NEURAL JUMP ORDINARY DIFFERENTIAL EQUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Combinations of neural ODEs with recurrent neural networks (RNN), like GRU-ODE-Bayes or ODE-RNN are well suited to model irregularly-sampled time series. While those models outperform existing discrete-time approaches, no theoretical guarantees for their predictive capabilities are available. Assuming that the irregularly-sampled time series data originates from a continuous stochastic processes, the optimal on-line prediction is the conditional expectation given the currently available information. We introduce the *Neural Jump ODE* (NJ-ODE) that provides a data-driven approach to learn, continuously in time, the conditional expectation of a stochastic process. Our approach models the conditional expectation between two observations with a neural ODE and jumps whenever a new observation is made. We define a novel training framework, which allows us to prove theoretical convergence guarantees for the first time. In particular, we demonstrate the predictive capabilities of our model by proving that, under some regularity assumptions, the output process converges to the conditional expectation process. We provide experiments showing that the theoretical results also hold empirically. Moreover, we experimentally show that our model outperforms one state of the art model in more complex learning tasks and give comparisons on a real-world dataset.

1 INTRODUCTION

Stochastic processes are widely used in many fields to model time series that exhibit a random behaviour. Often such a process is the solution of a stochastic differential equation (SDE) of the form

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t \,,$$

with certain assumptions on the drift μ and the diffusion σ. The best prediction of a future value of the process is provided by the conditional expectation given the current value. If the drift and diffusion are known or a good estimation is available, the conditional expectation can be approximated by a Monte Carlo (MC) simulation. However, since μ and σ are usually unknown, this approach strongly depends on the assumptions made on their parametric form. A more flexible approach is given by *neural SDEs*, where the drift μ and diffusion σ are modelled by neural networks (Tzen & Raginsky, 2019; Li et al., 2020; Jia & Benson, 2019). Nevertheless, modelling the diffusion can be avoided if one is only interested in estimating conditional expectations.

An alternative approach is to use Recurrent Neural Networks (RNN), where a neural network dynamically updates a latent variable with the observations of a discrete input time-series. RNNs are successfully applied to tasks for which time-series are regularly sampled, as for example speech or text recognition. However, input as well as the output time-series of RNNs are discrete, while stochastic processes are continuous in time. Moreover, often observations are irregularly sampled in time. The solution of dividing the time-line into equally-sized intervals and imputing or aggregating quotations is again making assumptions and information is lost. Frameworks that overcome this issue are the GRU-ODE-Bayes (Brouwer et al., 2019) and the ODE-RNN (Rubanova et al., 2019), which combine a RNN with a *neural ODE* (Chen et al., 2018). In standard RNNs, the hidden state is updated at each observation and constant in between. Conversely, in the GRU-ODE-Bayes and ODE-RNN framework, a neural ODE is trained to model the continuous evolution of the hidden state of the RNN between two observations. While GRU-ODE-Bayes and ODE-RNN both provide convincing empirical results, they lack a thorough theoretical justification.

Contribution. In this paper, we introduce a mathematical framework to precisely describe the problem statement of on-line prediction of a stochastic process with temporal irregularly sampled observations. Based on this rigorous mathematical description, we introduce the *Neural Jump ODE* (NJ-ODE), together with a novel training framework, which allows us to prove convergence guarantees for the first time. Moreover, we demonstrate empirically the capabilities of our model.

Outline of the paper:

- We suppose that we have data coming from a continuous stochastic process X described by a general SDE, only with weak regularity assumptions on the drift μ and diffusion σ, i.e. without imposing any particular parametrization (Section 2.1).
- We want to treat irregularly sampled time-series, i.e. without a fixed number of observations and not necessarily equally distributed. Therefore, we suppose that we have a possibly random number of n observations and that those observations are made at random times $t_1, \dots, t_n$ (Section 2.2).
- We describe the conditional expectation process $\hat{X}$ and show that it is the optimal prediction of X given the available information. Since the information on which we condition is constant in time except for "jumps" whenever a new observation is made, $\hat{X}$ is a jump process (Section 2.4).
- We introduce NJ-ODE, which models the conditional expectation between two observations with a neural ODE and jumps whenever a new observation is made (Section 3.1). Moreover, we defined a novel loss function, such that the NJ-ODE output approximates $\hat{X}$, when trained with it (Section 3.2).
- In particular, we prove that the NJ-ODE output converges to $\hat{X}$ if the number of training samples and the neural network size go to infinity (Section 4).
- Finally, we confirm these theoretical results with an empirical convergence study on a synthetic datasets, show that our model heavily outperforms others on more complex synthetic datasets and provide comparisons on real world data (Section 5).

2 Setup

2.1 Stochastic process X

Let $d_X, d_W \in \mathbb{N}$ and $T > 0$ be the fixed time horizon. Consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F} := \{\mathcal{F}_t\}_{0 \le t \le T}, \mathbb{P})$, on which an adapted d_W-dimensional Brownian motion $\{W_t\}_{t \in [0,T]}$ is defined. We define the stochastic process[1] $X := (X_t)_{t \in [0,T]}$ as the solution of the stochastic differential equation (SDE)

$$dX_t = \mu(t, X_t)dt + \sigma(t, X_t)dW_t \,, \tag{1}$$

for all $0 \le t \le T$, where $X_0 = x \in \mathbb{R}^{d_X}$ is the starting point and the measurable functions $\mu : [0,T] \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}$ and $\sigma : [0,T] \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_X \times d_W}$ are the *drift* and the *diffusion* respectively. We assume that X is continuous and square integrable, μ and σ are both globally Lipschitz continuous and their growth is at most linear in the second component, μ is bounded and continuous in its first component (t) uniformly in its second component (x) and that σ is càdlàg in the first component and satisfies $\sigma \in L^2(W)$. The exact definitions are recalled in Appendix A.

2.2 Random observation dates

Here we describe the observation dates. We consider a second probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, on which the observation times of the stochastic process X are defined. More precisely, we assume that:

- $n : \tilde{\Omega} \to \mathbb{N}_{>0}$ is a random variable with $\mathbb{E}_{\tilde{\mathbb{P}}}[n] < \infty$, the amount of observations.
- $t_i : \tilde{\Omega} \to [0,T]$ for $0 \le i \le n$ are *sorted* [2] random variables, the observation times.

[1] A stochastic process is a collection of random variables $X_t : \Omega \to \mathbb{R}^{d_X}, \omega \mapsto X_t(\omega)$ for $0 \le t \le T$.
[2] For all $\tilde{\omega} \in \tilde{\Omega}$, $0 = t_0 < t_1(\tilde{\omega}) < \dots < t_{n(\tilde{\omega})}(\tilde{\omega}) \le T$.

We denote the joint pushforward measure of n and $\{t_i\}_{0 \leq i \leq n}$ as $\tilde{\mathbb{P}}_t := (n, t_0, \ldots, t_n)\#\tilde{\mathbb{P}}$. The random variable n can but does not have to be unbounded. If it is bounded, we define $K := \max\left\{k \in \mathbb{N} \,|\, \tilde{\mathbb{P}}(n \geq k) > 0\right\}$ to be the maximal value of n, which otherwise is infinity. Then we can define for each $1 \leq k \leq K$ a measure on the time interval

$$\lambda_k : \mathcal{B}([0,T]) \to [0,1], \quad B \mapsto \lambda_k(B) := \frac{\tilde{\mathbb{P}}(n \geq k, (t_k -) \in B)}{\tilde{\mathbb{P}}(n \geq k)},$$ [3]

for which it is shown in Lemma B.2 that it defines a probability measure. Moreover, we define τ as the time of the last observation before a certain time t,

$$\tau : [0,T] \times \tilde{\Omega} \to [0,T], \quad (t, \tilde{\omega}) \mapsto \tau(t, \tilde{\omega}) := \max\{t_i(\tilde{\omega}) | 0 \leq i \leq n(\tilde{\omega}), t_i(\tilde{\omega}) \leq t\}.$$

2.3 Information σ-algebra

In the following, we leave away $\tilde{\omega} \in \tilde{\Omega}$ whenever the meaning is clear. We define the filtration of the currently available information $\mathbb{A} := (\mathcal{A}_t)_{t \in [0,T]}$ by

$$\mathcal{A}_t := \boldsymbol{\sigma}(X_{t_i} | t_i \leq t),$$

where t_i are the observation times and $\boldsymbol{\sigma}(\cdot)$ denotes the generated σ-algebra. By the definition of τ we have $\mathcal{A}_t = \mathcal{A}_{\tau(t)}$ for all $t \in [0,T]$. $(\Omega \times \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$ is the filtered product probability space which, intuitively speaking, combines the randomness of the stochastic process with the randomness of the observations. Here, $\mathbb{F} \otimes \tilde{\mathcal{F}}$ consists of the tensor-product σ-algebras $(\mathcal{F} \otimes \tilde{\mathcal{F}})_t := \mathcal{F}_t \otimes \tilde{\mathcal{F}}$ for $t \in [0,T]$. As explained in Remark B.4, $\mathcal{A}_t$ can be identified with a sub-σ-algebra of $(\mathcal{F} \otimes \tilde{\mathcal{F}})_t$.

2.4 Optimal approximation $\hat{X}$ of the stochastic process X

We are interested in the "best" approximation (or prediction) $\hat{X}_t$ of the process X that one can make at any time $t \in [0,T]$, given the currently available information $\mathcal{A}_t$. For us "best" refers to the $L^2(\Omega \times \tilde{\Omega}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$-minimizer, therefore, this approximation is given by the conditional expectation. Indeed, if we define $\Delta := \{(t,r) \in [0,T]^2 | t + r \leq T\}$, and the function

$$\tilde{\mu} : \Delta \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}, \quad (t, r, \xi) \mapsto \mathbb{E}\left[\mu(t+r, X_{t+r}) | X_t = \xi\right],$$

this is proven in the following proposition.

Proposition 2.1. *The optimal $\mathbb{A}$-adapted process in $L^2(\Omega \times \tilde{\Omega}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$ approximating $(X_t)_{t \in [0,T]}$ is given by $\hat{X} := (\hat{X}_t)_{t \in [0,T]}$ with $\hat{X}_t := \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[X_t | \mathcal{A}_t]$. Moreover, this process is unique up to $(\mathbb{P} \times \tilde{\mathbb{P}})$-null-sets. In addition we have ($\tilde{\omega}$-wise for $\tilde{\omega} \in \tilde{\Omega}$) that*

$$\hat{X}_t = X_{\tau(t)} + \int_{\tau(t)}^{t} \tilde{\mu}\left(\tau(t), s - \tau(t), X_{\tau(t)}\right) ds, \tag{2}$$

Proposition 2.2. *The function $\tilde{\mu}$ is (jointly) continuous.*

A thorough derivation of Proposition 2.1 with an in-depth discussion is provided in Appendix C. The proof of Proposition 2.2 is given in Appendix D.

3 Neural Jump ODE

We propose a model framework which can be interpreted as a simplification of the ODE-RNN (Rubanova et al., 2019) and the GRU-ODE-Bayes (Brouwer et al., 2019) model. A thorough recall of ODE-RNN and an explanation of our way of combining the continuous neural ODE part and the jumps in just one equation is given in Appendix E.

[3] Here, $(t_k -)$ means the left-point of t_k, for example, if $t_0 = 0, t_1 = 1$ we have $t_0 \in [0,1), t_1 \notin [0,1)$ but $t_0 - \notin [0,1), t_1 - \in [0,1)$.

3.1 THE MODEL FRAMEWORK: NEURAL ODE BETWEEN JUMPS

We define $\mathcal{X} \subset \mathbb{R}^{d_X}$ and $\mathcal{H} \subset \mathbb{R}^{d_H}$ to be the observation and latent space for $d_X, d_H \in \mathbb{N}$. Moreover, we define the following feed-forward neural networks with sigmoid activation functions:

- $f_{\theta_1} : [-1,1]^{d_H} \times [-1,1]^{d_X} \times [0,T] \times [0,T] \to \mathbb{R}^{d_H}$ modelling the ODE dynamics,
- $\rho_{\theta_2} : [-1,1]^{d_X} \to \mathbb{R}^{d_H}$ modelling the jumps when new observations are made, and
- $g_{\theta_3} : \mathbb{R}^{d_H} \to \mathbb{R}^{d_Y}$ the readout map, mapping into the target space $\mathcal{Y} \subset \mathbb{R}^{d_Y}$ for $d_Y \in \mathbb{N}$.

The trainable parameters of the neural networks are $\theta := (\theta_1, \theta_2, \theta_3) \in \Theta := \cup_{M \geq 1} \Theta_M$. Here, for every $M \in \mathbb{N}$, Θ_M is defined to be the set of all parameters such that f_{θ_1} and ρ_{θ_2} have hidden dimension M. We define the pure jump stochastic process (cf. Appendix E)

$$u : \tilde{\Omega} \times [0,T] \to \mathbb{R}, (\tilde{\omega}, t) \mapsto u_t(\tilde{\omega}) := \sum_{i=1}^{n(\tilde{\omega})} \mathbf{1}_{[t_i(\tilde{\omega}), \infty)}(t). \tag{3}$$

Then the *neural jump ODE* (NJ-ODE) is defined by the latent process $H := (H_t)_{t \in [0,T]}$ and the output process $Y := (Y_t)_{t \in [0,T]}$ defined as solutions of the SDE system (cf. Appendix E)

$$\begin{aligned} H_0 &= \rho_{\theta_2}\left(\tanh(X_0)\right), \\ dH_t &= f_{\theta_1}\left(\tanh(H_{t-}), \tanh(X_{\tau(t)}), \tau(t), t - \tau(t)\right) dt + \left(\rho_{\theta_2}\left(\tanh(X_t)\right) - H_{t-}\right) du_t, \\ Y_t &= g_{\theta_3}(H_t). \end{aligned} \tag{4}$$

We apply the hyperbolic tangent function to each component of every input variable except for the time variables, before passing them to the network.[4] This ensures that every input to the neural networks is bounded in the compact sets where f_{θ_1} and ρ_{θ_2} are defined. Compactness of the domain is an assumption needed to apply the universal approximation theorem by Hornik et al. (1989).

Note that, only the values of X at the times $t_i, 0 \leq i \leq n$ are used as inputs. We will use the notation $H_t^\theta(X)$ and $Y_t^\theta(X)$ to emphasize the dependence of the latent process H and the output process Y on the model parameters θ and the input X.

Existence and uniqueness. We note that u is an adapted process starting at 0 with finite variation on the product probability space $(\Omega \times \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$, since the variation of u up to time T is n and $\mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[n] < \infty$. Hence, since f_{θ_1} and ρ_{θ_2} are Lipschitz continuous (as composition of Lipschitz continuous functions) a unique càdlàg solution H^θ exists, once an initial value is fixed (Protter, 2005, Thm. 7, Chap. V). Moreover, the resulting process $(Y_t^\theta)_{t \in [0,T]}$ is also càdlàg and $\mathbb{A}$-adapted.

3.2 OBJECTIVE FUNCTION

We introduce a loss function, such that the output Y^θ of the NJ-ODE can be trained to approximate $\hat{X}$, i.e. to model the best *on-line* prediction of the stochastic process X. We define the theoretical loss function and its Monte Carlo approximation. Moreover, we provide its Ergodic approximation in Appendix F.

Objective function. Let us define $\mathbb{D}$ to be the set of all $\mathbb{R}^{d_X}$-valued $\mathbb{A}$-adapted processes on the probability space $(\Omega \times \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$. Then we define our objective functions

$$\Psi : \mathbb{D} \to \mathbb{R}, \qquad Z \mapsto \Psi(Z) := \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(|X_{t_i} - Z_{t_i}|_2 + |Z_{t_i} - Z_{t_i -}|_2 \right)^2 \right], \tag{5}$$

$$\Phi : \Theta \to \mathbb{R}, \qquad \theta \mapsto \Phi(\theta) := \Psi(Y^\theta(X)), \tag{6}$$

where Φ will be our (theoretical) loss function. Remark that from the definition of Y^θ, it directly follows that it is an element of $\mathbb{D}$, hence Φ is well-defined.

Monte Carlo approximation of the objective function. Let us assume, that we observe $N \in \mathbb{N}$ independent realisations of the path X at times $(\tilde{t}_1^j, \cdots, \tilde{t}_{n(j)}^j)$, $1 \leq j \leq N$, which are themselves

[4] By abuse of notation we will write $\tanh(x)$ when applying the $\tanh$ element wise to a vector x.

independent realisations of the random vector $(n, t_1, \cdots, t_n)$. In particular, let us assume that $X^j \sim X$ and $(n^j, t_1^j, \cdots, t_{n^j}^j) \sim \mathbb{P}_t$ are i.i.d. random processes (respectively variables) for $1 \leq j \leq N$ and that our training data is one realisation of them. Then

$$\hat{\Phi}_N(\theta) := \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n^j} \sum_{i=1}^{n^j} \left(\left| X_{t_i^j}^j - Y_{t_i^j}^\theta(X^j) \right|_2 + \left| Y_{t_i^j}^\theta(X^j) - Y_{t_i^j -}^\theta(X^j) \right|_2 \right)^2, \tag{7}$$

converges $(\mathbb{P} \times \tilde{\mathbb{P}})$-a.s. to $\Phi(\theta)$ as $N \to \infty$, by the law of large numbers (cf. Theorem 4.2). Therefore, (7) is a good approximation of Φ (6) that can be used in practice. However, our theoretical result makes use of the limiting loss function Φ, i.e. we assume to be in the case of having infinitely many independent training samples.

4 Theoretical convergence results

Our main results show, that the output of NJ-ODE converges to the conditional expectation when the size of the neural networks and the number of samples go to infinity. For each network size (and number of samples) we assume to have the weights minimizing the (Monte Carlo approximation of the) loss function. In particular, we do not consider the problem of finding those optimal weights, and therefore also do not analyse backpropagation through our model. In a similar setting, backpropagation was studied by Jia & Benson (2019). The proof of the following theorems, further remarks and discussions are give in Appendix G and H.

4.1 Convergence with respect to theoretical objective function

Theorem 4.1. *Let $\theta_M^{\min} \in \Theta_M^{\min} := \operatorname{argmin}_{\theta \in \Theta_M} \{\Phi(\theta)\}$ for every $M \in \mathbb{N}$. Then, for $M \to \infty$, the value of the loss function Φ (6) converges to the minimal value of Ψ (5) which is uniquely achieved by $\hat{X}$, i.e.*

$$\Phi(\theta_M^{\min}) \xrightarrow{M \to \infty} \min_{Z \in \mathbb{D}} \Psi(Z) = \Psi(\hat{X}).$$

Furthermore, for every $1 \leq k \leq K$ we have that $Y^{\theta_M^{\min}}$ converges to $\hat{X}$ as random variable in $L^1(\Omega \times [0,T], \mathbb{P} \times \lambda_k)$. In particular, the limit process $Y := \lim_{M \to \infty} Y^{\theta_M^{\min}}$ equals $\hat{X}$ $(\mathbb{P} \times \lambda_k)$-almost surely as a random variable on $\Omega \times [0,T]$.

4.2 Convergence of the Monte Carlo approximation

In the following, we assume the size of the neural network M is fixed and we study the convergence with respect to the number of samples N. Moreover, we show that both types of convergence can be combined. To do so, we define $\tilde{\Theta}_M := \{\theta \in \Theta_M \mid |\theta|_2 \leq M\}$, which is a compact subspace of Θ_M. It is straight forward to see, that Θ_M in Theorem 4.1 can be replaced by $\tilde{\Theta}_M$. Indeed, if the needed neural network weights for an ε-approximation have too large norm, then one can increase M until it is sufficiently big. Furthermore, we make the additional assumption that $\mathbb{E}_{\tilde{\mathbb{P}}}[n^2] < \infty$. The following convergence analysis is based on (Lapeyre & Lelong, 2019, Chapter 4.3).

Theorem 4.2. *Let $\theta_{M,N}^{\min} \in \Theta_{M,N}^{\min} := \arg\inf_{\theta \in \tilde{\Theta}_M} \{\hat{\Phi}_N(\theta)\}$ for every $M, N \in \mathbb{N}$. Then, for every $M \in \mathbb{N}$, $(\mathbb{P} \times \tilde{\mathbb{P}})$-a.s.*

$$\hat{\Phi}_N \xrightarrow{N \to \infty} \Phi \quad \textit{uniformly on } \tilde{\Theta}_M.$$

Moreover, for every $M \in \mathbb{N}$, $(\mathbb{P} \times \tilde{\mathbb{P}})$-a.s.

$$\Phi(\theta_{M,N}^{\min}) \xrightarrow{N \to \infty} \Phi(\theta_M^{\min}) \quad \textit{and} \quad \hat{\Phi}_N(\theta_{M,N}^{\min}) \xrightarrow{N \to \infty} \Phi(\theta_M^{\min}).$$

In particular, one can define an increasing sequence $(N_M)_{M \in \mathbb{N}}$ in $\mathbb{N}$ such that for every $1 \leq k \leq K$ we have that $Y^{\theta_{M,N_M}^{\min}}$ converges to $\hat{X}$ for $M \to \infty$ as random variable in $L^1(\Omega \times [0,T], \mathbb{P} \times \lambda_k)$. In particular, the limit process $Y := \lim_{M \to \infty} Y^{\theta_{M,N_M}^{\min}}$ equals $\hat{X}$ $(\mathbb{P} \times \lambda_k)$-almost surely as a random variable on $\Omega \times [0,T]$.

5 EXPERIMENTS

5.1 TRAINING ON SYNTHETIC DATASETS

We test our algorithm on three scalar stochastic models, Black-Scholes, Ornstein-Uhlenbeck and Heston, with fixed parameters. For each model, we generate a dataset by sampling $N = 20'000$ paths on the time interval $[0, 1]$ using the Euler-scheme with 100 time steps. Independently for each path, on average 10% of the grid points are randomly chosen as observation times. The NJ-ODE is trained on 80% of the data. On the remaining 20% the model is tested, by comparing the loss function (5) computed with the NJ-ODE to the loss function computed with the true conditional expectation (Figure 2). During training, the relative difference becomes very small, hence the true conditional expectation is nearly replicated. See Appendix J.1 and J.2 for further details and results.

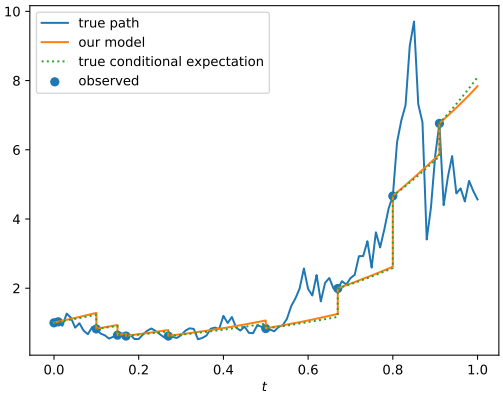

Figure 1: Predicted and true conditional expectation on a test sample of the Heston dataset.

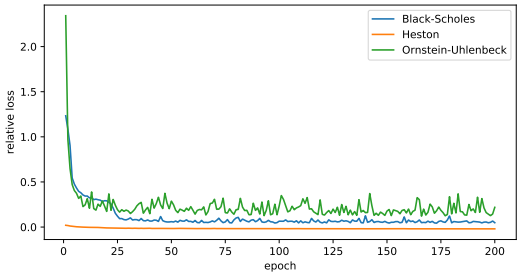

Figure 2: The relative difference between the loss function computed with the predicted conditional expectation and the loss function computed with the true conditional expectation, $\frac{\hat{\Psi}_N(Y^\theta)-\hat{\Psi}_N(\hat{X})}{\hat{\Psi}_N(\hat{X})}$ is shown with respect to the number of epochs.

5.2 EMPIRICAL CONVERGENCE STUDY

We confirm the theoretical results of Theorem 4.2 by an empirical convergence study for growing numbers of training samples N_1 and network size M. Performance is measured by the evaluation metric (41), which computes the mean squared error between the predicted and the true conditional expectation at each time point of the sampling grid. For each combination of N_1 and M, the NJ-ODE is trained 5 times on the Black-Scholes dataset with different random network initializations and randomly chosen training sets (of the given size) for a total of 100 epochs. For each of the trained models, the epoch with the smallest evaluation metric is chosen and the means over the 5 trials with their standard deviations are shown in Figure 9. Already 200 training samples lead to a small evaluation metric if the network size is big enough, suggesting that the MC approximation of the loss function is good, already with a few samples. Further details are given in Appendix J.3.

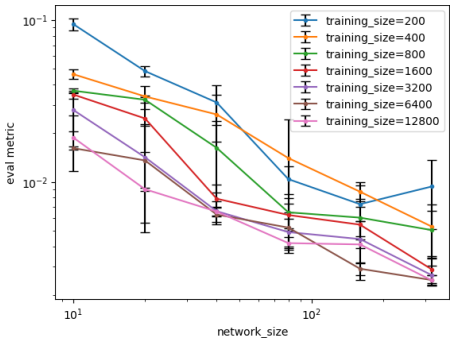

Figure 3: Mean $\pm$ standard deviation (black bars) of the evaluation metric for varying number of training samples N_1 and network size M.

5.3 COMPARISON TO GRU-ODE-BAYES ON SYNTHETIC DATA

We keep using the three synthetic datasets (Black-Scholes, Ornstein-Uhlenbeck and Heston), because they make it easy to evaluate the performance of models in settings of different complexity using the evaluation metric (41). We compare our model to GRU-ODE-Bayes (Brouwer et al., 2019), which is, to the best of our knowledge, the neural network based method with the most similar task to ours.

Results of our comparison are shown in Table 5.3. Further details on GRU-ODE-Bayes and the implementation are given in Appendix J.4 and a more in depth discussion of the results is provided in Appendix J.4.1. On the Black-Scholes and Ornstein-Uhlenbeck dataset, our model performs similarly to GRU-ODE-Bayes. However, on the Heston dataset, the training of GRU-ODE-Bayes is unstable and does not converge. On the other hand, our model converges during training. Although the value of the evaluation metric is much higher than for Black-Scholes and Ornstein-Uhlebeck, the resulting model output is still meaningful, which is not the case for GRU-ODE-Bayes (cf. Figure 14, 17 in Appendix J.4.1). Hence we conclude, that GRU-ODE-Bayes can not be applied reasonably for the Heston dataset, while our method works as the theoretical results suggest.

Table 1: The minimal, last and average value of the evaluation metric (smaller is better) throughout the 100 epochs of training are shown for GRU-ODE-Bayes and our method, together with the number of trainable parameters.

	Black-Scholes		Ornstein-Uhlenbeck		Heston	
	GRU	ours	GRU	ours	GRU	ours
min	$7 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$3 \cdot 10^{-5}$	$5 \cdot 10^{-4}$	4.24	1.33
last	$3 \cdot 10^{-3}$	$8 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$6 \cdot 10^{-4}$	30.9	1.35
average	$3 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$7 \cdot 10^{-4}$	$8 \cdot 10^{-4}$	56.2	1.65
params	$112'602$	$10'071$	$112'602$	$10'071$	$112'602$	$10'071$

5.4 Real world datasets with incomplete observations

Self-imputation for incomplete observations. In the theoretical description of our model, we always assumed that at each observation time all coordinates of X are observed. However, in many real world dataset, the observations are incomplete, i.e. only for some of the coordinates an observation is available. To deal with this, we propose the following self-imputation method. Whenever an observation $X_t \in \mathbb{R}^{d_X}$ is made the mask $m_t \in \{0,1\}^{d_X}$ tells whether the i-th coordinate of X_t is observed ($m_t^i = 1$) or not ($m_t^i = 0$). Then we use the prediction Y_{t-} of our model to generate the imputed observation vector

$$\tilde{X}_t := m_t \odot X_t + (\mathbf{1}_{d_X} - m_t) \odot Y_{t-},$$

where $\odot$ is the element-wise multiplication (Hadamar product) and $\mathbf{1}_{d_X} \in \mathbb{R}^{d_X}$ is the one-vector. Instead of X_t we use $(\tilde{X}_t, m_t)$ as an input for the jump part ρ_{θ_2}. The intuition behind this definition is the following. In the one dimensional case, if we do not make an observation, but input Y_{t-} as if it was an observation, we expect that this does not change the output Y_s for $s \geq t$. From this point of view, Y_{t-} does not provide any additional information for the model. Similarly, we expect that imputing Y_{t-} for unobserved coordinates does not provide any information about this coordinate to the model. However, since the model might learn to transfer the information about an observed coordinate to an unobserved one, we extend the input to also include the information which coordinates were observed. For the ODE part, we use $Y_{\tau(t)}$, which is the prediction after processing the input $\tilde{X}_t$, instead of $X_{\tau(t)}$. Here the intuition is that if the model learns how to best use the incomplete observation, i.e. if the jump is good, then this is the best approximation for $X_{\tau(t)}$. In accordance, the loss function is adjusted to only use the non-imputed coordinates, i.e. if m is the random process in $\{0,1\}^{d_X}$ describing which coordinates are observed we have

$$\Psi(Z) := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(|m_{t_i} \odot (X_{t_i} - Z_{t_i})|_2 + |m_{t_i} \odot (Z_{t_i} - Z_{t_i-})|_2\right)^2\right].$$

Climate forecast. We compare our model to GRU-ODE-Bayes on the USHCN daily dataset (Menne et al., 2016), using the same experimental setting as was used by Brouwer et al. (2019). In particular, we use the same indices for train, validation and test sets on the same 5 folds for cross validation. Once we train the NJ-ODE of the same size as in Section 5.1 and 5.3 (S) and once a bigger version (L). The validation set was used for early stopping after the first 100 epochs, where we trained for a total of 200 epochs. We see in Table 5.4 that our small version performs slightly worse than GRU-ODE-Bayes while our large version slightly outperforms it. For further details see Appendix J.5.

Table 2: Mean and standard deviation of MSE on the test sets of USHCN.

	USHCN - MSE
neural ODE-VAE	0.96 ± 0.11
neural ODE-VAE-MASK	0.83 ± 0.10
sequential VAE	0.83 ± 0.07
GRU-SIMPLE	0.75 ± 0.12
GRU-D	0.53 ± 0.06
T-LSTM	0.59 ± 0.11
GRU-ODE-Bayes	0.43 ± 0.07
NJ-ODE (S)	0.45 ± 0.06
NJ-ODE (L)	$\mathbf{0.40 \pm 0.07}$

6 Related work

Stochastic filtering theory. Our main theorem is similar to the theoretical results of *neural filtering* (Lo, 2009), which is a neural network approach to stochastic filtering theory in discrete time. In stochastic filtering, potentially incomplete and noisy observations of X are available continuously in time. This observation process Y is usually described by the dynamics $dY_t = h(X_t)dt + dB_t$, where h is a measurable function and B is a Brownian motion. Stochastic filtering then estimates the conditional law of X_t given the noisy observations $(Y_s)_{0 \leq s \leq t}$ (Bain & Crisan, 2008, Def. 3.2) and therefore can provide conditional expectations. Comparably, we directly compute the conditional expectation given the last observation. Similar to our assumptions, in the neural filtering approach of GRU-ODE-Bayes (Brouwer et al., 2019) and (Ryder et al., 2018) observations are only available at irregular discrete time points. They approximate the conditional law of X given the last observation by a Gaussian distribution and learn its mean and variance parameters. In particular, the conditional expectation is then given by this mean parameter. In contrast, we do not make normality assumptions about the conditional distribution and we theoretically prove convergence to the true conditional expectation.

Neural ODEs with jumps. Except for GRU-ODE-Bayes (Brouwer et al., 2019) and ODE-RNN (Rubanova et al., 2019) another work studying a neural ODE with jumps is *neural jump SDE* (NJSDE) (Jia & Benson, 2019). Similar to the NJ-ODE framework (4), the latent process of NJSDE is described by a neural ODE with jumps at random times. This model is used to describe hybrid systems which evolve continuously in time but may also be interrupted by stochastic events. In contrast to that, we model the conditional expectation of a continuous stochastic process.

7 Discussion

We presented theoretical guarantees for the NJ-ODE, a general data-driven framework for modelling the conditional expectation of a stochastic process given the previous observation. To the best of our knowledge, this is the first time that convergence of a neural network based model to the conditional expectation process is proven theoretically. We provided extensive experiments, to confirm our theoretical results and to show the advantage of of our model over other state-of-the-art methods for the same task. Moreover, we provided experiments on real world datasets proving the applicability of our model with competitive performance, even beyond the setting of our theoretical results. The theoretical analysis in this more complicated setting with incomplete observations will be addressed in future work.

References

Alan Bain and Dan Crisan. *Fundamentals of stochastic filtering*, volume 60. Springer Science & Business Media, 2008.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *NeurIPS*, 2019.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018.

Christa Cuchiero, Martin Larsson, and Josef Teichmann. Deep neural networks, generic universal interpolation, and controlled ODEs, 2019. URL http://arxiv.org/abs/1908.07838.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, New York, NY, USA, 4th edition, 2010. ISBN 0521765390, 9780521765398.

Massimiliano Gubinelli. Stochastic analysis - course note 4. *https://www.iam.uni-bonn.de/fileadmin/user_upload/gubinelli/stochastic-analysis-ss16/sa-note-4.pdf*, 2016. URL https://www.iam.uni-bonn.de/fileadmin/user_upload/gubinelli/stochastic-analysis-ss16/sa-note-4.pdf.

Calypso Herrera, Florian Krach, and Josef Teichmann. Estimating full lipschitz constants of deep neural networks. 2020. URL https://arxiv.org/abs/2004.13135.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359–366, July 1989. ISSN 0893-6080.

K. Islam and C. Shelton. Neural stochastic differential equations with bayesian jumps for marked temporal point process. 2020.

Junteng Jia and Austin R Benson. Neural jump stochastic differential equations. pp. 9847–9858, 2019. URL http://papers.nips.cc/paper/9177-neural-jump-stochastic-differential-equations.pdf.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

Bernard Lapeyre and Jérôme Lelong. Neural network regression for bermudan option pricing. *arXiv preprint arXiv:1907.06474*, 2019.

Mathias Lechner and Ramin Hasani. Learning long-term dependencies in irregularly-sampled time series, 2020.

Michel Ledoux and Michel Talagrand. Probability in banach spaces. *Springer-Verlag*, 62:67–69, 1991.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations, 2020.

J. Ting-Ho Lo. Neural Filtering. *Scholarpedia*, 4(8):7868, 2009. doi: 10.4249/scholarpedia.7868. revision #91563.

M. J. Menne, C. N. Williams, Jr., and R. S. Vose. Long-term daily and monthly climate records from stations across the contiguous united states (u.s.historical climatology network) (ndp-019). 1 2016. doi: 10.3334/CDIAC/CLI.NDP019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Philip Protter. *Stochastic integration and differential equations*. Springer-Verlag, 2nd edition, 2005.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent odes for irregularly-sampled time series. 2019.

Reuven Y Rubinstein and Alexander Shapiro. *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method*. John Wiley & Sons Inc, 1993.

Sanae Rujivan and Song-Ping Zhu. A simplified analytical approach for pricing discretely-sampled variance swaps with stochastic volatility. *Applied Mathematics Letters*, 25(11):1644 – 1650, 2012. ISSN 0893-9659. doi: https://doi.org/10.1016/j.aml.2012.01.029. URL http://www.sciencedirect.com/science/article/pii/S0893965912000468.

Thomas Ryder, Andrew Golightly, A. Stephen McGough, and Dennis Prangle. Black-box variational inference for stochastic differential equations. In *ICML*, 2018.

Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit, 2019.

Boxuan Yue, Junwei Fu, and Jun Liang. Residual recurrent neural networks for learning sequential representations. *Information*, 9(3), 2018. ISSN 2078-2489. doi: 10.3390/info9030056. URL https://www.mdpi.com/2078-2489/9/3/56.

APPENDIX

A DEFINITIONS

Definition A.1. *A process $\{X_t\}_{t\in[0,T]}$ is **square integrable** if we have $\mathbb{E}[X_t^2] < \infty$ for every $t \in [0,T]$.*

Definition A.2. *Let $d_X, d_2 \in \mathbb{N}$. A function $\varphi : [0,T] \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_2}$ is **globally Lipschitz continuous** if there exists a constant $M > 0$ for which we have for all $t \in [0,T]$*

$$|\varphi(t,x) - \varphi(t,y)|_2 \leq M|x-y|_2 \quad \text{and} \quad |\varphi(x,t)|_2 \leq (1+|x|_2)M. \tag{8}$$

where $|\cdot|_2$ denotes the L^2 norm. In particular, this means that the growth of ϕ is at most linear in x.

Definition A.3. *Let $d_X, d_2 \in \mathbb{N}$. $\varphi : [0,T] \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}$ is **continuous in its first component** (t) **uniformly in its second component** (x) i.e. for every $t \in [0,T]$ and $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $s \in [0,T]$ with $|t-s| < \delta$ and all $x \in \mathbb{R}^{d_X}$ we have $|\varphi(t,x) - \varphi(s,x)| < \epsilon$*

Definition A.4. *Let $d_X, d_W \in \mathbb{N}$. We say that $\varphi : [0,T] \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_X \times d_W}$ is L^2 **integrable with respect** W, i.e $\phi \in L^2(W)$, when we have*

$$\mathbb{E}\left[\sum_{i=1}^{d_X}\sum_{j=1}^{d_W}\int_0^T \sup_x \varphi_{i,j}(x,t)^2 d[W^j,W^j]_t\right] = \int_0^T |\sup_x \varphi(x,t)|_F^2 \, dt < \infty, \tag{9}$$

where $|\cdot|_F$ denotes the Frobenius matrix norm. We remark that this is in particularly implied when φ is bounded.

B AUXILIARY RESULTS

For completeness we restate the straight forward generalization of the universal approximation result (Hornik et al., 1989, Thm. 2.4) to multidimensional output.

Theorem B.1 (Hornik)**.** *Let $r, d \in \mathbb{N}$ and σ be a sigmoid function. Let $\mathcal{NN}_{r,d}^{\sigma}$ be the set of all 1-hidden layer neural networks mapping $\mathbb{R}^r$ to $\mathbb{R}^d$. Then for every compact subset $K \subset \mathbb{R}^r$, every $\epsilon > 0$ and every $f \in C(\mathbb{R}^r, \mathbb{R}^d)$ there exists a neural network $g \in \mathcal{NN}_{r,d}^{\sigma}$ such that $\sup_{x\in K}|f(x) - g(x)|_2 < \epsilon$.*

Lemma B.2. *For $1 \leq k \leq K$ the map*

$$\lambda_k : \mathcal{B}([0,T]) \to [0,1], \quad B \mapsto \lambda_k(B) := \tfrac{\tilde{\mathbb{P}}(n\geq k,(t_k-)\in B)}{\tilde{\mathbb{P}}(n\geq k)}.$$

Proof. First we see that $\lambda_k([0,T]) = 1$ since t_k maps to $[0,T]$. Furthermore, for any disjoint sets $B_i \in \mathcal{B}([0,T])$, $i \in \mathbb{N}$, we have $\{n \geq k, t_k- \in \dot{\cup}_{i\geq 1} B_i\} = \dot{\cup}_{i\geq 1}\{n \geq k, t_k- \in B_i\}$ and these sets are $\tilde{\mathcal{F}}$-measurable, since they are defined through pre-images of random variables. Therefore, the additivity of $\tilde{\mathbb{P}}$ implies that $\lambda_k\left(\dot{\cup}_{i\geq 1} B_i\right) = \sum_{i\geq 1}\lambda_k\left(B_i\right)$. □

The following Lemmas are used in the Proof of Theorem 4.1.

Lemma B.3. *Let $1 \leq k \leq K$ and let $Z \in \mathbb{D}$ be a process such that $(\mathbb{P}\times\lambda_k)[\hat{X} \neq Z] > 0$. Then there exists an $\varepsilon > 0$ such that $\tilde{B} := \{t \in [0,T] \,|\, \mathbb{E}_{\mathbb{P}}[|X_t - Z_{t-}|_2^2] \geq \varepsilon + \mathbb{E}_{\mathbb{P}}[|X_t - \hat{X}_{t-}|_2^2]\}$ satisfies $\lambda_k(\tilde{B}) > 0$.*

Proof. First remark that since X is continuous, we have $X_t = X_{t-}$. Let us define $C := \{(\omega,t) \in \Omega \times [0,T] \,|\, \hat{X}_{t-}(\omega) \neq Z_{t-}(\omega)\}$ and for each $t \in [0,T]$ let $C_t := \{\omega \in \Omega \,|\, (\omega,t) \in C\}$. Then we have for $B := \{t \in [0,T] \,|\, \mathbb{P}(C_t) > 0\}$ that $\lambda_k(B) > 0$, since otherwise by Fubini's theorem

$$0 < (\mathbb{P}\times\lambda_k)[C] = \int_{[0,T]} \mathbb{P}(C_t) d\lambda_k(t) = 0,$$

which is a contradiction. Now Proposition C.1 yields that for each $t \in B$ there exists some $\varepsilon_t > 0$ such that $\mathbb{E}_{\mathbb{P}}[|X_{t-} - Z_{t-}|_2^2] \geq \varepsilon_t + \mathbb{E}_{\mathbb{P}}[|X_{t-} - \hat{X}_{t-}|_2^2]$. This implies

the claim. Indeed, assume no such $\varepsilon > 0$ exists, then we have for each $n \in \mathbb{N}$ that $\lambda_k\left(\{t \in [0,T] \,|\, \mathbb{E}_{\mathbb{P}}[|X_{t-} - Z_{t-}|_2^2] \geq \frac{1}{n} + \mathbb{E}_{\mathbb{P}}[|X_{t-} - \hat{X}_{t-}|_2^2]\}\right) = 0$. Therefore,

$$\begin{aligned}
&\lambda_k\left(\{t \in [0,T] \,|\, \mathbb{E}_{\mathbb{P}}[|X_{t-} - Z_{t-}|_2^2] > \mathbb{E}_{\mathbb{P}}[|X_{t-} - \hat{X}_{t-}|_2^2]\}\right) \\
&\quad \leq \sum_{n\in\mathbb{N}} \lambda_k\left(\{t \in [0,T] \,|\, \mathbb{E}_{\mathbb{P}}[|X_{t-} - Z_{t-}|_2^2] \geq \tfrac{1}{n} + \mathbb{E}_{\mathbb{P}}[|X_{t-} - \hat{X}_{t-}|_2^2]\}\right) = 0,
\end{aligned}$$

which is a contradiction to $\lambda_k(B) > 0$. □

Remark B.4. *The σ-fields $\mathcal{A}_t$ depend on $\tilde{\omega} \in \tilde{\Omega}$ as well. If we look at the product probability space $(\Omega \times \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$, where $\mathbb{F} \otimes \tilde{\mathcal{F}}$ consists of the tensor-product σ-algebras $(\mathcal{F} \otimes \tilde{\mathcal{F}})_t := \mathcal{F}_t \otimes \tilde{\mathcal{F}}$ for $t \in [0,T]$, then*

$$\begin{aligned}
&\boldsymbol{\sigma}(X_{t_i} | t_i \leq t) \\
&\quad := \boldsymbol{\sigma}\left(\left\{\left\{(\omega,\tilde{\omega}) \in \Omega \times \tilde{\Omega} \middle| X_{t_i(\tilde{\omega})}(\omega) \in A, n(\tilde{\omega}) \geq i, t_i(\tilde{\omega}) \leq t\right\} \middle| A \in \mathcal{B}(\mathbb{R}^{d_X}), i \in \mathbb{N}\right\}\right)
\end{aligned}$$

is a well defined sub-σ-algebra of $(\mathcal{F} \otimes \tilde{\mathcal{F}})_t$. Furthermore, we can recover the $\tilde{\omega}$-wise defined version of $\mathcal{A}_t$ by intersecting each set in it with $\Omega \times \{\tilde{\omega}\}$ and subsequently projecting the intersection on its first component. We use the notation $\tilde{\mathcal{A}}_t := \tilde{\mathcal{A}}_t(\tilde{\omega}) = \boldsymbol{\sigma}(X_{t_i(\tilde{\omega})} | t_i(\tilde{\omega}) \leq t)$ to distinguish this $\tilde{\omega}$-wise definition from the definition as sub-σ-algebra of the product space given above. However, Lemma C.2 implies that for our considerations, both versions of this σ-algebra have the same effect, therefore we will simply write $\mathcal{A}_t$ for both versions, by abuse of notation.

Lemma B.5. *For any $\mathbb{A}$-adapted process Z it holds that*

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n |X_{t_i} - Z_{t_i-}|_2^2\right] = \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left|X_{t_i} - \hat{X}_{t_i-}\right|_2^2\right] + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left|\hat{X}_{t_i-} - Z_{t_i-}\right|_2^2\right].$$

Proof. First recall that by continuity $X_{t_i} = X_{t_i-}$. Then the statement is a consequence of Proposition C.1, Lemma C.2 and Fubini's theorem, which imply

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n |X_{t_i-} - Z_{t_i-}|_2^2\right] &= \mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}}\left[|X_{t_i-} - Z_{t_i-}|_2^2\right]\right] \\
&= \mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left(\mathbb{E}_{\mathbb{P}}\left[\left|X_{t_i-} - \hat{X}_{t_i-}\right|_2^2\right] + \mathbb{E}_{\mathbb{P}}\left[\left|\hat{X}_{t_i-} - Z_{t_i-}\right|_2^2\right]\right)\right] \\
&= \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left|X_{t_i-} - \hat{X}_{t_i-}\right|_2^2\right] + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left|\hat{X}_{t_i-} - Z_{t_i-}\right|_2^2\right].
\end{aligned}$$

□

C BEST ESTIMATOR $\hat{X}$ — PROOF OF PROPOSITION 2.1

We are interested in the "best" approximation (or prediction) of the process X that one can make at each time t, given the currently available information $\mathcal{A}_t$. If "best" refers to the $L^2(\Omega \times \tilde{\Omega}, \mathbb{P} \times \tilde{\mathbb{P}})$-minimizer, then this approximation is given by the conditional expectation, as stated in the elementary Proposition below (which is proven for example in (Durrett, 2010, Thm. 5.1.8) for $\mathbb{R}$-valued random variables and can easily be extended to $\mathbb{R}^d$-valued random variables when using the 2-norm).

Proposition C.1. *Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sub-σ-algebra $\mathcal{A} \subset \mathcal{F}$, the orthogonal projection of $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ on $L^2(\Omega, \mathcal{A}, \mathbb{P})$ is given by $\hat{X} := \mathbb{E}[X|\mathcal{A}]$. In particular, for every $Z \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(Z \neq \hat{X}) > 0$ we have*

$$E[|X - Z|_2^2] = \mathbb{E}[|X - \hat{X}|_2^2] + \mathbb{E}[|Z - \hat{X}|_2^2] > \mathbb{E}[|X - \hat{X}|_2^2].$$

In our case this means, that the optimal $\mathbb{A}$-adapted process in $L^2(\Omega \times \tilde{\Omega}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$ approximating X is given by $(\hat{X}_t)_{t \in [0,T]}$ with $\hat{X}_t := \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[X_t|\mathcal{A}_t]$. Here, $\mathcal{A}_t$ is meant as a sub-σ-algebra of $\mathcal{F}_t \times \tilde{\mathcal{F}}$. This process is unique up to $(\mathbb{P} \times \tilde{\mathbb{P}})$-null-sets. Moreover, the following lemma shows that it coincides $\tilde{\omega}$-wise with $\mathbb{E}_{\mathbb{P}}[X_t|\tilde{\mathcal{A}}_t](\tilde{\omega})$, where $\tilde{\mathcal{A}}_t = \tilde{\mathcal{A}}_t(\tilde{\omega})$ is defined in Remark B.4.

Lemma C.2. *For $\tilde{\mathbb{P}}$-almost-every $\tilde{\omega} \in \tilde{\Omega}$ we have $\mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[X_t|\mathcal{A}_t](\tilde{\omega}) = \mathbb{E}_{\mathbb{P}}[X_t|\tilde{\mathcal{A}}_t](\tilde{\omega})$ $\mathbb{P}$-almost-surely.*

Proof. Otherwise we have by Fubini's theorem, Proposition C.1 and an argument similar to the one in Lemma B.3

$$\begin{aligned}\mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}\left[\left|X_t - \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[X_t|\mathcal{A}_t]\right|_2^2\right] &= \mathbb{E}_{\tilde{\mathbb{P}}}\left[\mathbb{E}_{\mathbb{P}}\left[\left|X_t - \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[X_t|\mathcal{A}_t]\right|_2^2\right]\right] \\ &> \mathbb{E}_{\tilde{\mathbb{P}}}\left[\mathbb{E}_{\mathbb{P}}\left[\left|X_t - \mathbb{E}_{\mathbb{P}}[X_t|\tilde{\mathcal{A}}_t]\right|_2^2\right]\right],\end{aligned}$$

which is a contradiction to Proposition C.1. □

This proves the first part of Proposition 2.1. The second part of this Proposition, i.e. (2), should be understood $\tilde{\omega}$-wise, for $\tilde{\omega} \in \tilde{\Omega}$. This is justified by Lemma C.2 and derived below. In particular, for the remainder of this section, all statements are meant $\tilde{\omega}$-wise.

With the assumption that μ and σ are Lipschitz, (Protter, 2005, Thm. 7, Chap. V) implies that a unique solution of (1) exists. Furthermore, this solution is a Markov process as soon as the starting point x is fixed (Protter, 2005, Thm. 32, Chap. V). Hence, one can define a transition function P (compare (Protter, 2005, Chap. V.6)) such that for all $s < t$ and ϕ bounded and measurable,

$$P_{s,t}(X_s, \phi) := \mathbb{E}[\phi(X_t)|\boldsymbol{\sigma}(X_s)] = \mathbb{E}_{\mathbb{P}}[\phi(X_t)|\mathcal{F}_s].$$

We have that $X_{\tau(s)}$ is $\mathcal{A}_{\tau(s)}$-measurable and therefore, since $\mathcal{A}_s = \mathcal{A}_{\tau(s)} \subset \mathcal{F}_{\tau(s)}$,

$$P_{\tau(s),t}(X_{\tau(s)}, \phi) = \mathbb{E}_{\mathbb{P}}[\phi(X_t)|\mathcal{A}_s]. \tag{10}$$

By our additional assumption on σ it follows from (Protter, 2005, Lem. before Thm. 28, Chap. IV) that

$$M_t := \int_0^t \sigma(s, X_s) dW_s, \quad 0 \leq t \leq T,$$

is a square integrable martingale, since the Browian motion W is square integrable. In particular, for $0 \leq s \leq t \leq T$ we have $\mathbb{E}[\int_s^t \sigma(r, X_r) dW_r|\mathcal{F}_s] = \mathbb{E}[M_t - M_s|\mathcal{F}_s] = 0$. Moreover, the same is true when conditioning on $\mathcal{A}_s$ [5].

[5] To see this, we choose a localizing sequence $(\tau_n)_{n \in \mathbb{N}}$ such that M^{τ_n} is bounded by n (works since M is continuous). Then the Markov property implies that $\mathbb{E}[M_t^{\tau_n} - M_s^{\tau_n}|\mathcal{A}_s] = \mathbb{E}[M_t^{\tau_n} - M_s^{\tau_n}|\mathcal{F}_s] = 0$. Since $M_t^{\tau_n} \xrightarrow{n \to \infty} M_t$ $\mathbb{P}$-a.s. and since this sequence is dominated by the integrable random variable $1 + \sup_{r \leq t} \|M_r\|^2$ (by Doob's inequality and square integrability of M), dominated convergence implies that $\mathbb{E}[M_t - M_s|\mathcal{A}_s] = 0$.

Using the martingale property of M, we have ($\tilde{\omega}$-wise) for every $t \in [0,T]$

$$\begin{aligned}\hat{X}_t &= \mathbb{E}_{\mathbb{P}}[(X_t - X_{\tau(t)}) + X_{\tau(t)}|\mathcal{A}_{\tau(t)}] \\ &= X_{\tau(t)} + \mathbb{E}_{\mathbb{P}}\left[\int_{\tau(t)}^t \mu(r, X_r)dr\Big|\mathcal{A}_{\tau(t)}\right] + \mathbb{E}_{\mathbb{P}}\left[\int_{\tau(t)}^t \sigma(r, X_r)dW_r\Big|\mathcal{A}_{\tau(t)}\right] \\ &= X_{\tau(t)} + \int_{\tau(t)}^t \mathbb{E}_{\mathbb{P}}\left[\mu(r, X_r)|\mathcal{A}_{\tau(t)}\right] dr,\end{aligned} \tag{11}$$

where we used Fubini's Theorem (for conditional expectations) in the last step. This is justified because $\mathbb{E}_{\mathbb{P}}[\int_0^T |\mu(r, X_r)|_2 dr] < \infty$ follows from μ being bounded. Let us define $\Delta := \{(t,r) \in [0,T]^2 | t + r \leq T\}$ and the function

$$\tilde{\mu} : \Delta \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}, \quad (t, r, \xi) \mapsto P_{t,t+r}(X_t, \mu)\big|_{X_t = \xi} = \mathbb{E}_{\mathbb{P}}\left[\mu(t+r, X_{t+r})|X_t = \xi\right],$$

then we can use (10) to rewrite (11) as

$$\hat{X}_t = X_{\tau(t)} + \int_{\tau(t)}^t \tilde{\mu}\left(\tau(t), s - \tau(t), X_{\tau(t)}\right) ds. \tag{12}$$

This proves the second part of Proposition 2.1.

D Proof of Proposition 2.2

Proof of Proposition 2.2. For any fixed $s \in [0,T], x \in \mathbb{R}^{d_X}$, we define

$$\zeta_{s,\cdot}(x) : [0,T] \times \Omega \to \mathbb{R}^{d_X}, (t, \omega) \mapsto \zeta_{s,t}(x)(\omega)$$

to be the solution of the SDE

$$\zeta_{s,t}(\xi) = \xi + \int_s^t \mu(r, \zeta_{s,r})dr + \int_s^t \sigma(r, \zeta_{s,r})dW_r.$$

This solution exists and is unique by (Protter, 2005, Chap. V, Thm. 7), therefore we have that $X_t = \zeta_{0,t}(x)$ $\mathbb{P}$-almost surely. Furthermore, (Gubinelli, 2016, Thm. 4) implies that for $s \leq t$ we have $X_t = \zeta_{s,t}(\zeta_{0,s}(x))$. Hence, for $t = s + r$, we have the identity $\tilde{\mu}(s, r, \xi,) = \mathbb{E}[\mu(t, \zeta_{s,t}(\xi))]$.

Furthermore, by (Gubinelli, 2016, Thm. 8) we have for any $\xi, \xi' \in \mathbb{R}^{d_X}$ and $(s,r), (s',r') \in \Delta$ with $t := s + r, t' := s' + r' \in [0,T]$ that there exists some constant C such that

$$\mathbb{E}\left[|\zeta_{s,t}(\xi) - \zeta_{s',t'}(\xi')|_2^2\right] \leq C\left[|\xi - \xi'|_2^2 + (1 + |\xi|_2 + |\xi'|_2)^2 (|t - t'| + |s - s'|)\right]. \tag{13}$$

Therefore, we have that

$$\begin{aligned}|\tilde{\mu}(s,r,\xi) - \tilde{\mu}(s',r',\xi')|_2 &= |\mathbb{E}\left[\mu(t, \zeta_{s,t}(\xi))\right] - \mathbb{E}\left[\mu(t', \zeta_{s',t'}(\xi'))\right]|_2 \\ &\leq |\mathbb{E}\left[\mu(t, \zeta_{s,t}(\xi))\right] - \mathbb{E}\left[\mu(t', \zeta_{s,t}(\xi))\right]|_2 \\ &\quad + |\mathbb{E}\left[\mu(t', \zeta_{s,t}(\xi))\right] - \mathbb{E}\left[\mu(t', \zeta_{s',t'}(\xi'))\right]|_2 \\ &\leq \mathbb{E}\left[|\mu(t, \zeta_{s,t}(\xi)) - \mu(t', \zeta_{s,t}(\xi))|_2^2\right]^{1/2} \\ &\quad + \mathbb{E}\left[|\mu(t', \zeta_{s,t}(\xi)) - \mu(t', \zeta_{s',t'}(\xi'))|_2^2\right]^{1/2} \\ &\leq \mathbb{E}\left[|\mu(t, \zeta_{s,t}(\xi)) - \mu(t', \zeta_{s,t}(\xi))|_2^2\right]^{1/2} \\ &\quad + M\,\mathbb{E}\left[|\zeta_{s,t}(\xi) - \zeta_{s',t'}(\xi')|_2^2\right]^{1/2},\end{aligned} \tag{14}$$

where we used Jensen's inequality in the second last and (8) in the last step. Hence, for $(s', r', \xi') \to (s, r, \xi)$ we have that the first term of (14) goes to zero due to continuity of μ in its first component uniformly in the second component. Moreover, the second term of (14) converges to zero by (13). Together, this proves continuity of $\tilde{\mu}$. □

E RECALL: RNN, NEURAL ODE AND ODE-RNN - EQUIVALENT WAYS OF WRITING

We describe our model as the solution of the SDE (4), which is a compact way of writing. In the following, we first shortly recall recurrent neural networks (RNN) and the neural ODE and then recall the ODE-RNN. Furthermore, we give a step-by-step explanation, how the way ODE-RNN was formalized, can equivalently be written in terms of an SDE similar to (4). Finally we give the alternative way of writing our model and a short comparison to ODE-RNN.

Recurrent Neural Network. The input to a RNN is a discrete time series of observations $\{x_{t_1}, \cdots, x_{t_n}\}$. At each time t_{i+1}, the latent variable h is updated using the previous latent variable h_{t_i} and the input $x_{t_{i+1}}$ as

$$h_{t_{i+1}} := \text{RNNCell}(h_{t_i}, x_{t_{i+1}}), \tag{15}$$

where RNNCell is a neural network.

Neural Ordinary Differential Equation. Neural ODEs (Chen et al., 2018) are a family of continuous-time models defining a latent variable $h_t := h(t)$ to be the solution to an ODE initial-value problem (IVP):

$$h_t := h_{t_0} + \int_{t_0}^{t} f(h_s, s, \theta) ds, \quad t \geq t_0, \tag{16}$$

where $f(\cdot, \cdot, \theta) = f_\theta$ is a neural network with weights θ. Therefore, the latent variables can be updated continuously by solving this ODE (16). We can emphasize the dependence of h_t on a numerical ODE solver by rewriting (16) as

$$h_t := \text{ODESolve}(f_\theta, h_{t_0}, (t_0, t)) \,. \tag{17}$$

ODE-RNN. ODE-RNN (Rubanova et al., 2019) is a mixture of a RNN and a neural ODE. In contrast to a standard RNN, we are not only interested in an output at the observation times t_i, but also in between those times. In particular, we want to have an output stream that is generated continuously in time. This is achieved by using a neural ODE to model the latent dynamics between two observation times, i.e. for $t_i < t < t_{i+1}$ the latent variable is defined as in (16) and (17), with h_{t_0} and t_0 replaced by h_{t_i} and t_i. At the next observation time t_{i+1}, the latent variable is then updated by a RNN. Rubanova et al. (2019) write this as

$$\begin{cases} h'_{t_{i+1}} & := & \text{ODESolve}(f_\theta, h_{t_i}, (t_i, t_{i+1})) \\ h_{t_{i+1}} & := & \text{RNNCell}(h'_{t_{i+1}}, x_{t_{i+1}}) \,. \end{cases} \tag{18}$$

Therefore, fixing $h_{t_0} := h_0$, the entire latent process can be computed by iteratively solving an ODE followed by applying a RNN.

ODE-RNN as càdlàg process. Thinking about the process $h := (h_t)_{t \geq t_0}$ defined in (18) as a (stochastic) process in time, it is defined to evolve continuously for $t_i \leq t < t_{i+1}$ according to the ODE dynamics f_θ and jumps at time t_{i+1} according to the RNN cell. In particular, it is defined to be a *càdlàg*[6] process, for which $h_{t_{i+1}-}$ is the standard notation for the left limit, i.e. the last point before the jump at time t_{i+1}. According to this notation we have $h_{t_{i+1}-} = h'_{t_{i+1}}$, hence, we can rewrite (18) as

$$\begin{cases} h_{t_{i+1}-} & := & \text{ODESolve}(f_\theta, h_{t_i}, (t_i, t_{i+1})) \\ h_{t_{i+1}} & := & \text{RNNCell}(h_{t_{i+1}-}, x_{t_{i+1}}) \,. \end{cases} \tag{19}$$

E.1 RESIDUAL ODE-RNN AS A SPECIAL CASE OF CONTROLLED ODE

Residual ODE-RNN. We replace the standard RNN cell by a *residual RNN cell* (rRNN), as it was described e.g. in Yue et al. (2018). In particular, instead of applying the RNN cell such that $h_{t_i} = \text{RNNCell}(h_{t_i-}, x_i)$ we use a residual RNN cell to have $h_{t_i} = h_{t_i-} + \text{rRNNCell}(h_{t_i-}, x_i)$. The residual RNN is as expressive as the standard RNN and was empirically shown to perform very similarly or even better than the standard framework (Yue et al., 2018). This way, the residual RNN cell models exactly the jump of the latent variable (i.e. the differences) that occurs at the time t_{i+1}

[6] i.e. right-continuous with existing left limits, also denoted as *RCLL*

when taking into account the next observation $x_{t_{i+1}}$. Therefore, we can rewrite the ODE-RNN (19) as

$$\begin{cases} h_{t_{i+1}-} & := & \text{ODESolve}(f_\theta, h_{t_i}, (t_i, t_{i+1})) \\ h_{t_{i+1}} & := & h_{t_{i+1}-} + \text{rRNNCell}(h_{t_{i+1}-}, x_i)\,. \end{cases} \tag{20}$$

Controlled Ordinary Differential Equation. We briefly recall the definition of *controlled ODEs* as it was given in (Herrera et al., 2020, Section 4.1) and used in (Cuchiero et al., 2019; Herrera et al., 2020) to describe neural networks. We fix $\ell, d \in \mathbb{N}$ and define for $1 \leq i \leq d$ the vector fields

$$V_i : \Theta \times \mathbb{R}_{\geq 0} \times \mathbb{R}^\ell \to \mathbb{R}^\ell, (\theta, t, x) \mapsto V_i^\theta(t, x),$$

which are càglàd[7] in t and Lipschitz continuous in x. Furthermore, we define the scalar càdlàg *control* functions

$$u_i : \mathbb{R}_{\geq 0} \to \mathbb{R}, t \mapsto u_i(t),$$

which have finite variation and satisfy $u_i(0) = 0$. Then we define the process $Z := (Z_t)_{t \geq 0}$ as the solution of the controlled ODE

$$dZ_t = \sum_{i=1}^{d} V_i^\theta(t, Z_{t-}) du_i(t), \;\; Z_0 = z, \tag{21}$$

where $z \in \mathbb{R}^\ell$ is some starting point. (21) is written in Itô's differential notation for (stochastic) integrals. The solution of (21) exists and is unique under much more general assumptions than we made here (Protter, 2005, Chap. V, Thm. 7).

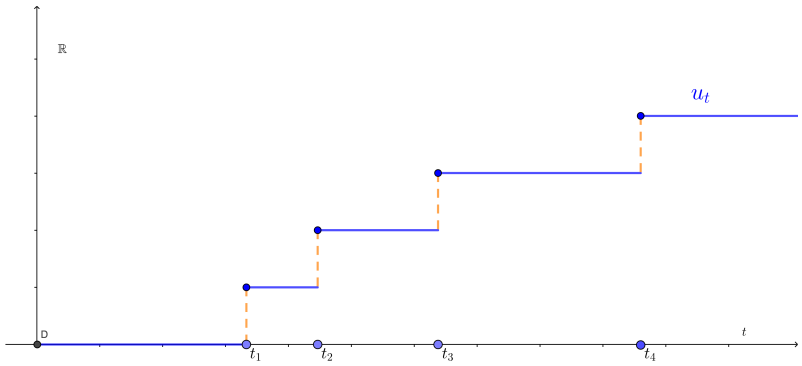

Figure 4: Schematic representation of the stochastic control u_t (22).

Special Case: ODE-RNN. To write the ODE-RNN as controlled ODE, we set $d = 2$ and define $V_1^\theta := f_\theta$, $V_2 :=$ rRNNCell and $u_1(t) := t$. As pointed out by Herrera et al. (2020), one can take the u_i to be semimartingales instead of deterministic functions. In line with this, we define $u_2 := u$ as the pure jump stochastic control process

$$u : \tilde{\Omega} \times [0, T] \to \mathbb{R}, (\tilde{\omega}, t) \mapsto u_t(\tilde{\omega}) := \sum_{i=1}^{n(\tilde{\omega})} \mathbf{1}_{[t_i(\tilde{\omega}), \infty)}(t). \tag{22}$$

We note that u is an adapted process starting at 0 with finite variation on the product probability space $(\Omega \times \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$, since the variation of u up to time T is n and $\mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[n] < \infty$. The following result shows that the residual ODE-RNN can compactly be described as a controlled ODE.

Proposition E.1. *Using the vector fields and controls defined above, the latent variable process $h = (h_t)_{t \geq t_0}$ of the residual ODE-RNN can equivalent be written as the solution of the controlled ODE*

$$dh_t = f_\theta(h_{t-}, t-)dt + \text{rRNNCell}(h_{t-}, x_t)du(t), \quad h_{t_0} = h_0. \tag{23}$$

[7] i.e. left continuous with existing right limits

Proof of Proposition E.1. By (Protter, 2005, Chap. II, Thm. 17), the stochastic integral is indistinguishable from the path-wise Lebesgue-Stieltjes integral if the integrator is of finite variation. Hence, we can assume that some $\tilde{\omega} \in \tilde{\Omega}$ is fixed and that the following expressions are evaluated at this $\tilde{\omega}$ whenever applicable. First note, that u is constant except at the t_i where it increases by 1 (cf. Figure 4). In particular, the Lebesgue-Stieltjes integral of some càdlàg function g with respect to u is a sum, i.e. $\int_0^t g(s-)du_s = \sum_{t_i \leq t} g(t_i-)\Delta u_{t_i}$, where $\Delta u_{t_i} = u_{t_i} - u_{t_i-} = 1$. Therefore, integrating (23) from t_0 to t we get

$$\begin{aligned} h_t &= h_{t_0} + \int_{t_0}^t f_\theta(h_{s-}, s-)ds + \int_{t_0}^t \text{rRNNCell}(h_{s-}, x_s)du(s) \\ &= h_{t_0} + \int_{t_0}^t f_\theta(h_{s-}, s-)ds + \sum_{t_i \leq t} \text{rRNNCell}(h_{t_i-}, x_{t_i}). \end{aligned}$$

In particular, since the first integral is continuous in t, we have for every $1 \leq k \leq n$

$$\begin{aligned} h_{t_k} &= h_{t_0} + \int_{t_0}^{t_k-} f_\theta(h_{s-}, s-)ds + \sum_{t_i < t_k} \text{rRNNCell}(h_{t_i-}, x_{t_i}) + \text{rRNNCell}(h_{t_k-}, x_{t_k}) \\ &= h_{t_k-} + \text{rRNNCell}(h_{t_k-}, x_{t_k}), \end{aligned} \tag{24}$$

and for $t_k < t < t_{k+1}$

$$\begin{aligned} h_t &= h_{t_0} + \int_{t_0}^{t_k} f_\theta(h_{s-}, s-)ds + \sum_{t_i \leq t_k} \text{rRNNCell}(h_{t_i-}, x_{t_i}) + \int_{t_k}^t f_\theta(h_{s-}, s-)ds \\ &= h_{t_k} + \int_{t_k}^t f_\theta(h_{s-}, s-)ds. \end{aligned} \tag{25}$$

Together, (24) and (25) prove that (23) is equivalent to (20). We also emphasize that x_t is used as input only for $t = t_i$, $1 \leq i \leq n$. □

Equivalent way of writing NJ-ODE. By applying the steps that were explained above backwards to (4) (but without introducing the residual version of the RNN), it is easy to see that our model can equivalently be written (similar to (19)) as

$$\begin{cases} h_{t_{i+1}-} &:= \text{ODESolve}(f_{\theta_1}, (\tanh(h_t), \tanh(x_{t_i}), t_i, t - t_i), (t_i, t_{i+1})) \\ h_{t_{i+1}} &:= \rho_{\theta_2}(\tanh(x_{t_{i+1}})) \,. \end{cases} \tag{26}$$

In particular, the main difference to (19) is that we use a modified input for f_{θ_1} and that the neural network performing the jumps does not take $h_{t_{i+1}-}$ as an input, i.e. it is not a RNN.

F Ergodic approximation of the objective function

If we only observe one realization of the path X at times $(\tilde{t}_1, \cdots, \tilde{t}_N)$, we can still approximate the objective function by assuming that μ and σ are time-independent and that the stochastic process X is ergodic in the following sense. We fix $n = 1$ and assume that the time increments $\Delta\tilde{t}_j := \tilde{t}_j - \tilde{t}_{j-1}$ are i.i.d. realizations of the probability distribution λ_1. Furthermore, we consider each observation $X_{\tilde{t}_j}$ as one sample with initial condition $X_{\tilde{t}_{j-1}}$ for which $Y^{\theta,j}$ is the realization of Y^θ. Then we approximate the objective function by

$$\hat{\Phi}_N(\theta) := \frac{1}{N} \sum_{j=1}^N \left(\left| X_{\tilde{t}_j} - Y^\theta_{\Delta\tilde{t}_j}(X) \right|_2 + \left| Y^\theta_{\Delta\tilde{t}_j}(X) - Y^\theta_{\Delta\tilde{t}_j-}(X) \right|_2 \right)^2, \tag{27}$$

which is assumed to converge by the ergodicity assumption for $N \to \infty$ to

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}} \left[\left(\left| X_{t_1} - Y^\theta_{t_1} \right|_2 + \left| Y^\theta_{t_1} - Y^\theta_{t_1-} \right|_2 \right)^2 \right]. \tag{28}$$

Instead of setting the random variable $n = 1$, one could similarly fix the time horizon T, and take for each sample all subsequent observations that lie in the time interval $[t_{\text{start}}, t_{\text{start}} + T]$, where t_{start} is the date of the first observation of this sample. The next sample would then start with the first observation after $t_{\text{start}} + T$.

G PROOF OF THEOREM 4.1

The idea of the proof is to split up the target jump process into its continuous-in-time parts and into the jumps. Both parts are continuous functions of their inputs and can therefore be approximated by neural networks (Hornik et al., 1989). More precisely, by Proposition 2.1 and 2.2 the continuous-in-time part can be written as an integral over time of a function which is jointly continuous in its inputs. This jointly continuous function is approximated by a neural network, which is itself integrated over time – that is the neural ODE part of the NJ-ODE. The remainder of the proof shows L^1-convergence of the output of NJ-ODE with optimal weights (with respect to the loss function) to the conditional expectation process.

Proof of Theorem 4.1. We start by showing that $\hat{X} \in \mathbb{D}$ is the unique minimizer of Ψ up to $(\mathbb{P} \times \lambda_k)$-null-sets for any $k \leq K$. First, we recall that for every t_i we have $\hat{X}_{t_i} = X_{t_i}$ and by continuity of X that $X_{t_i} = X_{t_i-}$. Therefore,

$$
\begin{aligned}
\Psi(\hat{X}) &= \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left|X_{t_i} - \hat{X}_{t_i-}\right|_2^2\right] \\
&= \mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbb{P}}\left[\left|X_{t_i} - \hat{X}_{t_i-}\right|_2^2\right]\right] \\
&= \min_{Z\in\mathbb{D}}\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}|X_{t_i} - Z_{t_i-}|_2^2\right] \\
&\leq \min_{Z\in\mathbb{D}}\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(|X_{t_i} - Z_{t_i}|_2 + |Z_{t_i} - Z_{t_i-}|_2\right)^2\right] \\
&= \min_{Z\in\mathbb{D}}\Psi(Z),
\end{aligned}
$$

where we used Fubini's theorem for the second line, Proposition C.1 for the third line and the triangle inequality for the fourth line. Hence, $\hat{X}$ is a minimizer of Ψ. To see that it is unique $(\mathbb{P} \times \lambda_k)$-a.s., let $Z \in \mathbb{D}$ be a process such that $(\mathbb{P} \times \lambda_k)[\hat{X} \neq Z] > 0$. By Lemma B.3, this implies that there exists an $\varepsilon > 0$ such that $B := \{t \in [0,T] \,|\, \mathbb{E}_{\mathbb{P}}[|X_{t-} - Z_{t-}|_2^2] \geq \epsilon + \mathbb{E}_{\mathbb{P}}[|X_{t-} - \hat{X}_{t-}|_2^2]\}$ satisfies $\lambda_k(B) > 0$.

Now recall that by definition of λ_k we have $\lambda_k(B) = \tilde{\mathbb{P}}(\dot{\cup}_{j\geq k}\{n = j, t_k- \in B\})/\tilde{\mathbb{P}}(n \geq k) > 0$. This implies that there exists $j \in \mathbb{N}_{\geq k}$ such that $\tilde{\mathbb{P}}(n = j, t_k- \in B) > 0$. Therefore,

$$
\begin{aligned}
\mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{t_i-\in B\}}\right] &\geq \mathbb{E}_{\tilde{\mathbb{P}}}\left[\mathbf{1}_{\{n=j\}}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{t_i-\in B\}}\right] \\
&\geq \mathbb{E}_{\tilde{\mathbb{P}}}\left[\mathbf{1}_{\{n=j\}}\tfrac{1}{j}\mathbf{1}_{\{t_k-\in B\}}\right] \\
&= \tfrac{1}{j}\tilde{\mathbb{P}}(n = j, t_k- \in B) > 0.
\end{aligned}
$$

This inequality implies now that Z is not a minimizer of Ψ, because

$$\begin{aligned}
\Psi(Z) &= \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(|X_{t_i}-Z_{t_i}|_2+|Z_{t_i}-Z_{t_i-}|_2\right)^2\right] \\
&\geq \mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbb{P}}\left[|X_{t_i}-Z_{t_i-}|_2^2\right]\right] \\
&= \mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{\{t_i-\in B\}}+\mathbf{1}_{\{t_i-\in B^C\}}\right)\mathbb{E}_{\mathbb{P}}\left[|X_{t_i}-Z_{t_i-}|_2^2\right]\right] \\
&\geq \mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon\,\mathbf{1}_{\{t_i-\in B\}}+\mathbb{E}_{\mathbb{P}}\left[\left|X_{t_i}-\hat{X}_{t_i-}\right|_2^2\right]\right)\right] \\
&= \varepsilon\,\mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\{t_i-\in B\}}\right]+\min_{Z\in\mathbb{D}}\Psi(Z) \\
&> \min_{Z\in\mathbb{D}}\Psi(Z).
\end{aligned}$$

Next we show that (4) can approximate $\hat{X}$ arbitrarily well. Since the dimension d_H can be chosen freely, let us fix it to $d_H := d_X$. Furthermore, let us fix θ_3^* such that $g_{\theta_3^*} = \mathrm{id}$. From Theorem B.1 it follows that for any $\varepsilon > 0$ there exist $M \in \mathbb{N}$ and θ_1^*, θ_2^* with $(\theta_1^*, \theta_2^*, \theta_3^*) \in \Theta_M$ such that

$$\sup_{\tilde{u}\in d_X}\left|\rho_{\theta_2^*}\left(\tanh(\tilde{u})\right)-\tilde{u}\right|_2 = \sup_{u\in[-1,1]^{d_X}}\left|\rho_{\theta_2^*}\left(u\right)-\tanh^{-1}(u)\right)|_2 \leq \varepsilon, \tag{29}$$

and

$$\begin{aligned}
&\sup_{(\tilde{u},\tilde{v},t,r)\in\mathbb{R}^{d_H\times d_X}\times\Delta}\left|f_{\theta_1^*}\left(\tanh(\tilde{u}),\tanh(\tilde{v}),t,r\right)-\tilde{\mu}\left(\tilde{v},t,r\right)\right|_2 \\
&\qquad = \sup_{(u,v,t,r)\in[-1,1]^{d_H\times d_X}\times\Delta}\left|f_{\theta_1^*}\left(u,v,t,r\right)-\tilde{\mu}\left(\tanh^{-1}(v),t,r\right)\right|_2 \leq \varepsilon,
\end{aligned} \tag{30}$$

where we used that $\tilde{\mu}$ is continuous by Proposition 2.2 and substituted $u = \tanh(\tilde{u}), v = \tanh(\tilde{v})$. Using this, we can bound the distance between $Y_t^{\theta_M^*}$ and $\hat{X}$. In particular, if $t \in \{t_1, \cdots, t_n\}$, (29) yields

$$\left|Y_t^{\theta_M^*}-\hat{X}_t\right|_2 = \left|\left(H_{t-}+\left(\rho_{\theta_M^*}(\tanh(X_t))-H_{t-}\right)\right)-X_t\right|_2 \leq \varepsilon,$$

and if t not in $\{t_1, \cdots, t_n\}$, then (2), (30) and the previous bound yield

$$\begin{aligned}
\left|Y_t^{\theta_M^*}-\hat{X}_t\right|_2 &\leq \left|Y_{\tau(t)}^{\theta_M^*}-\hat{X}_{\tau(t)}\right|_2 \\
&\quad+\int_{\tau(t)}^{t}\left|f_{\theta_1^*}\left(\tanh(H_{s-}),\tanh(X_{\tau(s)}),\tau(s),s-\tau(s)\right)-\tilde{\mu}\left(X_{\tau(s)},\tau(s),s-\tau(s)\right)\right|_2 ds \\
&\leq \varepsilon+\varepsilon T.
\end{aligned}$$

Now we can show the convergence of $\Phi(\theta_M^{\min})$ using these two bounds and $X_{t_i} = \hat{X}_{t_i}$. Indeed,

$$\begin{aligned}
\min_{Z\in\mathbb{D}} \Psi(Z) &\leq \Phi(\theta_M^{\min}) \leq \Phi(\theta_M^*) \\
&= \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left(\left|X_{t_i} - Y_{t_i}^{\theta_M^*}\right|_2 + \left|Y_{t_i}^{\theta_M^*} - Y_{t_i-}^{\theta_M^*}\right|_2\right)^2\right] \\
&\leq \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left(\left|\hat{X}_{t_i} - Y_{t_i}^{\theta_M^*}\right|_2 + \left|Y_{t_i}^{\theta_M^*} - \hat{X}_{t_i}\right|_2 + \left|\hat{X}_{t_i} - \hat{X}_{t_i-}\right|_2 + \left|\hat{X}_{t_i-} - Y_{t_i-}^{\theta_M^*}\right|_2\right)^2\right] \\
&\leq \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left(\varepsilon + \varepsilon + \left|X_{t_i} - \hat{X}_{t_i-}\right|_2 + \varepsilon\,(T+1)\right)^2\right] \\
&\leq \varepsilon^2\,(T+3)^2 + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left(2\varepsilon\,(T+3)\left|X_{t_i} - \hat{X}_{t_i-}\right|_2 + \left|X_{t_i} - \hat{X}_{t_i-}\right|_2^2\right)\right] \\
&\leq \varepsilon^2\,(T+3)^2 + 2\varepsilon\,(T+3)\,\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left|X_{t_i} - \hat{X}_{t_i-}\right|_2\right] + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \left|X_{t_i} - \hat{X}_{t_i-}\right|_2^2\right] \\
&\leq \varepsilon^2\,(T+3)^2 + 2\varepsilon\,(T+3)\left(\min_{Z\in\mathbb{D}}\Psi(Z) + 1\right) + \min_{Z\in\mathbb{D}}\Psi(Z) \xrightarrow{\varepsilon\to 0} \min_{Z\in\mathbb{D}}\Psi(Z),
\end{aligned}$$

where we used $\Psi(\hat{X}) = \min_{Z\in\mathbb{D}}\Psi(Z)$ and that for $a \in \mathbb{R}$, $a \leq a^2 + 1$ in the last step. It is clear from the construction before, that $\varepsilon \to 0$ as $M \to \infty$.

In the last step we show that the limits $\lim_{M\to\infty} Y^{\theta_M^{\min}}$ and $\lim_{M\to\infty} Y^{\theta_M^*}$ exist as limits in the Banach space $\mathbb{L} := L^1(\Omega \times [0,T], \mathbb{F}\otimes\mathcal{B}([0,T]), \mathbb{P}\times\lambda_k)$, for every $k \leq K$, and that they are both equal to $\hat{X}$. Let us fix $k \leq K$. First we note that for every $B \in \mathcal{B}([0,T])$ we have

$$\mathbb{E}_{\lambda_k}[\mathbf{1}_B] = \lambda_k(B) = \frac{\tilde{\mathbb{P}}(n\geq k, t_k - \in B)}{\tilde{\mathbb{P}}(n\geq k)} = \frac{\mathbb{E}_{\tilde{\mathbb{P}}}[\mathbf{1}_{\{n\geq k\}}\mathbf{1}_{\{t_k-\in B\}}]}{\tilde{\mathbb{P}}(n\geq k)}.$$

Using "measure theoretic induction" (Durrett, 2010, Case 1-4 of Proof of Thm. 1.6.9) this yields for $c := (\mathbb{P}(n \geq k))^{-1}$ and a $\mathcal{B}([0,T])$-measurable function $Z : [0,T] \to \mathbb{R}, t \mapsto Z_t := Z(t)$ that

$$\mathbb{E}_{\lambda_k}[Z] = c\,\mathbb{E}_{\tilde{\mathbb{P}}}[\mathbf{1}_{\{n\geq k\}} Z_{t_k-}]. \tag{31}$$

Moreover, the triangle inequality and Lemma B.5 yield

$$\begin{aligned}
\Phi(\theta_M^*) - \Psi(\hat{X}) &\geq \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^n \left|X_{t_i} - Y_{t_i-}^{\theta_M^*}\right|_2^2\right] - \Psi(\hat{X}) \\
&= \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^n \left|\hat{X}_{t_i-} - Y_{t_i-}^{\theta_M^*}\right|_2^2\right].
\end{aligned} \tag{32}$$

For any $\mathbb{R}^{d_X}$-valued $Z \in \mathbb{L}$ the Hölder inequality, together with the fact that $n \geq 1$, yields

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[|Z|_2\right] = \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{\sqrt{n}}{\sqrt{n}}|Z|_2\right] \leq \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[n]^{1/2}\,\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}|Z|_2^2\right]^{1/2}. \tag{33}$$

Together, this implies that $\lim_{M\to\infty} Y^{\theta_M^*} = \hat{X}$ as a $\mathbb{L}$-limit. Indeed, with $\tilde{c} := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[n]^{1/2} < \infty$ we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}\times\lambda_k}\left[\left|\hat{X} - Y^{\theta_M^*}\right|_2\right] &= c\,\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\mathbf{1}_{\{n\geq k\}}\left|\hat{X}_{t_k-} - Y^{\theta_M^*}_{t_k-}\right|_2\right] \\
&\leq c\,\tilde{c}\,\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\mathbf{1}_{\{n\geq k\}}\frac{1}{n}\left|\hat{X}_{t_k-} - Y^{\theta_M^*}_{t_k-}\right|_2^2\right]^{1/2} \\
&\leq c\,\tilde{c}\,\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\mathbf{1}_{\{n\geq k\}}\frac{1}{n}\sum_{i=1}^{n}\left|\hat{X}_{t_i-} - Y^{\theta_M^*}_{t_i-}\right|_2^2\right]^{1/2} \\
&\leq c\,\tilde{c}\,\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left|\hat{X}_{t_i-} - Y^{\theta_M^*}_{t_i-}\right|_2^2\right]^{1/2} \\
&\leq c\,\tilde{c}\left(\Phi(\theta_M^*) - \Psi(\hat{X})\right)^{1/2} \xrightarrow{M\to\infty} 0,
\end{aligned}$$

where we used first (31) and (33) followed by two simply upper bounds and (32) in the last step. The same argument can be applied to show that $\lim_{M\to\infty} Y^{\theta_M^{\min}} = \hat{X}$ as a $\mathbb{L}$-limit. In particular, this proves that the limit $Y := \lim_{M\to\infty} Y^{\theta_M^{\min}}$ exists as $\mathbb{L}$-limit and equals $\hat{X}$ $(\mathbb{P}\times\lambda_k)$-almost surely, for any $k \leq K$. □

Remark G.1. *This result can be extended to any other neural network architecture for which a universal approximation theorem equivalent to (Hornik et al., 1989, Thm. 2.4) exists. Moreover, the stochastic process X defined in (1) can be chosen more general, in particular, the diffusion part $\int \sigma dW$ can be replaced by any martingale, as long as the resulting process still is a Markov process and $\tilde{\mu}$ stays continuous.*

Remark G.2. *If we used the modified loss function $\tilde{\Psi}$ which is identical to Ψ except that we drop the factor $\frac{1}{n}$, everything would work similarly and we could show L^2-convergence instead of L^1-convergence of $Y^{\theta_M^{\min}}$ to $\hat{X}$. However, we remark that there might exists $Z \in \mathbb{D}$, such that $\Psi(Z) < \infty$ while $\tilde{\Psi}(Z) = \infty$. In particular, if some moment of n does not exist, such a process can be constructed.*

Remark G.3. *It follows directly from Theorem 4.1 that $\lim_{M\to\infty} Y^{\theta_M^{\min}} = \hat{X}$ as random variables on $\Omega \times [0,T]$ except on sets which are null sets with respect to every product measure $\mathbb{P}\times\lambda_k$ for $1 \leq k \leq K$.*

Remark G.4. *The result of Theorem 4.1 does not imply that $\hat{X}$ and $\lim_{M\to\infty} Y^{\theta_M^{\min}}$ are modifications or indistinguishable. For example, if $B \subset [0,T]$ is a subset such that no left-point of the observation times (t_k-) lies in B with probability greater 0, i.e. $\lambda_k(B) = 0$ for $1 \leq k \leq K$, then Theorem 4.1 does not tell us how close $(\lim_{M\to\infty} Y^{\theta_M^{\min}})_t$ is to $\hat{X}_t$ for $t \in B$. In particular, it does not tell us whether they are equal $\mathbb{P}$-almost surely. Furthermore, such a set B always exists, since there has to exists $t \in [0,T]$ such that $B := \{t\}$ has measure 0 for all k.*

In the following corollary we show, that Theorem 4.1 can be extended to show convergence to the conditional expectation of $\varphi(X)$, for some function $\varphi \in C^{2,b}(\mathbb{R}^{d_X}, \mathbb{R})$, i.e. a function that is twice continuously differentiable with bounded derivatives.

Corollary G.5. *Let $\varphi \in C^{2,b}(\mathbb{R}^{d_X}, \mathbb{R})$, then the statement of Theorem 4.1 holds equivalently, when replacing X in the loss functions Ψ and Φ by $\Gamma = \varphi(X)$ and $\hat{X}$ by the conditional expectation $\hat{\Gamma}$, where $\hat{\Gamma}_t := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[\varphi(X_t)|\mathcal{A}_t]$.*

Corollary G.5 combined with the monotone convergence theorem for conditional expectation theoretically enables us to make statements about the conditional law and conditional moments of X under some a priori integrability assumptions.

Proof of Corollary G.5. We first remark that Proposition C.1 and Lemma B.3, B.5, C.2 hold similarly for the conditional expectation $\hat{\Gamma}$. Hence, the same argument as in the proof of Theorem 4.1 implies that $\hat{\Gamma}$ is the unique $\mathbb{A}$-adapted minimizer of Ψ.
For simplicity of the notation we assume that $d_X = d_Y = 1$, i.e. that the process X and the Brownian

motion W are 1-dimensional. However, the following works as well in the general case, where the correlations of the Brownian motion components have to be taken into account. By Itô's Formula (Protter, 2005, Chap. II, Thm. 32), $\Gamma = \varphi(X)$ is the solution of the SDE

$$\begin{aligned} d\varphi(X)_t &= \varphi'(X_t)dX_t + \tfrac{1}{2}\varphi''(X_t)d[X,X]_t \\ &= \varphi'(X_t)\mu(t,X_t)dt + \varphi'(X_t)\sigma(t,X_t)dW_t + \tfrac{1}{2}\varphi''(X_t)\sigma(t,X_t)^2dt \\ &= \alpha(t,X_t)dt + \beta(t,X_t)dW_t, \end{aligned}$$

for $\alpha(t,X_t) := \varphi'(X_t)\mu(t,X_t) + \frac{1}{2}\varphi''(X_t)\sigma(t,X_t)^2$ and $\beta(t,X_t) := \varphi'(X_t)\sigma(t,X_t)$. Defining $\tilde{\alpha}$ similar to $\tilde{\mu}$ as

$$\tilde{\alpha} : \Delta \times \mathbb{R}^{d_X} \to \mathbb{R}^{d_X}, \quad (t,r,\xi) \mapsto P_{t,t+r}(X_t,\alpha)\big|_{X_t=\xi} = \mathbb{E}_{\mathbb{P}}\left[\alpha(t+r,X_{t+r})|X_t=\xi\right],$$

one can use the boundedness of of φ' and φ'' to show that it is continuous and that

$$\hat{\Gamma}_t = \mathbb{E}[\varphi(X)_t|\mathcal{A}_t] = \Gamma_{\tau(t)} + \int_{\tau(t)}^{t} \tilde{\alpha}\left(\tau(t), s-\tau(t), X_{\tau(t)}\right) ds.$$

In particular, Proposition 2.1 and 2.2 hold equivalently for Γ.
Similar to (29) and (30), the neural network parameters can be chosen such that

$$\sup_{\tilde{u}\in\mathbb{R}^{d_X}} \left|\rho_{\theta_2^*}\left(\tanh(\tilde{u})\right) - \varphi(\tilde{u})\right|_2 = \sup_{u\in[-1,1]^{d_X}} \left|\rho_{\theta_2^*}\left(u\right) - \varphi(\tanh^{-1}(u))\right|_2 \leq \varepsilon, \tag{34}$$

and

$$\begin{aligned} &\sup_{(\tilde{u},\tilde{v},t,r)\in\mathbb{R}^{d_H\times d_X}\times\Delta} \left|f_{\theta_1^*}\left(\tanh(\tilde{u}),\tanh(\tilde{v}),t,r\right) - \tilde{\alpha}\left(\tilde{v},t,r\right)\right|_2 \\ &\qquad = \sup_{(u,v,t,r)\in[-1,1]^{d_H\times d_X}\times\Delta} \left|f_{\theta_1^*}\left(u,v,t,r\right) - \tilde{\alpha}\left(\tanh^{-1}(v),t,r\right)\right|_2 \leq \varepsilon, \end{aligned} \tag{35}$$

which then implies the statement of the Corollary similar as in the proof of Theorem 4.1. □

H PROOF OF THEOREM 4.2

The following Monte Carlo convergence analysis is based on (Lapeyre & Lelong, 2019, Section 4.3). In comparison to them, we do not need the additional assumptions that were essential in (Lapeyre & Lelong, 2019, Section 4.3), i.e. that all minimizing neural network weights generate the same neural network output. This assumption is not needed, because we do not aim to show that $Y^{\theta^{\min}_{M,N}}$ converges to $Y^{\theta^{\min}_{M}}$.

We define the separable Banach space $\mathcal{S} := \{x = (x_i)_{\in\mathbb{N}} \in \ell^1(\mathbb{R}^{d_X}) \mid \exists n \in \mathbb{N} \text{ s.t. } \forall i > n : x_i = 0\}$ with the norm $\|x\|_{\ell^1} := \sum_{i\in\mathbb{N}} |x_i|_2$.

H.1 CONVERGENCE OF OPTIMIZATION PROBLEMS

Consider a sequence of real valued function $(f_n)_n$ defined on a compact set $K \subset \mathbb{R}^d$. Define, $v_n = \inf_{x\in K} f_n(x)$ and let x_n be a sequence of minimizers $f_n(x_n) = \inf_{x\in K} f_n(x)$.

From (Rubinstein & Shapiro, 1993, Theorem A1 and discussion thereafter) we have the following Lemma.

Lemma H.1. *Assume that the sequence $(f_n)_n$ converges uniformly on K to a continuous function f. Let $v^* = \inf_{x\in K} f(x)$ and $\mathcal{S}^* = \{x \in K : f(x) = v^*\}$. Then $v_n \to v^*$ and $d(x_n, \mathcal{S}^*) \to 0$ a.s.*

The following lemma is a consequence of (Ledoux & Talagrand, 1991, Corollary 7.10) and (Rubinstein & Shapiro, 1993, Lemma A1).

Lemma H.2. *Let $(\xi_i)_{i\geq 1}$ be a sequence of i.i.d random variables with values in $\mathcal{S}$ and $h : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ be a measurable function. Here ℓ^1 denotes the space of sequences with absolutely convergent series. Assume that a.s, the function $\theta \in \mathbb{R} \mapsto h(\theta, \xi_1)$ is continuous and for all $C > 0$, $\mathbb{E}(\sup_{|\theta|_2\leq C} |h(\theta, \xi_1)|) < +\infty$. Then, a.s $\theta \in \mathbb{R}^d \mapsto \frac{1}{N}\sum_{i=1}^N h(\theta, \xi_i)$ converges locally uniformly to the continuous function $\theta \in \mathbb{R}^d \mapsto \mathbb{E}(h(\theta, \xi_1))$.*

$$\lim_{N\to\infty} \sup_{|\theta|_2\leq C} \left| \frac{1}{N}\sum_{i=1}^N h(\theta, \xi_i) - \mathbb{E}(h(\theta, \xi_1)) \right| = 0 \qquad a.s.$$

H.2 STRONG LAW OF LARGE NUMBERS

Let us define

$$F(\theta, x, y, z) := |x - y|_2 + |y - z|_2$$

and $\xi_j := (X^j_{t^j_1}, \dots, X^j_{t^j_{n^j}}, 0, \dots)$, where $X^j_{t^j_i}$ are random variables describing the realizations of the training data, as defined in Section 3.2. By this definition we have $n^j := n^j(\xi_j) := \max_{i\in\mathbb{N}}\{\xi_{j,i} \neq 0\}$ $\mathbb{P}$-almost-surely and we know that ξ_j are i.i.d. random variables taking values in $\mathcal{S}$. Furthermore, let us write $Y^\theta_t(\xi)$ to make the dependence of Y on the input explicit. Then we define

$$h(\theta, \xi_j) := \frac{1}{n^j}\sum_{i=1}^{n^j} F\left(\theta, X_{t^j_i}, Y_{t^j_i}(\xi_j), Y_{t^j_i -}(\xi_j)\right)^2.$$

Lemma H.3. *The following properties are satisfied.*

- ($\mathcal{P}_1$) *There exists $\kappa > 0$ such that for all $S \in \mathcal{S}$ and $\theta \in \tilde{\Theta}_M$ we have $|Y^\theta_t(S)|_2 \leq \kappa\, n(S)$ for all $t \in [0, T]$, where $n(S) = \max_{i\in\mathbb{N}}\{S_i \neq 0\}$.*
- ($\mathcal{P}_2$) *Almost-surely the random function $\theta \in \tilde{\Theta}_M \mapsto Y^\theta_t$ is uniformly continuous for every $t \in [0, T]$.*
- ($\mathcal{P}_3$) *We have $\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n |X_{t_i}|_2^2\right] < \infty$ and $\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n |X_{t_i-}|_2^2\right] < \infty$.*

Proof. By definition of the neural networks with sigmoid activation functions (in particular having bounded outputs), all neural network outputs are bounded in terms of the norm of the network weights,

which is assumed to be bounded, not depending on the norm of the input. Since Y jumps at most $n(S)$ times, we can find κ depending on T, such that the claimed bound is satisfied for all t, proving $(\mathcal{P}_1)$.

Since the activation functions are continuous, also the neural networks are continuous with respect to their weights θ, which implies that also $\theta \in \tilde{\Theta}_M \mapsto Y_t^\theta$ is continuous. Since $\tilde{\Theta}_M$ is compact, this automatically yields uniform continuity and therefore finishes the proof of $(\mathcal{P}_2)$.

From our assumption that X is square integrable on the closed interval $[0,T]$, it follows that $\max_{0\le t\le T} \mathbb{E}_{\mathbb{P}}[|X_t|^2] < \infty$. Hence, by independence we have

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n |X_{t_i}|_2^2\right] = \mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}}[|X_{t_i}|_2^2]\right] \le \mathbb{E}_{\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n \max_{0\le t\le T} \mathbb{E}_{\mathbb{P}}[|X_t|_2^2]\right] = \max_{0\le t\le T} \mathbb{E}_{\mathbb{P}}[|X_t|_2^2] < \infty,$$

and the same arguments shows $\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n |X_{t_i-}|_2^2\right] < \infty$, therefore also $(\mathcal{P}_2)$ is proven. □

Proof of Theorem 4.2. We apply Lemma H.2 to the sequence of $i.i.d$ random function $h(\theta,\xi_j)$. From $(\mathcal{P}_1)$ of Lemma H.3 we have that

$$\begin{aligned} F(\theta, X_{t_i^j}, Y_{t_i^j}, Y_{t_i^j-})^2 &= \left(\left|X_{t_i^j} - Y^\theta_{t_i^j}\right|_2 + \left|Y^\theta_{t_i^j} - Y^\theta_{t_i^j-}\right|_2\right)^2 \\ &\le 4\left(|X_{t_i^j}|_2^2 + |Y^\theta_{t_i^j}|_2^2 + |Y^\theta_{t_i^j}|_2^2 + |Y^\theta_{t_i^j-}|_2^2\right) \\ &\le 4\left(|X_{t_i^j}|_2^2 + 3\kappa(n^j)^2\right). \end{aligned}$$

Hence, we obtain that

$$h(\theta,\xi_j) = \frac{1}{n^j}\sum_{i=1}^{n^j} F(\theta, X_{t_i^j}, Y_{t_i^j}, Y_{t_i^j-})^2 \le 12\kappa\,(n^j)^2 + \frac{4}{n^j}\sum_{i=1}^{n^j}\left|X_{t_i^j}\right|_2^2,$$

implying that

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\sup_{\theta\in\tilde{\Theta}_M} h(\theta,\xi_j)\right] \le 12\kappa\,\mathbb{E}_{\tilde{\mathbb{P}}}\left[n^2\right] + 4\,\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^n |X_{t_i}|_2^2\right] < \infty, \tag{36}$$

using $(\mathcal{P}_3)$ of Lemma H.3 and the stronger assumption that n is square integrable. By $(\mathcal{P}_2)$ of Lemma H.3, the function $\theta \mapsto h(\theta)$ is continuous. Therefore, we can apply Lemma H.2, yielding that almost-surely for $N \to \infty$ the function

$$\theta \mapsto \frac{1}{N}\sum_{j=1}^N h(\theta,\xi_j) = \hat{\Phi}_N(\theta) \tag{37}$$

converges uniformly on $\tilde{\Theta}_M$ to

$$\theta \mapsto \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[h(\theta,\xi_1)] = \Phi(\theta). \tag{38}$$

We deduce from Lemma H.1 that $d(\theta^{\min}_{M,N}, \Theta^{\min}_M) \to 0$ a.s. when $N \to \infty$. Then there exists a sequence $(\hat{\theta}^{\min}_{M,N})_{N\in\mathbb{N}}$ in $\Theta^{\min}_M$ such that $|\theta^{\min}_{M,N} - \hat{\theta}^{\min}_{M,N}|_2 \to 0$ a.s. for $N \to \infty$. The uniform continuity of the random functions $\theta \mapsto Y_t^\theta$ on $\tilde{\Theta}_M$ implies that $|Y_t^{\theta^{\min}_{M,N}} - Y_t^{\hat{\theta}^{\min}_{M,N}}|_2 \to 0$ a.s. when $N \to \infty$ for all $t \in [0,T]$. By continuity of F this yields $|h(\theta^{\min}_{M,N},\xi_1) - h(\hat{\theta}^{\min}_{M,N},\xi_1)|_2 \to 0$ a.s. as $N \to \infty$. With (36) we can apply dominated convergence which yields

$$\lim_{N\to\infty} \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[|h(\theta^{\min}_{M,N},\xi_1) - h(\hat{\theta}^{\min}_{M,N},\xi_1)|_2\right] = 0.$$

Since for every integrable random variable Z we have $0 \le |\mathbb{E}[Z]|_2 \le \mathbb{E}[|Z|_2]$ and since $\hat{\theta}^{\min}_{M,N} \in \Theta^{\min}_M$ we can deduce

$$\lim_{N\to\infty} \Phi(\theta^{\min}_{M,N}) = \lim_{N\to\infty} \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[h(\theta^{\min}_{M,N},\xi_1)\right] = \lim_{N\to\infty} \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[h(\hat{\theta}^{\min}_{M,N},\xi_1)\right] = \Phi(\theta^{\min}_M). \tag{39}$$

Now by triangle inequality,

$$|\hat{\Phi}_N(\theta^{\min}_{M,N}) - \Phi(\theta^{\min}_M)| \leq |\hat{\Phi}_N(\theta^{\min}_{M,N}) - \Phi(\theta^{\min}_{M,N})| + |\Phi(\theta^{\min}_{M,N}) - \Phi(\theta^{\min}_M)|. \tag{40}$$

(37), (38) and (39) imply that both terms on the right hand side converge to 0 when $N \to \infty$, which finishes the proof of the first part of the Theorem.

We define $N_0 := 0$ and for every $M \in \mathbb{N}$

$$N_M := \min\left\{N \in \mathbb{N} \mid N > N_{M-1}, |\Phi(\theta^{\min}_{M,N}) - \Phi(\theta^{\min}_M)| \leq \tfrac{1}{M} + |\Phi(\theta^{\min}_M) - \Psi(\hat{X})|\right\},$$

which is possibly due to (39). Then Theorem 4.1 implies that

$$|\Phi(\theta^{\min}_{M,N_M}) - \Psi(\hat{X})| \leq \tfrac{1}{M} + 2|\Phi(\theta^{\min}_M) - \Psi(\hat{X})| \xrightarrow{M\to\infty} 0.$$

Therefore, we can apply the same arguments as in the proof of Theorem 4.1 (starting from (32)) to show that

$$\mathbb{E}_{\mathbb{P}\times\lambda_k}\left[\left|\hat{X} - Y^{\theta^{\min}_{M,N_M}}\right|_2\right] \leq c\,\tilde{c}\,\left(\Phi(\theta^{\min}_{M,N_M}) - \Psi(\hat{X})\right)^{1/2} \xrightarrow{M\to\infty} 0,$$

for every $1 \leq k \leq K$. □

Corollary H.4. *In the setting of Theorem 4.2, we also have that* $(\mathbb{P}\times\tilde{\mathbb{P}})$*-a.s.*

$$\Phi(\theta^{\min}_{M,N_M}) \xrightarrow{N\to\infty} \Psi(\hat{X}) \quad \textit{and} \quad \hat{\Phi}_{\tilde{N}_M}(\theta^{\min}_{M,\tilde{N}_M}) \xrightarrow{N\to\infty} \Psi(\hat{X}),$$

where $(\tilde{N}_M)_{M\in\mathbb{N}}$ *is another increasing sequence in* $\mathbb{N}$.

Proof. The first convergence result was already shown in the proof of Theorem 4.2 and the second one can be shown similarly, when defining $\tilde{N}_M$ by $\tilde{N}_0 := 0$ and for every $M \in \mathbb{N}$

$$\tilde{N}_M := \min\left\{N \in \mathbb{N} \mid N > \tilde{N}_{M-1}, |\hat{\Phi}_N(\theta^{\min}_{M,N}) - \Phi(\theta^{\min}_M)| \leq \tfrac{1}{M} + |\Phi(\theta^{\min}_M) - \Psi(\hat{X})|\right\},$$

which is possibly due to (40). □

I ALGORITHM

The proposed NJ-ODE can be trained with the empirical loss function presented in Section 3.2. Once the training is done, the following algorithm can be used to compute the conditional expectation on the fly. In particular, the conditional expectation can be computed at arbitrary time points while the process X is not observed.

Algorithm 1 The NJ-ODE – evaluation of a trained model

Input: Data points with timestamps $\{(x_i, t_i)\}_{i=0\dots n}$, intermediate evaluation times $t_{i-1} < s_1^i < \dots < s_{k_i}^i < t_i$, for $k_i \in \mathbb{N}$, $1 \leq i \leq n$
Initialize $\mathbf{h_{t_0}} = 0$
for $i = 1$ **to** n **do**
  for $j = 1$ **to** k_i **do**
    $\mathbf{h_{s_j^i}} = \text{ODESolve}(f_\theta, \tanh(\mathbf{h_{t_{i-1}}}), \tanh(x_{i-1}), (t_{i-1}, s_j^i))$
  end for
  $\mathbf{h_{t_i-}} = \text{ODESolve}(f_\theta, \tanh(\mathbf{h_{t_{i-1}}}), \tanh(x_{i-1}), (t_{i-1}, t_i))$
  $\mathbf{h_{t_i}} = \rho_\theta(\tanh(x_i))$
end for
$o_{i,j} = \text{outputNN}(\mathbf{h_{s_j^i}})$, $o_i = \text{outputNN}(\mathbf{h_{t_i}})$ for all $1 \leq j \leq k_i$, $1 \leq i \leq n$
return $\{o_i\}_{i=1\dots n}$

Figure 5: Pseudocode of the neural jump ODE. Differences from the ODE-RNN (Rubanova et al., 2019) are highlighted in blue.

J EXPERIMENTAL DETAILS

All implementations were done using PyTorch (Paszke et al., 2019). The code is available.

J.1 IMPLEMENTATION DETAILS

Dataset. For each of the SDE models (Black-Scholes, Ornstein-Uhlenbeck, Heston) a dataset was generated by sampling $N = 20'000$ paths of the SDE using the Euler-scheme. We used an equidistant time grid of mesh 0.01 between time 0 and $T = 1$. Independently for each path, observation times were sampled from $\mathbb{P}_t$, by using each of the grid points with probability 0.1 as an observation time. In particular, $n \sim \mathrm{Bin}(100, 0.1)$, $t_0 = 0$ and the observation times $\{t_i\}_{1 \leq i \leq n}$ were chosen uniformly on the time grid. Hence, 10% of the grid points were used on average. For each of these datasets, the samples were used in a $80\%/20\%$ split for training and testing. The SDEs of the dataset models and the chosen parameters are described below.

- **Black-Scholes:**
  - SDE: $dX_t = \mu X_t dt + \sigma X_t dW_t$, where W is a 1-dimensional Brownian motion
  - conditional expectation: $E(X_{t+s}|X_t) = X_t e^{\mu s}$
  - used parameters: $\mu = 2$, $\sigma = 0.3$, $X_0 = 1$
- **Ornstein-Uhlenbeck:**
  - SDE: $dX_t = -k(X_t - m)dt + \sigma dW_t$, where W is a 1-dimensional Brownian motion
  - conditional expectation: $E(X_{t+s}|X_t) = X_t e^{-ks} + m\left(1 - e^{-ks}\right)$
  - used parameters: $k = 2, m = 4, \sigma = 0.3, X_0 = 1$
- **Heston:**
  - SDE: for W and Z 1-dimensional Brownian motions

$$dX_t = \mu X_t dt + \sqrt{v_t} X_t dW_t$$
$$dv_t = -k(v_t - m)dt + \sigma\sqrt{v_t} dZ_t$$

  - conditional expectation[8]: $E(X_{t+s}|X_t) = X_t e^{\mu s}$
  - used parameters: $\mu = 2$, $\sigma = 0.3$, $X_0 = 1$ $k = 2$, $m = 4$, $v_0 = 4$, $\rho = \mathrm{Corr}(W, Z) = 0.5$

Architecture. In our experiments we choose the dimension of the latent variable to be $d_H = 10$. For f_{θ_1}, $\tilde{g}_{\theta_3}$ and $\tilde{\rho}_{\theta_2}$ we use 2-hidden-layer feed-forward neural networks, with 50 nodes in each hidden layer and $\tanh$ activation functions. Then the neural networks g_{θ_3} and ρ_{θ_2} are defined as residual versions of $\tilde{g}_{\theta_3}$ and $\tilde{\rho}_{\theta_2}$, by adding a residual shortcut between the input and the output of the neural networks. Moreover, also to the input of g_{θ_3} a $\tanh$ was applied, for reasons of symmetry. Dropout was applied after each non-linearity with a rate of 0.1. As we applied $\tanh$ to the possibly unbounded inputs x and h in every neural network, there was no further need to normalize or scale the data. To solve the ODE of the neural ODE part, the simple Euler-method was used.

Training. The neural networks were trained using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 and weight decay 0.0005 for 200 epochs using a batch size of 200. A random initialization was used and no hyper-parameter optimization was needed.

[8] see (Rujivan & Zhu, 2012, Equation 2.9)

J.2 FURTHER RESULTS

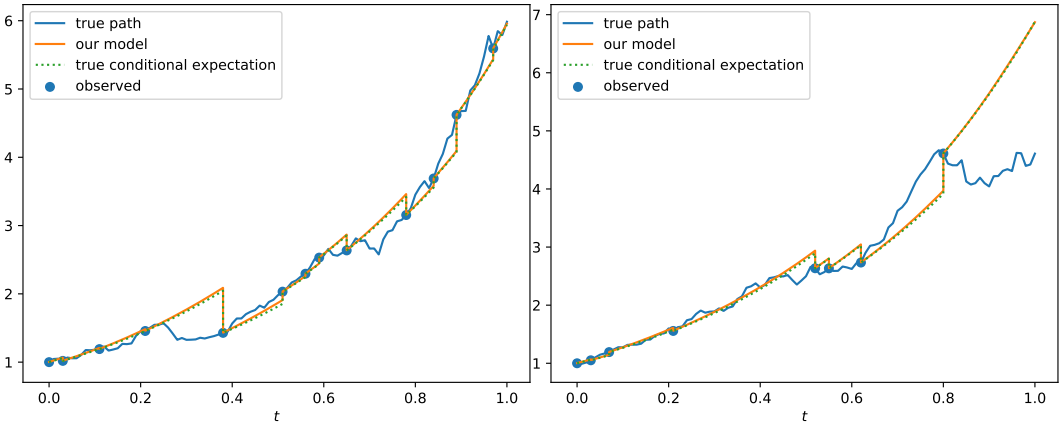

Figure 6: Black-Scholes

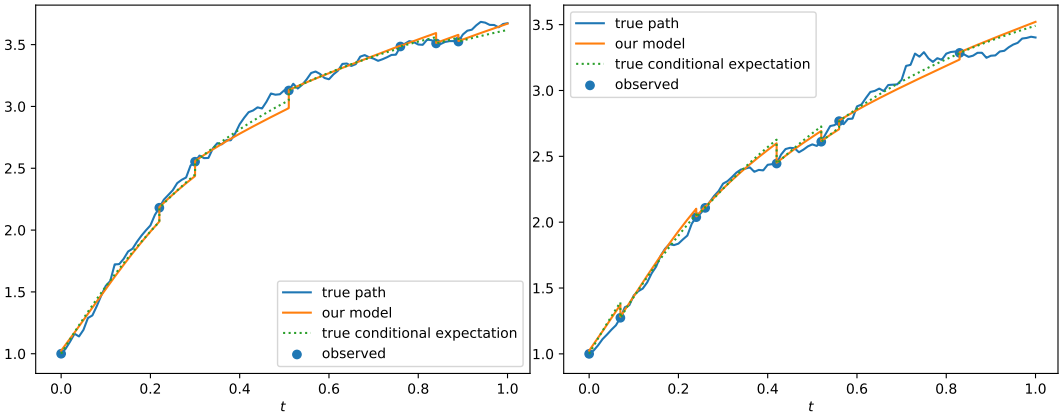

Figure 7: Ornstein-Uhlenbeck

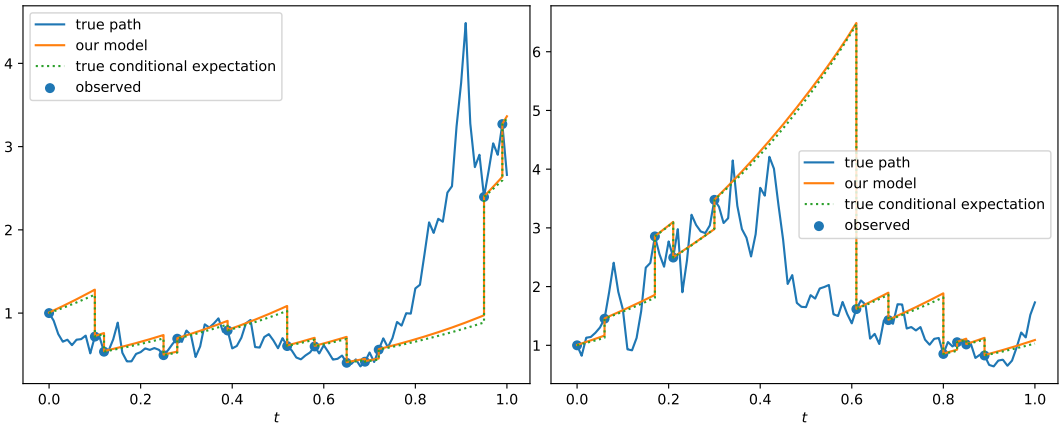

Figure 8: Heston

J.3 DETAILS ON CONVERGENCE STUDY

Evaluation metric. For the sampling time grid with equidistant step size $\Delta_t := \frac{T}{\nu}$, $\nu \in \mathbb{N}$, on $[0, T]$ and the true and predicted conditional expectation for path $j \in \mathbb{N}$, $\hat{X}^j$ and Y^j respectively, we define the evaluation metric as

$$\text{eval}(\hat{X}, Y) := \frac{1}{N_2} \sum_{j=1}^{N_2} \frac{1}{\nu + 1} \sum_{i=0}^{\nu} \left(\hat{X}^j_{i\Delta_t} - Y^j_{i\Delta_t} \right)^2, \tag{41}$$

where N_2 is the number of test samples, and j accordingly iterates over the paths in the test set.

Increasingly big training sets. We use the following procedure to create increasingly big training sets, while keeping the *exactly same* test set for evaluation. Out of the initial $20'000$ paths, we take $N_2 := 4'000$ paths, which are fixed as the testing set. Out of the remaining $16'000$ paths, we randomly choose N_1 training paths for $N_1 \in \{200, 400, 800, 1'600, 3'200, 6'400, 12'800\}$.

Increasing neural network sizes. The increasingly big neural networks are defined as follows. For all involved networks, we use the feed-forward 2-layer architecture with $\tanh$ activations (cf. Appendix J.1), where each hidden layer has the same size M for $M \in \{10, 20, 40, 80, 160, 320\}$.

Results. In accordance with the theoretical results in Theorem 4.1 and 4.2, we see that the evaluation metric decreases when N_1 and M increase. It is important to notice, that already a quite small number of samples can lead to a good approximation of the conditional expectation, if the network is big enough. In particular, the Monte Carlo approximation of the theoretical loss function is good already with a few samples. On the other hand, even with a large number of samples, the evaluation metric does not become so small, if the network size is not big enough. From a practitioners point of view, this is good news, since increasing the network size is often much easier than collecting more training data.

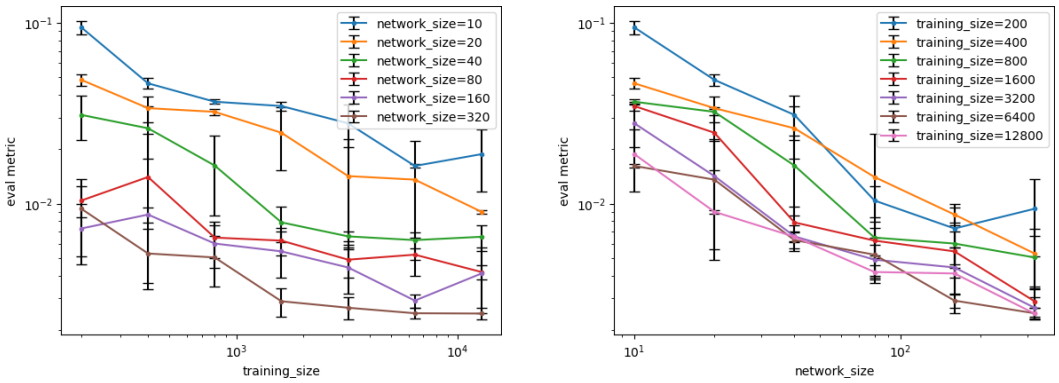

Figure 9: Mean $\pm$ standard deviation (black bars) of the evaluation metric for varying N_1 and M.

J.4 DETAILS ON COMPARISON TO GRU-ODE-BAYES (BROUWER ET AL., 2019)

GRU-ODE-Bayes (Brouwer et al., 2019). To the best of our knowledge, this is the neural network based method with the most similar task to ours. In particular, this continuous time model is trained to learn the unknown temporal parameters of a normal distribution, best describing the conditional distribution of X given the previous observations. This distribution is given by the Fokker-Planck equation. Brouwer et al. (2019) outlined, that their model can exactly represent the Fokker-Planck dynamics of the Ornstein-Uhlenbeck process, since the corresponding distribution is Gaussian.

Implementation of GRU-ODE-Bayes. We use the code of the official implementation of (Brouwer et al., 2019)[9] and slightly adjust it for our purpose. In particular, we do not use incomplete observations, hence the input mask used for this task has always only 1-entries. Furthermore, we slightly changed the scheme how the time steps are taken, to be the same as in our implementation,

[9] https://github.com/edebrouwer/gru_ode_bayes

so that comparisons can be made. Besides these minor changes, the original model is used and trained on all 3 datasets (Black-Scholes, Ornstein-Uhlenbeck and Heston). We tried out all combinations of the following parameters and always chose the best performing one for comparison to our model: $\text{impute} \in \{\text{True}, \text{False}\}$, $\text{logvar} \in \{\text{True}, \text{False}\}$, $\text{mixing} \in \{0.0001, 0.5\}$ and $\text{hiddensize} \in \{50, 100\}$, whereby phidden and prephidden were chosen to be equal to hiddensize. The first parameter choices were the ones used in the official implementation. The model was always trained using the Euler-method for solving ODEs and with $\text{dropout} = 0.1$, to be comparable to our implementation. Furthermore, we always used the full GRU-ODE-Bayes implementation, since it should be the more powerful one. For the comparison, we only use the estimated mean of the normal distribution (the estimated variance is not used), which is precisely the estimated conditional expectation of the model.

Implementation of NJ-ODE. Same as described in Section J.1. For each dataset only one model was trained, since we already saw the convergence properties before and wanted to have a qualitative comparison to GRU-ODE-Bayes. In particular, we did not try to optimize our hyper-parameters for best performance (e.g. by choosing different number of layers or neurons or different activation functions) and used about $10K$ trainable parameters compared to the best performing GRU-ODE-Bayes models of our study which used $112K$ trainable parameters.

Datasets. Same as described in Section J.1. The train and test sets were fixed to be the same for all trained models.

Training. All models were trained using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 and weight decay 0.0005 for 100 epochs using a batch size of 20. A random initialization was used.

J.4.1 Further results and discussion of the comparison

Evolution of losses and evaluation metric during training. In Figure 10 and 11 we see, that already after a few epochs the NJ-ODE model finds close to optimal weights for the given network size on the Black-Scholes and Orstein-Uhlenbeck dataset and oscillates around this optimum. The Heston dataset is considerably more difficult and the model slowly converges to close-to-optimal weights for the given network size. In comparison to this, it takes the GRU-ODE-Bayes model longer to converge to its close-to-optimal weights on the Black-Scholes and Orstein-Uhlenbeck dataset. Moreover, the model does not converge to close-to-optimal weights on the Heston dataset, but rather oscillates between bad weights. Due to some very large outliers, this is not directly visible in the plot, but can be deduced from Table 5.3.

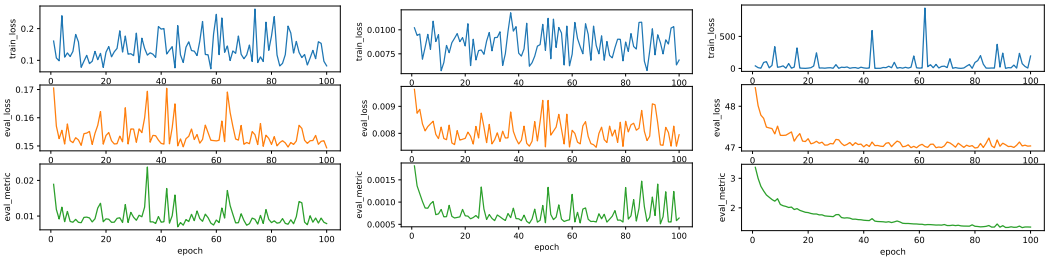

Figure 10: NJ-ODE on Black-Scholes, Ornstein-Uhlenbeck and Heston (from left to right). Blue (1st row): training loss, orange (2nd row): evaluation loss, green (3rd row): evaluation metric.

Comparison of predicted paths. For each dataset we show 5 paths that were predicted with NJ-ODE and with GRU-ODE-Bayes, first at the optimal epoch, i.e. where the test loss was minimal during training (Figures 12, 13, 14), and then at the last epoch (Figures 15, 16, 17). For the sake of comparison, in each row the performance on the same test sample is shown. The results are very similar for NJ-ODE and GRU-ODE-Bayes on the Black-Scholes and Ornstein-Uhlenbeck dataset, with sometimes the one and sometimes the other having a slightly better prediction. On the Heston dataset, the results of NJ-ODE are good predictions, being correct whenever a new observation is made and not too far away even after longer periods without observations. On the other hand, GRU-ODE-Bayes is not even correct at times of new observations and learns an incorrect behaviour

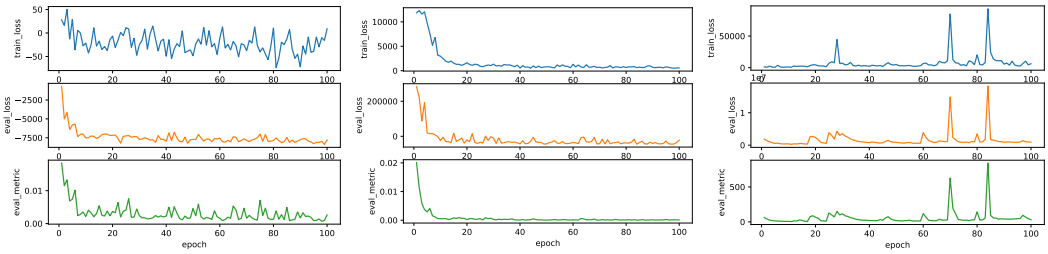

Figure 11: GRU-ODE-Bayes' best performing models on Black-Scholes, Ornstein-Uhlenbeck and Heston (from left to right). Blue (1st row): training loss, orange (2nd row): evaluation loss, green (3rd row): evaluation metric.

in between observations (e.g. making kinks where there should not be any). For the last epoch, we see that this malfunction is amplified.

J.5 Climate forecasting details

Dataset. We use the publicly available United State Historical Climatology Network (USHCN) daily dataset (Menne et al., 2016) together with all pre-processing steps as they were provided by Brouwer et al. (2019). In particular, there are 5 sporadically observed (i.e. incomplete observations) climate variables (daily temperatures, precipitation, and snow) measured at $1'114$ stations scattered over the United States during an observation window of 4 years (between 1996 and 2000) where each station has an average of 346 observations over those 4 years. For 5 folds, the data is split into train (70%), validation (20%) and test (10%) sets. The task is to predict the next 3 measurements after the first 3 years of observation. The mean squared error between the prediction and the correct values is computed on the validation and test set.

Baselines. We use the results reported in (Brouwer et al., 2019, Table 1) as baselines for our comparison and perform the exact same 5-fold cross validation using the same folds with the same train, validation and test sets. For completeness, we give all results of the table together with our results in Table **??**. We only show the mean squared error (MSE), since our model does not provide the negative log-likelihood.

Implementation of NJ-ODE. We once use the architecture described in Section J.1 (S) and once use the exact same architecture, but with hidden size $d_H = 50$ and 400 instead of 50 nodes in each hidden layer (L). To deal with the incomplete observations, the self-imputation scheme described in Section 5.4 is used. In particular, the data is self-imputed and passed together with the observation mask as input to the network $\tilde{\rho}_{\theta_2}$. Moreover, the network f_{θ_1} uses the NJ-ODE output at the last observation time instead of the last observation as input.

Training. Both versions of NJ-ODE were trained using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001 and weight decay 0.0005 for 100 epochs using a batch size of 100. A random initialization was used. The performance on the validation set was used for early stopping after the first 100 epochs, i.e. early stopping was possible at any epoch between $\{101, \dots, 200\}$.

Results. The results of our model and all models reported in (Brouwer et al., 2019, Table 1) are shown in Table 5.4.

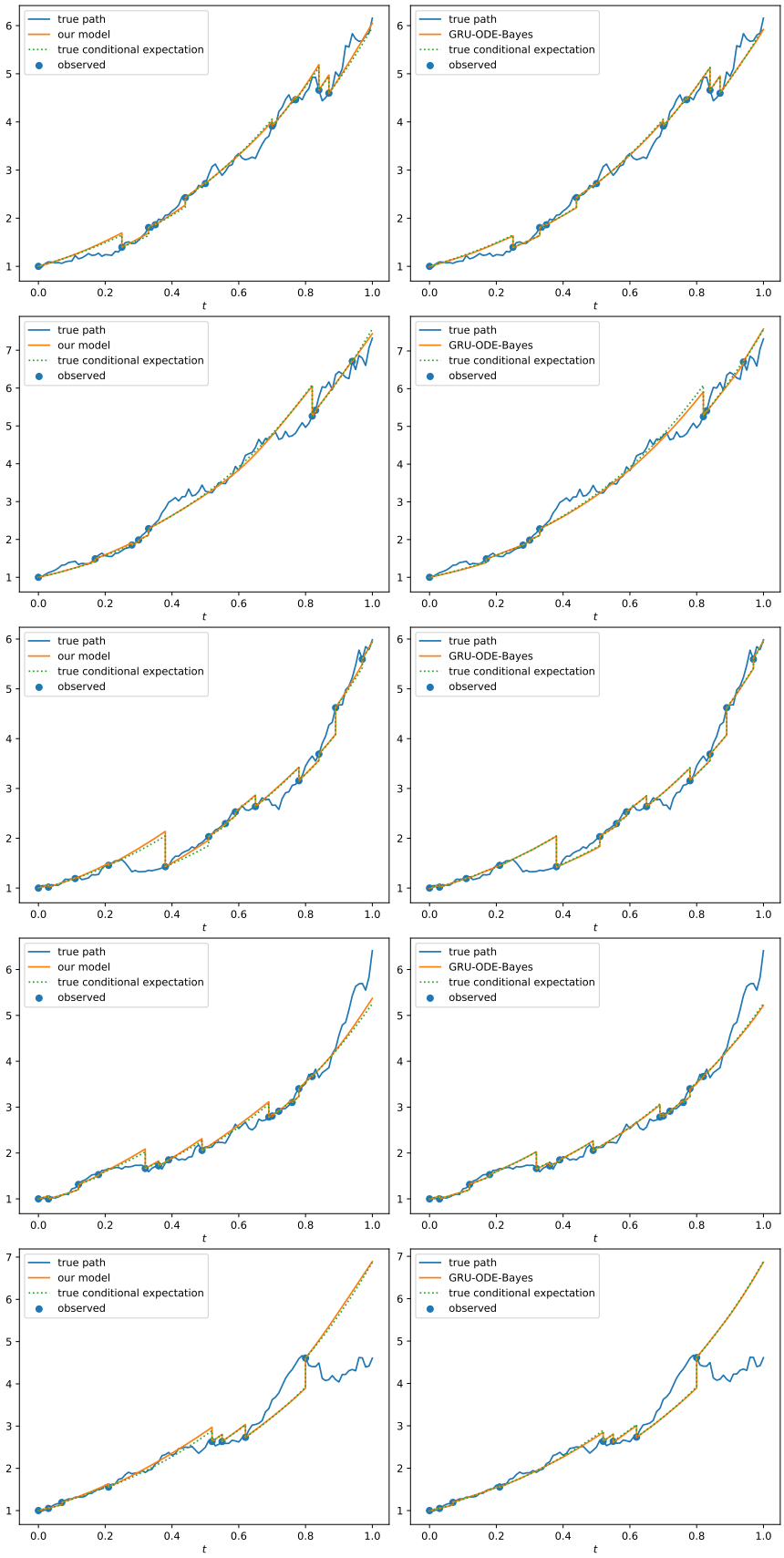

Figure 12: Comparison of predictions for Black-Scholes paths at best epoch.

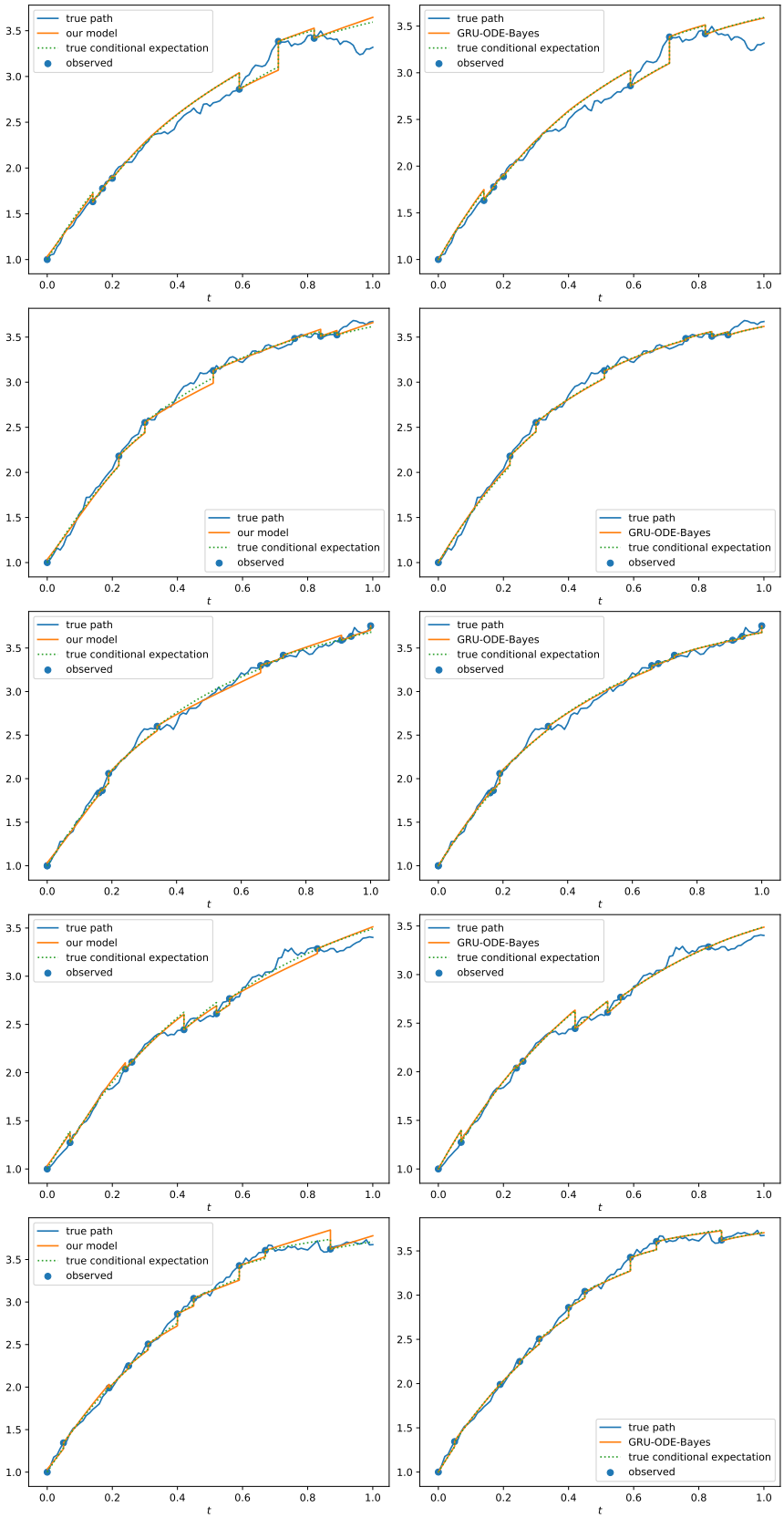

Figure 13: Comparison of predictions for Ornstein-Uhlenbeck paths at best epoch.

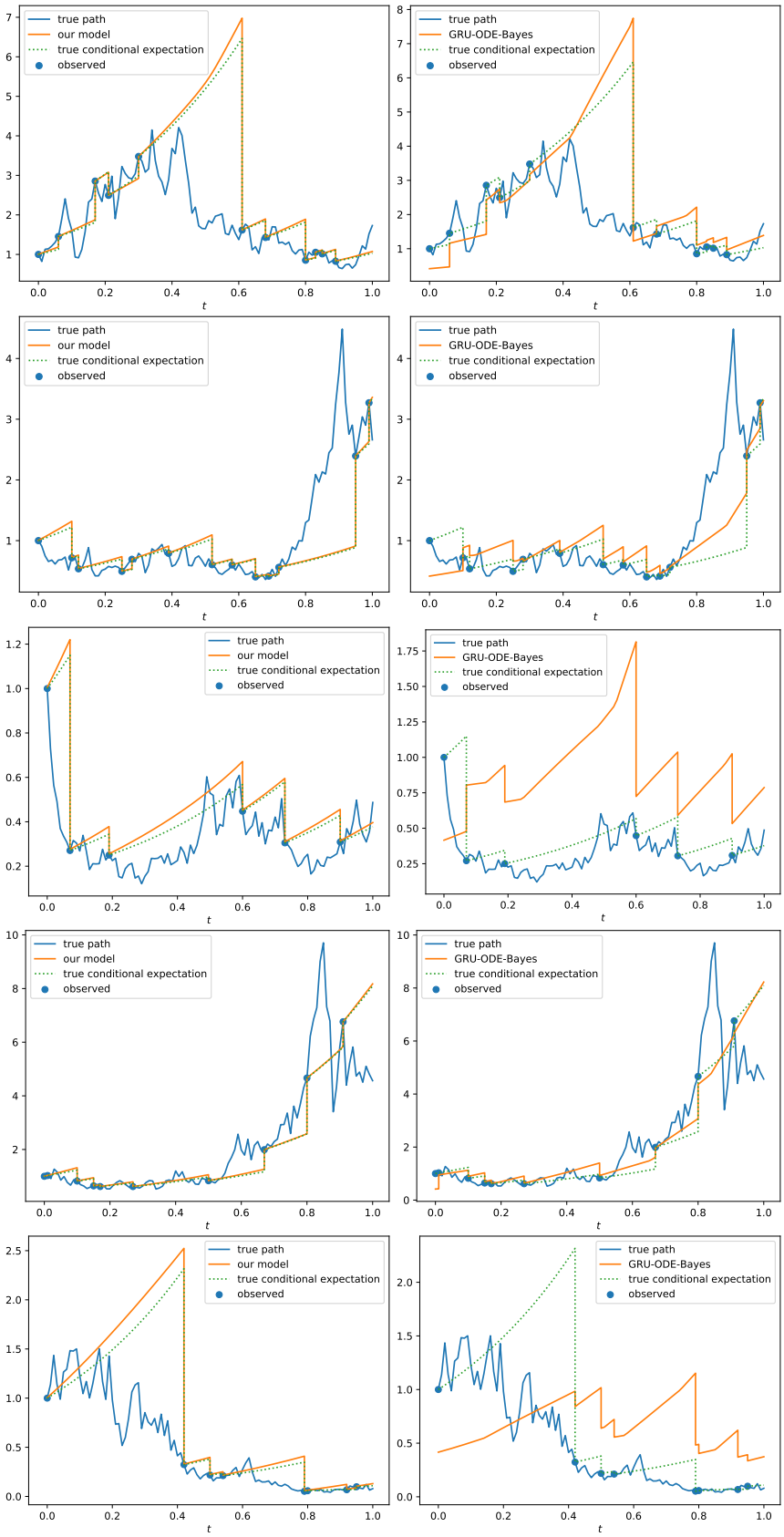

Figure 14: Comparison of predictions for Heston paths at best epoch.

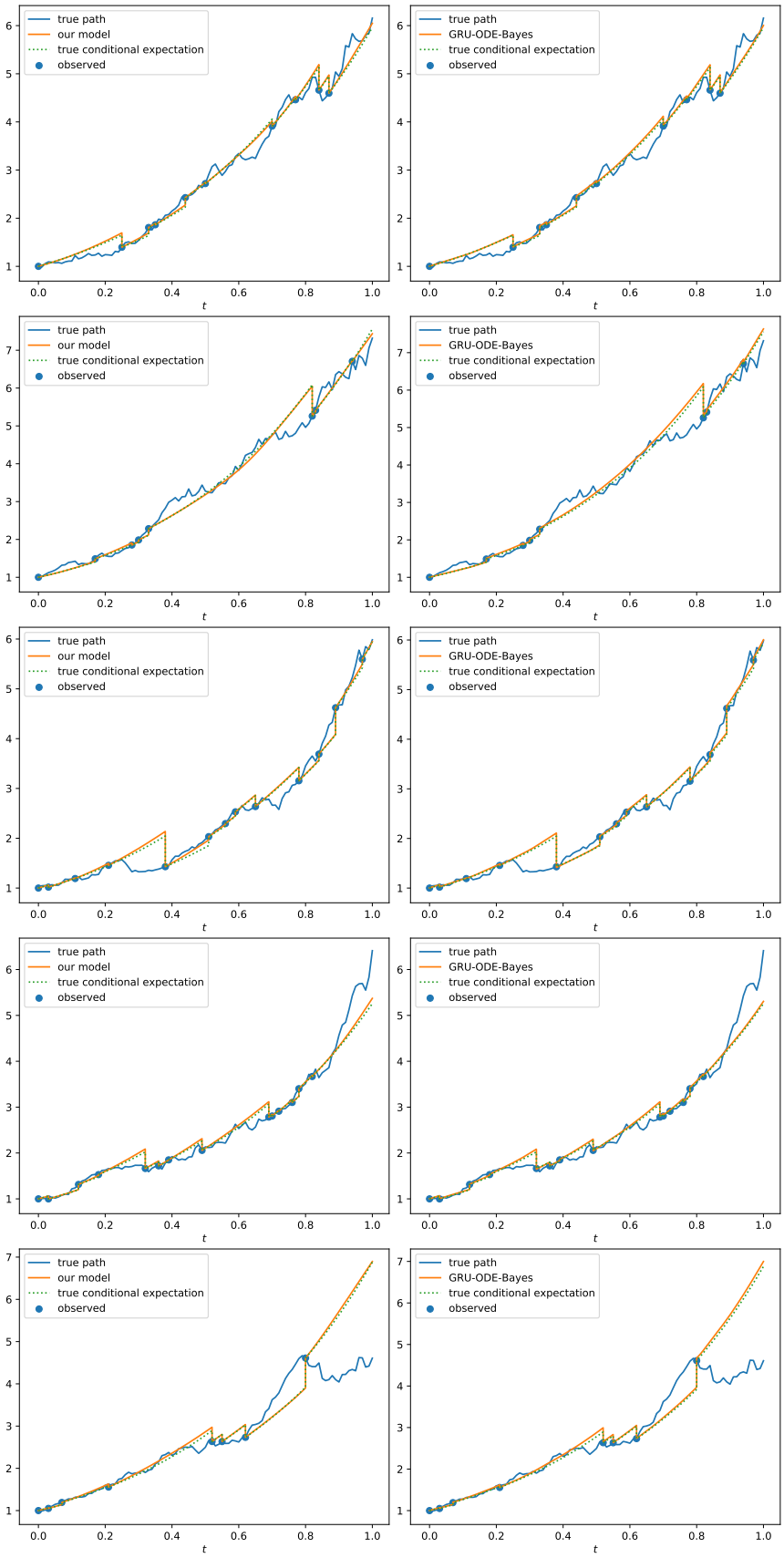

Figure 15: Comparison of predictions for Black-Scholes paths at last epoch.

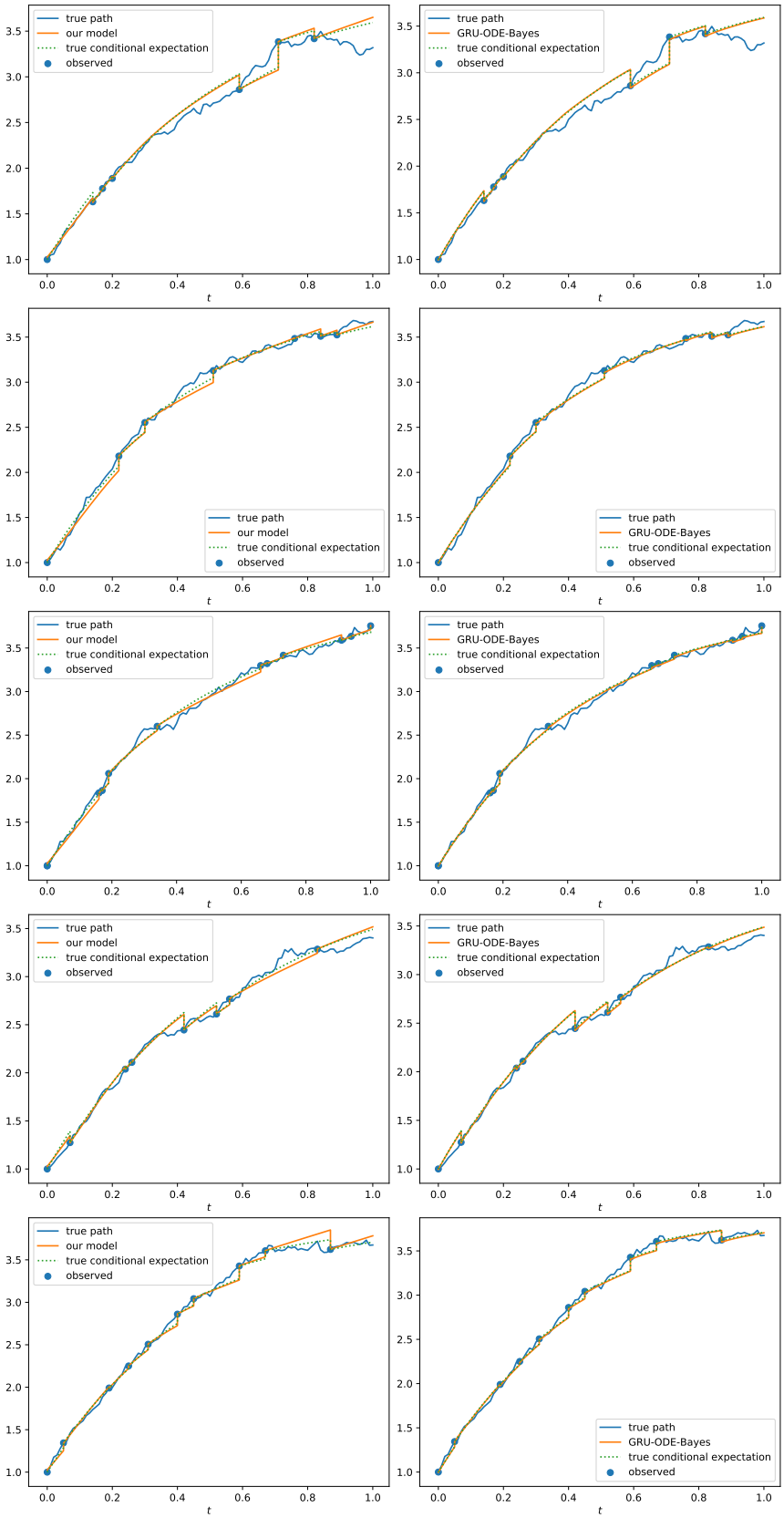

Figure 16: Comparison of predictions for Ornstein-Uhlenbeck paths at last epoch.

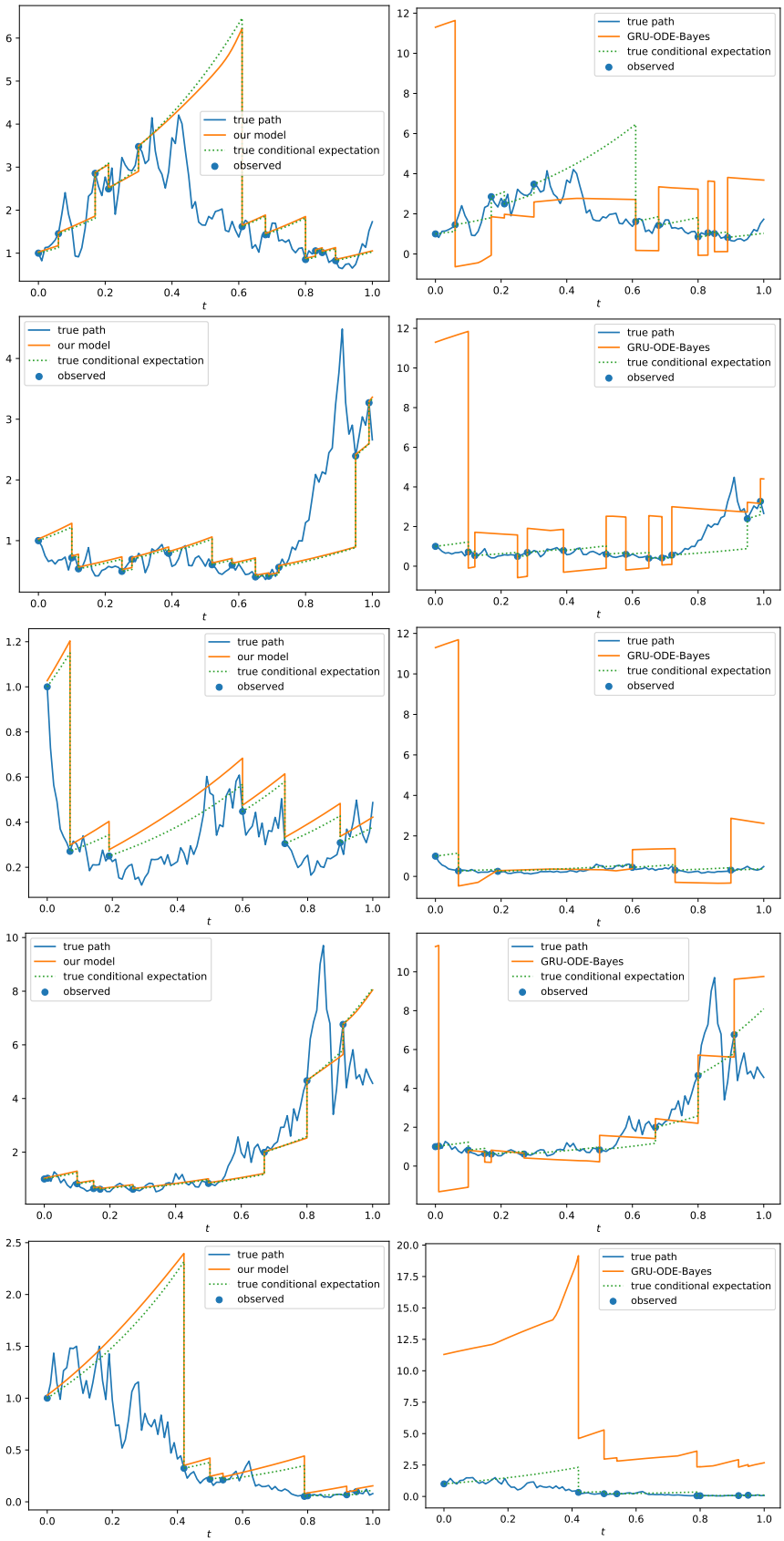

Figure 17: Comparison of predictions for Heston paths at last epoch.