]})\\&\qquad-[(b\circ B)(s,y_{[0:s]})+\Sigma_s A_1^\top(s,y_{[0:s]})](B\circ B)^{-1}(s,y_{[0:s]})\\&\qquad\qquad\times[(b\circ B)(s,y_{[0:s]})+\Sigma_s A_1^\top(s,y_{[0:s]})]^\top ds.\end{aligned}$$

Let $\Sigma_t^{(1)}$, $\Sigma_t^{(2)}$ be the covariance matrices corresponding to paths $y_{\cdot}^{(1)},y_{\cdot}^{(2)}\in C^1([0:T],\mathbb{R}^{d_Y})$ where $y_{\cdot}^{(1)}$ is arbitrary and $y_{\cdot}^{(2)}$ s.t. $\|y_{[0:t]}^{(1)}-y_{[0:t]}^{(2)}\|_t<\epsilon$. The difference $\Sigma_t^{(1)}-\Sigma_t^{(2)}$ satisfies the integral representation

$$\begin{aligned}\Sigma_t^{(1)}-\Sigma_t^{(2)}=\Sigma_0^{(1)}-\Sigma_0^{(2)}&+\int_0^t\Xi_1(s,y_{[0:s]}^{(1)},y_{[0:s]}^{(2)},\Sigma_s^{(1)},\Sigma_s^{(2)})ds\\&+\int_0^t(\Sigma_s^{(1)}-\Sigma_s^{(2)})\Xi_2(s,y_{[0:s]}^{(1)},y_{[0:s]}^{(2)},\Sigma_s^{(1)},\Sigma_s^{(2)})+\Xi_2^\top(s,y_{[0:s]}^{(1)},y_{[0:s]}^{(2)},\Sigma_s^{(1)},\Sigma_s^{(2)})(\Sigma_s^{(1)}-\Sigma_s^{(2)})ds,\end{aligned}$$

where we denote

$$\begin{aligned}&\Xi_1(s,y_{[0:s]}^{(1)},y_{[0:s]}^{(2)},\Sigma_s^{(1)},\Sigma_s^{(2)})\\&\quad\stackrel{\text{def.}}{=}(a_1(s,y_{[0:s]}^{(1)})-a_1(s,y_{[0:s]}^{(2)}))\Sigma_s^{(1)}+\Sigma_s^{(1)}(a_1^\top(s,y_{[0:s]}^{(1)})-a_1^\top(s,y_{[0:s]}^{(2)}))\\&\qquad+(b\circ b)(s,y_{[0:s]}^{(1)})-(b\circ b)(s,y_{[0:s]}^{(2)})\\&\qquad-\Sigma_s^{(2)}[A_1^\top(B\circ B)^{-1}(b\circ B)(s,y_{[0:s]}^{(1)})-A_1^\top(B\circ B)^{-1}(b\circ B)(s,y_{[0:s]}^{(2)})]\\&\qquad-[(b\circ B)(B\circ B)^{-1}A_1(s,y_{[0:s]}^{(1)})-(b\circ B)(B\circ B)^{-1}A_1(s,y_{[0:s]}^{(2)})]\Sigma_s^{(2)}\\&\qquad-\Sigma_s^{(2)}[A_1^\top(B\circ B)^{-1}A_1(s,y_{[0:s]}^{(1)})-A_1^\top(B\circ B)^{-1}A_1(s,y_{[0:s]}^{(2)})]\Sigma_s^{(1)}\\&\qquad-[(b\circ B)(B\circ B)^{-1}(b\circ B)^\top(s,y_{[0:s]}^{(1)})-(b\circ B)(B\circ B)^{-1}(b\circ B)^\top(s,y_{[0:s]}^{(2)})]\end{aligned}\tag{15}$$

and

$$\begin{aligned}\Xi_2(s,y_{[0:s]}^{(1)},y_{[0:s]}^{(2)},\Sigma_s^{(1)},\Sigma_s^{(2)})&\stackrel{\text{def.}}{=}a_1^\top(s,y_s^{(2)})+A_1^\top(s,y_{[0:s]}^{(1)})(B\circ B)^{-1}(s,y_{[0:s]}^{(1)})\\&\qquad\times(b\circ B)(s,y_{[0:s]}^{(1)})+A_1^\top(s,y_{[0:s]}^{(1)})(B\circ B)^{-1}(s,y_{[0:s]}^{(1)})A_1(s,y_{[0:s]}^{(1)})\Sigma_s^{(1)}.\end{aligned}\tag{16}$$

By the triangular inequality, Jensen's inequality, and Cauchy-Schwarz inequality, we have

$$\big\|\Sigma_t^{(1)}-\Sigma_t^{(2)}\big\|_2\leq\big\|\Sigma_0^{(1)}-\Sigma_0^{(2)}\big\|_2+\int_0^t\big\|\Xi_1(s,y_{[0:s]}^{(1)},y_{[0:s]}^{(2)},\Sigma_s^{(1)},\Sigma_s^{(2)})\big\|_2ds$$

$$+ 2\int_0^t \big\|\Sigma_s^{(1)} - \Sigma_s^{(2)}\big\|_2 \cdot \big\|\Xi_2(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma_s^{(1)}, \Sigma_s^{(2)})\big\|_2 ds.$$

By Grönwall's inequality, it holds that

$$\begin{aligned}\big\|\Sigma_t^{(1)} - \Sigma_t^{(2)}\big\|_2 \le \big\|\Sigma_0^{(1)} - \Sigma_0^{(2)}\big\|_2 + \int_0^t \big\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma_s^{(1)}, \Sigma_s^{(2)})\big\|_2 ds \\ + \int_0^t \Big(\big\|\Sigma_0^{(1)} - \Sigma_0^{(2)}\big\|_2 + \int_0^s \big\|\Xi_1(r, y^{(1)}_{[0:r]}, y^{(2)}_{[0:r]}, \Sigma_r^{(1)}, \Sigma_r^{(2)})\big\|_2 dr\Big)\cdot \\ \cdot 2\big\|\Xi_2(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma_s^{(1)}, \Sigma_s^{(2)})\big\|_2 \cdot \\ \cdot \exp\Big(2\int_s^t \big\|\Xi_2(r, y^{(1)}_{[0:r]}, y^{(2)}_{[0:r]}, \Sigma_r^{(1)}, \Sigma_r^{(2)})\big\|_2 dr\Big) ds. \quad (17)\end{aligned}$$

From Lemma 6, we know that there exists a constant $K \in \mathbb{R}$ such that

$$\big\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma_s^{(1)}, \Sigma_s^{(2)})\big\|_2 \le K\big\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\big\|_s. \quad (18)$$

Combining (17) and (18), we obtain

$$\begin{aligned}\big\|\Sigma_t^{(1)} - \Sigma_t^{(2)}\big\|_2 \le \big(\big\|\Sigma_0^{(1)} - \Sigma_0^{(2)}\big\|_2 + KT\big\|y^{(1)} - y^{(2)}\big\|_T\big) \cdot \Big[1 + \int_0^t 2\big\|\Xi_2(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma_s^{(1)}, \Sigma_s^{(2)})\big\|_2 \cdot \\ \cdot \exp\Big(2\int_s^t \big\|\Xi_2(r, y^{(1)}_{[0:r]}, y^{(2)}_{[0:r]}, \Sigma_r^{(1)}, \Sigma_r^{(2)})\big\|_2 dr\Big) ds\Big]. \quad (19)\end{aligned}$$

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 (16). Therefore, from (19) we obtain a constant $C \in \mathbb{R}$ such that

$$\big\|\Sigma_t^{(1)} - \Sigma_t^{(2)}\big\|_2 \le C(\big\|\Sigma_0^{(1)} - \Sigma_0^{(2)}\big\|_2 + \big\|y^{(1)} - y^{(2)}\big\|_T).$$

We conclude the proof due to Assumption 1 (xiii). □

Lemma 6. *Let Ξ_1 be defined as in (15), and let Assumption 1 (viii) hold. Then, for every path $y^{(1)} \in C^1([0:T], \mathbb{R}^{d_Y})$ there exists constant $K \in \mathbb{R}$, such that for all $\epsilon > 0$, paths $y^{(2)} \in C^1([0:T], \mathbb{R}^{d_Y})$ with $\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t < \epsilon$, and times $s \in [0:T]$ holds*

$$\big\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma_s^{(1)}, \Sigma_s^{(2)})\big\|_2 \le K\big\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\big\|_s.$$

Proof. **of Lemma 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 $K_1 \in \mathbb{R}$

$$\begin{aligned}&\big\|\Xi_1(s, y^{(1)}_{[0:s]}, y^{(2)}_{[0:s]}, \Sigma_s^{(1)}, \Sigma_s^{(2)})\big\| \\ &\le 2K_1\big\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\big\|_s\Big(\big\|\Sigma_s^{(1)}\big\|_2 + \big\|\Sigma_s^{(2)}\big\|_2\Big) + K_1\big\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\big\|_s\big\|\Sigma_s^{(1)}\big\|_2\big\|\Sigma_s^{(2)}\big\|_2 \\ &\qquad + K_1\big\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\big\|_s \\ &= K_1\big\|y^{(1)}_{[0:s]} - y^{(2)}_{[0:s]}\big\|_s\Big(1 + 2\big\|\Sigma_s^{(1)}\big\|_2 + 2\big\|\Sigma_s^{(2)}\big\|_2 + \big\|\Sigma_s^{(1)}\big\|_2\big\|\Sigma_s^{(2)}\big\|_2\Big).\end{aligned}$$

With that, it is left to show that $\|\Sigma_s^{(1)}\|_2 < \infty$ and $\|\Sigma_s^{(2)}\|_2 < K_2$ 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 $\|\Sigma_s^{(2)}\|_2$ is bounded. □

6.1.2 Lipschitz-Continuity in the Time-Component of the Optimal Filter

Lemma 7 (Global Lipschitz-Cont. in the Time-Comp. of the Optimal Filter)**.** *Under Assumption 1, f_t from (1) 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 7** As in the proof of Lemma 1, we argue that by (10), Lemma 10, and Lemma 16 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 8 and Lemma 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 8 (Global Lipschitz-continuity of μ in time)**.** *Let the Assumption 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 (11).*

Proof. **of Lemma 8** With the notation of Lemma 2, we have

$$\|\mu_t - \mu_s\|_2 \leq \underbrace{\Big\| \int_s^t a_0(u, y_{[0:u]}) du \Big\|_2}_{\text{(I)}} + \underbrace{\Big\| \int_s^t \Phi_2(u, y_{[0:u]}) dy_{[0:u]} \Big\|_2}_{\text{(II)}} + \underbrace{\Big\| \int_s^t \Phi_1(u, y_{[0:u]}) \mu_u du \Big\|_2}_{\text{(III)}}.$$

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,

$$\begin{aligned}
\text{(II)} &= \Big\| \Phi_2(t, y_{[0:t]}) y_{[0:t]} - \Phi_2(s, y_{[0:s]}) y_{[0:s]} - \int_0^t y_{[0:u]} \, d\Phi_2(u, y_{[0:u]}) \Big\|_2 \\
&\leq \underbrace{\Big\| \Phi_2(t, y_{[0:t]}) y_{[0:t]} - \Phi_2(0, y_0) y_0 \Big\|_2}_{\text{(IV)}} + \underbrace{\Big\| \int_0^t y_{[0:u]} \, d\Phi_2(u, y_{[0:u]}) \Big\|_2}_{\text{(V)}},
\end{aligned}$$

we see that (IV) $\leq K\|y^{(1)}_{[0:t]} - y^{(2)}_{[0:t]}\|_t$ due to (path-)Lipschitz continuity of Φ_2. Let $\{[u_{i-1} : u_i] | i \in \{1, \ldots, I\}\}, I \in \mathbb{N}$ 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

$$\text{(V)} \leq \epsilon_I + \sum_{i=1}^I \|y_{[0:u_i]}\|_2 \big\| \Phi_2(u_i, y_{[0:u_i]}) - \Phi_2(u_{i-1}, y_{[0:u_{i-1}]}) \big\|_2 \leq \epsilon_I + \|y_{[0:t]}\|_t \sum_{i=1}^I K|u_i - u_{i-1}|$$

and, since $\epsilon_I \to 0$ for $I \to \infty$, it follows (II) $\leq K|t-s|$. Since (I) $< \infty$ and (II) $< \infty$, we can apply Grönwall's inequality to

$$\|\mu_t\|_2 \leq \overbrace{\Big\| \int_0^t a_0(u, y_{[0:u]}) du \Big\|_2 + \Big\| \int_0^t \Phi_2(u, y_{[0:u]}) dy_{[0:u]} \Big\|_2}^{\stackrel{\text{def.}}{=} \alpha(t)} + \int_0^t \big\| \Phi_1(u, y_{[0:u]}) \big\|_2 \big\| \mu_u \big\|_2 du,$$

which yields

$$\|\mu_t\|_2 \leq \alpha(t) + \int_0^t \alpha(s) \big\| \Phi_1(s, y_{[0:s]}) \big\|_2 \exp\left(\int_s^t \big\| \Phi_1(u, y_{[0:u]}) \big\|_2 du \right) ds \leq Kt$$

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

$$(\text{III}) \leq \int_s^t \big\|\Phi_1(u, y_{[0:u]})\big\|_2 \big\|\mu_u\big\|_2 du \leq \max_{u\in[0:t]} \|\Phi_1(u, y_{[0:u]})\|_2 KT|t-s|$$

what proves the claim. ☐

Lemma 9 (Global Lipschitz-continuity of Σ in time)**.** *Let Assumption 1 (viii) hold. Then, having Σ is defined as in (12), 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*

$$\|\Sigma_t - \Sigma_s\|_2 \leq K|t-s|.$$

Proof. **of Lemma 9** Note that $\Sigma_\cdot$ is continuous in time. From (12), 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 $[0:T]$. ☐

6.1.3 Lower Loewner order bound

Lemma 10 (Lower Loewner order bound)**.** *Let Σ be defined as in Equation (12) and Assumption 1 (ix), (x), hold. There is 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 $\Sigma_t \preccurlyeq rI_d$.*

Proof. **of Lemma 10** By the Courant-Fisher Theorem, see [65], 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 [89, 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 $\tilde{a}_1^\top(t, y_{[0:t]}) \stackrel{\text{def.}}{=} a_1(t, y_{[0:t]}) - (b\circ B)(B\circ B)A_1(t, y_{[0:t]})$.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 [89, 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 $\operatorname{tr}\left(G_s(y_{[0:s]})A_1^\top(B\circ B)^{-1}A_1(s, y_{[0:s]})G_s(y_{[0:s]})^\top\right) \leq \operatorname{tr}\left(G_s(y_{[0:s]})\right)^2 \operatorname{tr}\left(A_1^\top(B\circ B)^{-1}A_1(s, y_{[0:s]})\right)$. 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. ☐

6.2 Stable and Lossless Encoding by Pathwise Attention - Proof of Proposition 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 $y^{(1)}, \dots, y^{(N_{\mathrm{ref}})}$ in $C([0:T], \mathbb{R}^{d_Y})$, and quantitative estimates on how many paths must be generated.

Lemma 11 (Stable Lossless Feature Maps – Riemannian Case)**.** *Under Assumption 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 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 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)* $\delta \in \mathcal{O}(\operatorname{Vol}(\mathcal{M}, g)^{1/d_K})$
- *(ii)* $\|\theta\|_0 \in \mathcal{O}(N^2)$,
- *(iii)* $C_{d_K} \in \mathcal{O}({d_K}^{3(d_K+1)/2})$

The attention mechanism attn_T satisfies [78, Setting 3.6 (i)] for any Borel probability measure $\mathbb{P}_{\mathrm{in}}$ on K.

Proof. **of Lemma 11**

Step 1: Estimating N_{ref} 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 [59, Theorem 0.1.A] implies that the systole of $\mathcal{M}$ denoted by $\operatorname{sys}(\mathcal{M})$ satisfies

$$\operatorname{sys}(\mathcal{M}) \leq \tilde{C}_{d_K} \operatorname{Vol}(\mathcal{M}, g)^{1/\dim(\mathcal{M})}, \tag{20}$$

where $\operatorname{Vol}(\mathcal{M}, g)$ 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 [73, Theorem 1] (circa [73, Equation (1.1)]) we see that if $\delta = \operatorname{Sys}(\mathcal{M})/10$ 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,

$$\delta \leq C_{d_K} \operatorname{Vol}(\mathcal{M}, g)^{1/d_K}.$$

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 $\varphi_1 : (\mathcal{M}, d_g) \to (K, \|\cdot\|_T)$ by an any isometry, which we have postulated to exist. Set $\mathbb{X} \overset{\text{def.}}{=} \{\varphi(u) : u \in \widetilde{\mathbb{X}}\} = \{y^{(n)}\}_{n=1}^{N_{\mathrm{ref}}}$ where, for $n = 1, \dots, N_{\mathrm{ref}}$ we define $y^{(n)} \overset{\text{def.}}{=} \varphi_1(u_n)$. Define $\varphi : C([0:T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N_{\mathrm{ref}}}$ by $y_\cdot \mapsto (\|y_\cdot - y_\cdot^{(n)}\|_T)_{n=1}^{N_{\mathrm{ref}}}$. Observe that, for every path $y \in K$, the following holds

$$\begin{aligned}
\varphi(y_\cdot) \overset{\text{def.}}{=} \Big(\|y_\cdot - y_\cdot^{(n)}\|_T\Big)_{n=1}^{N_{\mathrm{ref}}} &= \Big(\big\|\varphi_1 \circ \varphi_1|_{\varphi_1(K)}^{-1}(y_\cdot) - \varphi_1 \circ \varphi_1|_{\varphi_1(K)}^{-1}(y_\cdot^{(n)})\big\|_T\Big)_{n=1}^{N_{\mathrm{ref}}} \\
&= \Big(d_g\big(\varphi_1|_{\varphi_1(K)}^{-1}(y_\cdot), \varphi_1|_{\varphi_1(K)}^{-1}(y_\cdot^{(n)})\big)\Big)_{n=1}^{N_{\mathrm{ref}}} \qquad (21) \\
&= \Big(d_g\big(\varphi_1|_{\varphi_1(K)}^{-1}(y_\cdot), u^n\big)\Big)_{n=1}^{N_{\mathrm{ref}}} = \varphi_2 \circ \varphi_1|_{\varphi_1(K)}^{-1}(y_\cdot), \qquad (22)
\end{aligned}$$

where (21) holds since φ_2 is an isometry and (22) 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 $\varphi|_K = \varphi_2 \circ \varphi_1|_{\varphi(K)}^{-1}$ 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 [25, 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 $\operatorname{Softmax} : u \mapsto (e^{u_j} / \sum_{k=1}^{N_{\text{ref}}} e^{u_k})_{j=1}^{N_{\text{sim}}}$, maps H surjectively and continuously onto the image set

$$\operatorname{Softmax}(\mathbb{R}^{N_{\text{sim}}}) \subseteq \operatorname{int}(\Delta_{N_{\text{sim}}}) \overset{\text{def.}}{=} \Big\{ v \in [0, 1]^{N_{\text{sim}}} : \sum_{i=1}^{N_{\text{sim}}} v_i = 1 \Big\};$$

i.e. of the interior of the N_{sim}-simplex. Since, $R : v \mapsto \big(\ln(v_i) - \ln(v_{N_{\text{sim}}}) + 1\big)_{i=1}^{N_{\text{sim}}}$ is a continuous right-inverse of Softmax on the interior of the image set $\operatorname{Softmax}(\mathbb{R}^{N_{\text{sim}}})$, the softmax function Softmax defines a continuous bijection from H onto $\operatorname{Softmax}(\mathbb{R}^{N_{\text{sim}}})$. Therefore, the map $\Psi \overset{\text{def.}}{=} \operatorname{Softmax} \circ \psi \circ \varphi$ 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 $\operatorname{int}(\Delta_{N_{\text{sim}}})$ of the N_{sim}-simplex, then, $C^\star \overset{\text{def.}}{=} \max_{u \in \Phi(K)} \ \min_{v \in \Delta_{N_{\text{sim}}} \setminus \operatorname{int}(\Delta_{N_{\text{sim}}})} \|u - v\| > 0$.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 R is locally Lipschitz on $\operatorname{int}(\Delta_{N_{\text{sim}}})$, 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 $\theta_0 = (A, B, a, b, \{y^{(n)}_\cdot\}_{n=1}^{N_{\text{ref}}})$ and observe that the map Ψ is of the form (5). We continue by choosing a map of the form (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 $y_\cdot \in K$. 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, $\|\theta\|_0 \in \mathcal{O}(N_{\text{pos}} \cdot d_Y + N_{\text{ref}})$. ☐

Lemma 12 (Stable Lossless Feature Maps – Finite Case)**.** *Under Assumption 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 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 [78, Setting 3.6 (i)].*

Proof. **of Lemma 12** We argue similarly to the proof of Lemma 11, with only Steps 1 and 2 of its proof being replaced with the following argument.

Steps 1-2 (Modified): Building the Feature Map.
Enumerate $K = \{x_n\}_{n=1}^{N_{\mathrm{ref}}}$. The map $\varphi_1 : (K, \|\cdot\|_T) \to (\mathbb{R}^{N_{\mathrm{ref}}}, \|\cdot\|_\infty)$ given for every $x \in K$ by $x \mapsto \big(\|x - x_n\|_t\big)_{n=1}^{N_{\mathrm{ref}}}$ is an isometric embedding, called the *Kuratowski embedding* (see [61, page 99]). The optimal constants for the equivalence of the Euclidean $\|\cdot\|_2$ and max $\|\cdot\|_\infty$ norms are given by

$$\|u\|_\infty \leq \|u\|_2 \quad \text{and} \quad \|u\|_2 \leq {N_{\mathrm{ref}}}^{1/2}\,\|u\|_\infty \tag{23}$$

for each $u \in \mathbb{R}^{N_{\mathrm{ref}}}$. Therefore, the "set theoretic identity map" $\varphi_2 : (\mathbb{R}^{N_{\mathrm{ref}}}, \|\cdot\|_\infty) \to (\mathbb{R}^{N_{\mathrm{ref}}}, \|\cdot\|_2)$ is bi-Lipschitz with optimal (shrinking and expansion) constants given by (23). Define $N_{\mathrm{sim}} \overset{\text{def.}}{=} \lceil 48\, \ln(N_{\mathrm{ref}}) \rceil$. By the Johnson-Lindenstrauss Lemma, with "small" constant given in the derivation of [47, Theorem 2.1][3], exists a linear map $\varphi_3 : (\mathbb{R}^{N_{\mathrm{ref}}}, \|\cdot\|_2) \to (\mathbb{R}^{N_{\mathrm{sim}}}, \|\cdot\|_2)$, i.e. $\varphi_3(u) = A_0\, u$ for some $N_{\mathrm{sim}} \times N_{\mathrm{ref}}$-matrix A_0, satisfying: For every $u, v \in \mathbb{R}^{N_{\mathrm{ref}}}$

$$2^{-1/2}\,\|u - v\|_2 \leq \|A_0\, u - A_0\, v\|_2 \leq (3/2)^{1/2}\,\|u - v\|_2.$$

Consequentially, the map $\varphi : (K, \|\cdot\|_T) \to (\mathbb{R}^{N_{\mathrm{sim}}}, \|\cdot\|_2)$ given by $\varphi \overset{\text{def.}}{=} \varphi_3 \circ \varphi_2 \circ \varphi_1$ is $(2^{-1/2}, (3N_{\mathrm{ref}}/2)^{1/2})$-bi-Lipschitz, since the Kuratowksi embedding φ_1 is an isometry and φ_2 satisfies (23); this is because for each $x, \tilde{x} \in K$ we have

$$\begin{aligned}
2^{-1/2}\,\|x - \tilde{x}\|_T =& 2^{-1/2}\,\|\varphi_1(x) - \varphi_1(\tilde{x})\|_\infty \\
\leq& 2^{-1/2}\,\|\varphi_2 \circ \varphi_1(x) - \varphi_2 \circ \varphi_1(\tilde{x})\|_2 \\
\leq& \|\varphi_3 \circ \varphi_2 \circ \varphi_1(x) - \varphi_3 \circ \varphi_2 \circ \varphi_1(\tilde{x})\|_2 \\
\overset{\text{def.}}{=}& \|\varphi(x) - \varphi(\tilde{x})\|_2 \\
\overset{\text{def.}}{=}& \|\varphi_3 \circ \varphi_2 \circ \varphi_1(x) - \varphi_3 \circ \varphi_2 \circ \varphi_1(\tilde{x})\|_2 \\
\leq& (3/2)^{1/2}\,\|\varphi_2 \circ \varphi_1(x) - \varphi_2 \circ \varphi_1(\tilde{x})\|_2 \\
\leq& (3/2)^{1/2}\,{N_{\mathrm{ref}}}^{1/2}\,\|\varphi_1(x) - \varphi_1(\tilde{x})\|_\infty = (3N_{\mathrm{ref}}/2)^{1/2}\,\|x - \tilde{x}\|_T.
\end{aligned}$$

Remaining Steps. The rest of the proof is identical to Steps 3 to 5[4] of the proof of Lemma 11 but with B defined instead as $B \overset{\text{def.}}{=} B_1 A_0$, where $B_1 = (I_{N_{\mathrm{sim}}}\ I_{N_{\mathrm{sim}}})^\top$.

This concludes the proof. ◻

Lemma 13 (Stable Lossless Feature Maps – Linear Case)**.** *Fix $N_{\mathrm{p}} \in \mathbb{N}_+$, $0 = t_0 < \cdots < t_{N_{\mathrm{p}}} = T$, and a constant $C_K > 0$. Let $K \subset C([0:T], \mathbb{R}^{d_Y})$ satisfy Assumption 2 (ii). Then, there is a parameter θ, as in Definition 3, defining a pathwise attention mechanism $\mathrm{attn}_T^\theta : C([0:T], \mathbb{R}^{d_Y}) \to \mathbb{R}^{N+1}$, with $N \overset{\text{def.}}{=} N_{\mathrm{p}} \cdot d_Y$ which restricts to a bi-Lipschitz embedding of $(K, \|\cdot\|_T)$ into $(\mathbb{R}^{N+1}, \|\cdot\|_2)$. Furthermore, $\|\theta\|_0 \in \mathcal{O}(N_{\mathrm{p}} \cdot d_Y)$.*

In particular, attn_T satisfies [78, Setting 3.6 (i)].

Proof. **of Lemma 13**
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 1. By fixing $\theta_0 = (A, B, a, b)$, we observe that: for all $y \in C([0:T], \mathbb{R}^{d_Y})$

$$\mathrm{sim}_T^{\theta_0}(y_\cdot) = 1 \in \mathbb{R}^{N_{\mathrm{p}}}. \tag{24}$$

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 2[5]. Set $\theta_1 = (U, V, \{t_n\}_{n=1}^{N_{\mathrm{p}}})$ and with that,

$$\mathrm{pos}_T^{\theta_1}(y_\cdot) = (y_1, \ldots, y_{N_{\mathrm{p}}}). \tag{25}$$

[3]We use this formulation since the constant in $4*48$ as opposed to $4*200$ in the standard derivation; found for example in [69].
[4]Here N_{sim} corresponds to $N_{\mathrm{sim}} - 1$ in Step 3.
[5]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.

In particular, the map $\mathrm{pos}_T^{\theta_1}$ in (25) 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 $N \stackrel{\text{def.}}{=} N_\mathrm{p} \cdot d_Y$ and let C be the $N \times N$-dimensional matrix with $C_{i,j} = 1$ for $i, j = 1, \dots, N$. Set $\theta = (\theta_0, \theta_1, C)$. By construction $\|\theta\|_0 \in \mathcal{O}(N_\mathrm{p} \cdot d_Y)$. Then,

$$\mathrm{attn}_T^{\theta}(t, y_\cdot) \stackrel{\text{def.}}{=} \Big(t, C \ \mathrm{vec}\left(\mathrm{sim}_T^{\theta_0}(y_\cdot) \odot \mathrm{post}_T^{\theta_1}(y_\cdot)\right)\Big) = \big(t, (y_1, \dots, y_{N_\mathrm{p}})\big). \tag{26}$$

Together, (24) and (25) imply that attn_T^{θ} defines a bi-Lipschitz embedding of K into the $N+1$-dimensional Euclidean space. □

Proof. **of Proposition 1**
Follows directly from Lemmata 11 to 13. □

6.3 Approximability of Locally Lipschitz Maps By FFs - Proof of Proposition 3

We now prove our main universal approximation theorem, Proposition 3. We show that the target space/-codomain of any considered FF is geometrically regular, in the sense of [1, 78]. Using this fact, we then combine and apply the results of [76, 78] to deduce the result.

6.3.1 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.[6] 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 $\mathrm{d}_{2,F}\left((m^{(1)}, A), (m^{(2)}, B)\right)^2 \stackrel{\text{def.}}{=} \|m^{(1)} - m^{(2)}\|^2 + \|A - B\|_F^2$. We show that $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathrm{d}_{2,F})$ satisfies the conditions of [78, Theorem 3.7]; namely [78, 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 $\mathcal{Q}_q((m, A)) \stackrel{\text{def.}}{=} (m, A^\top A)$. The family $\mathcal{Q} \stackrel{\text{def.}}{=} (\mathcal{Q}_q)_{q\in\mathbb{N}_+}$ 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 $(m^{(1)}, A^{(1)}), \dots, (m^{(N)}, A^{(N)}) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d_X^2}$ by $\eta\big(w, \{(m^{(n)}, A^{(n)}\}_{n=1}^N\big) \stackrel{\text{def.}}{=} \sum_{n=1}^N w_n \cdot (m^{(n)}, A^{(n)})$. 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 14. $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, \mathcal{Q}, \eta)$ *is a barycentric QAS space, in the sense of [1, Definition 3.4]. In particular, it satisfies [78, Setting 3.6 (iii)].*

Proof. **of Lemma 14**
We first show 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 [13, 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 [14, 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 $(\mathbb{R}^{d+d^2}, d_{2,F})$. Now, 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 $\mathcal{Q}_q((m, A)) \stackrel{\text{def.}}{=} (m, A^\top A)$.The map $\mathcal{Q}_q$ is a surjection since every symmetric matrix is the square of some $d \times d$ matrix [112]. Therefore, $\mathcal{Q}_\cdot = (\mathcal{Q}_q)_{q\in\mathbb{N}_+}$ trivially satisfies [1, Definition 3.2] with modulus of quantizability,

[6] The 0 emphasizes that there is no rank-restriction on the matrices in Sym_{0,d_X} unlike, for example, in [63, 96].

defined on [1, page 12], given by $\mathcal{Q}_K(\varepsilon) = D_1 \stackrel{\text{def.}}{=} d(d+1)$ for every compact subset K of $\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}$ and for every $\varepsilon > 0$. For any $N \in \mathbb{N}_+$, $w \in \Delta_N$, and $(m^{(1)}, A^{(1)}), \dots, (m^{(N)}, A^{(N)}) \in \mathbb{R}^{d_X} \times \mathbb{R}^{d\times d}$ we have

$$
\begin{aligned}
& d_{2,F}\Big(\eta\big(w, \{(m^{(n)}, A^{(n)}\}_{n=1}^N), y_i\Big) \\
&= \Big(\big\|\big(\sum_{n=1}^N w_n\, m^{(n)}\big) - m^{(i)}\big\|^2 + \big\|\big(\sum_{n=1}^N w_n\, A^{(n)}\big) - A^{(i)}\big\|_F^2\Big)^{1/2} \\
&= \Big(\big\|\sum_{n=1}^N w_n\, m^{(n)} - \sum_{n=1}^N w_n\, m^{(i)}\big\|^2 + \big\|\sum_{n=1}^N w_n\, A^{(n)} - \sum_{n=1}^N w_n\, A^{(i)}\big\|_F^2\Big)^{1/2} \qquad (27) \\
&\leq \Big(\sum_{n=1}^N w_n\Big)^2 \big\|m^{(n)} - m^{(i)}\big\|^2 + \Big(\sum_{n=1}^N w_n\Big)^2 \big\|A^{(n)}) - A^{(i)})\big\|_F^2\Big)^{1/2} \\
&\leq \sum_{n=1}^N w_n \Big(\big\|m^{(n)} - m^{(i)}\big\|^2 + \big\|A^{(n)}) - A^{(i)})\big\|_F^2\Big)^{1/2} \\
&= 1 \cdot \sum_{n=1}^N w_n\, d_{2,F}\big((m^{(n)}, A^{(n)}), (m^{(i)}, A^{(i)})\big)^1,
\end{aligned}
$$

for $i = 1, \dots, N$, where (27) 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 [21] 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 [48, 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, [78, 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 $(\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}, d_{2,F}, \eta, \mathcal{Q}.)$,

Lemma 15. *Fix an activation function σ satisfying Assumption 3. For every $K \subset C([0:T], \mathbb{R}^{d_X})$ satisfying Assumption 2, each $0 < \delta, \alpha \leq 1$, $0 \leq L$, and every (L,α)-Hölder[7] 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*

$$\max_{(t,y_\cdot)\in[0:T]\times K} d_{2,F}(\hat{g}(t,y_\cdot), f(t,y_\cdot)) < \delta \tag{28}$$

with representation

$$\hat{g}(t,y_\cdot) = \sum_{i=1}^{N'} P_{\Delta_{N'}}(\hat{f} \circ \mathrm{attn}^\theta(t,y_\cdot))_i \cdot (m^{(i)}, (A^{(i)})^\top A^{(i)}), \tag{29}$$

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 4.

[7] That is, f is α Hölder with optimal Hölder coefficient L.

Table 4: Complexity Estimates for transformer-type model $\hat{g}$ in Lemma 15.

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 15**
We work in the notation of [78, Theorem 3.7], or rather, its quantitative version [78, Lemma 5.10]. Our objective is to apply [78, Theorem 3.7] by verifying each of the conditions of [78, Setting 3.6].

Step 1: Implementing a Bi-Lipschitz Feature Map with the Attention Layer.

Remark. We first show that the conditions of [78, Setting 3.6 (i)] are met, by verifying that the parameters of the attention layer (7) can be chosen such that $\operatorname{attn}^\theta$ is a suitable feature map.

When convenient, let $\operatorname{attn}^\theta$ be as in either of Lemmata 11 to 13 depending on which assumption of Assumptions 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[8], as defined on [61, page 78]. Thus, [61, Theorem 12.1] implies that $(\mathcal{M},d_g)$ is a doubling metric space, as defined on [61, 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 [102, Lemma 9.6 (v)]). Thus, $([0:T]\times K,\|\cdot\|\times\|\cdot\|_T)$ is a doubling metric space. We have thus verified [78, Setting 3.6 (i)].

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, [78, Setting 3.6 (ii)] holds.

Step 3: Geometry of $(\mathbb{R}^{d_X}\times\operatorname{Sym}_{0,d_X},\mathcal{Q},\eta)$. Lemma 14 shows that $(\mathbb{R}^{d_X}\times\operatorname{Sym}_{0,d_X},\mathcal{Q},\eta)$ is barycentric and it is a QAS space with quantized mixing function, see [78, page 7], given for any $N'\in\mathbb{N}_+$, $u\in\mathbb{R}^{N'}$, and $(m^{(1)},A^{(1)}),\dots,(m^{(N)},A^{(N')})\in\mathbb{R}^{d_X}\times\mathbb{R}^{d_X^2}$ by

$$\hat{\eta}\big(w,\{(m^{(n)},A^{(n)}\}_{n=1}^{N'}\big)\overset{\text{def.}}{=}\sum_{n=1}^{N'}P_{\Delta_{N'}}(w_n)\cdot(m^{(n)},(A^{(n)})^\top A^{(n)}). \tag{30}$$

This verifies [78, Setting 3.6 (iii)].

Step 4: Determining The Euclidean Universal Approximator.

Remark. We now verify that the class of all MLPs with activation function σ satisfying Assumption 3, thus assumption [78, Setting 3.6 (iv)] holds. The case where $\sigma=\text{ReLU}$ and $\sigma\neq\text{ReLU}$ are treated separately.

First, consider the case where $\sigma\neq\text{ReLU}$. 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 (8) and satisfying

$$J\le c\text{ and }\max_{j=0,\dots,J}d_j\le N+N'+3. \tag{31}$$

[8] See [61, page 78].

Since σ was assumed to satisfy Assumption 3 then by [79, Theorem 9], as formulated in [79, Proposition 53], implies that $\mathcal{F}.$ is a universal approximator in the sense of [78, Definition 2.11]. Moreover, its rate function is recorded in [79, Proposition 53]. Therefore, $\mathcal{F}.$, as defined in (31), verifies [78, 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 (8) and satisfying

$$J \le c \text{ and } \max_{j=0,\dots,J} d_j \le c. \tag{32}$$

By [74, Theorem 1.1], $\mathcal{F}.$ is a universal approximator in the sense of [78, Definition 2.11]. Moreover, its rate function is given in [54, Theorem 1], as recorded in [54, Table 1]. In either case, $\mathcal{F}.$, as defined in (32), verifies [78, Setting 3.6 (iv)].

Step 5: Applying [78, Theorem 3.7]. Steps 1 though 5 verify that the conditions of [78, 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 [78, Definition 3.4], is a singleton. Therefore, by [78, 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[9]

$$\hat{F}(y.) = \beta_{\mathbb{R}^{d_X} \times \mathrm{Sym}_{0,d_X}} \big(\delta_{\hat{\eta}(P_{\Delta_{N'}} \circ \hat{f}_n \circ \varphi(y.), (Z_n)_{n=1}^N)}\big) \tag{33}$$

$$= \hat{\eta}(P_{\Delta_{N'}} \circ \hat{f}_n \circ \varphi(y.), (Z_n)_{n=1}^N) \tag{34}$$

$$= \sum_{n=1}^{N'} \big(P_{\Delta_{N'}} \circ \hat{f}_n \circ \varphi(y.)\big) \cdot (m^{(n)}, (A^{(n)})^\top A^{(n)}) \tag{35}$$

for each $y. \in K$ satisfying

$$\sup_{y. \in K} d_{2,F}\big(\hat{F}(y.), f(y.)\big) < \varepsilon_A + \varepsilon_Q + \varepsilon_E = \varepsilon,$$

where $\varepsilon_E = 0$ and $\varepsilon_A \stackrel{\text{def.}}{=} \varepsilon_Q \stackrel{\text{def.}}{=} \varepsilon/2$ 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 [78, 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 14). We observe that that (34) 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 (35) follows from the expression for $\hat{\eta}$ given in (30). Consequentially, (33) to (35) reduces to (29).

Step 6: Tallying Parameters. Since the quantitative version of [78, Theorem 3.7] held, namely [78, Lemma 5.10], then we obtain the following parameter estimates

(a) $N = \mathcal{N}_{\text{pack}}\big(K, C_{d_K} \ \mathrm{Vol}(\mathcal{M}, g)^{1/d_K}\big) \in \mathcal{O}(1)$,

(b) c is recorded in Table 4 as the depth of the network $\hat{f}$,

(c) The expression of N' is recorded, in detail, atop [78, page 46] and is

$$N' \le \big(C_{K,1}^{\lceil \frac{1}{4\alpha} \rceil}\big)^{\log_2(\mathrm{diam}(K)) - \frac{1}{\alpha} \log_2(\epsilon_A/(2LC_{K:2}))} \in \mathcal{O}_K\big(L/\varepsilon\big).$$

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. □

[9]Since the barycentric decomposition $\{(K, \varphi)\}$ of $(K, \|\cdot\|_T, \mu)$ has exactly one part then the partition of unity [78, Setting 3.7] is trivial and $\psi_1(x.) = 1$ for each $x. \in K$.

6.3.2 Proof of the Main Approximation Lemma

For any $d \in \mathbb{N}_+$, let $\overline{\mathcal{N}_d}$ denotes 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 [76] and [78], are used to deduce our main approximation theoretic tool, namely, Proposition 3.

Proof. **of Proposition 3**

Step 1: Bounded The Local Lipschitz Stability of ρ on $f(K)$. By Lemma 16, the map $\varrho : (\mathcal{N}_{d_X}, \mathcal{W}_2) \to (\mathbb{R}^{d_X} \times \operatorname{Sym}_{0,d_X}, d_{2,F})$ is continuous. By definition of the product topology, see [95, page 114], the projection map $\pi : (\mathbb{R}^{d_X}, \operatorname{Sym}_{0,d_X}, d_{2,F}) \ni (\mu, \Sigma) \to \Sigma \in (\operatorname{Sym}_{0,d_X}, \|\cdot\|_F)$ is continuous. Since the composition of continuous functions is again continuous, then the map

$$g : ([0,T] \times K, |\cdot| \times \|\cdot\|_T) \ni (t, y_\cdot) \mapsto \pi \circ \varrho \circ f(t, y_\cdot) \in (\mathbb{R}, |\cdot|)$$

is continuous. By [95, Theorem 26.5], $\tilde{K} \stackrel{\text{def.}}{=} g(K)$ is a compact subset of $(\mathbb{R}, |\cdot|)$.

By [86, Theorem 9.2.6 - page 130] the map $\lambda_{\min} : (\operatorname{Sym}_{0,d_X}, \|\cdot\|_F) \to (\mathbb{R}, |\cdot|)$ which sends any $d_X \times d_X$ square matrix Σ to its minimal eigenvalue $\lambda_{\min}(\Sigma)$ 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

$$\lambda_{\min}(\Sigma_0) = \min_{\Sigma \in g(K)} \lambda_{\min}(\Sigma) < \infty. \tag{36}$$

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, (36) implies that

$$0 < r \stackrel{\text{def.}}{=} \lambda_{\min}(\Sigma_0) = \min_{\Sigma \in g(K)} \lambda_{\min}(\Sigma) < \infty. \tag{37}$$

By Lemma 10, we have that there is an $r > 0$ such that $r = \lambda_{\min}(\Sigma_t)$ for all $0 \le t \le T$. Consequentially, Lemma 16 implies that the map $\varrho : (\mathcal{N}_{d_X}, \mathcal{W}_2) \to (\mathbb{R}^{d_X} \times \operatorname{Sym}_{0,d_X}, d_{2,F})$ is Lipschitz on $\tilde{K}$ with Lipschitz constant bounded-above by $\operatorname{Lip}\big(\varrho | f(K)\big)$. Consequentially, we have that

$$\operatorname{Lip}\big(\varrho \circ f | K\big) \le \operatorname{Lip}\big(\varrho | f(K)\big) \operatorname{Lip}\big(f | K\big) \le \max\big\{1, \frac{\sqrt{d}}{2\sqrt{r}}\big\} L. \tag{38}$$

Step 2: Approximating $\varrho \circ f$ on K. Fix $\varepsilon > 0$ and fix the "perturbed approximation error"

$$\delta \stackrel{\text{def.}}{=} \min\left\{1, \frac{\varepsilon}{\sqrt{d} 2 \sqrt{L \sqrt{T^2 + \operatorname{diam}(K)^2} + 1}}\right\}. \tag{39}$$

We apply Lemma 15 to deduce that there exists a map $\hat{F} : [0 : T] \times K \to \mathbb{R}^{d_X} \times \operatorname{Sym}_{0,d_X}$ with representation (29) satisfying the uniform estimate

$$\max_{(t,y_\cdot) \in [0:T] \times K} d_{2,F}(\hat{F}(t, y_\cdot), \varrho \circ f(t, y_\cdot)) < \delta. \tag{40}$$

Since Lemma 16 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

$$\bar{B}_1 \stackrel{\text{def.}}{=} \{(m, \Sigma) \in \mathbb{R}^{d_X} \times \operatorname{Sym}_{0,d_X} : (\exists (t, y_\cdot) \in [0:T] \times K)\, d_{2,F}(f(t, y_\cdot), (m, \Sigma)) \le 1\}.$$

Since δ was defined, in (39), to be at-most 1, then (39) implies that $\hat{F}(K) \subseteq \bar{B}_1$. From (40) and Lemma 16 we deduce the uniform estimate

$$\max_{(t,y_\cdot) \in [0:T] \times K} \mathcal{W}_2\big(\rho^{-1} \circ F(t, y_\cdot), f(t, y_\cdot)\big) \tag{41}$$

$$
\begin{aligned}
&= \max_{(t,y_\cdot)\in[0:T]\times K} \mathcal{W}_2\big(\varrho^{-1}\circ \hat{F}(t,y_\cdot), \varrho^{-1}\circ\varrho\circ f(t,y_\cdot)\big) \\
&\leq \mathrm{Lip}\left(\varrho^{-1}|\bar{B}_1\right) \max_{(t,y_\cdot)\in[0:T]\times K} d_{2,F}\left(\hat{F}(t,y_\cdot), \varrho\circ f(t,y_\cdot)\right) \\
&\leq \mathrm{Lip}\left(\varrho^{-1}|\bar{B}_1\right)\delta \\
&\leq \mathrm{Lip}\left(\varrho^{-1}|\bar{B}_1\right) \frac{\varepsilon}{\sqrt{d}2\sqrt{L\sqrt{T^2+\mathrm{diam}(K)^2}+1}} \qquad (42) \\
&\leq \varepsilon, \qquad (43)
\end{aligned}
$$

where (42) followed from Lemma 16 and the definition of $\bar{B}_1$ and bound (43) followed from the estimate (38) for $\mathrm{Lip}(\varrho^{-1}|\bar{B}_1)$. Since $\mathcal{W}_p \leq \mathcal{W}_2$ for all $1 \leq p \leq 2$, see [115, Remark 6.6] then, the estimates in (41)-(43) imply that

$$
\max_{(t,y_\cdot)\in[0:T]\times K} \mathcal{W}_p\big(\hat{F}(t,y), f(y)\big) \leq \varepsilon, \tag{44}
$$

for all $1 \leq p \leq 2$; as claimed.

Step 3: Counting Parameters. Using δ, as defined in (40), in the place of ε in Table 4 and noting that $\delta^{-1} \in \mathcal{O}(\varepsilon^{-1})$ yields the conclusion.

□

A Supplementary Material

Acknowledgments

AK acknowledges financial support from an NSERC Discovery Grant No. RGPIN-2023-04482 and their McMaster Startup Funds. AK and XY acknowledge that resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute https://vectorinstitute.ai/partnerships/current-partners/.

Proposition 4 (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 is a compact subset $K \subseteq C([0:T], \mathbb{R}^{d_Y})$ and an isometry from $\mathcal{X}$ onto K.*

A.1 Isometric Copies of Every Compact Riemannian Manifold in the Path Space – Proof of Proposition 4

Proof. **of Proposition 4**
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 $f \mapsto (f(\cdot/T), 0, \dots, 0)$, is an isometric embedding; then $\varphi \stackrel{\text{def.}}{=} \varphi_3 \circ \varphi_2 \circ \varphi_1$ is an isometric embedding of $\mathcal{X}$ into $C([0:T], \mathbb{R}^{d_Y})$. □

B An Auxiliary Lemma - I. Pinelis

This section records an auxiliary lemma due to Iosif Pinelis, [99]. We include the result and its proof here to keep our manuscript self-contained.

Lemma 16 ([99]). *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*[10] *then the following lower-bound holds*

$$\frac{1}{\min\{1, \sqrt{d}(2\sqrt{R})\}} \sqrt{\|m^{(1)} - m^{(2)}\|^2 + \|A - B\|_F^2} \leq \mathcal{W}_2\big(\mathcal{N}(m^{(1)}, A), \mathcal{N}(m^{(2)}, B)\big).$$

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 16**
By [55, 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

$$\|A\|_{\lambda:2} \overset{\text{def.}}{=} \sqrt{\max_{i=1,\dots,d} \lambda_{\max}(C^\top C)}$$

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. $\square$

[10] 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.

References

[1] Beatrice Acciaio, Anastasis Kratsios, and Gudmund Pammer. Designing universal causal deep learning models: The geometric (hyper)transformer. *Mathematical Finance*, n/a(n/a), 2023.

[2] Ben Adcock, Simone Brugiapaglia, Nick Dexter, and Sebastian Moraga. Near-optimal learning of banach-valued, high-dimensional functions via deep neural networks. *arXiv preprint arXiv: 2211.12633*, 2022.

[3] John Armstrong, Damiano Brigo, and Bernard Hanzon. Optimal projection filters with information geometry. *Information Geometry*, pages 1–16, 2023.

[4] Simon H Babbs and K Ben Nowman. Kalman filtering of generalized vasicek term structure models. *Journal of financial and quantitative analysis*, 34(1):115–130, 1999.

[5] Eviatar Bach, Tim Colonius, Isabel Scherl, and Andrew Stuart. Filtering dynamical systems using observations of statistics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(3), 2024.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2016.

[7] Alan Bain and Dan Crisan. *Fundamentals of stochastic filtering*, volume 60 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2009.

[8] Vlad Bally, Lucia Caramellino, Rama Cont, Frederic Utzet, and Josep Vives. *Stochastic integration by parts and functional Itô calculus*. Springer, 2016.

[9] Stefan Banach. *Théorie des opérations linéaires*. Chelsea Publishing Co., New York, 1955.

[10] Ole E. Barndorff-Nielsen and Neil Shephard. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(2):253–280, 2002.

[11] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *J. Mach. Learn. Res.*, 20:Paper No. 63, 17, 2019.

[12] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[13] Giuliano Basso. Fixed point theorems for metric spaces with a conical geodesic bicombing. *Ergodic Theory and Dynamical Systems*, 38(5):1642–1657, 2018.

[14] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, second edition, 2017. With a foreword by Hédy Attouch.

[15] Christian Beck, Sebastian Becker, Patrick Cheridito, Arnulf Jentzen, and Ariel Neufeld. An efficient Monte Carlo scheme for Zakai equations. *Commun. Nonlinear Sci. Numer. Simul.*, 126:Paper No. 107438, 2023.

[16] VE Beneš. Existence of optimal strategies based on specified information, for a class of stochastic decision problems. *SIAM Journal on control*, 8(2):179–188, 1970.

[17] Ramaprasad Bhar. *Stochastic filtering with applications in finance*. World Scientific, 2010.

[18] Tomas Björk, Mark H. A. Davis, and Camilla Landén. Optimal investment under partial information. *Math. Methods Oper. Res.*, 71(2):371–399, 2010.

[19] Damiano Brigo, Bernard Hanzon, and François Le Gland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, pages 495–534, 1999.

[20] Damiano Brigo, Bernard Hanzon, and Francois LeGland. A differential geometric approach to nonlinear filtering: the projection filter. *IEEE Transactions on Automatic Control*, 43(2):247–252, 1998.

[21] Bernard Bru, Henri Heinich, and Jean-Claude Lootgieter. Distances de Lévy et extensions des théoremès de la limite centrale et de Glivenko-Cantelli. *Publ. Inst. Statist. Univ. Paris*, 37(3-4):29–42, 1993.

[22] Umut Çetin. Financial equilibrium with asymmetric information and random horizon. *Finance and Stochastics*, 22:97–126, 2018.

[23] Ren-Raw Chen. Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. *Journal of Fixed Income*, 3:14–31, 1993.

[24] Patrick Cheridito and Balint Gersey. Computation of conditional expectations with guarantees. *J. Sci. Comput.*, 95(1):Paper No. 12, 30, 2023.

[25] Patrick Cheridito, Arnulf Jentzen, and Florian Rossmannek. Efficient approximation of high-dimensional functions with neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):3079–3093, 2021.

[26] J. M. C. Clark and D. Crisan. On a robust version of the integral representation formula of nonlinear filtering. *Probab. Theory Related Fields*, 133(1):43–56, 2005.

[27] JMC Clark. The design of robust approximations to the stochastic differential equations of nonlinear filtering. *Communication systems and random process theory*, 25:721–734, 1978.

[28] Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Optimal stable nonlinear approximation. *Found. Comput. Math.*, 22(3):607–648, 2022.

[29] Rama Cont and David-Antoine Fournié. Functional Itô calculus and stochastic integral representation of martingales. *Ann. Probab.*, 41(1):109–133, 2013.

[30] D. Crisan, J. Diehl, P. K. Friz, and H. Oberhauser. Robust filtering: correlated noise and multidimensional observation. *Ann. Appl. Probab.*, 23(5):2139–2160, 2013.

[31] D. Crisan, J. Diehl, P. K. Friz, and H. Oberhauser. Robust filtering: correlated noise and multidimensional observation. *Ann. Appl. Probab.*, 23(5):2139–2160, 2013.

[32] Christa Cuchiero, Martin Larsson, and Josef Teichmann. Deep neural networks, generic universal interpolation, and controlled ODEs. *SIAM J. Math. Data Sci.*, 2(3):901–919, 2020.

[33] Paresh Date and Ksenia Ponomareva. Linear and non-linear filtering in mathematical finance: a review. *IMA Journal of Management Mathematics*, 22(3):195–211, 2011.

[34] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear approximation and (deep) ReLU networks. *Constr. Approx.*, 55(1):127–172, 2022.

[35] M. H. A. Davis. On a multiplicative functional transformation arising in nonlinear filtering theory. *Z. Wahrsch. Verw. Gebiete*, 54(2):125–139, 1980.

[36] M. H. A. Davis and M. P. Spathopoulos. Pathwise nonlinear filtering for nondegenerate diffusions with noise correlation. *SIAM J. Control Optim.*, 25(2):260–278, 1987.

[37] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

[38] Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.

[39] Pierre Del Moral. Nonlinear filtering using random particles. *Theory of Probability & Its Applications*, 40(4):690–701, 1996.

[40] Pierre Del Moral. Nonlinear filtering: Interacting particle resolution. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 325(6):653–658, 1997.

[41] Pierre Del Moral. Measure-valued processes and interacting particle systems. application to nonlinear filtering problems. *The Annals of Applied Probability*, 8(2):438–495, 1998.

[42] Pierre Del Moral. *Mean field simulation for Monte Carlo integration*, volume 126 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2013.

[43] Pierre Del Moral and Jean Jacod. *Interacting particle filtering with discrete observations*, volume 8. Univ. Paul Sabatier, 1999.

[44] Petar M Djuric, Jayesh H Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F Bugallo, and Joaquin Miguez. Particle filtering. *IEEE signal processing magazine*, 20(5):19–38, 2003.

[45] D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.*, 12(3):450–455, 1982.

[46] Jin-Chuan Duan and Jean-Guy Simonato. Estimating and testing exponential-affine term structure models by kalman filter. *Review of quantitative finance and accounting*, 13:111–135, 1999.

[47] Devdatt P. Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge, 2009.

[48] Richard M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002.

[49] Bruno Dupire. Functional Itô calculus. *Quant. Finance*, 19(5):721–729, 2019.

[50] Robert J Elliott and Cody B Hyndman. Parameter estimation in commodity markets: A filtering approach. *Journal of Economic Dynamics and Control*, 31(7):2350–2373, 2007.

[51] David-Antoine Fournie. *Functional Ito calculus and applications*. ProQuest LLC, Ann Arbor, MI, 2010. Thesis (Ph.D.)–Columbia University.

[52] Rüdiger Frey and Wolfgang Runggaldier. Pricing credit derivatives under incomplete information: a nonlinear-filtering approach. *Finance and Stochastics*, 14(4):495–526, 2010.

[53] Rüdiger Frey and Thorsten Schmidt. Pricing and hedging of credit derivatives via the innovations approach to nonlinear filtering. *Finance Stoch.*, 16(1):105–133, 2012.

[54] Luca Galimberti, Giulia Livieri, and Anastasis Kratsios. Designing universal causal deep learning models: The case of infinite-dimensional dynamical systems from stochastic analysis. *arXiv preprint arXiv:2210.13300*, 2022.

[55] Clark R. Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Math. J.*, 31(2):231–240, 1984.

[56] Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Approximation bounds for random neural networks and reservoir systems. *The Annals of Applied Probability*, 33(1):28–69, 2023.

[57] Lukas Gonon and Josef Teichmann. Linearized filtering of affine processes using stochastic riccati equations. *Stochastic Processes and their Applications*, 130(1):394–430, 2020.

[58] Lukas Gonon and Josef Teichmann. Linearized filtering of affine processes using stochastic Riccati equations. *Stochastic Process. Appl.*, 130(1):394–430, 2020.

[59] Mikhael Gromov. Filling riemannian manifolds. *Journal of Differential Geometry*, 18(1):1–147, 1983.

[60] Andrew C Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge university press, 1990.

[61] Juha Heinonen. *Lectures on analysis on metric spaces*. Universitext. Springer-Verlag, New York, 2001.

[62] Jakob Heiss, Florian Krach, Thorsten Schmidt, and Félix B Tambe-Ndonfack. Nonparametric filtering, estimation and classification using neural jump odes. *arXiv preprint arXiv:2412.03271*, 2024.

[63] Calypso Herrera, Florian Krach, Anastasis Kratsios, Pierre Ruyssen, and Josef Teichmann. Denise: Deep robust principal component analysis for positive semidefinite matrices. *Transactions on Machine Learning Research*, 2023.

[64] 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.

[65] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

[66] Songyan Hou, Parnian Kassraie, Anastasis Kratsios, Jonas Rothfuss, and Andreas Krause. Instance-dependent generalization bounds via optimal transport. *arXiv preprint arXiv:2211.01258*, 2022.

[67] Alireza Javaheri, Delphine Lautier, and Alain Galli. Filtering in finance. *Wilmott*, 3:67–83, 2003.

[68] Alireza Javaheri, Delphine Lautier, and Alain Galli. Filtering in finance. *Wilmott*, 3:67–83, 2003.

[69] 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)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

[70] Olav Kallenberg. *Foundations of modern probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer, Cham, 2021.

[71] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[72] R. E. Kalman and R. S. Bucy. New Results in Linear Filtering and Prediction Theory. *Journal of Basic Engineering*, 83(1):95–108, 03 1961.

[73] Karin Usadi Katz and Mikhail G Katz. Bi-lipschitz approximation by finite-dimensional imbeddings. *Geometriae Dedicata*, 150(1):131–136, 2011.

[74] Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2306–2327. PMLR, 09–12 Jul 2020.

[75] Florian Krach, Marc Nübel, and Josef Teichmann. Optimal estimation of generic dynamics by path-dependent neural jump odes. *arXiv preprint arXiv:2206.14284*, 2022.

[76] Anastasis Kratsios. Universal regular conditional distributions via probabilistic transformers. *Constructive Approximation*, 57(3):1145–1212, 2023.

[77] Anastasis Kratsios, Valentin Debarnot, and Ivan Dokmanić. Small transformers compute universal metric embeddings. *J. Mach. Learn. Res.*, 24:Paper No. [170], 48, 2023.

[78] 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. *arXiv e-prints*, April 2023.

[79] Anastasis Kratsios and Leonie Papon. Universal approximation theorems for differentiable geometric deep learning. *Journal of Machine Learning Research*, 23(196):1–73, 2022.

[80] Rahul Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[81] Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv: 1511.05121*, 2015.

[82] Harold J. Kushner. A robust discrete state approximation to the optimal nonlinear filter for a diffusion. *Stochastics*, 3(2):75–83, 1979.

[83] Peter Lakner. Optimal trading strategy for an investor: the case of partial information. *Stochastic Processes and their Applications*, 76(1):77–97, 1998.

[84] Samuel Lanthaler and Andrew M Stuart. The curse of dimensionality in operator learning. *arXiv preprint arXiv:2306.15924*, 2023.

[85] Delphine Lautier and Alain Galli. Simple and extended kalman filters: an application to term structures of commodity prices. *Applied Financial Economics*, 14(13):963–973, 2004.

[86] Peter D. Lax. *Linear algebra and its applications*. Pure and Applied Mathematics (Hoboken). Wiley-Interscience, Hoboken, NJ, second edition, 2007.

[87] Gabriele Lillacci and Mustafa Khammash. Parameter estimation and model selection in computational biology. *PLoS computational biology*, 6(3):e1000696, 2010.

[88] Erik Lindström, Jonas Ströjby, Mats Brodén, Magnus Wiktorsson, and Jan Holst. Sequential calibration of options. *Computational Statistics & Data Analysis*, 52(6):2877–2891, 2008.

[89] Robert S. Liptser and Albert N. Shiryaev. *Statistics of random processes. II*, volume 6 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, expanded edition, 2001. Applications, Translated from the 1974 Russian original by A. B. Aries, Stochastic Modelling and Applied Probability.

[90] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM J. Math. Anal.*, 53(5):5465–5506, 2021.

[91] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

[92] Mihaela Manoliu and Stathis Tompaidis. Energy futures prices: term structure models with kalman filter estimation. *Applied mathematical finance*, 9(1):21–43, 2002.

[93] Carlo Marcati, Joost AA Opschoor, Philipp C Petersen, and Christoph Schwab. Exponential relu neural network approximation rates for point and edge singularities. *Foundations of Computational Mathematics*, 23(3):1043–1127, 2023.

[94] Hrushikesh N Mhaskar and Tomaso Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

[95] James R. Munkres. *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, 2000. Second edition of [MR0464128].

[96] A Martina Neuman, Yuying Xie, and Qiang Sun. Restricted riemannian geometry for positive semidefinite matrices. *Linear Algebra and its Applications*, 665:153–195, 2023.

[97] Sejun Park, Jaeho Lee, Chulhee Yun, and Jinwoo Shin. Provable memorization via deep neural networks using sub-linear parameters. In *Conference on Learning Theory*, pages 3627–3661. PMLR, 2021.

[98] Guergana Petrova and Przemysł aw Wojtaszczyk. Lipschitz widths. *Constr. Approx.*, 57(2):759–805, 2023.

[99] Iosif Pinelis. Local lipschitzness of parameterization of gaussians in wasserstein space. MathOverflow, 2020. URL:https://mathoverflow.net/q/449287 (version: 2023-06-22).

[100] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[101] Bruno Remillard. *Statistical methods for financial engineering*. CRC press, 2013.

[102] James C. Robinson. *Dimensions, embeddings, and attractors*, volume 186 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2011.

[103] Tom Ryder, Andrew Golightly, A. Stephen McGough, and Dennis Prangle. Black-box variational inference for stochastic differential equations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4423–4432. PMLR, 10–15 Jul 2018.

[104] Reza Sameni, MB Shamsollahi, and Christian Jutten. Filtering electrocardiogram signals using the extended kalman filter. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 5639–5642. IEEE, 2006.

[105] Eduardo Schwartz and James E Smith. Short-term variations and long-term dynamics in commodity prices. *Management Science*, 46(7):893–911, 2000.

[106] Eduardo S Schwartz. The stochastic behavior of commodity prices: Implications for valuation and hedging. *The Journal of finance*, 52(3):923–973, 1997.

[107] Jonathan W Siegel and Jinchao Xu. Characterization of the variation spaces corresponding to shallow neural networks. *Constructive Approximation*, pages 1–24, 2023.

[108] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020.

[109] Edwin H Spanier. *Algebraic topology*. Springer Science & Business Media, 1989.

[110] Ruslan Leontyevich Stratonovich. Optimum nonlinear systems which bring about a separation of a signal with constant parameters from noise. *Radiofizika*, 2(6):892–901, 1959.

[111] Ruslan Leontyevich Stratonovich. Application of the markov processes theory to optimal filtering. *Radio Engineering and Electronic Physics*, 5:1–19, 1960.

[112] TBD. Tbd. *TBD*, X.

[113] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[114] Roman Vershynin. Memory capacity of neural networks with threshold and rectified linear unit activations. *SIAM Journal on Mathematics of Data Science*, 2(4):1004–1033, 2020.

[115] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[116] A. N. Širjaev. Some explicit formulae in a problem on "disorder". *Teor. Verojatnost. i Primenen.*, 10:380–385, 1965.

[117] A. N. Širjaev. *Optimal stopping rules*, volume Vol. 8 of *Applications of Mathematics*. Springer-Verlag, New York-Heidelberg, 1978. Translated from the Russian by A. B. Aries.

[118] Curt Wells. *The Kalman filter in finance*, volume 32. Springer Science & Business Media, 1995.

[119] Curt Wells. *The Kalman filter in finance*, volume 32. Springer Science & Business Media, 2013.

[120] Magnus Wiktorsson and Erik Lindström. Fast simultaneous calibration and quadratic hedging under parameter uncertainty. In *8th World Congress of the Bachelier Finance Society*, 2014.

[121] W Murray Wonham. Some applications of stochastic differential equations to optimal nonlinear filtering. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 2(3):347–369, 1964.

[122] Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018.

[123] Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

[124] Moshe Zakai. On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11(3):230–243, 1969.

[125] Fengshan Zhang, Yongkui Zou, Shimin Chai, and Yanzhao Cao. Numerical analysis of a time discretized method for nonlinear filtering problem with lévy process observations. *Advances in Computational Mathematics*, 50(4):73, 2024.