Extending Path-Dependent NJ-ODEs to Noisy Observations and a Dependent Observation Framework

William Andersson  
Department of Computer Science  
ETH Zurich, Switzerland  
anwillia@ethz.ch

Jakob Heiss  
Department of Mathematics  
ETH Zurich, Switzerland  
jakob.heiss@math.ethz.ch

Florian Krach  
Department of Mathematics  
ETH Zurich, Switzerland  
florian.krach@math.ethz.ch

Josef Teichmann  
Department of Mathematics  
ETH Zurich, Switzerland  
josef.teichmann@math.ethz.ch

Reviewed on OpenReview: https://openreview.net/forum?id=0T2OTVCCC1

Abstract

The Path-Dependent Neural Jump Ordinary Differential Equation (PD-NJ-ODE) (Krach et al., 2022) is a model for predicting continuous-time stochastic processes with irregular and incomplete observations. In particular, the method learns optimal forecasts given irregularly sampled time series of incomplete past observations. So far the process itself and the coordinate-wise observation times were assumed to be independent and observations were assumed to be noiseless. In this work we discuss two extensions to lift these restrictions and provide theoretical guarantees as well as empirical examples for them. In particular, we can lift the assumption of independence by extending the theory to much more realistic settings of conditional independence without any need to change the algorithm. Moreover, we introduce a new loss function, which allows us to deal with noisy observations and explain why the previously used loss function did not lead to a consistent estimator.

1 Introduction

While the online prediction\footnote{With online prediction we mean that we use the currently available information to predict until we get new information. As soon as new information becomes available, it is part of the available information and therefore taken into account for subsequent predictions.} of regularly observed or sampled time series is a classical machine learning problem that can be solved with recurrent neural networks (RNNs) as proven e.g. by Schäfer & Zimmermann (2006), the forecasting of continuous-time processes with irregular observation has long been an unsolved problem. The Neural Jump ODE (NJ-ODE) (Herrera et al., 2021) was the first framework with theoretical guarantees to converge to the optimal prediction in this setting. However, it was restricted to Markovian Itô-diffusions with irregular but complete (i.e., all coordinates are observed at the same time) observations. This was heavily generalised with the Path-Dependent NJ-ODE (PD-NJ-ODE) (Krach et al., 2022), where the convergence guarantees hold for very general (non-Markovian) stochastic processes with irregular and incomplete observations. Still, the process itself and the observation framework were assumed to be independent and observations were assumed to be noisefree. In practice both of these assumptions are often unrealistic.
E.g., for medical patient data collected at a hospital irregularly over time such as Goldberger et al. (2000), measurements are never noise-free and the decision whether to make a measurement depends on the status of the patient. Therefore, the focus of this work is to lift those two restrictions. A detailed outline is given below.

1.1 Related Work

GRU-ODE-Bayes (Brouwer et al., 2019) and the latent ODE Rubanova et al. (2019) both use a model very similar to the NJ-ODE model, however, with different training frameworks. While the latent ODE can only be used for offline forecasting, GRU-ODE-Bayes is applicable to online forecasting as the NJ-ODE. However, in comparison to NJ-ODE, no theoretical guarantees exist for GRU-ODE-Bayes.

Neural controlled differential equations (NCDE) (Kidger et al., 2020; Morrill et al., 2022) and neural rough differential equations (Morrill et al., 2021) also use similar model frameworks, but their primary objective are labelling problems, i.e., the prediction a classification or regression label for the input of an irregularly sampled time series. For example, based on health parameters of a patient these models try to decide whether the patient will develop a certain disease in the future.

As explained in Krach et al. (2022), PD-NJ-ODEs can be used for stochastic filtering. Another well known model class for this problem are particle filters, also called sequential Monte Carlo methods (Maddison et al., 2017; Le et al., 2017; Corenflos et al., 2021; Lai et al., 2022). Particle filtering methods are applied in the context of state-space models (SSM), which are characterized by a discrete latent Markov process \((X_t)_{t=1}^T\) and a discrete observation process \((Z_t)_{t=1}^T\) defined on a fixed time-grid. Particle filters are used to approximate e.g. the conditional distribution of \(X_t\) given the observations \((Z_s)_{1\leq s\leq t}\), or the joint distribution of \((X_s, Z_s)_{1\leq s\leq t}\), for any \(t \geq 1\), using weighted sequential Monte Carlo samples. In our work, we allow for a much more general setting than the SSM. In particular, we allow for a continuous-time (instead of discrete-time), non-Markovian stochastic process. Since our setting allows for jumps of the process, this also includes the SSM case of a discrete-time Markov process. Moreover, in our setting, the underlying process can be observed at random, irregularly sampled, discrete observation times and the framework allows for incomplete observations, where some coordinates might not be observed. The primary goal of the PD-NJ-ODE method is to make optimal forecasts for \(X_t\) given all observations of \(X\) prior to time \(t\). As a special case (since the framework can deal with incomplete observations), this includes the filtering problem, however, in a more general setting, allowing for example to predict \(X_t\) while only having (discrete) observations of \(Z\) at randomly sampled observation times until time \(s < t\).

For further related work we refer the interested reader to the respective sections in Herrera et al. (2021) and Krach et al. (2022).

1.2 Outline of the Work

We introduce two extensions of the PD-NJ-ODE (Krach et al., 2022) that can be used separately or jointly. To highlight the needed adjustments for each of the extensions, we first recall the setup, model and results from Krach et al. (2022) (Section 2) and then introduce the respective changes in the assumptions and proofs for noisy observations (Section 3) and dependence between the underlying process and the observation framework (Section 4) separately. We focus on re-proving the main results (Krach et al., 2022, Theorem 4.1 and Theorem 4.4) in the new settings, by giving the arguments which need to be adjusted while skipping those which remain unchanged. In Appendix C we give the full proof for the most general result with both extensions, making the paper self-contained. We remark here that also the results for the conditional variance and for stochastic filtering (Krach et al., 2022, Section 5 and 6) follow in these extended settings similarly as the main results. Due to the similarity, we do not elaborate on this but leave the details to the interested reader. In Section 5 we discuss the practical implications of our main convergence result. Finally, in Section 6 we show empirically that the PD-NJ-ODE performs well in these generalised settings.
2 Recall: the PD-NJ-ODE

In this section we recall the PD-NJ-ODE framework together with the main result from Krach et al. (2022). The PD-NJ-ODE is a prediction model that can be used to learn the $L^2$-optimal prediction of a stochastic process (continuously in time), given its discrete, irregular and incomplete observations in the past. This means that the PD-NJ-ODE learns to compute the conditional expectation, which is the $L^2$-optimal prediction. Importantly, only data samples are needed to train the model. In particular, no knowledge about the dynamics of the underlying process is needed. We first give an intuitive example for the application of this model, which will be reused throughout the paper, and then discuss its technical details.

2.1 Intuitive Example Application

Suppose we have a dataset of $N$ patients $1 \leq j \leq N$. Any patient $j$ has $d_X$ medical values $X_{t,k}^j$ (such as body temperature or blood pressure) at any time $t \in [0,T]$. However, we only have (noisy) measurements of some of these coordinates at some irregular times $t_{(j)}^i$ (e.g., we measure the body temperature on one day and the blood pressure on another day and in-between we do not measure anything). Our training dataset consists of all these (noisy) measurements including their time-stamps. Based on this training dataset we train a PD-NJ-ODE which then allows us to make online forecasts for new patients based on their noisy incomplete regularly observed measurements. For example the Physionet dataset (Goldberger et al., 2000) is exactly such a dataset, but of course our method can also be applied in many other situations.

2.2 Technical Background and Mathematical Notation

We start by recalling the most relevant parts of the problem setting of the PD-NJ-ODE. For more details, please refer to Krach et al. (2022).

For $d_X \in \mathbb{N}$ and $T > 0$ we consider a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with an adapted càdlàg stochastic process $X := (X_t)_{t \in [0,T]}$ taking values in $\mathbb{R}^{d_X}$. The main goal in this paper is the predict $X$ optimally for future times, based on discrete observations of it in the past. We denote its running maximum as body temperature or blood pressure at any time $X$,

$$X_{t,k}^j = \sup_{s \leq t} X_{s,k}^j,$$

where $k$ is the $k$-th coordinate of the stochastic process $X_{t,k}^j$. The conditional expectation process of $X$, which is its $L^2$-optimal prediction (Krach et al., 2022, Proposition 2.5) and therefore the process we would like to compute, is defined as $\hat{X} = (\hat{X}_t)_{0 \leq t \leq T}$, with $\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t|A_t]$. Even though

$$\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t|A_t],$$

where $A_t := \sigma(X_{t,j},t_i,M_{t_k}|t_i \leq t, j \in \{1 \leq l \leq d_X|\mathcal{M}_{t_k,l} = 1\}),$

$$\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t|A_t],$$

where $\sigma(\cdot)$ denotes the generated $\sigma$-algebra. We note that $A_t = A_{t,t}$ for all $t \in [0,T]$. The conditional expectation process of $X$, which is its $L^2$-optimal prediction (Krach et al., 2022, Proposition 2.5) and therefore the process we would like to compute, is defined as $\hat{X} = (\hat{X}_t)_{0 \leq t \leq T}$, with $\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t|A_t]$. Even though

$$\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t|A_t],$$

where $\sigma(\cdot)$ denotes the generated $\sigma$-algebra. We note that $A_t = A_{t,t}$ for all $t \in [0,T]$. The conditional expectation process of $X$, which is its $L^2$-optimal prediction (Krach et al., 2022, Proposition 2.5) and therefore the process we would like to compute, is defined as $\hat{X} = (\hat{X}_t)_{0 \leq t \leq T}$, with $\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t|A_t]$. Even though

$$\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t|A_t],$$

where $\sigma(\cdot)$ denotes the generated $\sigma$-algebra. We note that $A_t = A_{t,t}$ for all $t \in [0,T]$. The conditional expectation process of $X$, which is its $L^2$-optimal prediction (Krach et al., 2022, Proposition 2.5) and therefore the process we would like to compute, is defined as $\hat{X} = (\hat{X}_t)_{0 \leq t \leq T}$, with $\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t|A_t]$. Even though
this solution to our prediction problem is abstractly well understood, it is not clear how to actually compute it, especially if one only observes data samples without knowing the distribution of $X$. The PD-NJ-ODE is designed to do exactly this. To compute the conditional expectation, it uses the powerful framework of signatures to convert all information available from the past observations into a tractable feature-vector. This is needed, since the underlying process is not assumed to be Markovian. Since the signature is defined for (continuous) paths, the first step is to construct a continuous interpolation of the discrete observations, the \textit{interpolated observation process}, in such a way that it carries the same information as the discrete observations. In particular, for any $0 \leq t \leq T$ the $j$-th coordinate of the interpolated observation process $\tilde{X}^{\leq t} \in \mathbb{R}^{2dr+1}$ at time $0 \leq s \leq T$ is defined by

$$\tilde{X}^{\leq t}_{s,j} := \begin{cases} X_{t_{a}(s,t),j} \mathbb{I}_{t_{a}(s,t)-\Delta t_{a}(s,t)} \leq s + X_{t_{a}(s,t),j} \mathbb{I}_{t_{a}(s,t)-\Delta t_{a}(s,t)}-1, & \text{if } t_{b}(s,t)-1 < s \leq t_{b}(s,t) \\ X_{t_{a}(s,t),j}, & \text{if } 1 \leq j \leq d_{X} \\ \bar{u}_{t_{a}(s,t),j} - d_{X}, & \text{if } s \leq t_{b}(s,t)-1 \text{ and } 1 \leq j \leq d_{X} \\ \sum_{k=0}^{K} M_{k,j} \mathbb{I}_{t_{a}(s,t) \leq 1} - d_{X}, & \text{if } d_{X} < j \leq 2d_{X} \\ \bar{u}_{t_{a}(s,t),j} - d_{X}, & \text{if } s \leq t_{b}(s,t)-1 \text{ and } d_{X} < j \leq 2d_{X} \\ \bar{u}_{t_{a}(s,t),j} - d_{X}, & \text{if } j = 2d_{X} + 1, \end{cases}$$

where $\bar{u}_{t_{a},j} := \sum_{k=0}^{K} M_{k,j} \mathbb{I}_{t_{a} \leq 1}$ is the jump process that counts the coordinate-wise observations and

$$\begin{align*}
\hat{a}(s,t) &:= \hat{a}(s,t,j) := \max\{0 \leq a \leq n | t_{a} \leq \min(s,t), M_{a,j} = 1\} \\
\hat{b}(s,t) &:= \hat{b}(s,t,j) := \inf\{1 \leq b \leq n | s \leq t_{b} \leq t, M_{b,j} = 1\},
\end{align*}$$

with $t_{\infty} := \infty$. Simply put, $\tilde{X}^{\leq t}$ is a continuous version (without information leakage and with time-consistency) of the rectangular interpolation of the observations of $X$ and of $\bar{u}$. The paths of $\tilde{X}^{\leq t}$ belong to $BO^{c}([0, T])$, the set of continuous $\mathbb{R}^{d_X}$-valued paths of bounded variation on $[0, T]$. Most importantly, the way it is defined ensures that no information about the next observation is leaked (through the forward-looking interpolation) until the last observation time prior to it and that $\tilde{X}^{\leq t}$ is $\mathcal{A}_{t}$-measurable. Moreover, it is time-consistent in the sense that for all $r \geq t$ and $s \leq \tau(t)$ we have $\tilde{X}^{\leq s}_{s} = \tilde{X}^{\leq r}_{s}$. Clearly, $\tilde{X}_{t_{a},j}, t_{a}, M_{a,j}$ can be reconstructed from the coordinates of $\tilde{X}^{\leq t}$ for all $t_{a} \leq t$ and $j \in \{1 \leq i \leq d_{X} | M_{a,i} = 1\}$, hence, $\mathcal{A}_{t} = \sigma(\tilde{X}^{\leq t})$. Moreover, it is easy to see that $\tilde{X}^{\leq t} = \tilde{X}^{\leq \tau(t)}$, consequently $\tilde{X}_{t}$ is $\sigma(\tilde{X}^{\leq \tau(t)})$-measurable. Therefore, the Doob-Dynkin Lemma (Taraldsen, 2018, Lemma 2) implies the existence of measurable functions $F_{j} : [0, T] \times [0, T] \times BO^{c}([0, T]) \to \mathbb{R}$ such that $\tilde{X}_{t,j} = F_{j}(t, \tau(t), \tilde{X}^{\leq \tau(t)})$. In Krach et al. (2022) convergence of the PD-NJ-ODE model to the conditional expectation $\tilde{X}$ was shown under the following assumptions.

\textbf{Assumption 2.1.} We assume that:

(i) For every $1 \leq k, l \leq K$, $M_{k,l}$ is independent of $t_{l}$ and of $n$, $\hat{P}(M_{k,j} = 1) = 0$ and $M_{0,j} = 1$ for all $1 \leq j \leq d_{X}$ (every coordinate can be observed at any observation time and $X$ is completely observed at $0$) and $|M_{k}| > 0$ for every $1 \leq k \leq K$ $\hat{P}$-almost surely (at every observation time at least one coordinate is observed).

(ii) The probability that any two observation times are closer than $\epsilon > 0$ converges to 0 when $\epsilon$ does, i.e., if $\delta(\hat{\omega}) := \min_{0 \leq i < n(\hat{\omega})} |t_{i+1}(\hat{\omega}) - t_{i}(\hat{\omega})|$ then $\lim_{\epsilon \to 0} \hat{P}(\delta < \epsilon) = 0$.

(iii) Almost surely $X$ is not observed at a jump, i.e., $(P \times \hat{P})(t_{i} \in \mathcal{J} | i \leq n) = (P \times \hat{P})(A_{i} X_{i} \neq 0 | i \leq n) = 0$ for all $1 \leq i \leq K$.

(iv) $F_{j}$ are continuous and differentiable in their first coordinate $t$ such that their partial derivatives with respect to $t$, denoted by $f_{j}$, are again continuous and there exists $B > 0$ and $p \in \mathbb{N}_{\geq 1}$ such that for every $t \in [0, T]$ the functions $f_{j}, F_{j}$ are polynomially bounded in $X^{*}$, i.e.,

$$|f_{j}(\tau(t), \tilde{X}^{\leq \tau(t)})| + |F_{j}(t, \tau(t), \tilde{X}^{\leq \tau(t)})| \leq B(X^{*}_{t} + 1)^{p}.$$
(v) $X^*$ is $L^{2p}$-integrable, i.e., $\mathbb{E}[(X^*_T)^{2p}] < \infty$.

(vi) The random number of observations $n$ is integrable, i.e., $\mathbb{E}^\mathbb{P}[n] < \infty$.

**Remark 2.2.** The assumption that $|M_k| > 0$ for every $1 \leq k \leq K$ $\mathbb{P}$-almost surely is not needed in the proof. It is just added such that the denotation “observation time” is meaningful. When removing it, some of the observation times might be “pseudo” observation times, where no coordinate is actually observed.

**Remark 2.3.** It turns out that Assumption 2.1 (ii) is always satisfied in the described setting, hence, we can also leave it away. Indeed, $\lim_{\epsilon \to 0} \mathbb{P}(\delta < \epsilon) = 0$, holds, if $\mathbb{P}(\delta = 0) = 0$, which is satisfied by definition of the stopping times to be in strictly increasing order.

Moreover, relaxations on the assumption of observing $X_0$ completely were discussed in Krach et al. (2022, Remark 2.2). Convergence is defined with respect to the following distance.

**Definition 2.4.** Let $c_0 := c_0(k) := (\mathbb{P}(n \geq k))^{-1}$. A distance between càdlàg $\mathbb{A}$-adapted processes $Z, \xi : [0,T] \times (\Omega \times \tilde{\Omega}, \mathbb{P} \otimes \tilde{\mathbb{P}}^\mathbb{T}) \to \mathbb{R}^d$ is defined through the pseudo metrics

$$d_k(Z, \xi) = c_0(k) \mathbb{E}^\mathbb{P}_{X,\xi} \left[ \mathbb{1}_{\{n \geq k\}} |Z_{t_k} - \xi_{t_k}|^2 \right],$$

for $1 \leq k \leq K$ and two processes are called indistinguishable, if $d_k(Z, \xi) = 0$ for all $1 \leq k \leq K$.

In particular, this distance compares two càdlàg processes at their left limits at the observation times $t_k$.

The path-dependent generalisation of the Neural Jump ODE model (Herrera et al., 2021) uses the truncated signature transformation $\pi_m$ (Krach et al., 2022, Definition 3.4). The signature of a path is an (infinite) tuple of features, which allows to approximate any continuous function of the path by a linear combination of the signature terms. Therefore, it is a very powerful framework in learning theory. It is defined as follows.

**Definition 2.5.** Let $J$ be a closed interval in $\mathbb{R}$ and $X : J \to \mathbb{R}^d$ be a continuous path with finite variation. The signature of $X$ is defined as the collection of iterated integrals

$$S(X) = (1, X^1, X^2, \ldots),$$

where, for each $m \geq 1$,

$$X^m_J = \int_{u_1 < \cdots < u_m} dX_{u_1} \otimes \cdots \otimes dX_{u_m} \in (\mathbb{R}^d)^\otimes m.$$

The truncated signature of $X$ of order $m \in \mathbb{N}$ is defined as

$$\pi_m(X) = (1, X^1, X^2, \ldots, X^m),$$

i.e., the first $m + 1$ terms (levels) of the signature of $X$. For a path of dimension $d$, the dimension of the truncated signature of order $m$ is

$$\begin{cases} m + 1, & \text{if } d = 1, \\ \frac{d^{m+1}-1}{d-1}, & \text{if } d > 1. \end{cases}$$

Moreover, PD-NJ-ODE uses bounded output neural networks $f_{(\hat{\theta},\gamma)}$, where $\hat{\theta}$ are the weights of the standard neural network and $\gamma > 0$ is the trainable parameter of the bounded output activation function (Krach et al., 2022, Definition 3.12)

$$\Gamma_{\gamma} : \mathbb{R}^d \to \mathbb{R}^d, x \mapsto x \cdot \min \left( 1, \frac{\gamma}{|x|_2} \right),$$

applied to the output of the standard neural network. By $\mathcal{N}$ we denote the set of all bounded output neural networks based on a set $\hat{\mathcal{N}}$ of standard neural networks. In the following we assume that $\mathcal{N}$ is a set of standard feedforward neural networks with Lipschitz continuous activation functions, with id $\in \mathcal{N}$ that satisfies the standard universal approximation theorem with respect to the supremum-norm on compact sets, see for example Hornik (1991, Theorem 2).
Definition 2.6. The Path-Dependent Neural Jump ODE (PD-NJ-ODE) model is given by

\[
H_0 = \rho_{\theta_2}(0, 0, \pi_m(0), X_0),
\]

\[
dH_t = f_{\theta_1}(H_{t-}, t, \tau(t), \pi_m(\tilde{X}^{\leq \tau(t)} - X_0), X_0) \, dt
\]

\[
+ (\rho_{\theta_2}(H_{t-}, t, \pi_m(\tilde{X}^{\leq \tau(t)} - X_0), X_0) - H_{t-}) \, d\tilde{u}_t,
\]

\[
Y_t = \tilde{g}_{\theta_3}(H_t).
\]

The functions \(f_{\theta_1}, \rho_{\theta_2} \in \mathcal{N}\) are bounded output feedforward neural networks and \(\tilde{g}_{\theta_3} \in \tilde{\mathcal{N}}\) is a feedforward neural network with trainable parameters \(\theta = (\theta_1, \theta_2, \tilde{\theta}_3) \in \Theta\), where \(\theta_i = (\tilde{\theta}_i, \gamma_i)\) for \(i \in \{1, 2\}\) and \(\Theta\) is the set of all possible weights for the PD-NJ-ODE; \(m \in \mathbb{N}\) is the signature truncation level and \(u\) is the jump process counting the observations defined as \(u_t := \sum_{k=1}^{N} 1_{t_k \leq t}\).

In the PD-NJ-ODE model, \(H_t\) has the role of a latent process and \(Y_t\) is the model output (similar to the latent variable and model output of an RNN). The goal of the model output \(Y_t\) is to approximate the true conditional expectation \(\tilde{X}_t\). \(H_t\) evolves continuously between any two observations, according to the neural ODE defined by \(f_{\theta_1}\). At an observation time, when there is a “jump” in the available information, \(H_t\) jumps according to \(\rho_{\theta_2}\), which can be interpreted as an RNN-cell. The latent process \(H_t\) is mapped to the output process \(Y_t\) through the readout map \(\tilde{g}_{\theta_3}\). The existence and uniqueness of a solution \((H, Y)\) of (2) for a fixed \(\theta\) is implied by Cohen & Elliott (2015, Theorem 16.3.11)\(^3\). To emphasize the dependence of the PD-NJ-ODE output \(Y\) on \(\theta\) and \(X, (t_k)_{1 \leq k \leq K}\) and \(M\) we write \(Y^\theta(\tilde{X}^{\leq \tau(\cdot)})\) (since \(\tilde{X}^{\leq \tau(\cdot)}\) summarizes \(X, (t_k)_{1 \leq k \leq K}\) and \(M\)).

The objective function (cf. equivalent objective function from Remark 4.8 & Appendix D.1.5 of Krach et al. (2022)) for the training of the PD-NJ-ODE is defined as

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

\[
\Phi : \Theta \to \mathbb{R}, \quad \theta \mapsto \Phi(\theta) := \Psi(Y^\theta(\tilde{X}^{\leq \tau(\cdot)})�),
\]

where \(\odot\) is the element-wise multiplication (Hadamard product) and \(\mathbb{D}\) the set of all càdlàg \(\mathbb{R}^{d_X}\)-valued \(\tilde{A}\)-adapted processes on the product probability space \(\tilde{\Omega} \times \tilde{\Omega}\). The two terms in (4) compare the distances between \(X_{t_k}\) and \(Z\) after and before its jump at observation times \(t_k\). Intuitively speaking, to minimize the second term, \(Z_{t_i -}\) has to be the best possible prediction of \(X_{t_i -}\) given the information available before the new observation becomes available; and to minimize the first term, \(Z\) has to jump to the new observation \(X_{t_k}\) after it became available.

For \(N \in \mathbb{N}\) the number of training paths and for every \(1 \leq j \leq N\), let \((X^{(j)}, M^{(j)}, n^{(j)}, t_1^{(j)}, \ldots, t_n^{(j)}) \sim (X, M, n, t_1, \ldots, t_n)\) be independent and identically distributed (i.i.d.) random processes and variables with the same distribution as our initially introduced process \(X\) together with its random observation framework. Then the Monte Carlo approximation of (4) is

\[
\hat{\Phi}_N(\theta) := \frac{1}{N} \sum_{j=1}^{N} \frac{1}{m^{(j)}} \sum_{i=1}^{n^{(j)}} \left( |M_i^{(j)} \odot (X_{t_i^{(j)}} - Y_{\theta, j}^{(j)})|_2 + |M_i^{(j)} \odot (X_{t_i^{(j)}} - Y_{\theta, j}^{(j)})|_2 \right)^2,
\]

where \(Y_{\theta, j} := Y^\theta(\tilde{X}^{\leq \tau(\cdot), j})\).

Based on these loss functions, the following convergence guarantees can be derived, where \(\Theta_{m} \subset \Theta\) is defined as the set of possible parameters for the 3 (bounded output) neural networks, such that their widths and

\(^3\)\(f_{\theta_1}\) and \(\rho_{\theta_2}\) are Lipschitz continuous as neural networks with Lipschitz continuous activation functions, hence, the stochastic functions \(((\tilde{\omega}, \tilde{\tau}, \tilde{H}), t) \mapsto f_{\theta_1}(H_{t-}, t, \tau(t), \pi_m(\tilde{X}^{\leq \tau(t)} - X_0), X_0)\), and similarly for \(\rho_{\theta_2}\), are uniformly Lipschitz according to Cohen & Elliott (2015, Definition 16.3.2). Moreover, it is immediate to see that these functions are \textit{coefficients} according to Cohen & Elliott (2015, Definition 16.0.3), since they are continuous, hence predictable, and since we integrate with respect to finite variation processes. In particular, integrability (Cohen & Elliott, 2015, Definition 12.3.10) is trivially satisfied (path-wise), because any continuous bounded function is Stieltjes-integrable.
depths are at most $m$ and such that the truncated signature of level $m$ or smaller is used and such that
the norms of the weights $\tilde{\theta}_i$ and the bounds $\gamma_i$ are bounded by $m$ (i.e., $|\tilde{\theta}_i|_2 \leq m, \gamma_i \leq m$). Thus, $\Theta_m$ is a compact subset of $\Theta$.

Theorem 2.7. Let $\theta^\text{min}_m \in \Theta^\text{min}_m := \arg\min_{\theta \in \Theta_m} \{\Phi(\theta)\}$ for every $m \in \mathbb{N}$. If Assumption 2.1 is satisfied, then,
for $m \to \infty$, the value of the loss function $\Phi$ converges to the minimal value of $\Psi$ which is uniquely achieved
by $\hat{X}$ up to indistinguishability, i.e.,
\[
\Phi(\theta^\text{min}_m) \xrightarrow{m \to \infty} \min_{Z \in \mathcal{D}} \Psi(Z) = \Psi(\hat{X}).
\]
Furthermore, for every $1 \leq k \leq K$ we have that $Y^\text{min}_m$ converges to $\hat{X}$ in the metric $d_k$ as $m \to \infty$.
Let $\theta^\text{min}_{m,N} \in \Theta^\text{min}_{m,N} := \arg\min_{\theta \in \Theta_m} \{\hat{\Phi}_N(\theta)\}$ for every $m, N \in \mathbb{N}$. Then, for every $m \in \mathbb{N}$, $(\mathbb{P} \times \tilde{\mathbb{P}})$-a.s.
\[
\hat{\Phi}_N \xrightarrow{N \to \infty} \Phi \quad \text{uniformly on } \Theta_m.
\]
Moreover, for every $m \in \mathbb{N}$, we have a.s.,
\[
\hat{\Phi}(\theta^\text{min}_{m,N}) \xrightarrow{N \to \infty} \hat{\Phi}(\theta^\text{min}_m) \quad \text{and} \quad \hat{\Phi}_N(\theta^\text{min}_{m,N}) \xrightarrow{N \to \infty} \hat{\Phi}(\theta^\text{min}_m).
\]
In particular, one can define an increasing random sequence $(N_m)_{m \in \mathbb{N}}$ in $\mathbb{N}$ such that for every $1 \leq k \leq K$ we have a.s. that $Y^\text{min}_{m,N_m}$ converges to $\hat{X}$ in the metric $d_k$ as $m \to \infty$.

This result shows that, given we find the minimizer $\theta^\text{min}_{m,N_m}$ of the loss function, the output $Y^\text{min}_{m,N_m}$ of the
PD-NJ-ODE model converges to the optimal prediction, i.e., to the true conditional expectation $\hat{X}$, in the metrics $d_k$. This convergence holds for almost every realization of the training data, which is used to derive the minimizer $\theta^\text{min}_{m,N_m}$, when evaluated on independent test data. In particular, this result verifies that the model can approximate $\hat{X}$ arbitrarily well and that minimizing the loss function is expedient to find such an approximation of $\hat{X}$. In Section 5, we further discuss the practical implication of this convergence result. In this work, we do not focus on the task of finding the minimizer for the loss function, which is an independent and well studied problem on its own. Different optimization schemes exists, which yield convergence to global or local optima. These can be combined with our results as was further discussed in Herrera et al. (2021, Appendix E.2). In our experiments we use Adam (Kingma & Ba, 2014), an stochastic gradient descent (SGD) version, which yields good empirical results.

3 PD-NJ-ODE with Noisy Observations

So far the PD-NJ-ODE model was only applicable when having noise-free observations of $X$. In particular, we demonstrate that the PD-NJ-ODE with the original loss yields incorrect predictions in the presence of measurement noise, i.e., i.i.d. noise terms that are added to each observation of $X$. Using the stochastic filtering approach described in Krach et al. (2022, Section 6) would be a possibility to include (discrete or continuous) noise. However, this requires strong additional assumptions (i.e., the knowledge of the joint distribution of the process and the noise or equivalently training samples split up into the noise-free observations and the noise terms) which are not satisfied easily. Therefore, we want to adapt our framework, such that it can be applied to noisy observations, while only imposing weak assumptions that are easily satisfied. In particular, we introduce a method that can deal with noisy observations, even if we have never seen any noise-free observation during training. In contrast to classical filtering, there is no need to have a strong prior knowledge of the underlying dynamics.

In this section, we introduce observation noise (e.g., measurement-noise), i.e., i.i.d. noise terms $\epsilon_i$ that are added to the process $X$ at each observation time $t_i$, leading to the noisy observations $O_{t_i} := X_{t_i} + \epsilon_i$. Even though we only observe the $O_{t_i}$, the goal still is to predict $X$, in particular to compute $\mathbb{E}[X_{t_i} | O_{t_0}, \ldots, O_{t(i)}]$.

Inspecting the PD-NJ-ODE model and its loss function (replacing $X$ by $O$), we notice two things. First, that nothing in the model’s architecture prevents it from learning this modified objective, which should still have the same properties. Second, that the loss function needs to be modified. Indeed, the first term of the
loss function would train the model to jump to the noisy observation \( O_t \). This would be incorrect, since in general \( \mathbb{E}[X_t | O_{t_0}, \ldots, O_{t}] \neq O_t \) because the conditional expectation of \( X \) filters out the noise as well as possible. We therefore drop the first term of the loss function.

On the other hand, it is easy to see that the conditional expectations of \( X \) and \( O \) coincide in between observation times\(^4\) if the observation noise \( \epsilon_t \) is independent of the observations and has mean 0. Therefore, the second term of the loss function is minimised if the model learns the conditional expectation \( \mathbb{E}[X_t | O_{t_0}, \ldots, O_{\tau(t)}] \) between observation times.

Along these lines, it turns out that it suffices to omit the first term of the loss function to recover the original results of Theorem 2.7 under noisy observations. In particular, to optimize the loss, the model learns to jump to the conditional expectation of \( X \) at observation times even without the respective loss term. Indeed, since it evolves continuously after an observation, it would otherwise be different from the optimal prediction right after the observation time and therefore would not optimize the loss. In the following this is formalised.

### 3.1 Setting with Noisy Observations

The process \( X \) as well as the \( n, K, t, \tau, M \) are defined as in Section 2. Additionally, we define

- \((\epsilon_k)_{0 \leq k \leq K}\), the observation noise, which is a sequence of i.i.d. random variables on \((\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})\) taking values in \( \mathbb{R}^{d_X}\).
- \( O_{t_k} := X_{t_k} + \epsilon_k \) for \( 0 \leq k \leq n \), the noisy observation sequence.

Since the goal is to predict \( X \) given the observations \( O_t \), we redefine the filtration of the currently available information via

\[ \mathcal{A}_t := \sigma(O_{t_{i}}, t_i, M_{t_i} | t_i \leq t, j \in \{1 \leq l \leq d_X | M_{t_i,l} = 1\}) \]

such that \( \tilde{X}_t = \mathbb{E}_{\tilde{\mathbb{P}}}[X_t | A_t] \) is the conditional expectation of \( X \) given the noisy observations. We define \( \tilde{O}^{\leq t}_i \) in the same way as \( \tilde{X}^{\leq t}_i \) and note that similarly as before there exist measurable functions \( F_j \) such that \( \tilde{X}_{t,j} = F_j(t, \tau(t), \tilde{O}^{\leq \tau(t)}) \). We need the following slight modification of Assumption 2.1.

**Assumption 3.1.** We assume that Assumption 2.1 items (i) to (iii) and (v) hold and additionally that:

- (iv) \( F_j \) are continuous and differentiable in their first coordinate \( t \) such that their partial derivatives with respect to \( t \), denoted by \( f_j \), are again continuous and there exists a \( B > 0 \) and \( p \in \mathbb{N} \) such that for every \( t \in [0, T] \) the functions \( f_j, F_j \) are polynomially bounded in \( X^* \), i.e.,

\[
|F_j(\tau(t), \tau(t), \tilde{O}^{\leq \tau(t)})| + |f_j(t, \tau(t), \tilde{O}^{\leq \tau(t)})| \leq B(X_t^* + 1)^p + B \sum_{i=0}^{n} |\epsilon_i|.
\]

- (vi) \( n \) is square-integrable, i.e., \( \mathbb{E}_{\tilde{\mathbb{P}}}[n^2] < \infty \).

- (vii) The i.i.d. random noise variables \( \epsilon_k \) are independent of \( X, n, M, (t_i)_{1 \leq i \leq K} \), are centered and square-integrable, i.e., \( \mathbb{E}_{\tilde{\mathbb{P}}}[\epsilon_k] = 0 \) and \( \mathbb{E}_{\tilde{\mathbb{P}}}[|\epsilon_k|^2] < \infty \).

**Remark 3.2.** The relaxations on the assumption of observing \( X_0 \) completely discussed in Krach et al. (2022, Remark 2.2) can equivalently be applied in this setting here.

In this setting, the PD-NJ-ODE uses the noisy observations \( O_{t_i} \) and \( \tilde{O}^{\leq \tau(t)} \) as inputs instead of \( X_{t_i} \) and \( \tilde{X}^{\leq \tau(t)} \). Moreover, we define the new *noise-adapted* objective function as described before as

\[
\Psi : \mathbb{D} \rightarrow \mathbb{R}, Z \mapsto \Psi(Z) := \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}} \left[ \frac{1}{n} \sum_{i=1}^{n} |M_i \odot (O_{t_i} - Z_{t_i})|^2 \right],
\]

\[
\Phi : \Theta \rightarrow \mathbb{R}, \theta \mapsto \Phi(\theta) := \Psi(Y^\theta(X)),
\]

and its Monte Carlo approximation \( \hat{\Phi}_N \) accordingly.

---

\(^4\)Note that, in general we have \( \mathbb{E}[X_t | O_{t_0}, \ldots, O_{t}] \neq \mathbb{E}[O_t | O_{t_0}, \ldots, O_{t}] = O_t \) at observation times, in contrast to their equality between observation times.
3.2 Convergence Theorem with Noisy Observations

In the setting defined in Section 3.1, Theorem 2.7 holds equivalently as before.

**Theorem 3.3.** If Assumption 3.1 is satisfied and using the definitions of Section 3.1, the claims of Theorem 2.7 hold equivalently, upon replacing the original loss functions and their Monte Carlo approximations by their noise-adapted versions. In particular, we obtain convergence of our estimator \( Y^{\theta_{\min}}_n \) to the true conditional expectation \( \hat{X} \) in \( d_h \).

To prove this, we first need to adjust the orthogonal projection result (Krach et al., 2022, Lemma 4.2) for this setting. At this point, we want to highlight the relevance of this result as the backbone of the proof of Theorem 3.3, which is first used to show that \( \hat{X} \) is a minimizer of \( \Psi \), then to show that it is unique and finally to bound the distance between \( \hat{X} \) and \( Y^{\theta_{\min}}_n \) in \( d_h \) through the difference of their loss function values.

**Lemma 3.4.** For any \( \bar{X} \)-adapted process \( Z \) it holds that

\[
E_{P_{X|\bar{X}}} \left[ \frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \odot (O_{t_i} - Z_{t_i}) \right|^2 \right] = E_{P_{X|\bar{X}}} \left[ \frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \odot (\hat{X}_{t_i} - Z_{t_i}) \right|^2 \right].
\]

**Proof.** First note that by Assumption 3.1 point (iii) we have that \( X_{t_i} = X_{t_i} - \epsilon_i \) almost surely and when defining \( O_{t_i} := X_{t_i} - \epsilon_i \) we therefore also have that \( O_{t_i} = O_{t_i} - \epsilon_i \) almost surely. Similarly as in Krach et al. (2022, Lemma 4.2), we can derive for \( \hat{O}_{t_i} := E_{P_{X|\bar{X}}} [O_{t_i} | \bar{A}_{t_i}] \) that

\[
E_{P_{X|\bar{X}}} \left[ \frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \odot (O_{t_i} - Z_{t_i}) \right|^2 \right] = E_{P_{X|\bar{X}}} \left[ \frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \odot (\hat{O}_{t_i} - Z_{t_i}) \right|^2 \right].
\]

To conclude the proof, it is enough to note that

\[
\hat{O}_{t_i} = \hat{X}_{t_i} - \epsilon_i \left| A_{t_i} \right| = \hat{X}_{t_i} - \epsilon_i = \hat{X}_{t_i} - \epsilon_i,
\]

using that \( \epsilon_i \) has expectation 0 and is independent of \( A_{t_i} \).

In the following we sketch the proof of Theorem 3.3, by only outlining those parts of it that need to be changed in comparison with the original proof of Theorem 2.7 in Krach et al. (2022). The main differences are that the loss function needs to be adjusted whenever used, and when showing integrability, we additionally have to account for the noise terms \( \epsilon_k \). A full proof is given in Appendix C.

**Sketch of Proof of Theorem 3.3.** First, it follows directly from Lemma 3.4 that \( \Psi(\hat{X}) = \Psi(X) = \min_{Z \in \mathbb{D}} \Psi(Z) \), i.e., that \( \hat{X} \) is a minimizer of the redefined objective function \( \Psi \). Secondly, again by Lemma 3.4, we have for any process \( Z \in \mathbb{D} \) that

\[
\Psi(Z) = \Psi(\hat{X}) + E_{P_{X|\bar{X}}} \left[ \frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \odot (\hat{X}_{t_i} - Z_{t_i}) \right|^2 \right],
\]

hence \( \Psi(Z) > \Psi(\hat{X}) \) follows as before if \( Z \) is not indistinguishable from \( \hat{X} \), meaning that \( \hat{X} \) is the unique minimizer of \( \Psi \).

Under Assumption 3.1 the approximation of the functions \( f_j, F_j \) by bounded output feedforward neural networks works similarly as before, with the slight adjustment that their differences are now upper bounded by

\[
U := 3B \left( (X_1^p + 1)^{p} + \sum_{i=0}^{n} |\epsilon_i| \right).
\]
Defining
\[ c_m := c \varepsilon (T + 1) dX + c (T + 1) dX U \left( \mathbb{1}_{\{X_t^q \geq 1/\varepsilon\}} + \mathbb{1}_{\{n \geq 1/\varepsilon\}} + \mathbb{1}_{\{\delta \leq \varepsilon\}} \right) \]

it follows that there exists \( \theta_m^* \in \Theta_m \) such that \( Y_t^\theta_m^* - \hat{X}_t \leq c_m \) for all \( t \in [0, T] \). Convergence of \( \Phi(\theta_m^*) \) to \( \Psi(\hat{X}) \) then follows similarly as before, when noting that by Assumption 3.1

\[ \mathbb{E} \left[ (X_t^* + 1)^{2p} + \left( \sum_{i=0}^n |\varepsilon_i| \right)^2 \right] \leq \mathbb{E} \left[ (X_t^* + 1)^{2p} \right] + \mathbb{E}[n^2] \mathbb{E} \left[ |\varepsilon_0|^2 \right] < \infty, \quad (9) \]

using Cauchy–Schwarz and that the \( \varepsilon_i \) are i.i.d. and independent of \( n \) for the first step and the integrability of \( X^* \), \( \varepsilon_0 \) and \( n^2 \) for the upper bound. Moreover, the convergence of \( d_k(\hat{X}, Y^{\theta_m^*}) \rightarrow 0 \) follows as before.

Finally, the remaining claims of the theorem (including the Monte Carlo convergence) also hold similarly as before upon replacing \( X_{t_i} \) by \( O_{t_i} \) and noting that the integrability of \( \sup_0 h(\theta, \xi) \) follows from (9). \( \Box \)

### 3.3 More General Noise Structure & Conditional Moments

Revisiting the proof in Section 3.2, we see that the noise terms need neither be independent nor centered. If we assume that the conditional bias of the noise,

\[ \beta_i(\hat{\theta}(t)) := \mathbb{E}[\varepsilon_i | A_{t_i -}], \]

is a known function of the observations (using Doob-Dynkin Lemma (Taraldsen, 2018, Lemma 2) for its existence), then we can modify the objective function by subtracting it. This leads to

\[ \Psi : \mathbb{D} \rightarrow \mathbb{R}, Z \mapsto (Z) := \mathbb{E}_{\mathbb{P} \times K} \left[ \frac{1}{n} \sum_{i=1}^n |M_i \cap \{(O_{t_i} - \beta_i(\hat{\theta}(t))) - Z_{t_i -}\}|^2 \right]. \quad (10) \]

Revisiting (8), which is the only part of the proof where we needed the noise terms to be centered, we see that

\[ \mathbb{E} \left[ (O_{t_i -} - \beta_i(\hat{\theta}(t))) | A_{t_i -} \right] = \hat{X}_{t_i -} + \mathbb{E}[\varepsilon_i | A_{t_i -}] - \beta_i = \hat{X}_{t_i -}. \quad (11) \]

This implies that the statement of Lemma 3.4 holds equivalently under the reduced assumption of a known conditional bias function, when using the adjusted loss (10). Additionally assuming that \( \mathbb{E} \left[ (\sum_{i=0}^n |\varepsilon_i|^2 \right] < \infty, \) the following result follows as before.

**Corollary 3.5.** In the setting described in this subsection (i.e., arbitrary known mean of the noise and no independence assumption on the noise), which is a generalisation of the setting in Section 3.1, Theorem 2.7 holds equivalently as before when using the objective function (10).

The following remark explains how Corollary 3.5 can be used to predict conditional higher moments (instead of only the conditional expectation) under certain assumptions.

**Remark 3.6.** This result makes it possible to compute the conditional moments of \( X \) given the noisy observations, which doesn’t work in the setting of Section 3.1. In particular, we consider observations \( O_{t_i -} = X_{t_i -} + \varepsilon_i \), where we assume that

- \( \varepsilon_i \) is independent of \( A_{t_i -} \),
- \( \varepsilon_i \) is conditionally independent of \( X_{t_i -} \) given \( A_{t_i -} \),
- and \( \varepsilon_i \) have known finite moments.

Remark that Proposition A.8 implies that the first two assumptions are in particular satisfied if \( \varepsilon_i \) is independent of \( \sigma(A_{t_i -}, X_{t_i -}) \). The binomial theorem implies for any \( q \in \mathbb{N} \)

\[ O_{t_i -}^q = X_{t_i -}^q + \sum_{j=1}^q \binom{q}{j} X_{t_i -}^{q-j} \varepsilon_i^j. \]

Note that \( q, j \) and \( q - j \) denote exponents here rather than superscripts.
We interpret the entire sum as the observation noise and accordingly define the conditional bias of the observation noise of the $q$-th moment as

$$
\beta^q_i := \mathbb{E} \left[ \sum_{j=1}^{q} \binom{q}{j} X_{t_i}^{q-j} \epsilon_j^q \mid \mathcal{A}_{t_i} \right] = \sum_{j=1}^{q} \binom{q}{j} \mathbb{E}[X_{t_i}^{q-j} \mid \mathcal{A}_{t_i}] \mathbb{E}[\epsilon_j^q],
$$

where we use the assumptions on $\epsilon_i$ together with Proposition A.7 for the second equality.

Then an inductive argument shows that $\beta^q_i$ is a known function of the observations, using the assumption that the moments of $\epsilon_i$ are known. Indeed, to compute $\beta^q_i$ the conditional expectations of smaller moments $\mathbb{E}[X_{t_i}^{q-j} \mid \mathcal{A}_{t_i}]$ need to be computed, which can be done according to the induction hypothesis (note that the base case follows directly from Corollary 3.5 and the assumptions on $\epsilon_i$). Therefore, Corollary 3.5 implies that we can compute $\mathbb{E}[X_{t_i}^{q} \mid \mathcal{A}_{t_i}]$ (assuming that we reach the limit where the PD-NJ-ODE output equals the conditional expectation). In case of an exponential moment assumption $\mathbb{E}[^{\lambda}X_{t_i}^{-}\mid \mathcal{A}_{t_i}] < \infty$ for some $\lambda > 0$ we can therefore infer the conditional law of $X_{t_i}$.

In the remainder of this section, we discuss the cases of linear and non-linear (noisy) observation models.

**Remark 3.7 (Linear Observation Model).** Let us consider a linear observation process

$$
O_{t_k} := AX_{t_k} + \epsilon_k,
$$

for some fixed matrix $A \in \mathbb{R}^{d \times d_x}$, for $d \in \mathbb{N}$, where we have observations of $O$, but are ultimately interested in predicting $X$ without having access to any observations of $X$. For this to be suitable for our framework, we need that Assumption 3.1 is satisfied for $Z := AX$.\(^6\) Then we can apply our method to approximate $\hat{Z}_t = \mathbb{E}[Z_t \mid \mathcal{A}_t]$, as discussed before. To be able to infer predictions for $X$ from $\hat{Z}$, we need to further assume that $A$ defines an injective map, hence, a left inverse $A^{-1}$ with $A^{-1}A = \text{id}$ exists. Under this assumption, $\mathbb{E}[X_t \mid \mathcal{A}_t] = A^{-1}\hat{Z}_t$, by linearity. It is important to note that in an incomplete observation setting, it would in general not be possible to apply our method to directly approximate $X$, since (noisy) observations of $X$ cannot be computed from (noisy) observations of $Z$, if they are incomplete (i.e., $A^{-1}O$ might not be well defined). Therefore, the detour via first computing $\hat{Z}$ is essential. Moreover, it should be apparent that in the case the map defined by $A$ is not injective (or at least injective on the support of $X$), information about $X$ will be lost when only observing $O$. Nevertheless, any process of the form $BZ$ for another matrix $B$ can be predicted similarly as described above when observing $O$.

**Remark 3.8 (Non-linear Observation Model & Beyond).** Let us consider the observation process

$$
O_{t_k} := \phi(X_{t_k}) + \epsilon_k,
$$

for some measurable function $\phi$, where we again have observations of $O$ (in particular we assume that $\phi(X)$ satisfies the assumptions), but are ultimately interested in predicting $X$. Then, even if a left inverse $\phi^{-1}$ with $\phi^{-1} \circ \phi = \text{id}$ exists, the technique from Remark 3.7 does not work, since $\mathbb{E}[X_t \mid \mathcal{A}_t] \neq \phi^{-1}(\mathbb{E}[\phi(X)_t \mid \mathcal{A}_t])$ in general. However, under the stronger assumption that we have access to (noisy) samples of the joint process $\xi := (X, \phi(X))$ for training and that this $\xi$ satisfies Assumption 3.1, we can compute $\hat{\xi} = (\hat{X}, \phi(\hat{X}))$, which can then also be evaluated for samples where only (noisy) observations of the $\phi(X)$-coordinates are available (cf. Krach et al. (2022, Corollary 6.2)). Moreover, we note that the same is true, when replacing $\phi(X)$ by a general process $Z$.

### 3.4 Examples of Processes Satisfying the Assumptions

In principle, all the examples presented in Krach et al. (2022, Section 7) are valid examples for this setting when adding some type of i.i.d. observation noise satisfying our assumptions, as e.g. Gaussian or uniform noise. However, it is important to note that the (true) conditional expectation is not the same, since we now condition on the noisy observations $O_t$, instead of the original observations $X_t$. Therefore, we give one explicit example where we compute the conditional expectation in the noisy observation setting.

\(^6\)We note that by linearity of the (conditional) expectation we have $\mathbb{E}[Z_t \mid \mathcal{A}_t] = A \mathbb{E}[X_t \mid \mathcal{A}_t]$, hence, $Z$ satisfies Assumptions 3.1 (iii) to (v), if $X$ does.
3.4.1 Brownian Motion with Gaussian Observation Noise

Let $X := W$ be a standard Brownian motion and let $\epsilon_0 = 0$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i \geq 1$ be the i.i.d. noise terms for some $\sigma > 0$. Then $O_{tk} = X_{tk} + \epsilon_i$ are the observations. Clearly, all integrability assumptions are satisfied by $X$ and $\epsilon_i$ (cf. Krach et al. (2022, Section 7.5)). To compute the true conditional expectations we first note that the independent increments property of the Brownian motion imply for $t_k \leq t < t_{k+1}$

$$
E[X_t | A_t] = E[W_t - W_{t_k} | A_{t_k}] + E[W_{t_k} | A_{t_k}] = E[W_{t_k} | O_{t_k}, \ldots, O_{t_1}],
$$

and therefore, $f(s, \tau(t), \hat{O} \leq \tau(t)) = 0$. Since $W$ is a Brownian motion and $\epsilon_i$ are independent i.i.d. Gaussian noise terms, we know that

$$
v := (O_{t_1}, \ldots, O_{t_k}, W_{t_k})^T \sim \mathcal{N}(0, \Sigma)
$$

where

$$
\Sigma = \begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix} \in \mathbb{R}^{(k+1) \times (k+1)},
$$

with $\Sigma_{11} \in \mathbb{R}^{k \times k}$ and $(\Sigma_{11})_{i,j} = \min(t_i, t_j) + \sigma^2 I_{i = j}$, $\Sigma_{12} = \Sigma_{21} = (t_1, \ldots, t_k) \in \mathbb{R}^{1 \times k}$ and $\Sigma_{22} = t_k$. Then the conditional distribution of $(W_{t_k} | O_{t_1}, \ldots, O_{t_k})$ is again normal with mean $\hat{\mu} := \Sigma_{21} \Sigma_{11}^{-1} (O_{t_1}, \ldots, O_{t_k})^T$ and variance $\hat{\Sigma} := \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$ (Eaton, 2007, Proposition 3.13). In particular we have

$$
E[X_t | A_t] = E[W_{t_k} | O_{t_1}, \ldots, O_{t_k}] = \hat{\mu}.
$$

3.5 A Practical Note on Using the Noise-Adapted Loss Function

We have seen in Section 3.2 and 3.3, considering the limit where network size $n$ and training samples $N$ go to infinity, that there is no disadvantage of the noise-adapted loss function (10) compared to the original loss function (3). In particular, the noise-adapted training framework yields the optimal solution with noisy observations, but also if there is no observation noise. Indeed, setting $\epsilon_k = 0$ for all $k$, we have $O_{tk} = X_{tk}$ and therefore the original result also holds with the noise-adapted loss function. Hence, the question arises, why one should use the original loss function (3) at all anymore?

To answer this question, we first note that in practice we are not in the limit case, but we only have access to a finite training set (i.e., we only observe a finite amount of observations $n(3)$ from finite number of paths $N$). Hence, the inductive bias when training the model becomes more important. See Appendix B for more details on the inductive bias. In particular, using the original loss function (3) in a setting without observation noise is preferable, since it directly penalizes the model for not jumping to observations and therefore making it easier for the model to learn this behaviour. Indeed, in the noise-adapted loss function, the penalization for not jumping to an observation is more indirect, since it will only be punished at the following observation time until which the output of the model further evolves through the neural ODE. Therefore, the feedback signal is weaker, i.e., the respective gradients are smaller.

In a setting with observation noise, the practical relevance of the noise-adapted loss function depends on the size of the variance of the noise, $\text{Var}_{\text{noise}}$, compared to the size of the variance of the data, $\text{Var}_{\text{data}}$. Clearly, if the variance ratio $\frac{\text{Var}_{\text{noise}}}{\text{Var}_{\text{data}}}$ is small, the noise does not have a large impact on the observations and therefore neither on the training of the model when using the original loss function (3). Due to the better inductive bias during training, it might therefore be beneficial to use (3) instead of the noise-adapted loss (10). On the other hand, if the variance ratio is large, the impact of the noise on the observations is substantial and therefore the noise-adapted training framework leads to better results. This behaviour is well visible in the experiment presented in Section 6.3. The turning point, where training with the noise-adapted loss function becomes better, is problem specific (e.g., the number of training samples has a big influence) and, if in doubt, we suggest to train the PD-NJ-ODE model with both loss functions and compare their results.

4 PD-NJ-ODE with Dependence between $X$ and the Observation Framework

Recall that the observation mask process is given by $M$ and the underlying process by $X$. In this section, we remove the assumptions that the observation times are independent of $X$, and that $M$ is independent.
of the observation times and of $X$. In essence, the model is now defined on only one probability space $\mathbb{P}$ and no independence assumptions between the random variables are made. Instead, we need some weaker conditional independence assumptions to recover the results of Theorem 2.7.

### 4.1 Intuition on independence assumptions

In many real-world applications, the independence of the process $X$ and the observation framework (i.e., $(t_i)_{i \in \{1, \ldots, n\}$ and $M$) is heavily violated. For example, consider the irregular measurements of a patient’s health parameters. A nurse or doctor will only take (expensive) measurements if information on the patient’s state, $X$, hints that the measurement is necessary. In practice, different measurements are taken from different patients depending on observations of their state. This motivates the crucial importance of lifting the independence assumption for real-world applications.

However, even in this paper, we cannot completely remove any independence assumption; we still need conditional independence of the process $X$ and the observation framework, given all past observations (as we make precise in Assumption 4.1 items (vii) and (ix) and Proposition 4.9). This is a realistic assumption, which we show by continuing our hospital example. If we assume that every piece of information the hospital gets from the patient is immediately logged as an observation (perhaps noisy or incomplete) of $X$, then all information about $X$ that the hospital has is contained entirely within these observations. Thus measurement decisions are conditionally independent of $X$, given the past observations (if there are no further hidden confounders).

On the other hand, it is easy to violate this assumption. Imagine that the patient tells the nurse and doctor she feels feverish and they subsequently take her temperature. If they log her temperature, but not her feeling, the assumption is violated, as the hospital has a piece of information that is not logged as an observation of $X$. In such cases where information is not logged as an observation, our model (and most other classical forecasting methods) would learn a biased forecast of the body temperature. In the extreme case that body temperature is only measured if patients feel very feverish, measurements will record a high temperature, which leads our model to always predict a high body temperature (i.e., the expected body temperature conditioned on feeling feverish and all other past observations) even if the patient does not feel feverish.

There are often ways to mitigate such issues. In our case, for example, by first logging the feverish feeling and then the actual measurement at a later time stamp. Moreover, a non-feverish feeling has to be logged whenever no feverish feeling is reported during a regular nurse visit. Nevertheless, this discussion shows that it is important to be aware that even our weakened, more realistic assumption of conditional independence is often not fully satisfied in practical situations. Hence, one must be careful when verifying the assumptions and potentially adapt the experimental setting such that they are satisfied.

### 4.2 Setting with Dependence

To allow for dependence, we only consider the probability measure $\mathbb{P}$ and define $X, n, K, t, \tau, M, A$ similar as before, but all on the same (filtered) probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ associated with $\mathbb{P}$. For a random variable $Z$ and a family of sets $\mathcal{B}$ we use the natural notation for their smallest jointly generated sigma algebra $\sigma(Z, \mathcal{B}) := \sigma(Z) \vee \sigma(\mathcal{B})$.

Then we need the following assumptions, most of which are borrowed from Assumption 2.1. Only (vii) and (viii) are new, while (i) is a strict generalisation of Assumption 2.1 (i).

**Assumption 4.1.** We assume that Assumption 2.1 items (ii) to (vi) hold, with all instances of $\tilde{\mathbb{P}}$ and $\mathbb{P} \times \tilde{\mathbb{P}}$ replaced by $\mathbb{P}$. Additionally, we assume that:

1. $M_{0,j} = 1$ for all $1 \leq j \leq d_X$ ($X$ is completely observed at 0) and $|M_k|_1 > 0$ for every $1 \leq k \leq K \mathbb{P}$-almost surely (at every observation time at least one coordinate is observed).

2. For every $1 \leq k \leq n$, $X_{t_k-}$ is conditionally independent of $\sigma(n, M_{t_k})$ given $A_{t_k-}$. 

13
(viii) For all $1 \leq k \leq K$, $1 \leq j \leq d_X$ there is some $\eta_{k,j} > 0$ such that $\mathbb{P}(M_{k,j} = 1 \mid \sigma(n, A_{t_k}^-)) > \eta_{k,j}$ (i.e., given the currently known information and $n$, for each coordinate the probability of observing it at the next observation time is positive).

(ix) For every $1 \leq k \leq K$ the process $X$ is conditionally independent of $t_k$ given $A_{t_{k-1}}$.

**Remark 4.2.** The relaxations on the assumption of observing $X_0$ completely discussed in Krach et al. (2022, Remark 2.2) can equivalently be applied here.

We can use the original objective function (4) and its Monte Carlo approximation (5).

### 4.3 Convergence Theorem with Dependence

In the setting defined in Section 4.2, Theorem 2.7 holds equivalently.

**Theorem 4.3.** If Assumption 4.1 is satisfied and using the definitions of Section 4.2, the claims of Theorem 2.7 hold equivalently. In particular, we obtain convergence of our estimator $Y_{\theta_{\min}, N_m}$ to the true conditional expectation $\hat{X}$ in $d_k$.

The main technical challenge when generalising Theorem 2.7 to Assumption 4.1 is to replace the arguments using independence and e.g. Fubini’s theorem, by arguments using conditional independence. This difficulty is well visible when comparing the proof of the orthogonal projection Lemma 4.6 below, with the original proof of Krach et al. (2022, Lemma 4.2). Before, the claim followed relatively immediately from Fubini’s theorem and the standard orthogonal projection result for conditional expectations of random variables. In contrast to this, now we have to condition first on the known information and the additional random variables of the observation framework, then we have to argue by conditional independence that this is the same as conditioning only on the known information and only after that we can apply the conditional version Proposition 4.4 of the orthogonal projection result. The relevance of Lemma 4.6 as the backbone of the proof of Theorem 4.3 was already discussed in Section 3.2. Finally, in Lemma 4.8 we show how Assumption 4.1(viii) can replace independence between $M$ and $X$, when deriving that the distance between $\hat{X}_{t_k^-}$ and $Y_{t_k^-}$ can be bounded through their distance when multiplied with the observation mask.

First, we prove the extension of the standard $L^2$-orthogonality result that was stated in Herrera et al. (2021, Proposition B.2).

**Proposition 4.4.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a fixed probability space, and $\mathcal{A}, \mathcal{B}$ be sub-$\sigma$-algebras such that $\mathcal{B} \subseteq \mathcal{A} \subseteq \mathcal{F}$. For some random variable $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ we define $\hat{X} := \mathbb{E}[X \mid \mathcal{A}]$. Then for every random variable $Z \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(Z \neq \hat{X}) > 0$ we have

$$\mathbb{E}\left[|X - Z|^2 \mid \mathcal{B}\right] = \mathbb{E}\left[|X - \hat{X}|^2 \mid \mathcal{B}\right] + \mathbb{E}\left[|Z - \hat{X}|^2 \mid \mathcal{B}\right] \geq \mathbb{E}\left[|X - \hat{X}|^2 \mid \mathcal{B}\right],$$

with strict inequality with positive probability.

The proof is based on Durrett (2010, Theorem 5.1.8). We focus on the one-dimensional case, though this can easily be generalised to multiple dimensions via the 2-norm, as in Herrera et al. (2021).

**Proof.** We begin by expanding the left hand side.

$$\mathbb{E}\left[|X - Z|^2 \mid \mathcal{B}\right] = \mathbb{E}\left[|(X - \hat{X}) - (Z - \hat{X})|^2 \mid \mathcal{B}\right]$$

$$= \mathbb{E}\left[|X - \hat{X}|^2 \mid \mathcal{B}\right] + \mathbb{E}\left[|Z - \hat{X}|^2 \mid \mathcal{B}\right] - 2\mathbb{E}\left[(X - \hat{X})(Z - \hat{X}) \mid \mathcal{B}\right]$$

We now analyse the cross term, which expands to $\mathbb{E}[Z(X - \hat{X}) \mid \mathcal{B}] - \mathbb{E}[\hat{X}(X - \hat{X}) \mid \mathcal{B}]$. Focusing on the first term, we note that since $Z \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, it holds that $\mathbb{E}[Z \mid \mathcal{A}] = \mathbb{E}[ZX \mid \mathcal{A}]$. By taking expectation (conditioned on $\mathcal{B}$) of both sides, we get

$$\mathbb{E}\left[Z \hat{X} \mid \mathcal{B}\right] = \mathbb{E}\left[Z\mathbb{E}[X \mid \mathcal{A}] \mid \mathcal{B}\right] = \mathbb{E}\left[Z\mathbb{E}[X \mid \mathcal{A}] \mid \mathcal{B}\right] = \mathbb{E}[ZX \mid \mathcal{B}]$$
via the tower property, as $\mathcal{B} \subseteq \mathcal{A}$. Hence, $\mathbb{E}[Z|X - \hat{X})|\mathcal{B}] = 0$. Note that showing this only required that $Z$ is $\mathcal{A}$-measurable. Since this is also satisfied by $\hat{X}$ we directly have $\mathbb{E}[X|X - \hat{X})|\mathcal{B}] = 0$ and therefore the cross term vanishes and the equality follows. The inequality holds since $\mathbb{E}[Z|X - \hat{X})|\mathcal{B}]$ is non-negative. We therefore just need to show that the inequality is strict with positive probability. To this end, assume for the sake of contradiction that $\mathbb{E}[Z|X - \hat{X})^2|\mathcal{B}] = 0$ $\mathbb{P}$-a.s., which implies (by the tower property) that $\mathbb{E}[Z|X - \hat{X})^2] = 0$. This is only possible if $Z = \hat{X}$ $\mathbb{P}$-a.s., which contradicts the assumption that $\mathbb{P}(Z \neq \hat{X}) > 0$. \hfill $\Box$

As a next step we show that under Assumption 4.1(ix) we can recover that $\hat{X}$ is the optimal predictor.

**Proposition 4.5.** Under Assumption 4.1(ix) we have

$$\hat{X}_{t_k} = \mathbb{E}[X_t | A_{t_k-1}] |_{t = t_k} = \mathbb{E}[X_{t_k} | A_{t_k-1}].$$

**Proof.** The result is a direct consequence of Proposition A.9. \hfill $\Box$

Next we re-prove Krach et al. (2022, Lemma 4.2) under the relaxed assumptions with the help of Proposition 4.5.

**Lemma 4.6.** For any $\mathcal{A}$-adapted process $Z \in L^2(\Omega, \mathcal{A}, \mathbb{P})$ it holds that

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{ti} \odot (X_{ti} - Z_{t_i})|^2 \right] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{ti} \odot (X_{ti} - \hat{X}_{t_i})|^2 \right] + \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{ti} \odot (\hat{X}_{t_i} - Z_{t_i})|^2 \right].$$

**Proof.**

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{ti} \odot (X_{ti} - Z_{t_i})|^2 \right] = \sum_{i=1}^{K} \mathbb{E} \left[ \frac{1}{n} \sum_{i \leq n} |M_{ti} \odot (X_{ti} - Z_{t_i})|^2 \right] = \sum_{i=1}^{K} \mathbb{E} \left[ \frac{1}{n} \sum_{i \leq n} |M_{ti} \odot (X_{ti} - Z_{t_i})|^2 \right] \sigma(n, M_{ti}, A_{t_i-1}) \right]$$

$$= \sum_{i=1}^{K} \mathbb{E} \left[ \frac{1}{n} \sum_{i \leq n} \sum_{j=1}^{d_{X}} |M_{ti,j} \mathbb{E} \left[ |X_{t_i,j} - Z_{t_i,j}|^2 \right] \sigma(n, M_{ti}, A_{t_i}) \right]$$

$$= \sum_{i=1}^{K} \mathbb{E} \left[ \frac{1}{n} \sum_{i \leq n} \sum_{j=1}^{d_{X}} |M_{ti,j} \mathbb{E} \left[ |X_{t_i,j} - Z_{t_i,j}|^2 \right] A_{t_i-1} \right].$$

The first step follows by monotone convergence, the last by Lemma 4.7 below. Now we can conclude by first applying the equality from Proposition 4.4 with $\mathcal{A}, \mathcal{B} = \mathcal{A}_{t_i-1}$, which yields that $\mathbb{E}[X_{t_i} | A_{t_i-1}]$ minimises the expression, and then using Proposition 4.5 to conclude that $\hat{X}_{t_i}$ equals this minimizer. Finally, we can reverse the above steps to arrive at the desired form. \hfill $\Box$

**Lemma 4.7.** Assume the context of Lemma 4.6. Then for all $i, j$ it holds that

$$\mathbb{E} \left[ |X_{t_i,j} - Z_{t_i,j}|^2 \sigma(n, M_{ti}, A_{t_i}) \right] = \mathbb{E} \left[ |X_{t_i,j} - Z_{t_i,j}|^2 A_{t_i-1} \right].$$

**Proof.** We prove this by showing that if we expand the $|X_{t_i,j} - Z_{t_i,j}|^2$ term, all three resulting terms can just be conditioned on $A_{t_i-1}$. This is a valid argument as $X$ and $Z$ are both assumed to be square-integrable. Note that squaring a random variable plays no role in the information given by the $\sigma$-algebra it is being conditioned on, and so we only need to analyse the terms $Z_{t_i,j}, X_{t_i,j}$, and $X_{t_i,j}Z_{t_i,j}$. See Appendix A for an overview of conditional independence and how it’s used here.

CASE $Z_{t_i,j}$: $Z$ is $\mathcal{A}$-adapted, and so $Z_{t_i,j}$ is $\mathcal{A}_{t_i-1}$-measurable. Thus we have

$$\mathbb{E} [Z_{t_i,j} | \sigma(n, M_{ti}, A_{t_i-1})] = Z_{t_i,j} = \mathbb{E} [Z_{t_i,j} | A_{t_i-1}]$$

$\mathbb{E} [Z_{t_i,j} | A_{t_i-1}]$
as desired.

CASE $X_{t_i - j}$: Assumption 4.1 point (vii) implies that $X_{t_i - j}$ is conditionally independent of $\sigma(n, M_{t_i})$ given $\mathcal{A}_{t_i}$. We therefore have by Proposition A.4

$$E[X_{t_i - j} \mid \sigma(n, M_{t_i}, \mathcal{A}_{t_i})] = E[X_{t_i - j} \mid \mathcal{A}_{t_i}].$$

CASE $X_{t_i - j}Z_{t_i - j}$: We combine the previous two ideas, namely that $Z_{t_i - j}$ is $\mathcal{A}_{t_i}$-measurable and that $X_{t_i - j}$ is conditionally independent of $\sigma(n, M_{t_i})$ given $\mathcal{A}_{t_i}$. Thus

$$E[X_{t_i - j}Z_{t_i - j} \mid \sigma(n, M_{t_i}, \mathcal{A}_{t_i})] = Z_{t_i - j}E[X_{t_i - j} \mid \sigma(n, M_{t_i}, \mathcal{A}_{t_i})] = Z_{t_i - j}E[X_{t_i - j} \mid \mathcal{A}_{t_i}] = E[Z_{t_i - j}X_{t_i - j} \mid \mathcal{A}_{t_i}].$$

Combining these 3 cases proves the claim.

The following lemma shows how Assumption 4.1(viii) can replace the independence between $M$ and $X$.

**Lemma 4.8.** There exists some $c_2(k) > 0$ such that for any $\mathbb{A}$-adapted process $Z \in L^2(\Omega, \mathbb{A}, \mathbb{P})$ we have

$$E\left[\mathbbm{1}_{\{n \geq k\}} \left| M_{t_k} \circ (\hat{X}_{t_k} - Z_{t_k})\right|_1\right] \leq \frac{1}{c_2(k)} E\left[\mathbbm{1}_{\{n \geq k\}} \left| M_{t_k} \circ (\hat{X}_{t_k} - Z_{t_k})\right|_1\right].$$

**Proof.** Assumption 4.1 point (viii) states that $0 < \eta_{k,j} < \mathbb{P}(M_{k,j} = 1 \mid \sigma(n, A_{t_k})) = E[M_{t_k,j} \mid \sigma(n, A_{t_k})]$ for all $k, j$. Let $c_2 := c_2(k) := \min_{1 \leq j \leq d_X} \eta_{k,j}$, then $c_2 > 0$ and

$$E\left[\mathbbm{1}_{\{n \geq k\}} \left| M_{t_k} \circ (\hat{X}_{t_k} - Z_{t_k})\right|_1\right] \leq \sum_{j=1}^{d_X} E\left[\mathbbm{1}_{\{n \geq k\}} M_{t_k,j} \left| \hat{X}_{t_k,j} - Z_{t_k,j}\right|\right]$$

$$\leq \sum_{j=1}^{d_X} E\left[\mathbbm{1}_{\{n \geq k\}} \left| \hat{X}_{t_k,j} - Z_{t_k,j}\right| E[M_{t_k,j} \mid \sigma(n, A_{t_k})]\right]$$

$$\geq c_2 E\left[\mathbbm{1}_{\{n \geq k\}} \left| \hat{X}_{t_k,j} - Z_{t_k,j}\right|\right].$$

where we used that $\mathbbm{1}_{\{n \geq k\}} \left| \hat{X}_{t_k,j} - Z_{t_k,j}\right|$ is $\sigma(n, A_{t_k})$-measurable in the second line and the definition of $c_2$ in the last line.

With the help of these lemmas we are now ready to prove Theorem 4.3. In the following we again give a sketch of the proof, by only outlining those parts of it that need to be changed in comparison with the original proof of Theorem 2.7 in Krach et al. (2022). We refer the interested reader to the full proof in Appendix C for all details.

**Sketch of Proof of Theorem 4.3.** As before it follows that $\hat{X}$ is a minimizer of $\Psi$. To show its uniqueness, we first note that

$$E\left[\mathbbm{1}_{\{n \geq k\}} \left| \hat{X}_{t_k} - Z_{t_k}\right|_2\right] \leq \frac{c_3}{c_2(k)} E\left[\mathbbm{1}_{\{n \geq k\}} \left| M_{t_k} \circ (\hat{X}_{t_k} - Z_{t_k})\right|_2\right]$$

(12)

is implied as before when using Lemma 4.8 instead of the independence. Then

$$E\left[\frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \circ (\hat{X}_{t_i} - Z_{t_i})\right|_2^2\right] \geq \left(\frac{c_2(k)}{c_0 c_1 c_3}\right)^2 d_k(\hat{X}, Z)^2 > 0,$$

follows as before, implying the uniqueness of $\hat{X}$ as minimizer of $\Psi$.

The approximation of the functions $f_j, F_j$ also works as before. With this we have

$$\min_{Z \in \mathbb{B}} \Psi(Z) \leq \Phi(\theta_{m}^{\text{min}}) \leq \Phi(\theta_{m}^{*})$$
dominated convergence can be used to show then Assumption 4.1 (vii) can be replaced by the weaker assumption (vii’-a) together with the additional integrability assumption (vii’-b) below.

where the triangle inequality was used in the third line, and we use Assumption 4.1 point (iv) (which implies is in general only known at final time \( T \)) observations in the loss by assumptions such that also the second part of the theorem follows.

If we change the weighting in the loss function to Proposition 4.9.

Integrability assumption as explained in the following proposition.

\[
\min_{Z \in D} \Psi(Z) = \Psi(\hat{X}) = \Phi(\theta^\text{min}_m) \leq \Phi(\theta^*_m) \xrightarrow{m \to \infty} \min_{Z \in D} \Psi(Z),
\]

since \( 2p \) is again the largest power of \( X_T^\ast \) in both terms (remember \( c_m \) has an \( (X_T^\ast)^p \) term). The convergence of \( d_k(X, Y^\ast_m) \to 0 \) follows as before, finishing the first part of the theorem. Finally, it is easy to see that the proof of the convergence of the Monte Carlo approximation is not affected by our more general dependence assumptions such that also the second part of the theorem follows.

The assumption that \( X_{t_{i-}} \) and \( n \) are conditionally independent given \( A_{t_{i-}} \) seems a bit odd, because \( n \) is a quantity that is not known at \( t_i \), i.e., not \( A_{t_{i-}} \)-measurable. This assumption is needed, because we weight the observations in the loss by \( 1/n \), which means that we weight the terms known at time \( t_i \) with a quantity that is in general only known at final time \( T \). In line with this, we can drop the assumption if instead we weight the terms with the time step \( \Delta t_i := t_i - t_{i-1} \), which is \( A_{t_{i-}} \)-measurable. However, this needs an additional integrability assumption as explained in the following proposition.

**Proposition 4.9.** If we change the weighting in the loss function to

\[
\Psi(Z) := \mathbb{E} \left[ \sum_{i=1}^{n} \Delta t_i \left( |M_t \circ (X_{t_i} - Z_{t_i})|_2 + |M_t \circ (X_{t_i} - Z_{t_i})|_2 \right)^2 \right],
\]

then Assumption 4.1 (vii) can be replaced by the weaker assumption (vii’-a) together with the additional integrability assumption (vii’-b) below.
Both loss functions yield the same unique optimizer, the choice of weighting can influence the training of the model for finite
more important for the model to predict well after a long time without observations, while it is less important
in the latter case, observations have more importance if the previous observation is farther away. Although
We need to ensure that Assumption 4.1 points (vii) and (viii) are satisfied (assuming that a dummy variable
A main problem when constructing examples where the observation times
Then we can control whether the process
4.4.1 Class of Examples Incorporating Dependence
Clearly, all the examples from Krach et al. (2022) are trivially valid for our generalised settings of Section 4.2
4.4 Examples of Processes Satisfying the Assumptions
A main problem when constructing examples where the observation times \( t_i \) have a dependence on the process \( X \), is that in general this will also lead to \( n \) having a dependence on \( X \) (since \( n \), as the amount of observations, increases when observations become more likely). This, in turn, might lead to a contradiction of Assumption 4.1 point (vii).
One way to circumvent this is to define \( n \) conditionally independent of \( X_{t_i} \) given \( A_{t_i} \) for all \( i \) and to allow for pseudo observation times, i.e., observation times at which no coordinate is observed (cf. Remark 2.2). Then we can control whether the process \( X \) is observed at an observation time \( t_i \) via the observation mask \( M_i \). In this way, the times at which \( X \) is actually observed can depend on \( X \) through the observation mask \( M_k \), even though the \( t_k \) do not depend on \( X \). Therefore, the original problem is replaced by having an observation mask depending on \( X \), which will be discussed in detail below.
When using a time grid on which the process \( X \) is sampled, one concrete example of defining \( n \) conditionally independent of \( X_{t_i} \) given \( A_{t_i} \) is to use the grid points as observation times \( t_i \) leading to \( n \) being the number of grid points. In this case, both \( n \) and the \( t_i \) are deterministic, hence, they satisfy the conditional independence assumptions.
We need to ensure that Assumption 4.1 points (vii) and (viii) are satisfied (assuming that a dummy variable is observed at every observation time), when defining the observation mask. One way to define \( M_i \) such that this point (vii) is satisfied is to define it as a function of random variables that are \( A_{t_i} \)-measurable and random variables that are independent of \( \sigma(A_{t_i} \cup X_{t_i}) \).
In particular, let \( M_i := g_i \left( (U^i_{j_1})_{j_1 \in J_1^i}, (V^i_{j_2})_{j_2 \in J_2^i} \right) \), where \( g_i \) is a measurable function mapping to \( \{0, 1\}^{d_x} \); \( J_1^i, J_2^i \subseteq \mathbb{N} \) are index sets; \( U^i_{j_1} \) is a \( A_{t_i} \)-measurable random variable; and \( V^i_{j_2} \) a random variable independent of \( \sigma(A_{t_i} \cup X_{t_i}, n) \) for all \( j \) in \( J_1^i \) and \( J_2^i \) respectively.
By Proposition A.5 we need to show that for any measurable function \( \phi \) we have \( \mathbb{E}[\phi(X_{t,-}) | \sigma(n, M_{t,-}, A_{t,-})] = \mathbb{E}[\phi(X_{t,-}) | A_{t,-}] \). Indeed, \( M_{t} \) is \( \sigma(A_{t,-}, (V_{j})_{j\in J_{t}}) \)-measurable by construction, therefore, we have for such a \( \phi \) that

\[
\mathbb{E}[\phi(X_{t,-}) | \sigma(n, M_{t,-}, A_{t,-})] = \mathbb{E}\left[ \mathbb{E}[\phi(X_{t,-}) | \sigma(n, (V_{j})_{j\in J_{t}}, A_{t,-})] | \sigma(n, M_{t,-}, A_{t,-}) \right],
\]

by the tower property. On the other hand,

\[
\mathbb{E}\left[ \phi(X_{t,-}) | \sigma(n, (V_{j})_{j\in J_{t}}, A_{t,-}) \right] = \mathbb{E}[\phi(X_{t,-}) | \sigma(n, A_{t,-})] = \mathbb{E}[\phi(X_{t,-}) | A_{t,-}],
\]

using the independence of \( V_{j} \) together with Corollary A.2 in the first equality and \( n \) being conditionally independent together with Proposition A.5 in the second equality. Together, this implies the claim and therefore Assumption 4.1 point (vii).

For Assumption 4.1 point (viii), we note that by Durrett (2010, Lemma 6.2.1) we have

\[
\mathbb{P}(M_{k} = 1 | \sigma(n, A_{tk,-})) = \mathbb{E}[M_{tk} | \sigma(n, A_{tk,-})] = \mathbb{E}\left[ g_{k} \left( (U_{j}^{k})_{j\in J_{t}^{k}}, (V_{j}^{k})_{j\in J_{t}^{k}} \right) | \sigma(n, A_{tk,-}) \right],
\]

for \( g_{k}(u_{j}^{k})_{j\in J_{t}^{k}} := \mathbb{E}\left[ g_{k} \left( (u_{j}^{k})_{j\in J_{t}^{k}}, (V_{j}^{k})_{j\in J_{t}^{k}} \right) \right] \). Hence, we need to define the \( g_{k} \) and \( V_{j}^{k} \) such that \( \tilde{g}_{k} > \eta_{k} \) (coordinate-wise) for some \( \eta_{k} > 0 \).

**Example 4.11** (Homogeneous Poisson Point Process with Dependent Observations). We use a 1-dimensional homogeneous Poisson point process \( X = N^{\lambda} \) (Krach et al., 2022, Section 7.3) and defined observations depending on its value following the instructions above. To begin with, we define the observation times to be the grid points of the sampling grid of the process and \( n \) accordingly to be the number of grid points. To permit observation times of the process depending on the process, we define the observation mask as

\[
M_{t} = \begin{cases} 
1_{\{x_{i} > \lambda t_{i,-} \}}, & x_{i} \sim N(N_{t_{i,-}}, \sigma^{2}) \\
ui \sim \text{Bernoulli}(p), & M_{i,-} = 1, \\
& \text{if } M_{i,-} = 1,
\end{cases}
\]

for some \( \sigma > 0 \) and \( p \in (0, 1) \). Thus the process is more likely to be observed if the previous value was observed and was large (note \( \mathbb{E}[N_{t_{i,-}}^{\lambda}] = \lambda t_{i} \) for all \( k \)). Otherwise the mask value is sampled from a Bernoulli distribution.

To satisfy Assumption 4.1 point (i), we define \( M_{0} := 1 \). To see that Assumption 4.1 point (vii) holds, let \( V_{1} \sim \text{Bernoulli}(p) \) and \( V_{2} \sim N(0, \sigma^{2}) \) be independent random variables independent of \( \sigma(A_{t,-}, N_{t_{i,-}}^{\lambda}) \). Then

\[
M_{t} = 1_{\{M_{i,-} = 1\}} 1_{\{N_{t_{i,-}^{\lambda} + V_{2} > \lambda t_{i,-} \}} + V_{1}^{i} 1_{\{M_{i,-} = 0\}} := g_{i}(M_{i,-}, N_{t_{i,-}}, t_{i}, V_{1}^{i}, V_{2}^{i}),
\]

and the claim follows as explained above. Moreover, Assumption 4.1 point (viii) is satisfied because

\[
\tilde{g}_{i}(M_{i,-}, N_{t_{i,-}}, t_{i} = \min \left( \mathbb{P}\left[ V_{2}^{i} > \lambda b - a \right] 1_{(a,b)} = (N_{t_{i,-}, t_{i}}, 0) \right), \mathbb{E}\left[ V_{1}^{i} \right] \right) \geq \min \left( \mathbb{P}\left[ V_{2}^{i} > \lambda b - a \right] 1_{(a,b)} = (N_{t_{i,-}, t_{i}}, 0) \right), \mathbb{E}\left[ V_{1}^{i} \right] \right) \geq \min \left( \mathbb{P}\left[ V_{2}^{i} > \lambda T \right], p \right) =: \eta_{i} > 0,
\]

using that \( \lambda t_{i} - N_{t_{i,-}}^{\lambda} \leq \lambda T, \) since \( N_{t} \geq 0, t_{i} \leq T \) and \( \lambda > 0 \), in the last line.

**Remark 4.12.** We note that the choice \( X = N^{\lambda} \) is only explicitly used for the fact that \( N^{\lambda} \) is lower bounded (by 0). Hence, the Poisson point process could be replaced by any other process that is lower bounded and satisfies Assumption 2.1 points (iv) \& (v).
Example 4.13 (Black–Scholes with Dependent Observations). We use a 1-dimensional Black–Scholes process (geometric Brownian motion) (Krach et al., 2022, Example 7.3) with constant drift and volatility $\mu, \sigma$ starting at $x_0$ and again define observations depending on its value. The observation times and $n$ are defined as in the previous example and we define the actual times when $X$ is observed via the mask $M$. We set $M_0 = 1$. Moreover, we redefine $\tau$ to be the last time before $t$ at which $X$ was observed, i.e., $\tau(t) = \max\{t_i | t_i < t, M_i = 1\}$. Let $V^i_1 \sim \text{Bernoulli} \left( \frac{t_i - t_{i-1}}{\tau(t_i)} \right)$, $V^i_2 \sim \mathcal{N}(0, \eta^2)$ and $V^i_3 \sim \text{Bernoulli}(p)$ be independent random variables for some $\eta > 0$ and $p \in (0, 1)$. Then we define

$$M_i := V^i_1 1_{\{X_{\tau(t_i)} + V^i_2 \geq x_0 e^{\eta t_i}\}} + (1 - V^i_1) V^i_3,$$

for all $i \geq 1$. In particular, if $X$ was observed at the previous observation time $V^i_1 = 1$ and the probability of observing $X$ increases with the size of the last observation of $X$ (compared to the current expected value of $X$ at the current time). The further the last observation of $X$ is in the past, the more likely $V^i_1 = 0$ in which case $X$ is observed with probability $p$. Upon noting that the $t_i$ are deterministic, it follows as in the previous example that Assumption 4.1 points (vii) and (viii) are satisfied.

5 Practical Implications of the Convergence Result

In this section we discuss which practically relevant conclusions we can draw from convergence in the metrics $d_k$ for $1 \leq k \leq K$. We mainly focus on the setting in Section 4.2, however, the same is also true in the combined setting of Section C.1. In Section 5.1, we study the practical meaning of Assumption 4.1(ix). We give an intuitive counterexample, where this assumption is not satisfied and our model does not converge to $\hat{X}$. Secondly, in Section 5.2, we discuss the implications of convergence in the (pseudo) metrics $d_k$ (which ensures a good approximation at left-limits of observation times) for general times $t$. In particular, we study two practically relevant examples for the conditional distribution of the observation times and show that in these cases, with high probability, our model approximates the conditional expectation well on the entire support of the observation times.

5.1 Practical meaning of Assumption 4.1(ix)

If Assumption 4.1(ix) is not satisfied, this can lead to situations where $\hat{X}$ is not the minimizer of $\Psi$ and therefore our model does not converge to it, as we explain in the following example. Assume that patients in a hospital get asked by a nurse every morning at 8 am, whether they feel feverish. If the answer is yes, then their temperature is measured and logged at 8 am; if the answer is no, their temperature is measured and logged at 4 pm. The (non-)feverish feeling is not logged at all. This data satisfies Assumption 4.1 (since $n$ is deterministic; $d_X = 1$ and $M_i = 1$) except for item (ix). It is clear that the model will predict higher temperatures in the morning than in the afternoon for any test sample, even if the patient always has the same temperature, since the model only saw this behaviour in the training data. If we assume that on average patients have the same (or even lower) temperature in the morning as in the afternoon, this prediction is not optimal in practice.

The mathematical reason for this discrepancy is that our model learns to approximate $\mathbb{E}[X_{t_k} \mid \mathcal{A}_{t_k}]$ (see proof of Lemma 4.6), which does not coincide with $\hat{X}$, i.e.,

$$\mathbb{E}[X_{t_k} \mid \mathcal{A}_{t_k}] \neq \mathbb{E}[X_t \mid \mathcal{A}_t] \bigg|_{t=t_k} = \mathbb{E}[X_t \mid \mathcal{A}_{t_k-1}] \bigg|_{t=t_k} = \hat{X}_{t_k},$$

(surprisingly) holds for this example. Indeed, Proposition 4.5 (which would turn this inequality into an equality) heavily relies on Assumption 4.1(ix). $\mathbb{E}[X_{t_k} \mid \mathcal{A}_{t_k}]$ minimizes the test error (if the joint distribution of $(X, t_k)$ is the same as during training), since it can exploit the knowledge that it is queried at an observation time $t_k$ (i.e., high temperature if $t_k$ is in the morning and low temperature if $t_k$ is in the afternoon). However, in this example we do not obtain the optimal forecast $\hat{X}_t = \mathbb{E}[X_t \mid \mathcal{A}_{t_k-1}]$ for any time $t$, which is the best prediction of $X$ at any time, given the information available at the previous observation time.

Mathematically, $\mathcal{A}_{t_k} = \sigma(\mathcal{A}_{t_k-1}, t_k)$ includes the information that a measurement is taken at time $t_k$. 

20
Assumption 4.1(ix) ensures (via Proposition 4.5) that our model converges to the practically meaningful version of the conditional expectation \( \hat{X}_{t_k} = E[X_t \mid \mathcal{A}_{t_{k-1}}]\) at observation times. In the following, we discuss what this implies for the approximation at any \( t \in [0, T] \).

5.2 From Approximations at Observation Times to Approximations at any Time

Under Assumption 4.1 we have established that our model output \( Y^\text{dmin}_t (\hat{X} \leq t_k^-) \) converges to \( E[X_t \mid \mathcal{A}_{t_{k-1}}] \) at every observation time \( t_k \). Both, \( Y^\text{dmin}_t (\hat{X} \leq t_k^-) \) and \( E[X_t \mid \mathcal{A}_{t_{k-1}}] \) are functions of the random variables summarized in \( \mathcal{A}_{t_{k-1}} \) (by definition for our model and by the Doob-Dynkin lemma for the conditional expectation) and therefore conditionally independent of \( t_k \) given \( \mathcal{A}_{t_{k-1}} \). We note in particular that \( \hat{X} \leq t_k^- \) carries exactly the information of \( \mathcal{A}_{t_{k-1}} \), i.e., it has no information about \( t_k \) (despite its notation). Hence, intuitively speaking, the approximation has to be good for every \( t \) in the support of \( t_k \) for the expectation to converge. In the following we formalize this.

Let \( \mathbb{P}_k \) be the probability measure conditioned on the event that \( n \geq k \), i.e.
\[
\mathbb{P}_k(\cdot) = \mathbb{P}(\cdot \mid \{n \geq k\}),
\]
and denote by \( \mathbb{E}_k \) the expectation operator with respect to this probability measure. Then the following lemma is a consequence of the definitions of \( \mathbb{P}_k \) and \( c_0(k) \).

**Lemma 5.1.** We have \( d_k(\hat{X}, Y^\text{dmin}_t) = \mathbb{E}_k \left[ \left\| \hat{X}_{t_k} - Y^\text{dmin}_t (\hat{X} \leq t_k^-) \right\|_2 \right] \) and the equivalent result for \( Y^\text{dmin}_{t_{m,Nm}} \).

**Proof.** It is enough to note that \( \mathbb{E}[1 \{n \geq k\}] = \mathbb{P}(n \geq k) \mathbb{E}_k[\cdot] \).

Next, we define for each \( 1 \leq k \leq K \) the (regular) conditional distribution of \( t_k \) given \( \hat{X} \leq t_k^- \) as the almost surely unique probability kernel (or random measure) \( \mu_{t_k}(\cdot; \hat{X} \leq t_k^-) \) satisfying
\[
\mu_{t_k}(B; \hat{X} \leq t_k^-) = \mathbb{P}_k \left(t_k \in B \mid \mathcal{A}_{t_{k-1}}\right) \quad \text{a.s.,} \quad \forall B \in \mathcal{B}([0, T]).
\]
Kallenberg (2021, Theorem 8.5) ensures its existence (since \([0, T]\) is Borel) and implies for any measurable function \( \phi \) that almost surely
\[
\mathbb{E}_k \left[ \phi (\hat{X} \leq t_k^-) \mid \mathcal{A}_{t_{k-1}} \right] = \int_0^T \phi (\hat{X} \leq t_k^-) \mu_{t_k}(dt; \hat{X} \leq t_k^-).
\]
In particular, this implies the following result.

**Proposition 5.2.** Under Assumption 4.1 we have, with the same notation as in Theorem 2.7, that for every \( 1 \leq k \leq K \)
\[
d_k(\hat{X}, Y^\text{dmin}_t) = \mathbb{E}_k \left[ \int_0^T \left\| E[X_t \mid \hat{X} \leq t_k^-] - Y^\text{dmin}_{t_k^-} (\hat{X} \leq t_k^-) \right\|_2 \mu_{t_k}(dt; \hat{X} \leq t_k^-) \right] \xrightarrow{m \to \infty} 0
\]
and the equivalent result for \( Y^\text{dmin}_{t_{m,Nm}} \).

**Proof.** We first apply Lemma 5.1 and then use the tower property to get a nested conditional expectation (conditioning on \( \mathcal{A}_{t_{k-1}} \)) inside the outer expectation, which can be rewritten by (14) as an integral. Finally, we use Lemma 4.5 and then the Doob-Dynkin lemma (Taraldsen, 2018, Lemma 2) to rewrite \( \mathbb{E}[X_t \mid \mathcal{A}_{t_{k-1}}] \) as a function of \( \hat{X} \leq t_k^- \), which we denote by \( \mathbb{E}[X_t \mid \hat{X} \leq t_k^-] \) (and we replace \( t \) by \( t^- \) for the left-continuous version of it).

Proposition 5.2 implies convergence of a conditional \( L^1 \)-norm (or equivalently, an \( L^1 \)-norm integrated against a random measure) nested inside an \( L^1 \)-norm. Since \( L^p \)-convergence implies convergence in probability, we have in particular that for every \( \epsilon > 0 \)
\[
\mathbb{P}_k \left( \int_0^T \left\| E[X_t \mid \hat{X} \leq t_k^-] - Y^\text{dmin}_{t_k^-} (\hat{X} \leq t_k^-) \right\|_2 \mu_{t_k}(dt; \hat{X} \leq t_k^-) > \epsilon \right) \xrightarrow{m \to \infty} 0.
\]

(15)
In the following, we study two examples of the conditional distribution \( \mu_t \) that are relevant in practice. We study them for \( Y^{\mu_m} \) but note that the hold equivalently for \( Y^{\mu_m,N_m} \). We first remark that in general the support of \( t_k \) is a \( A_{t_{k-1}} \)-measurable random set (in particular, it often depends at least on \( t_{k-1} \)). We denote this set-valued random variable with

\[
S_{t_k} = \text{supp} \left( t_k \mid A_{t_{k-1}} \right),
\]

which is defined to be the smallest closed random subset \( S_{t_k} \subset [0,T] \) such that \( \mu_{t_k}(S_{t_k} \mid X^{\leq t_k-}) = 1 \) \( \mathbb{P}_k \)-almost surely. By Doob-Dynkin’s lemma (Taraldsen, 2018, Lemma 2) we can write \( S_{t_k} \) as a function \( X^{\leq t_k-} \), i.e.,

\[
S_{t_k} = S_n(\hat{X}^{\leq t_k-}).
\]

**Example 5.3.** Assume that \( S_k := \cup_\omega S_{t_k}(\omega) \) is finite and that there exists some \( \alpha > 0 \) such that for all \( t \in S_k \) we have \( \mu_{t_k}(\{t\} \mid X^{\leq t_k-}) \geq \alpha \mathbb{1}_{\{t > t_{k-1}\}} \), \( \mathbb{P}_k \)-almost surely. This is the case if all observation times are sampled from a (finite) grid with positively lower bounded probability to take any value of the grid, which is larger than the previous observation time (as in the synthetic examples in this paper and in Krach et al. (2022)). Then Proposition 5.2 implies that for every \( 1 \leq k \leq K \)

\[
\sum_{t \in S_k} \mathbb{E}_k \left[ \mathbb{1}_{\{t > t_{k-1}\}} \right] \mathbb{E} \left[ X_t - \frac{\hat{X}^{\leq t_k-}}{2} - Y^{\mu_m}_{t_k} \left( \hat{X}^{\leq t_k-} \right) \right] \xrightarrow{m \to \infty} 0
\]

and (15) implies that for every \( \epsilon, \delta > 0 \) there exists \( m_0 \in \mathbb{N} \) such that for \( m \geq m_0 \) there is a subset \( \Omega_m \subset \Omega \) with \( \mathbb{P}_k(\Omega_m) > 1 - \delta \) and for every \( \omega \in \Omega_m \) with \( n(\omega) \geq k \) and every \( t \in S_k \) with \( t > t_{k-1}(\omega) \)

\[
\left| \mathbb{E} \left[ X_t - \frac{\hat{X}^{\leq t_k-}(\omega)}{2} - Y^{\mu_m}_{t_k}(\hat{X}^{\leq t_k-}(\omega)) \right] \right| < \epsilon.
\]

In particular, with high probability, our model is close to the optimal prediction at the left limit of every possible grid point if \( m \) is large enough. At the same time it is apparent that we cannot infer anything for \( t \notin S_k \). In our synthetic examples this is not a problem since we only plot our model and the optimal prediction (and measure the distance between them) on the grid points.

**Remark 5.4 (Convergence in the Evaluation Metric).** In the same setting as in Example 5.3, if there exists one finite grid from which all observation times are sampled, i.e., \( S \subset [0,T] \) finite, such that \( S_k \subset S \) for all \( k \) and \( T \in S \), then \( n \leq |S| \). Revisiting the proof of Theorem 2.7 in Appendix C, we note that (23) can therefore be replaced by

\[
\mathbb{E} \left[ \frac{|Z|^2}{2} \right] \leq |S| \mathbb{E} \left[ \frac{1}{n} |Z|^2 \right],
\]

and therefore we can show the convergence stated in Theorem 2.7 in the stronger \( L^2 \)-type (pseudo) metric

\[
d^H_k(Z, \xi) := c_0(k) \mathbb{E} \left[ \mathbb{1}_{(n \geq k)} |Z_{t_k} - \xi_{t_k}|^2 \right],
\]

for which Proposition 5.2 holds equivalently.

We assume that for \( \mathbb{P} \)-a.e. \( \omega \) and for every \( t \in S \) we have that \( t_{k-1}(\omega) < t \leq t_k(\omega) \) implies that \( t \in S_{t_k}(\omega) \). Then the evaluation metric (cf. Section 6) defined on \( S \) for \( N_{\text{test}} \) i.i.d. test samples is

\[
\text{eval}_{S,N_{\text{test}}} (\hat{X}, Y^{\mu_m,N_m}) := \frac{1}{|S|} \sum_{t \in S} \sum_{j=1}^{N_{\text{test}}} \mathbb{E} \left[ X_t - \frac{\hat{X}^{\leq t-}}{2} - Y^{\mu_m}_{t_k} \left( \hat{X}^{\leq t-} \right) \right] \| \frac{1}{|S|} \sum_{t \in S} \mathbb{E} \left[ X_t - \frac{\hat{X}^{\leq t-}}{2} - Y^{\mu_m}_{t_k} \left( \hat{X}^{\leq t-} \right) \right] \|^2\]

and by a similar argument as in the second part of the proof of Theorem 2.7 (cf. (31), (32)), we have

\[
\text{eval}_{S,N_{\text{test}}} (\hat{X}, Y^{\mu_m,N_m}) \xrightarrow{P-a.s., N_{\text{test}} \to \infty} \mathbb{E} \left[ \frac{1}{|S|} \sum_{t \in S} \mathbb{E} \left[ X_t - \frac{\hat{X}^{\leq t-}}{2} - Y^{\mu_m}_{t_k} \left( \hat{X}^{\leq t-} \right) \right] \| \frac{1}{|S|} \sum_{t \in S} \mathbb{E} \left[ X_t - \frac{\hat{X}^{\leq t-}}{2} - Y^{\mu_m}_{t_k} \left( \hat{X}^{\leq t-} \right) \right] \|^2 \right] =: (I).
\]

If we add pseudo observation times at \( T \) (by choosing the respective mask to be 0 in case there is no actual observation at \( T \), cf. Remark 2.2), we have \( t_n = T \).\textsuperscript{8} Then, using that \( S \) is finite and with our assumption

\textsuperscript{8}We need to do this because otherwise we have terms where \( t > t_n \) which cannot be upper bounded by the distances in \( d^H_k \).

However, we note that the objective function is the same except for dividing by \( n + 1 \) instead of \( n \) for the samples which do not have an observation at \( T \). Hence, also the trained model is nearly the same.
for \( S_{t_k} \), we have

\[
(I) = \sum_{k=1}^{|S|} \mathbb{E} \left[ \mathbb{I}_{\{n=k\}} \frac{1}{|S|} \sum_{t \in S} \mathbb{E} \left[ X_{t-} | \hat{X} \leq t- \right] - Y_{t-}^{\min} (\hat{X}_{t-}) \right]^2
\]

\[
= \frac{1}{|S|} \sum_{k=1}^{|S|} \mathbb{E} \left[ \mathbb{I}_{\{n=k\}} \sum_{t \in S_{t_k}} \mathbb{I}_{\{t_{n-1} < t \leq t_n\}} \right] \mathbb{E} \left[ X_{t-} | \hat{X} \leq t- \right] - Y_{t-}^{\min} (\hat{X}_{t-}) \right]^2
\]

\[
\leq \frac{1}{|S|} \sum_{k=1}^{|S|} \mathbb{E} \left[ \sum_{n=1}^{|S|} \mathbb{I}_{\{t_{n-1} < t \leq t_n\}} \right] \mathbb{E} \left[ X_{t-} | \hat{X} \leq t- \right] - Y_{t-}^{\min} (\hat{X}_{t-}) \right]^2
\]

which converges to 0 according to Example 5.3. In particular, we have \( \mathbb{P} \)-almost surely

\[
\lim_{m \to \infty} \lim_{N_{\text{test}} \to \infty} \text{eval}_{S,N_{\text{test}}} (\hat{X}, Y_{\min}^{m,N_m}) = 0,
\]

i.e., the evaluation metric converges to 0 when the number of evaluation samples and the network sizes increase. Moreover, if the number of evaluation samples \( N_{\text{test}} < \infty \) is fixed, the evaluation metric \( \text{eval}_{S,N_{\text{test}}} (\hat{X}, Y_{\min}^{m,N_m}) \) converges to zero in probability as \( m \) and \( N_m \) tend to infinity.

**Example 5.5.** Assume that for \( 1 \leq k \leq K \), \( \mathbb{P}_k \)-almost surely we have that \( S_{t_k} = (t_{k-1}, T) \), that \( \mu_{t_k}(\cdot; \hat{X} \leq t_k-) \) is absolutely continuous with respect to the Lebesgue measure on \([0, T]\), and that its Radon-Nikodym derivative w.r.t. the Lebesgue measure satisfies

\[
\frac{d\mu_{t_k}(\cdot; \hat{X} \leq t_k-)}{dt} \geq \alpha \mathbb{1}_{(t_{k-1}, T)}(\cdot)
\]

for some \( \alpha > 0 \). This is for example the case if the observation time \( t_k \) is uniformly or exponentially (with an upper bound) distributed on \((t_{k-1}, T)\). Then Proposition 5.2 implies that

\[
\mathbb{E}_k \left[ \int_{t_{k-1}}^T \mathbb{E} \left[ X_{t-} | \hat{X} \leq t_k- \right] - Y_{t-}^{\min} (\hat{X}_{t-}) \right] dt \xrightarrow{m \to \infty} 0
\]

and (15) implies that for every \( \epsilon, \delta > 0 \) there exists \( m_0 \in \mathbb{N} \) such that for \( m \geq m_0 \) there is a subset \( \Omega_m \subset \Omega \) with \( \mathbb{P}_k(\Omega_m) > 1 - \delta \) and for every \( \omega \in \Omega_m \) with \( n(\omega) \geq k \)

\[
\int_{t_{k-1}}^T \mathbb{E} \left[ X_{t-} | \hat{X} \leq t_k- (\omega) \right] - Y_{t-}^{\min} (\hat{X}_{t-} (\omega)) \right] dt < \epsilon.
\]

In particular, with high probability, the \( L^1 \)-distance on \((t_{k-1}, T)\) between our model and the optimal prediction is small if \( m \) is large enough. Since our model and the optimal prediction are both continuous in \( t \), we know that in the limit they have to be point-wise the same, i.e., for every \( t \in (t_{k-1}, T) \).

To summarize, under Assumption 4.1, at every time \( s \), for every future time \( t \in [s,T] \) that was not deterministically excluded from being the next observation time given the previous observations, our model \( Y_{t-}^{\min} (\hat{X}_{s}) \) converges in probability to the correct conditional expectation \( \mathbb{E}[X_t | A_s] \) as \( m \) tends to infinity.\(^9\)

\(^9\)Mathematically this just means, “for every \( t \in S_{t_k} \)” given that \( \tau(s) = t_{k-1} \) for some \( k \).
5.3 Implications for a Trained PD-NJ-ODE model

Finally, we discuss what we can say about the distance between the true conditional expectation \( \hat{X} \) and the output of the PD-NJ-ODE model \( Y^{\theta^*} \) for parameters \( \theta^* \) which are \( \varepsilon \)-optimal. This is an important practical question which arises, when training the model with some optimization scheme (e.g. a version of stochastic gradient descent) yields such parameters \( \theta^* \). We note that the existence of such parameters follows from Theorem 2.7 or its versions for noisy observations (Theorem 3.3), dependence (Theorem 4.3) or both extensions (Theorem C.3).

**Proposition 5.6.** Assume that \( \theta^* \) are \( \varepsilon \)-optimal parameters for the PD-NJ-ODE, i.e.,

\[
\Phi(\theta^*) \leq \inf_{\theta \in \Theta} \Phi(\theta) + \varepsilon,
\]

for some \( \varepsilon > 0 \).\(^{10}\) Then for every \( 1 \leq k \leq K \) there exists a constant \( c = c(k) > 0 \) independent of \( \theta^* \) and \( \varepsilon \) such that \( d_k(\hat{X}, Y^{\theta^*}) \leq c \sqrt{\varepsilon} \).

**Proof.** The claim follows from applying (1), (24), (23) and finally Lemma C.4, which was summarized in (29) in the full proof of the main theorem with both extensions.

We note that the results from Section 5.2 equivalently hold here. Therefore, we are able to make statements about the distance between the paths of \( \hat{X} \) and \( Y^{\theta^*} \).

**Remark 5.7.** Under the same assumption as in Proposition 5.6 as well as Assumption 4.1, we have that for every \( 1 \leq k \leq K \)

\[
d_k(\hat{X}, Y^{\theta^*}) = \mathbb{E}_k \left[ \int_0^T \left\| \mathbb{E} \left[ X_t \mid \hat{X} \leq t_k \right] - Y_t^{\theta^*} (\hat{X} \leq t_k) \right\|_2 \mu_t(dt ; \hat{X} \leq t_k) \right] \leq c(k) \sqrt{\varepsilon}.
\]

Moreover, if additionally the assumptions of Example 5.3 are satisfied, then we have

\[
\sum_{t \in S_k} \mathbb{E}_k \left[ \mathbb{1}_{\{t > t_{k-1}\}} \left\| \mathbb{E} \left[ X_t \mid \hat{X} \leq t_k \right] - Y_t^{\theta^*} (\hat{X} \leq t_k) \right\|_2 \right] \leq \frac{c(k)}{\alpha} \sqrt{\varepsilon}.
\]

Equivalently, if additionally the assumptions of Example 5.5 are satisfied, then we have

\[
\mathbb{E}_k \left[ \int_{t_{k-1}}^T \left\| \mathbb{E} \left[ X_t \mid \hat{X} \leq t_k \right] - Y_t^{\theta^*} (\hat{X} \leq t_k) \right\|_2 dt \right] \leq \frac{c(k)}{\alpha} \sqrt{\varepsilon}.
\]

6 Experiments

The code with all new experiments and those from Krach et al. (2022) is available at [https://github.com/FlorianKrach/PD-NJODE](https://github.com/FlorianKrach/PD-NJODE). Further details about the experiments can be found in Appendix D. In particular, in Appendix D.1 we give details on the slight deviation of the practical implementation from the theoretical description.

As in Krach et al. (2022) we use the following evaluation metric to quantify and compare the training success.

\[
eval(\hat{X}, Y^\theta) := \frac{1}{N_{\text{test}}} \sum_{j=1}^{N_{\text{test}}} \frac{1}{k+1} \sum_{i=0}^{k} \left\| \hat{X}_{\frac{i}{k}}^{(j)} - Y_{t_{\frac{i}{k}}}^{\theta^j} \right\|_2^2,
\]

where the outer sum runs over the test set of size \( N_{\text{test}} \) and the inner sum runs over the equidistant grid points on the time interval \([0, T]\).

In Sections 6.1 and 6.2 we provide two illustrative experiments on easy synthetic datasets for the extensions discussed in this paper. We note that the extension to a dependence between the observation framework and

\(^{10}\)Note that \( \inf_{\theta \in \Theta} \Phi(\theta) = \min_{Z \in \mathbb{Z}} \Psi(Z) = \Psi(\hat{X}) \) according to our main theorem.
the process $X$ is purely theoretical, not changing anything in the implementation of the model, hence, the experiments on real world datasets of Krach et al. (2022) are representative for this extension as well. In particular, it is very likely that there actually is such a dependence in the Physionet dataset (Goldberger et al., 2000), as discussed in the introduction and in Section 4.1. Hence, our paper provides the theoretical foundation for those empirical results of Krach et al. (2022) and, vice versa, those results show that the method discussed in our paper is applicable to complex high-dimensional settings. In Section 6.3 we provide further experiments on the Physionet dataset to compare the noise-adapted training framework of Section 3 with the original one.

6.1 Noisy Observations – Brownian Motion with Gaussian Observation Noise

We test the PD-NJ-ODE trained with the loss function adapted to noisy observations (6) in the context of Section 3.4.1. In particular, $X$ is a Brownian motion and we assume to have observation noise of a centered normal distribution with standard deviation $\sigma = 0.5$. Moreover, we compare these results to using the original loss function (3) with the noisy observation. PD-NJ-ODE adapted to noisy observations achieves a minimal evaluation metric of $1.1 \cdot 10^{-3}$ while using the original loss function leads to a nearly 20 times larger evaluation metric of $1.9 \cdot 10^{-2}$. Moreover, in Figure 1 we see that the noise-adapted method learns to correctly jump when new observations become available, while the original method jumps to the noisy observations and afterwards tries to get close to the true conditional expectation quickly. We note that this is the expected behaviour.

![Figure 1: A test sample of a Brownian motion X with noisy observations $O_{it} = X_{it} + \epsilon_i$ together with the true and predicted conditional expectation. The PD-NJ-ODE is trained with the noise-adapted loss (left) and the original loss (right).](image)

6.2 Dependent Observation Framework – a Black–Scholes Example

Based on Example 4.13 we train a PD-NJ-ODE on a 1-dimensional geometric Brownian motion with drift $\mu = 2$ and volatility $\sigma = 0.3$ and with observation probability depending on the last observation of $X$, the time since the last observation and on independent random variables $V_{(2,3)}$ for which we use the parameters $\eta = 3$ and $p = 0.1$. As our theoretical result suggests, our model learns to predict the conditional expectation well with a minimal evaluation metric of $1.1 \cdot 10^{-3}$ which is also visible in Figure 2.

6.3 Physionet with Observation Noise

To test the new training framework for noisy observations in a 41-dimensional, complex real world setting, we use the Physionet dataset (Goldberger et al., 2000). Even though the Physionet dataset, as it is, is
likely to have some observation noise, it turns out that this noise is too small to have a big impact on the training (cf. Figure 3 and Section 3.5). Therefore, we add synthetic noise to all observations that are used as input for the PD-NJ-ODE (but not to the observation on which the evaluation MSE is computed) and study its impact on the evaluation MSE depend on the size of its standard deviation. More precisely, we first compute each coordinate’s standard deviation $\sigma_{data,j}$ on the training set and then add i.i.d. noise samples $\epsilon_{i,j} \sim N(0, \zeta^2 \sigma_{data,j}^2)$, for $\zeta \geq 0$, to each observed coordinate $X_{t,i,j}$. In Figure 3 we show the minimal evaluation MSE results when training the PD-NJ-ODE with the original loss (5) and the noise-adapted loss (see Equations (6) and (7)) on these datasets for increasing standard deviation factor $\zeta$. Without additional noise ($\zeta = 0$), the original loss function leads to slightly better results than the noise-adapted loss function (which is approximately 10% larger), suggesting that in this case the potential intrinsic observation noise has a smaller impact than the inductive bias of the original loss function (cf. Section 3.5). The larger the
synthetically added noise gets, the larger is its impact. For the noise-adapted loss function the evaluation MSE only grows linearly (with a small slope) when $\zeta$ increases. In contrast to this, it grows more than linearly for the original loss function. For some $\zeta \in (0.4, 0.5)$ there is the turning point, where the noise-adapted framework becomes more efficient. Overall, the evaluation MSE grows less than 20% for the noise-adapted loss (from $\zeta = 0$ to $\zeta = 1$), while it grows by nearly 300% for the original loss function. Moreover, the evaluation MSE with $\zeta = 1$ is more than 3 times larger for the original PD-NJ-ODE than for the noise-adapted one. This example demonstrates well the importance of the noise-adapted training framework when the observation noise is large.

7 Conclusion

In this work we broadened the applicability of the PD-NJ-ODE of Krach et al. (2022) by extending the theoretical foundation to allow for the observation framework (i.e., observation times and masks) to depend on previous information and additionally proposed a new loss function that provably leads to optimal predictions even if observations are noisy. In particular, we showed that any centered i.i.d. observation noise satisfying some integrability conditions can be dealt with by switching to the noise-adapted objective function (6). Moreover, we showed that the proof of the main result can be retained when lifting the independence assumption between the process $X$ and the observation framework, by extensively working with conditional independence. Finally, we provided experiments showing empirically that the PD-NJ-ODE works well in those extended settings.

Acknowledgement

All authors would like to thank the anonymous reviewers for their detailed feedback and great suggestions that led to significant improvements of the paper.

References


Calypso Herrera, Florian Krach, and Josef Teichmann. Neural jump ordinary differential equations: Consistent continuous-time prediction and filtering. In International Conference on Learning Representations, 2021. 1, 2, 5, 7, 14, 35, 40, 43


Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. Neural Networks, 2, 1989. 38


Florian Krach, Marc Nübel, and Josef Teichmann. Optimal estimation of generic dynamics by path-dependent neural jump odes. arXiv, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17, 18, 19, 20, 22, 24, 25, 27, 34, 38, 40, 41, 42, 44, 46, 47


Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-encoding sequential monte carlo. arXiv, 2017. 2


James Morrill, Patrick Kidger, Lingyi Yang, and Terry Lyons. On the choice of interpolation scheme for neural CDEs. Transactions on Machine Learning Research, 2022. ISSN 2835-8856. 2

Behnam Neyshabur. Implicit regularization in deep learning, 2017. 35


Gunnar Taraldsen. Optimal learning from the doob-dynkin lemma. *arXiv*, 2018. 4, 10, 21, 22, 44

Appendix

A Conditional independence

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$ be two sigma-algebras and let $X$ be a random variable.

The assumption that $X$ is independent of $\mathcal{G}$ leads to the natural but incorrect conclusion that $E[X \mid \sigma(\mathcal{G}, \mathcal{H})] = E[X \mid \mathcal{H}]$. For this to hold, we actually need a stronger assumption as in the following proposition that is due to Hansen (2015); Yoo (2014).

**Proposition A.1.** If $\sigma(X, \mathcal{H})$ is independent of $\mathcal{G}$ then $E[X \mid \sigma(\mathcal{G}, \mathcal{H})] = E[X \mid \mathcal{H}]$.

**Proof.** We prove the desired equality in the context of basic measure theory. We first assume that $X$ is integrable on $\sigma(\mathcal{G}, \mathcal{H})$, since otherwise neither of the expectations are valid. Then we recall that conditional expectation is simply a random variable $Z$ that satisfies the following three properties

1. $Z$ is $\sigma(\mathcal{G}, \mathcal{H})$-measurable,
2. $Z$ is integrable,
3. $\int_A X \, dP = \int_A Z \, dP$ for all $A \in \sigma(\mathcal{G}, \mathcal{H})$.

By the definition of the conditional expectation we know that $E[X \mid \sigma(\mathcal{G}, \mathcal{H})]$ satisfies these properties. To show the claim, it is therefore enough to prove that also $Z = E[X \mid \mathcal{H}]$ satisfies them. The first two follow trivially since $\mathcal{H} \subseteq \sigma(\mathcal{G}, \mathcal{H})$. For the third, we note that by Dynkin’s $\pi$-$\lambda$ theorem it is enough to consider the $\cap$-stable generator $P = \{A \cap B \mid A \in \mathcal{G}, B \in \mathcal{H}\}$ of $\sigma(\mathcal{G}, \mathcal{H})$ and show that

$$\int_{A \cap B} X \, dP = \int_{A \cap B} Z \, dP \quad \forall A \in \mathcal{G}, B \in \mathcal{H},$$

or equivalently

$$E[X \cdot 1_{A \cap B}] = E[E[X \mid \mathcal{H}] \cdot 1_{A \cap B}] \quad \forall A \in \mathcal{G}, B \in \mathcal{H}. \quad (17)$$

Indeed, $P$ is a $\pi$-system (with $\sigma(P) = \sigma(\mathcal{G}, \mathcal{H})$) that is trivially included in the smallest Dynkin system $D$ including $P$ (this is the system that includes all sets in $P$, is closed under complements and under countable unions of disjoint sets). Moreover, if (17) holds for every set in $P$, it is easy to see (by linearity of the integral and dominated convergence) that the corresponding property also holds for every set in $D$, since for a set $A \in D$ we have $1_A^D = 1 - 1_A$ and for disjoint sets $(A_k)_{k \geq 1} \in D$ we have $1_{\cup_{k \geq 1} A_k} = \sum_{k \geq 1} 1_{A_k}$. Dynkin’s $\pi$-$\lambda$ theorem implies that $\sigma(P) \subset D$, therefore the third point follows.

To show (17), note that $\sigma(X, \mathcal{H})$ being independent of $\mathcal{G}$ implies that $\mathcal{H}$ is independent of $\mathcal{G}$. Therefore,

$$E[X \cdot 1_{A \cap B}] = E[X \cdot 1_B 1_A] = E[X \cdot 1_B] E[1_A] \quad (\sigma(X, \mathcal{H}) \text{ indep. of } \mathcal{G})$$

(tower property)

$$= E[E[X \cdot 1_B \mid \mathcal{H}] \cdot 1_A] \quad (\mathcal{H} \text{ indep. of } \mathcal{G})$$

$$= E[E[X \cdot 1_B \mid \mathcal{H}] \cdot 1_A] \quad (1_B \text{ is } \mathcal{H}\text{-measurable})$$

completing the proof.

**Corollary A.2.** If $\sigma(X, \mathcal{H})$ is independent of $\mathcal{G}$ then $E[\phi(X) \mid \sigma(\mathcal{G}, \mathcal{H})] = E[\phi(X) \mid \mathcal{H}]$ for all measurable and integrable functions $\phi$.

**Proof.** This follows directly from Proposition A.1 upon replacing $X$ by $\phi(X)$ and noting that $\sigma(\phi(X), \mathcal{H}) \subseteq \sigma(X, \mathcal{H})$, implying the needed condition. \qed
It should now be apparent that the assumption that $X$ and $G$ are independent is insufficient if we want to show $E[X \mid \sigma(G, H)] = E[X \mid H]$ (a counterexample is provided in Hansen (2015)). However, we can actually make a weaker assumption (though still stronger than the assumption that $X$ and $G$ are independent) to attain the same result, as shown in the next proposition.

**Definition A.3.** $X$ and $G$ are conditionally independent given $H$ if for all $x \in \sigma(X)$, $A \in G$,

$$E[1_x 1_A \mid H] = E[1_x \mid H] E[1_A \mid H].$$

**Proposition A.4.** If $X$ is conditionally independent of $G$ given $H$ then $E[X \mid \sigma(G, H)] = E[X \mid H]$.

**Proof.** Following the previous proof it is clear that we only need to show

$$E[X \cdot 1_{A \cap B}] = E[E[X \mid H] \cdot 1_{A \cap B}]$$

for all $A \in G$ and $B \in H$. We do this by measure-theoretic induction (as in Durrett (2010, Proof of Theorem 1.6.9)), in particular, we proceed in a four-part case distinction of $X$. In the first case, assume $X$ is an indicator function, i.e., $X = 1_x$ for some $x \in F$. Then

$$
E[X \cdot 1_{A \cap B}] = E[1_x 1_{A \cap B}]
= E[E[1_x 1_{A \cap B} \mid H]]
= E[E[1_x 1_A \mid H] 1_B]
= E[E[1_x \mid H] E[1_A \mid H] 1_B]
= E[E[1_x \mid H] E[1_A \mid H] 1_B]
= E[E[1_x \mid H] \cdot 1_{A \cap B}]
= E[E[X \mid H] \cdot 1_{A \cap B}].
$$

Thus in the most basic case we have the required property. In the second case we let $X = \sum_{i=1}^k c_i 1_{x_i}$ be a finite weighted sum of indicator random variables, where $k \in \mathbb{N}, c_i \in \mathbb{R}$ and $x_i \in F$. The result of the first case combined with linearity of expectation immediately shows that the property holds for $X$ in this form too.

In the third case, we assume $X$ is some non-negative function. We define a random variable $X_k$ that is simple and such that $X_k \uparrow X$ as $k \to \infty$. For example, we can take $X_k = \sum_{i=0}^{k^2-1} \frac{1}{2^k} 1_{\{\frac{i}{2^k} \leq X < \frac{i+1}{2^k}\}} + k 1_{X \leq k}$, as in Durrett (2010, Proof of Theorem 1.6.9). Then, by monotone convergence and the previous case, the property also holds when $X$ is an arbitrary non-negative function.

Finally, in the fourth case, we let $X$ be an arbitrary integrable function. Then we can write $X = X^+ - X^-$ where $f^+(x) := \max\{0, f(x)\}$ and $f^-(x) := \min\{0, f(x)\}$. Integrability of $X$ means that $X^+$ and $X^-$ are themselves integrable. We can use linearity of expectation and the previous cases to conclude that, in this general setting, the property still holds. This concludes the proof by measure-theoretic induction.

**Proposition A.5.** $X$ is conditionally independent of $G$ given $H$ if and only if $E[\phi(X) \mid \sigma(G, H)] = E[\phi(X) \mid H]$ for all measurable and integrable functions $\phi$.

**Proof.** The “$\Rightarrow$” direction is a simple corollary of Proposition A.4 that follows with the same proof upon replacing $X$ by $\phi(X)$ and noticing that $\phi(X)$ is a $\sigma(X)$-measurable random variable.

The “$\Leftarrow$” direction is also easy to see. Let $x \in \sigma(X)$, $A \in G$ and define $\phi$ such that $\phi(X) = 1_x$. Then we have

$$
E[1_x 1_A \mid H] = E[E[1_x 1_A \mid \sigma(G, H)] \mid H]
= E[E[1_x \mid \sigma(G, H)] 1_A \mid H]
= E[E[1_x \mid H] 1_A \mid H]
= E[1_x \mid H] E[1_A \mid H]
= E[1_x \mid H] E[1_A \mid H].
$$

This completes the proof.
Proposition A.6. If $X$ is conditionally independent of $G$ given $H$, and $Y$ is independent of $\sigma(X, G, H)$ then for any measurable and integrable function $f$ also $f(X, Y)$ is conditionally independent of $G$ given $H$.

Proof. By Proposition A.5 we have to show that for any measurable $\phi$ we have $E[\phi(f(X, Y)) | \sigma(G, H)] = E[\phi(f(X, Y)) | H]$. First note that it is enough to show this for $\phi$ being the identity, since both $\phi$ and $f$ are arbitrary measurable and integrable functions. Then we have for $g(x) := E_Y[f(X, Y)]$, which is a measurable and integrable function again, that

$$
E[f(X, Y) | \sigma(G, H)] = E[E[f(X, Y) | \sigma(X, G, H)] | \sigma(G, H)] 
= E[g(X) | \sigma(G, H)] 
= E[g(X) | H] 
= E[f(X, Y) | H],
$$

as wanted, proving the claim. □

Proposition A.7. If $X$ is conditionally independent of $Y$ given $H$ then

$$
E[XY | H] = E[X | H] E[Y | H].
$$

Proof. From Proposition A.4 we know that $E[X | \sigma(Y, H)] = E[X | H]$. Therefore,

$$
E[XY | H] = E[E[XY | \sigma(Y, H)] | H] 
= E[E[X | \sigma(Y, H)] Y | H] 
= E[E[X | H] Y | H] 
= E[X | H] E[Y | H],
$$

as wanted. □

Proposition A.8. If $X$ is independent of $\sigma(G, H)$ then $X$ is conditionally independent of $G$ given $H$.

Proof. Follows directly from Proposition A.5. □

Proposition A.9. Let $X, Y, Z$ be random variables such that $X$ and $Y$ are conditionally independent given $\sigma(Z)$. Then, for any measurable function $\phi$ we have

$$
E[\phi(X, Y) | \sigma(Z, Y)] = E[\phi(X, y) | \sigma(Z)] |_{y=Y}.
$$

We follow the proof of the somewhat less general result Durrett (2010, Lemma 6.2.1).

Proof. By its definition, it is clear that the r.h.s. is $\sigma(Y, Z)$-measurable. Hence, by the definition of the conditional expectation, we only need to show that

$$
E[\phi(X, Y) 1_C] = E[E[\phi(X, y) | \sigma(Z)] |_{y=Y} 1_C],
$$

(18)

for all $C \in \sigma(Y, Z)$.

Assume first that $\phi(x, y) = 1_A(x) 1_B(y)$ and let $C \in \sigma(Y, Z)$ be of the form $C = \{Y \in C_1, Z \in C_2\}$ for Borel sets $A, B, C_1, C_2$ in the images of $X, Y, Z$, respectively. Then,

$$
E[\phi(X, Y) 1_C] = E[E[1_A(X) 1_B(Y) 1_{C_1}(Y) 1_{C_2}(Z) | \sigma(Z)]]
= E[1_{C_2}(Z)] E[1_A(X) | \sigma(Z)] E[1_B(Y) 1_{C_1}(Y) | \sigma(Z)],
$$

(19)

using the tower property in the first and conditional independence in the second equality. Moreover, note that

$$
E[\phi(X, y) | \sigma(Z)] = E[1_A(X) 1_B(y) | \sigma(Z)] = 1_B(y) E[1_A(X) | \sigma(Z)],
$$

for all $C \subseteq \sigma(Y, Z)$.
hence,

\[
E \left[ \mathbb{E} \left[ \phi(X, y) \mid \sigma(Z) \right] \mathbb{I}_{y=Y} \mathbb{I}_C \right] = E \left[ \mathbb{I}_B(Y) \mathbb{E} \left[ \mathbb{I}_A(X) \mid \sigma(Z) \right] \mathbb{I}_C \right] \\
= E \left[ \mathbb{E} \left[ \mathbb{I}_B(Y) \mathbb{E} \left[ \mathbb{I}_A(X) \mid \sigma(Z) \right] \mathbb{I}_C(Z) \mathbb{I}_{C_1}(Z) \mid \sigma(Z) \right] \right] \\
= E \left[ \mathbb{E} \left[ \mathbb{I}_A(X) \mid \sigma(Z) \right] \mathbb{E} \left[ \mathbb{I}_B(Y) \mathbb{I}_C(Z) \mid \sigma(Z) \right] \right],
\]

using the tower property in the second and measurability in the third equality. (19) and (20) together show that (18) holds for the simple functions and the special \( C \). Next, we note that by Dynkin’s \( \pi\)-\( \lambda \) theorem, the same holds for general \( C \subset \sigma(Y, Z) \) (cf. proof of Proposition A.1). Then the claim follows for all bounded functions by the monotone class theorem (Durrett, 2010, Theorem 6.1.3), as outlined in the proof of Durrett (2010, Lemma 6.2.1). Moreover, it follows for general measurable functions by dominated convergence.

\[\square\]

**B Inductive Bias**

In this section we first discuss the inductive bias induced by the old loss function. Afterwards we discuss the general inductive bias introduced by the PD-NJ-ODE (2). In this paper, we have shown that under certain assumptions our estimator based on the new loss (10) is consistent, and thus asymptotically unbiased as the model complexity \( m \) and the number of training paths \( N \) tend to infinity. However in such an infinite hypothesis space it is impossible to obtain an unbiased estimator for any finite number of training paths \( N \in \mathbb{N} \). For this reason, it is helpful to have an appropriate inductive bias that guides the estimator in the right direction when there is a limited amount of training data (Mitchell, 1980).\(^{11}\)

### B.1 Inductive Bias of the Loss

While the new loss (10) is asymptotically unbiased, the old loss (3) is in general always biased. Even in the limit \( m, N \to \infty \) this bias induced by the old loss (3) does not vanish. On average, models obtained from the old loss will jump too closely to new observation \( O_{t_k} \) at observation times \( t_k \) as can be observed in the right subplot of Figure 1. However, in the case of no noise (i.e., \( X_{t_k} = O_{t_k} \)), the correct model should always exactly jump to new observations. Hence, teaching the model explicitly to jump to new observations via the term \( |M_k \odot (X_{t_k} - Z_{t_k})|^2 \) in the old loss (3) incorporates helpful prior knowledge which is particularly helpful for small training datasets. Even if the observations \( O_{t_k} \) are noisy (with a small noise-scale \( \sqrt{\mathbb{E}[\sigma^2]} \)), the term \( |M_k \odot (O_{t_k} - Z_{t_k})|^2 \) in the old loss (3) can be a beneficial inductive bias, if one only has access to a very small training dataset. If one has very little information from other observations about the dynamics of \( X \), then it might be a reasonable best guess to jump close to new observations \( O_{t_k} \) at observation times \( t_k \).

For example, if the model \( Y_{t_k}^{\text{lim}}(X \leq t_k - 1) \) obtained from the new loss (10) is further away from the true conditional expectation \( \mathbb{E}[X_t \mid A_{t_k-1}] \) than the noisy observation \( O_{t_k} \), the term \( |M_k \odot (O_{t_k} - Z_{t_k})|^2 \) in the old loss (3) would be helpful to push the model closer to the truth. This explains why we can see in Figure 3 that the old loss (3) outperforms the new loss (10) in the case of very small noise relative to the overall variability of the data.\(^{12}\)

Furthermore, different weighting schemes in the loss also influence the inductive bias (see Remark 4.10). The term \( |M_k \odot (O_{t_k} - Z_{t_k})|^2 \) in the old loss (3) can be particularly helpful if the loss is weighted via \( \Delta t_k \).

---

\(^{11}\)Occam’s razor suggests, that one should always pick the simplest (or most parsimonious) model that explains the data, i.e., the inductive bias should be directed towards simplicity/parsimony. Different machine learning (ML) models incorporate simplicity/parsimony in various different ways, so one should pick the right model according to the alignment of the inductive bias of the model with one’s prior belief.

\(^{12}\)In this experiment, we trained also only for 100 epochs. Additionally to the statistical learning argument given before, there is probably also an algorithmic effect, that the term \( |M_k \odot (O_{t_k} - Z_{t_k})|^2 \) in the old loss (3) helps the model to move directly in the right direction starting from the first epochs. At least during the first epochs the model \( Y_{t_k}^{\text{lim}}(X \leq t_k - 1) \) is usually very far off (at observation times \( t_k \)) compared to the noisy \( O_{t_k} \) observations thus the term \( |M_k \odot (O_{t_k} - Z_{t_k})|^2 \) in the old loss (3) pushes the model directly in the right direction at observation times, while the new loss (10) only has an indirect affect on the model at the observation times \( t_k \).
B.2 Inductive Bias of the PD-NJ-ODE architecture

As shown in our main theorem the PD-NJ-ODE architecture (2) is universal and when it is trained with the right loss, such as (10), it is asymptotically unbiased. However, for any finite number of training paths \( N \), there are infinitely many models (i.e., infinitely \( \theta \in \Theta \)) that all perfectly fit the data, i.e., achieve \( \hat{\Phi}_N(\theta) = 0 \), but correspond to very different predictions \( Y^\theta_0(X^{\leq s}) \) on unseen test data. Therefore, for \( N < \infty \), our architecture (including the training algorithm) necessarily needs to have an appropriate inductive bias to make a reasonable choice among all the models with reasonable low training loss \( \hat{\Phi}_N(\theta) \).

(Additionally we want this bias to vanish in the limit \( N \to \infty \), which is already given by our theory.) First, we give some intuition on our understanding of the important inductive bias of PD-NJ-ODE (2) and afterwards we will give some theoretical arguments that motivate these hypotheses. Most machine learning models have and inductive bias towards some form of simplicity for the function mapping inputs to predictions. E.g., some NN-architectures have an inductive bias towards low second derivative of this function w.r.t. the inputs. Therefore, we do not only need to choose among models that achieve equal training loss \( \rho \), but also need to choose among models that learn a function that is as simple as possible.

Example B.1. Consider the situation where the true underlying process of \( X \) follows a linear SDE (driven by a Brownian motion, i.e., a linear Itô-diffusion), then the conditional expectation \( \hat{X} \) follows a linear ODE (except the jumps at observation times \( t_k \), cf. Krach et al. (2022, Example B.2)). For simplicity consider the case of complete noiseless observations \( X_{t_k} \) (i.e., \( M = 1 \) and \( \epsilon = 0 \)). Then \( \hat{X} \) can be perfectly expressed by our PD-NJ-ODE (2) with 3 linear functions \( f_{\theta_1}, \rho_{\theta_2}, \hat{g}_{\theta_3} \) instead of simply favouring simplicity of the map \( (X^{\leq s}, t) \mapsto Y^\theta_0(X^{\leq s}) \). E.g., \( f_{\theta_1} \) is the right-hand-side of the ODE resulting in \( \hat{X} \), \( \rho_{\theta_2} \) maps new observations \( X_{t_k} \) simply to \( X_{t_k} \), and \( \hat{g}_{\theta_3} \) is simply the identity. As these 3 functions are linear, their second derivative is zero. Hence, for example, the inductive bias of L2-regularized ReLU-NNs \( f_{\theta_1}, \rho_{\theta_2}, \hat{g}_{\theta_3} \) strongly favours such simple linear functions. Therefore, the model can learn these linear dynamics already from a relatively small amount of training paths and extrapolate well far into the future. On the other hand, in this setting the map \( (X^{\leq s}, t) \mapsto X_{t_k}(X^{\leq s}) \) grows exponentially in the input \( t \) and thus has huge (second) derivative with respect to \( t \), which is considered the opposite of simple by most of the commonly used ML-models. Hence, if one trains, on the same dataset, a feed-forward neural network \( \bar{F}_0 : (X^{\leq s}, t) \mapsto \bar{F}_0(X^{\leq s}, t) \) that directly approximates \( F \), the model would find many other “simpler”/flatter functions \( \bar{F}_0 \) that fit the finite training dataset equally well. Even for larger \( N \), such a model, favouring a low Sobolev-norm of the map \( \bar{F}_0 \), would always have troubles to make predictions far ahead into the future \( t > \tau(t) \): For values of \( t - \tau(t) \) larger than the largest \( t_k - t_{k-1} \) ever observed in training, such a model has no reason to keep growing exponentially but would rather extrapolate as flat as possible. In contrast, \( Y^\theta_t(X^{\leq t\tau(t)}) \), obtained from our PD-NJ-ODE with linear functions \( f_{\theta_1}, \rho_{\theta_2}, \hat{g}_{\theta_3} \), would keep the exponential growth in \( t \) for arbitrarily large values of \( t - \tau(t) \) even far out of sample and even far out of distribution. In line with this, it is expected that the PD-NJ-ODE leads to better results than the model \( \bar{F}_0 \), when evaluated on previously unseen test data. This was confirmed by the experiment in Krach et al. (2022, Appendix E).

To summarize: While the PD-NJ-ODE (2) is able to learn arbitrarily complicated dynamics (fulfilling some mild regularity assumptions), it favours models \( \theta \) with “simple” functions \( f_{\theta_1}, \rho_{\theta_2}, \hat{g}_{\theta_3} \) among different models with equal training loss \( \hat{\Phi}_N(\theta) \), where the definition of “simplicity” depends on the architectures used for \( f_{\theta_1}, \rho_{\theta_2}, \hat{g}_{\theta_3} \). E.g., for L2-regularized ReLU-NNs the notion of simplicity relates to the second derivative of the functions.
This can lead to better generalization performance.

B.2.1 Inductive Bias towards Multi-task Learning

If one trains multiple tasks together (i.e., if one wants to forecast a multi-dimensional process $X$), many deep learning applications have shown that an inductive bias towards functions that use similar features among multiple tasks (rather than using very different features for each of these tasks) can be highly beneficial. This phenomenon is known as multi-task learning (Caruana et al., 1997). One can argue that multi-task learning (highly related to feature learning, representation learning, transfer learning, metric learning) is one of the biggest strengths of deep learning models compared to shallow learning models in terms of generalization (Heiss et al., 2022). The inductive bias of our PD-NJ-ODE (2) towards feature sharing (i.e., multi-task learning) is twofold:

1. Per definition the hidden state $H_t$ in (2) is shared among all tasks/outputs. I.e., each coordinate of the output $Y_t$ is a function of the same hidden state $H_t$. Therefore, the hidden state $H_t$ becomes more parsimonious, if it mainly contains features that are helpful for all coordinates of $Y_t$.

2. If the neural network architectures of $f_{\theta_1}, \rho_{\theta_2}, \tilde{g}_{\theta_3}$ contain trainable parameters in multiple hidden layers, then these functions favour feature sharing within their hidden layers, as shown by Heiss et al. (2022) in the case of L2-regularization. Caruana et al. (1997); Heiss et al. (2022) provide more intuition on multi-task learning.

B.2.2 Implicit Regularization

Even without any explicit L2-regularization and even for very large theoretical model complexity $m$ (i.e., large $\Theta_m$ in our setting), it is widely assumed that (stochastic) gradient descent (initialized close to zero) induces an implicit inductive bias towards simplicity (Neyshabur, 2017). In some settings this implicit regularization is extremely similar (or even exactly identical) to explicit L2-regularization (Heiss et al., 2019) and in other settings it can be (slightly) different. In our setting of training PD-NJ-ODEs, implicit regularization of the gradient descent does most probably not exactly correspond to explicit L2-regularization, but on a high level the qualitative behaviour might be similar.\footnote{Note that a classical gradient flow is always identical to a gradient flow with respect to the L2 norm and thus locally moves as little w.r.t. the L2-distance as possible. Therefore, it seems plausible that gradient descent does not increase the L2-norm of $\theta$ much more than necessary to achieve a certain training loss.}

In theory, $m$ could also be too large in relation to $N$, since too large $m$ allows for over-fitting. In contrast to this, $N$ can never be too large. In particular, if one picks an $\theta_{\min}^{m,N} \in \Theta_{m,N}$ with unnecessarily large L2-norm over-fitting can be a serious problem. However, in practice, too many neurons are usually not a problem due to implicit regularization of (stochastic) gradient descent as suggested empirically by Herrera et al. (2021, Figure 6) and theoretically by Neyshabur (2017); Savarese et al. (2019); Ongie et al. (2019); Heiss et al. (2019; 2023; 2022); Parhi & Nowak (2022).

B.2.3 Mathematical Theory on the Inductive Bias of PD-NJ-ODEs

For implicit regularization it seems difficult to get precise global results for PD-NJ-ODEs. However, for explicit L2-regularization of $\theta$, there is much more hope to get nice theoretical results on the inductive bias of the PD-NJ-ODEs (2). (Note that the bound $m$ on the L2-norm of $\theta$ in the definition of $\Theta_m$ corresponds to L2-regularization in a Lagrangian sense.\footnote{The L2-regularization regularization parameter $\lambda_{\text{reg}}$ can be seen as a KKT-multiplier of the constraint $|\theta|^2 < m$. However, if $m$ is too large the constraint is not active and $\theta_{\min}^{m,N}$ is highly undetermined. For L2-regularization with $\lambda_{\text{reg}} > 0$, the parameters $\theta_{\min}^{m,N}$ are usually also not uniquely defined, but the solution $Y_{\min}^{m,N}$ can be unique.}) The results by Savarese et al. (2019); Ongie 16\footnotetext{In the case of L2-regularization Heiss et al. (2022) shows (for certain architectures) that this multi-task learning effect does not vanish when the width of the network increases, while it would vanish with increasing width for some specific other training methods. For most deep learning methods used in practice, there is some multi-task learning effect assumed to be present even if it is theoretically not always as well understood as for L2-regularized wide ReLU-NNs.}
et al. (2019); Heiss et al. (2019; 2023; 2022); Parhi & Nowak (2022) strongly support the hypothesis that
\[ \arg\min_{\theta} \left( \hat{\Phi}_N(\theta) + \lambda_{\text{reg}} \left| \theta \right|^2 \right) \]
exactly converges to \( \arg\min_{f,\rho,\hat{g}} \left( \hat{\Psi}_N(Y_f^{\rho,\hat{g}}) + \lambda_{\text{reg}} \left( P_1(f) + P_2(\rho) + P_3(\hat{g}) \right) \)\),
where \( P_1 \) describe some regularization functionals (such as weighted \( L^p \) norms of the second derivative)
depending on the exact architectures used for \( f, \rho, \hat{g} \) given in Savarese et al. (2019); Ongie et al. (2019); Heiss et al. (2019; 2023; 2022); Parhi & Nowak (2022). However, (for rather technical reasons) their proofs do not
directly apply to PD-NJ-ODEs. The formulation of the main theorem in Heiss et al. (2023) could be
applied more directly to PD-NJ-ODE by considering the map \( (f, \rho, \hat{g}) \mapsto \hat{\Psi}_N(Y_f^{\rho,\hat{g}}) \) as the loss functional.\(^{19}\)

Proofing such results rigorously would be interesting future work.

C Combining the Two Extensions with Full Proof

In this section we prove convergence of the PD-NJ-ODE to the optimal prediction in the most general setting
that allows for non-Markovian processes with irregular incomplete noisy observations, where dependence
between the observation framework and the process is possible.

C.1 Setting

As in Section 4.2 we consider only the probability space \( (\Omega, \mathcal{F}, \mathbb{P}, \mathbb{P}) \) on which we define \( X, n, K, t_i, \tau, M \). As
in Section 3.1, we additionally define the observation noise \( (\epsilon_k)_{0 \leq k \leq K} \), the noisy observations \( O_{t_k} := X_{t_k} + \epsilon_k \) for \( 0 \leq k \leq n \), the filtration of the currently available information via
\[ \mathcal{A}_t := \sigma \left( O_{t_i, j, t_i}, M_{t_i}, t_i \leq t, j \in \{ 1 \leq l \leq d_X | M_{t_i, l} = 1 \} \right), \]
\( \hat{O}^{\leq t} \) and the functions \( F_j \) such that \( \hat{X}_{t, j} = F_j(t, \tau(t), \hat{O}^{\leq \tau(t)}) \). Then we make the following assumptions.

Assumption C.1. We assume that:

(i) \( M_{0, j} = 1 \) for all \( 1 \leq j \leq d_X \) (\( X \) is completely observed at 0) and \( |M_k|_1 > 0 \) for every \( 1 \leq k \leq K \)
\( \mathbb{P} \)-almost surely (at every observation time at least one coordinate is observed).

(ii) The probability that any two observation times are closer than \( \epsilon > 0 \) converges to 0 when \( \epsilon \) does, i.e.,
if \( \delta(\omega) := \min_{0 \leq i \leq n(\omega)} |t_{i+1}(\omega) - t_i(\omega)| \) then \( \lim_{\epsilon \to 0} \mathbb{P}(\delta < \epsilon) = 0 \).

(iii) Almost surely \( X \) is not observed at a jump, i.e., \( \mathbb{P}(t_i \in J | i \leq n) = \mathbb{P}(\Delta X_i \neq 0 | i \leq n) = 0 \) for all \( 1 \leq i \leq K \).

(iv) \( F_j \) are continuous and differentiable in their first coordinate \( t \) such that their partial derivatives with
respect to \( t \), denoted by \( f_j \), are again continuous and there exists a \( B > 0 \) and \( p \in \mathbb{N} \) such that for every \( t \in [0, T] \) the functions \( f_j, F_j \) are polynomially bounded in \( X^* \), i.e.,
\[ |F_j(\tau(t), \tau(t), \hat{O}^{\leq \tau(t)})| + |f_j(t, \tau(t), \hat{O}^{\leq \tau(t)})| \leq B(X^*_i + 1)^p + B \sum_{i=0}^{n} |\epsilon_i|. \]

(v) \( X^* \) is \( L^{2p} \)-integrable, i.e., \( \mathbb{E}[|X^*_j|^{2p}] < \infty \).

(vi) \( n \) is square-integrable, i.e., \( \mathbb{E}[n^2] < \infty \).

(vii) The i.i.d. \( \epsilon_k \) are independent of \( X, n, M, (t_i)_{1 \leq i \leq K}, \) are centered and square-integrable, i.e., \( \mathbb{E}[\epsilon_k] = 0 \)
and \( \mathbb{E}[|\epsilon_k|^2] < \infty \).

(viii) For every \( 1 \leq i \leq n \), \( X_{t_i, -} \) is conditionally independent of \( \sigma(n, M_{t_i}) \) given \( \mathcal{A}_{t_i, -}. \)

\(^{19}\) One would need to check if this loss functional meets the assumptions of Heiss et al. (2023) for a rigorous proof.

\(^{20}\) More precisely, one should formulate this assumption as “For every \( 1 \leq i \leq K \), \( X_{t_i, -} \) is conditionally independent of \( \sigma(n, M_{t_i}) \) given \( \mathcal{A}_{t_i, -}. \)”

36
(ix) For all $1 \leq k \leq K$, $1 \leq j \leq d_X$ there is some $\eta_{k,j} > 0$ such that $\mathbb{P}(M_{k,j} = 1 \mid \sigma(n, A_{t_k})) > \eta_{k,j}$ (i.e., given the currently known information and $n$, for each coordinate the probability of observing it at the next observation time is positive).

(x) We assume that for every $1 \leq k \leq K$ the process $X$ is conditionally independent of $t_k$ given $A_{t_{k-1}}$.

Remark C.2. All the relaxations and extensions discussed in Sections 2 to 5 hold also in this combined setting.

As in Section 3.1, the PD-NJ-ODE uses the noisy observations $O_{t_i}$ and $\tilde{O} \leq \tau(t)$ as inputs instead of $X_{t_i}$ and $\hat{X} \leq \tau(t)$ and is trained with the objective functions $(6)$ respectively $(7)$ and their Monte Carlo approximations.

C.2 Convergence Theorem

Theorem C.3. If Assumption C.1 is satisfied and using the definitions of Appendix C.1, the claims of Theorem 2.7 hold equivalently, upon replacing the original loss functions and their Monte Carlo approximations by their noise-adapted versions. In particular, we obtain convergence of our estimator $Y_{d_m,N_m}^{\min}$ to the true conditional expectation $\hat{X}$ in $d_k$.

We start with the following result which is a combination of Lemma 3.4 and Lemma 4.6.

Lemma C.4. For any $A_d$-adapted process $Z$ it holds that

$$
E \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{t_i} \odot (O_{t_i} - Z_{t_i})|^2 \right] = E \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{t_i} \odot (O_{t_i} - \hat{X}_{t_i})|^2 \right] + E \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{t_i} \odot (\hat{X}_{t_i} - Z_{t_i})|^2 \right].
$$

Proof. First note that by Assumption C.1 point (iii) we have that $X_{t_i} = X_{t_i}$ almost surely and when defining $O_{t_i} := X_{t_i} + \epsilon_i$ we therefore also have that $O_{t_i} = O_{t_i}$ almost surely. Next notice that Assumption C.1 point (viii) together with Proposition A.6 imply that $O_{t_j}$ is conditionally independent of $\sigma(n, M_{t_i})$ given $A_{t_j}$. Hence, for $\hat{O}_{t_j} := E[O_{t_j} \mid A_{t_j}]$ it follows as in Lemma 4.6 that

$$
E \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{t_i} \odot (O_{t_i} - Z_{t_i})|^2 \right] = E \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{t_i} \odot (O_{t_i} - \hat{O}_{t_j})|^2 \right] + E \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{t_i} \odot (\hat{O}_{t_j} - Z_{t_i})|^2 \right].
$$

Then we can conclude the proof as in Lemma 3.4, by noting that with Proposition 4.5 we obtain,

$$
\hat{O}_{t_j} = \hat{X}_{t_j} + E[\epsilon_i \mid A_{t_j}] = \hat{X}_{t_j} + E[\epsilon_i] = \hat{X}_{t_j},
$$

using that $\epsilon_i$ has expectation 0 and is independent of $A_{t_j}$.

In the following, we use the notation $\Theta_i^{m}$ and $\tilde{\Theta}_i^{m}$ if we speak of the projections of the sets on the weights $\theta_i$ and $\tilde{\theta}_i$ respectively.

Proof of Theorem C.3 - Part 1. We start by showing that $X \in \mathcal{D}$ is the unique minimizer of $\Psi$ up to indistinguishability (as defined in Definition 2.4). Note that for every $t_i$ we have $M_{t_i} \odot X_{t_i} = M_{t_i} \odot X_{t_i}$ and that $X_{t_i} = X_{t_i}$ if $t_i \notin J$, hence with probability 1. It follows directly from Lemma C.4 that $X$ is a minimizer of $\Psi$, since

$$
\Psi(Z) = \Psi(\hat{X}) + E \left[ \frac{1}{n} \sum_{i=1}^{n} |M_{t_i} \odot (\hat{X}_{t_i} - Z_{t_i})|^2 \right].
$$

21In particular, we replace the original empirical loss $\Phi_N(\theta) = (5)$ by the noise-adapted empirical loss function,

$$
\Phi_N(\theta) := \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} |M_{t_i}^{(j)} \odot (O_{t_i}^{(j)} - Y_{\theta_i}^{(j)})|^2.
$$

22For a convergence result at times $t$ that are not observation times, see Section 5.2.
Before we can show uniqueness of $\hat{X}$, we need some additional results. For those, let $Z \in \mathbb{D}$. Let $c_1 := \mathbb{E} [n]^{1/2} \in (0, \infty)$, then the Hölder inequality, together with the fact that $n \geq 1$, yields

$$\mathbb{E} \left[ |Z|_2 \right] = \mathbb{E} \left[ \frac{\sqrt{n}}{n} |Z|_2 \right] \leq c_1 \mathbb{E} \left[ \frac{1}{n} |Z|_2^2 \right]^{1/2}. \quad (23)$$

By Lemma 4.8 and by the equivalence of 1- and 2-norm we have for some constant $c_3 > 0$ and for any $1 \leq k \leq K$

$$\mathbb{E} \left[ \mathbb{I}_{\{n \geq k\}} \left| \hat{X}_{t_k} - Z_{t_k} \right|_2 \right] \leq \frac{c_3}{c_2(k)} \mathbb{E} \left[ \mathbb{I}_{\{n \geq k\}} \left| M_{t_k} \odot (\hat{X}_{t_k} - Z_{t_k}) \right|_2 \right]. \quad (24)$$

To see that $\hat{X}$ is the unique minimiser up to indistinguishability, let $Z \in \mathbb{D}$ be a process which is not indistinguishable from $X$. Hence, there exists some $1 \leq k \leq K$ such that $d_{k}(\hat{X}, Z) > 0$. We have

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \odot (\hat{X}_{t_i} - Z_{t_i}) \right|_2^2 \right] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{K} \mathbb{I}_{\{n \geq i\}} \left| M_{t_i} \odot (\hat{X}_{t_i} - Z_{t_i}) \right|_2^2 \right] \geq \mathbb{E} \left[ \frac{1}{n} \mathbb{I}_{\{n \geq k\}} \left| M_{t_k} \odot (\hat{X}_{t_k} - Z_{t_k}) \right|_2^2 \right] \geq c_1^{-2} \mathbb{E} \left[ \mathbb{I}_{\{n \geq k\}} \left| M_{t_k} \odot (\hat{X}_{t_k} - Z_{t_k}) \right|_2^2 \right]^2 \geq \left( \frac{c_2}{c_1 c_3} \right)^2 \mathbb{E} \left[ \mathbb{I}_{\{n \geq k\}} \left| \hat{X}_{t_k} - Z_{t_k} \right|_2 \right]^2 \geq \left( \frac{c_2}{c_0 c_1 c_3} \right)^2 d_{k}(\hat{X}, Z)^2 > 0,$$

where we used (23) for the 3rd, (24) for the 4th and (1) for the last line. Together with (22) this implies $\Psi(Z) > \Psi(\hat{X})$.

Next we show that (2) 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 $\hat{\theta}_3$ such that $\hat{g}_{\hat{\theta}_3} = \text{id}$, which is possible since we assumed that $\text{id} \in \mathcal{N}$. Let $\varepsilon > 0$, $N_{\varepsilon} := \left[ 2(T + 1) \varepsilon^{-4} \right]$ (implying that $\lim_{n \to 0} N_{\varepsilon} = \infty$) and $\mathcal{P}_{\varepsilon} \subseteq \text{BV}^c([0, T])$ be the closure of the set $A_{N_{\varepsilon}}$ of Krach et al. (2022, Remark 3.11), which is compact. For any $1 \leq j \leq d_X$, the function $f_j$ is continuous by Assumption C.1 and (by abuse of notation) equivalently be written as a (continuous) function $f_j(t, \tau(t), \hat{O} \leq \tau(t) - \hat{O}_0, \hat{O}_0)$. Therefore, Krach et al. (2022, Proposition 3.8) implies that there exists an $m_0 = m_0(\varepsilon) \in \mathbb{N}$ and a continuous function $\hat{f}_j$ such that

$$\sup_{(t, \tau, X) \in [0, T]^2 \times \mathcal{P}_{\varepsilon}} \left| f_j(t, \tau, X) - \hat{f}_j(t, \tau, \pi_{m_0}(X - X_0), X_0) \right| \leq \varepsilon/2.$$

Since the variation of functions in $\mathcal{P}_{\varepsilon}$ is uniformly bounded by a finite constant, the set of their truncated signatures $\pi_{m_0}(\mathcal{P}_{\varepsilon})$ is a bounded subset in $\mathbb{R}^d$ for some $d \in \mathbb{N}$ (depending on $d_X$ and $m_0$), hence its closure, denoted by $\Pi_{\varepsilon}$, is compact. Therefore, the universal approximation theorem for neural networks (Hornik et al., 1989, Theorem 2.4) implies that there exists an $m_1 = m_1(\varepsilon) \in \mathbb{N}$ and neural network weights $\hat{\theta}_1^{m_1} \in \hat{\Theta}_1^{m_1}$ such that for every $1 \leq j \leq d_X$ the function $\hat{f}_j$ is approximated up to $\varepsilon/2$ by the $j$-th coordinate of the neural network $\hat{\theta}_1^{m_1} \in \mathcal{N}$ (denoted by $\hat{\theta}_1^{m_1,j}$) on the compact set $[0, T]^3 \times \Pi_{\varepsilon}$. Hence, combining the two approximations we get (by triangle inequality)

$$\sup_{(t, \tau, X) \in [0, T]^2 \times \mathcal{P}_{\varepsilon}} \left| f_j(t, \tau, X) - \hat{f}_1^{m_1,j}(t, \tau, \pi_{m_0}(X - X_0), X_0) \right| \leq \varepsilon.$$

Obviously, extending the input of the neural network does not make the approximation worse, by simply setting the corresponding weights to 0, hence, also $H_{t, -}$ can be used as additional input. Similarly we get that there exists an $m_2 = m_2(\varepsilon) \in \mathbb{N}$ and neural network weights $\hat{\theta}_2^{m_2} \in \hat{\Theta}_2^{m_2}$ such that for every $1 \leq j \leq d_X$

$$\sup_{(t, X) \in [0, T] \times \mathcal{P}_{\varepsilon}} \left| F_j(t, t, X) - \hat{\rho}_{2}^{m_2,j}(t, \pi_{m_1}(X - X_0), X_0) \right| \leq \varepsilon.$$
As before, $H_{t-}$ can be used as additional input without worsening the approximation.

Next we define the bounded output neural networks based on these neural networks. For this let us define

$$\gamma_1 := \max_{(t,\tau,\xi) \in [0,T]^2 \times \mathcal{F}_\tau} \left| \hat{\theta}_{t-1}^{\star, m_1}(t, \tau, \pi_{m_0}(X - X_0), X_0) \right|$$

and $\gamma_2$ equivalently for $\hat{\theta}_{m_2}$. Since the neural networks are continuous functions they take a finite maximum on the compact sets, hence $\gamma_1, \gamma_2$ are finite. Then we define the bounded output neural networks $f_{\hat{\theta}_{t-1}^{\star, m_1}, \rho_{\hat{\theta}_{m_2}}} \in \mathcal{N}$ with $\theta_{t-1}^{\star, m_1} := (\hat{\theta}_{t-1}^{\star, m_1}, \gamma_1)$. Clearly, these bounded output neural networks coincide with the neural networks on the compact sets. Therefore, they satisfy the same $\varepsilon$-approximation and since $F_j, f_j$ are bounded by $U := B((X_T^T + 1)^p + \sum_{j=0}^n |\epsilon_j|)$ (Assumption C.1(iv)), it follows that $f_{\hat{\theta}_{t-1}^{\star, m_1}, \rho_{\hat{\theta}_{m_2}}} \in \mathcal{N}$ by $U + \varepsilon$. In particular, we have for $\varepsilon < B$ the global bounds $|f_j - f_{\hat{\theta}_{t-1}^{\star, m_1}, \gamma_2}| \leq 3U$ and $|F_j - f_{\hat{\theta}_{m_1}, \rho_{\hat{\theta}_{m_2}}}| \leq 3U$. Setting $m := \max(m_0, m_1, m_2, \gamma_1, \gamma_2, |\hat{\gamma}_1^{\star, m_1}|, |\hat{\gamma}_2|)$, it follows that $\hat{\theta}_{m}^{\star, m_1}, \hat{\theta}_{m_2}^{\star, m_2}, \hat{\theta}_{m}^{\star} \in \Theta_m$.

Now we can bound the distance between $Y_{t-1}^{\theta_{m_1}}(\hat{\Omega}^{\leq t})$ and $\hat{X}_t$. Whenever $X_T^T < 1/\varepsilon$, the number of observations satisfies $n < 1/\varepsilon$, the minimal difference between any two consecutive observation times $\delta > \varepsilon$ and all noise terms satisfy $|\epsilon_j| < 1/\varepsilon$ we know that the corresponding path $\hat{\Omega}^{\leq \tau(t)} - O_0$ is an element of $A_N$, and therefore the neural network approximations up to $\varepsilon$ hold. Otherwise, one of those conditions is not satisfied and the global upper bound can be used. Hence, if $t \in \{t_1, \ldots, t_n\}$, we have for $F = (F_j)_{1 \leq j \leq d_X}$ and $f = (f_j)_{1 \leq j \leq d_X}$

$$\left| Y_{t-1}^{\theta_{m_1}} - \hat{X}_t \right|_{1} = \left| \rho_{\hat{\theta}_{m_2}} \left( H_{t-}, t, \pi_{m_0}(\hat{\Omega}^{\leq t} - O_0), O_0 \right) - F(t, t, \hat{\Omega}^{\leq t}) \right|_{1}$$

$$\leq \varepsilon d_X \left( \mathbf{1}_{\{X_T^T < 1/\varepsilon\}} + \mathbf{1}_{\{|\epsilon_j| < 1/\varepsilon\}} \right) + d_X 3U \left( \mathbf{1}_{\{X_T^T \geq 1/\varepsilon\}} + \mathbf{1}_{\{|\epsilon_j| \leq 1/\varepsilon\}} \right),$$

and if $t \notin \{t_1, \ldots, t_n\}$,

$$\left| Y_{t-1}^{\theta_{m_1}} - \hat{X}_t \right|_{1} = \left| Y_{t-1}^{\theta_{m_1}} - \hat{X}_t \right|_{1} + \left| \int_{t-1}^{t} f_{\hat{\theta}_{t-1}^{\star, m_1}} \left( H_{s-}, s, \tau(s), \pi_{m_0}(\hat{\Omega}^{\leq \tau(s)} - O_0), O_0 \right) - f(s, \tau(s), \hat{\Omega}^{\leq \tau(s)}) \right|_{1} ds$$

$$\leq \varepsilon (T + 1) d_X \left( \mathbf{1}_{\{X_T^T < 1/\varepsilon\}} + \mathbf{1}_{\{|\epsilon_j| < 1/\varepsilon\}} \right) + (T + 1) d_X 3U \left( \mathbf{1}_{\{X_T^T \geq 1/\varepsilon\}} + \mathbf{1}_{\{|\epsilon_j| \leq 1/\varepsilon\}} \right).$$

Moreover, by equivalence of the 1- and 2-norm, there exists a constant $c > 0$ such that for all $t \in [0, T]$

$$\left| Y_{t-1}^{\theta_{m_1}} - \hat{X}_t \right|_{2} \leq c \varepsilon (T + 1) d_X + c(T + 1) d_X 3U \left( \mathbf{1}_{\{X_T^T \geq 1/\varepsilon\}} + \mathbf{1}_{\{|\epsilon_j| \leq 1/\varepsilon\}} \right) =: c_m.$$

So far, we have fixed an $\varepsilon > 0$ and argued that there exists some $m \in \mathbb{N}$ such that the neural network approximation bounds hold with $\varepsilon$-error. However, what we actually need to show is that this error converges to 0 when increasing the truncation level and network size $m$. Therefore, we define $\varepsilon_m \geq 0$ to be the smallest number such that the above bounds hold with error $\varepsilon_m$ when using an architecture with signature truncation level $m \in \mathbb{N}$ and weights in $\Theta_m$. Since increasing $m$ can only make the approximations better (by setting the new weights to 0, the same approximation error as before is achieved, but potentially there exists a better choice), we have $\varepsilon_m \leq \varepsilon_{m_2}$ for any $m_1 \leq m_2$. In particular $(\varepsilon_m)_{m \geq 0}$ is a decreasing sequence, hence, our derivations before proof that $\lim_{m \to \infty} \varepsilon_m = 0$. In the following we denote by $\hat{\theta}_{m}^{\star} \in \Theta_m$ the best choice for the weights within the set $\Theta_m$ to approximate the functions $F_j, f_j$.

Since $\hat{\theta}_{m}^{\min} \in \arg\min_{\theta \in \Theta_m} \{ \Phi(\theta) \}$ (note that at least one minimum exists in the compact set $\Theta_m$ since $\Phi$ is continuous), we get with Lemma C.4 that

$$\min_{Z \in B} \Psi(Z) \leq \Phi(\hat{\theta}_{m}^{\min}) \leq \Phi(\hat{\theta}_{m}^{\star}) = E \left[ \frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \odot (O_{t_i} - Y_{t_i-}^{\theta_{m}^{\star}}) \right|_{2}^{2} \right]$$

$$= \Psi(\hat{X}) + E \left[ \frac{1}{n} \sum_{i=1}^{n} \left| M_{t_i} \odot (\hat{X}_{t_i-} - Y_{t_i-}^{\theta_{m}^{\star}}) \right|_{2}^{2} \right] \leq \Psi(\hat{X}) + E \left[ c_{m}^{2} \right].$$

(26)
Integrability of $|X^*_2|_2$, $|n|$ and $|\epsilon_i|$ together with Assumption C.1(ii) on $\delta$ imply that
\[
\mathbb{E}[|X^*_2|_2^2 + |n| + |\epsilon_i|] < \infty.
\]
Therefore, we have for a suitable constant $c > 0$ (not depending on $\varepsilon_m$ and $m$),
\[
\mathbb{E}[\epsilon_m^2] \leq c\varepsilon_m^2 + c\mathbb{E}\left[U^2 \left( \mathbb{E}\left[\left|X^*_2 + 1\right|^{2p} + n \sum_{i=0}^{n} |\epsilon_i|^2 \right] \right) \right] \xrightarrow{m \to \infty} 0,
\]
by dominated convergence, since $U$ is $L^2$-integrable. Indeed,
\[
\mathbb{E}[U^2] \leq 8B^2\mathbb{E}\left[(X^*_2 + 1)^{2p} + n \sum_{i=0}^{n} |\epsilon_i|^2 \right] = 8B^2\left( \mathbb{E}\left[(X^*_2 + 1)^{2p} \right] + \mathbb{E}[n^2]\mathbb{E}\left[|\epsilon_0|^2 \right] \right) < \infty,
\]
using Cauchy-Schwarz inequality for the first step, that the $\epsilon_i$ are i.i.d. and independent of $n$ for the equality and the integrability of $X^*$, $\epsilon_0$ and $n^2$ for the upper bound. Using this and $\Psi(X) = \min_{z \in \mathbb{B}} \Psi(Z)$, we get from (26)
\[
\min_{Z \in \mathbb{B}} \Psi(Z) \leq \Phi(\theta_{m}^{\min}) \leq \Phi(\theta_{m}^{*}) \xrightarrow{m \to \infty} \min_{Z \in \mathbb{B}} \Psi(Z).
\]
Finally, we show that $\lim_{m \to \infty} d_k\left(\hat{X}, Y^{\theta_{m}^{*}}\right) = 0$ for all $1 \leq k \leq K$. Applying (23), (24) and (1) in reverse order than it was done in (25) and finally Lemma C.4, yields
\[
d_k\left(\hat{X}, Y^{\theta_{m}^{*}}\right) \leq \frac{c_0 c_1 c_3}{c_2} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \left| M_{t_{\hat{X}}} \circ (\hat{X}_{t_i} - Y_{t_i}^{\theta_{m}^{*}}) \right|^2 \right]^{1/2} = \frac{c_0 c_1 c_3}{c_2} \left( \Phi(\theta_{m}^{*}) - \Psi(\hat{X}) \right)^{1/2} \xrightarrow{m \to \infty} 0,
\]
which completes the first part of the proof.

Next we assume the size $m$ of the neural network and of the signature truncation level is fixed and we study the convergence of the Monte Carlo approximation when the number of training paths $N$ increases. The convergence analysis is based on Lapeyre & Lelong (2021, Chapter 4.3) and follows Herrera et al. (2021, Theorem E.13). We define the separable Banach space $S := \{x = (x_i)_{i \in \mathbb{N}} \in \ell^2(\mathbb{R}^d) \mid \|x\|_{\ell^2} < \infty \}$ for a suitable $d$ (see below) with the norm $\|x\|_{\ell^2} := \sum_{i \in \mathbb{N}} |x_i|/2$, the function
\[
G(x,y,m) := |m \circ (x - y)|_2
\]
and $\xi_j := (\xi_{j,0}, \ldots, \xi_{j,n(j)}, 0, \ldots)$, where $\xi_{j,k} := (t_{k}^{(j)}, O_{t_{k}^{(j)}}^{(j)}, M_{t_{k}^{(j)}}^{(j)}, \pi_{m}(\hat{O}_{t_{k}^{(j)}, (j)}^{(j)})) \in \mathbb{R}^d$ and $t_{k}^{(j)}, M_{t_{k}^{(j)}}^{(j)}$ and $O_{t_{k}^{(j)}}^{(j)}$ (with 0 entries for coordinates which are not observed) are random variables describing the $j$-th realization of the training data (cf. Section 2). Let $n^j(\xi_j) := \max_{k \in \mathbb{N}} (\xi_{j,k} \neq 0)$, $t_{k}(\xi_j) := t_{k}^{(j)}$, $O_{k}(\xi_j) := O_{t_{k}^{(j)}}^{(j)}$, and $M_{k}(\xi_j) := M_{t_{k}^{(j)}}^{(j)}$. By this definition we have $n^j(\xi_j) = n^j(\xi_j) \mathbb{P}$-almost-surely. Moreover, we have that $\xi_j$ are i.i.d. random variables taking values in $S$. Let us write $Y_{t_i}^{\theta}(\xi_j)$ to make the dependence of $Y$ on the input and the weight $\theta$ explicit. Then we define
\[
h(\theta, \xi_j) := \frac{1}{n^j(\xi_j)} \sum_{i=1}^{n^j(\xi_j)} G\left(O_{i}(\xi_j), Y_{t_i}^{\theta}(\xi_j) - (\xi_j), M_{i}(\xi_j)\right)^2.
\]

The following lemma is known from Krach et al. (2022).

**Lemma C.5.** Almost-surely the random function $\theta \in \Theta_m \to Y_{t}^{\theta}$ is uniformly continuous for every $t \in [0, T]$.
Proof of Theorem C.3 – Part 2. First we note that, $Y_t^\theta$ is the (integration over the) output of (bounded output) neural networks and therefore bounded in terms of the input, the weights (which are bounded by $m$), $T$ and some constant depending on the architecture and the activation functions of the neural network. In particular we have that $|Y_t^\theta(\xi_j)| \leq \tilde{B} \leq \tilde{B} \left( (X_T^{(j)} + 1)^p + \sum_{i=0}^{n_t(\xi_j)} |\xi_i^{(j)}| \right)$ for all $t \in [0, T]$ and $\theta \in \Theta_m$ for some constant $\tilde{B}$ (possibly depending on $m$), where $X^{(j)}$, $\xi_i^{(j)}$ corresponds to the input $\xi_j$. Hence,

$$G \left( O_t(\xi_j), Y_{t,i}(\xi_j) - (\xi_j), M_t(\xi_j) \right)^2 = \left| M_t(\xi_j) \circ (O_t(\xi_j) - Y_{t,i}(\xi_j)) \right|^2 \leq \left( (B + \tilde{B}) \left( (X_T^{(j)} + 1)^p + \sum_{i=0}^{n_t(\xi_j)} |\xi_i^{(j)}| \right) \right)^2 = \left( \frac{B + \tilde{B}}{B} U^{(j)} \right)^2,$$

where $U^{(j)}$ is as defined before corresponding to the input $\xi_j$. Hence,

$$\mathbb{E} \left[ \sup_{\theta \in \Theta_m} h(\theta, \xi_j) \right] \leq \mathbb{E} \left[ \frac{1}{n_t(\xi_j)} \sum_{i=1}^{n_t(\xi_j)} \left( \frac{B + \tilde{B}}{B} U^{(j)} \right)^2 \right] < \infty,$$

by Assumption 2.1 and (27). By Lemma C.5, the function $\theta \mapsto h(\theta, \xi_1)$ is continuous, hence, we can apply Krach et al. (2022, Lemma 4.6), yielding that almost-surely for $N \to \infty$ the function

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

converges uniformly on $\Theta_m$ to

$$\theta \mapsto \mathbb{E}[h(\theta, \xi_1)] = \Phi(\theta).$$

Moreover, we deduce from Krach et al. (2022, Lemma 4.6) that $d(\hat{\theta}^{\min}_{m,N}, \Theta^{\min}_m) \to 0$ 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 $|\hat{\theta}^{\min}_{m,N} - \hat{\theta}^{\min}_{m,N}|_2 \to 0$ a.s. for $N \to \infty$. The uniform continuity of the random functions $\theta \mapsto Y_t^\theta$ on $\Theta_m$ implies that for any fixed deterministic and bounded $\xi_0$ (taking values in the same space as the $\xi_j$),

$$|Y_{t,i}^{\hat{\theta}^{\min}_{m,N}}(\xi_0) - Y_{t,i}^{\hat{\theta}^{\min}_{m,N}}(\xi_0)|_2 \to 0 \text{ a.s. for all } t \in [0, T] \text{ as } N \to \infty.$$ 

By continuity of $G$ this yields $|h(\hat{\theta}^{\min}_{m,N}(\omega), \xi_0) - h(\hat{\theta}^{\min}_{m,N}(\omega), \xi_0)| \to 0 \text{ a.s. as } N \to \infty$. Let $\xi_0$ now be a random variable which is independent of and identically distributed as the $\xi_j$ defined on a copy $(\Omega_0, \mathbb{F}_0, \mathbb{F}, \mathbb{P})$ of the filtered probability space $(\Omega, \mathbb{F}, \mathbb{F}, \mathbb{P})$. Then the above statements hold for $\xi_0(\omega)$ for $\mathbb{P}$-a.e. $\hat{\theta}^{\min}_{m,N}$. Hence, we have for $\mathbb{P}$-a.e. fixed $\omega$, that $|h(\hat{\theta}^{\min}_{m,N}(\omega), \xi_0) - h(\hat{\theta}^{\min}_{m,N}(\omega), \xi_0)| \to 0 \mathbb{P}$-a.s. as $N \to \infty$. With (30) we can apply dominated convergence which yields

$$\lim_{N \to \infty} \mathbb{E}_{\xi_0} \left[ h(\hat{\theta}^{\min}_{m,N}(\omega), \xi_0) - h(\hat{\theta}^{\min}_{m,N}(\omega), \xi_0) \right] = 0 \text{ for } \mathbb{P}$-

Since for every integrable random variable $Z$ we have $0 \leq |\mathbb{E}[Z]| \leq \mathbb{E}[|Z|]$ and since $\hat{\theta}^{\min}_{m,N} \in \Theta^{\min}_m$ we can deduce that for $\mathbb{P}$-a.e. fixed $\omega$,

$$\lim_{N \to \infty} \Phi(\hat{\theta}^{\min}_{m,N}(\omega)) = \lim_{N \to \infty} \mathbb{E}_{\xi_0}(h(\hat{\theta}^{\min}_{m,N}(\omega), \xi_0)) = \lim_{N \to \infty} \mathbb{E}_{\xi_0} \left[ h(\hat{\theta}^{\min}_{m,N}(\omega), \xi_0) \right] = \Phi(\hat{\theta}^{\min}_{m,N}).$$

Now by triangle inequality, for $\mathbb{P}$-a.e. fixed $\omega$, we have for $\hat{\Phi}_N$ and $\Phi$ defined through test samples $\tilde{\xi}_j$ on $\Omega_0$, i.e., independent of and identically distributed as the training samples $\xi_j$ yielding $\theta^{\min}_{m,N}(\omega)$,

$$|\hat{\Phi}_N(\hat{\theta}^{\min}_{m,N}(\omega)) - \Phi(\hat{\theta}^{\min}_{m,N}(\omega))| \leq |\hat{\Phi}_N(\theta^{\min}_{m,N}(\omega)) - \Phi(\theta^{\min}_{m,N}(\omega))| + |\Phi(\theta^{\min}_{m,N}(\omega)) - \Phi(\hat{\theta}^{\min}_{m,N}(\omega))|.$$  

(31) and (32) imply that the first term on the right hand side converges to 0 when $N \to \infty$ and (33) implies that the second term on the right hand side converges to 0 a.s. when $N \to \infty$. Moreover, the uniform
convergence in (31) and (32) yields the same result when setting \( \hat{N} = N \). Furthermore, Krach et al. (2022, Lemma 4.6) yields the same result for \( \Phi_N(\theta_{m,N}^\text{min})(\omega) \), i.e., when \( \Phi_N \) and \( \Phi \) are defined through the \( \xi_j \) on the probability space corresponding to the training data. This finishes the proof of the convergence with respect to \( N \).

Finally, we want to show the joint convergence. We define \( N_0 := 0 \) and for every \( m \in \mathbb{N} \)
\[
N_m(\omega) := \min \{ N \in \mathbb{N} \mid N > N_{m-1}(\omega), |\Phi(\theta_{m,N}^\text{min}(\omega)) - \Phi(\theta_{m,N}^\text{min})| \leq \frac{1}{m} \},
\]
which is possible due to (33) for \( \mathbb{P} \)-a.e. \( \omega \). Then (28) implies that for \( \mathbb{P} \)-a.e. \( \omega \)
\[
|\Phi(\theta_{m,N_m}^\text{min}(\omega)) - \Psi(\hat{X})| \leq \frac{1}{m} + |\Phi(\theta_{m,N}^\text{min}) - \Psi(\hat{X})| \xrightarrow{m \to \infty} 0.
\]
Therefore, we can apply the same arguments as in the first part of the proof (cf. (29)) to show that
\[
d_k(\hat{X}, Y_{m,N_m}^\text{min}(\omega)) \leq \frac{c_0 c_1 c_3}{c_2} \left( \Phi(\theta_{m,N_m}^\text{min}(\omega)) - \Psi(\hat{X}) \right)^{1/2} \xrightarrow{m \to \infty} 0,
\]
for every \( 1 \leq k \leq K \) and for \( \mathbb{P} \)-a.e. \( \omega \).

The following corollary follows as in Krach et al. (2022).

**Corollary C.6.** In the setting of Theorem C.3, we also have that \( \mathbb{P} \)-a.s.
\[
\Phi(\theta_{m,N_m}^\text{min}) \xrightarrow{m \to \infty} \Psi(\hat{X}) \quad \text{and} \quad \hat{N}_m(m,N_m) \xrightarrow{m \to \infty} \Psi(\hat{X}),
\]
for a suitable increasing random sequence \((\hat{N}_m(m)) \in \mathbb{N}\) in \( \mathbb{N} \).

## D Experimental Details

Our experiments are based on the implementation used by Krach et al. (2022), which is available at https://github.com/FlorianKrach/PD-NJODE. Therefore, we refer the reader to its Appendix for any details that are not provided here.

### D.1 Differences between the Implementation and the Theoretical Description of the PD-NJ-ODE

Since we use the same implementation of the PD-NJ-ODE, all differences between the implementation and the theoretical description listed in Krach et al. (2022, Appendix D.1.1) also apply here. In particular, we use standard neural networks for \( f_{\tilde{u}_t} \) and \( \rho_{\tilde{u}_t} \) with additional inputs. In contrast to the original objective function, we do not need to add a regularizing constant for the noise-adapted objective functions (6) and (7), since no square-root needs to be computed. We do not use the \( \tilde{u} \)-coordinates of \( \tilde{X} \leq t \) (i.e., we only use the first \( d_X \) coordinates and omit the remaining coordinates from \( d_X + 1 \) to \( 2d_X \)) in the implementation, since we anyways use the observation times and masks as additional inputs for \( \rho_{\tilde{u}_t} \).

### D.2 Details for Noisy Observations

**Dataset.** We sample paths from a standard 1-dimensional Brownian motion on the interval \([0, 1]\), i.e., with \( T = 1 \), and discretisation time grid with step size 0.01. At each time point we observe the process with probability \( p = 0.1 \). Whenever the process is observed, an independent observation noise is sampled from a centered normal distribution with standard deviation \( \sigma = 0.5 \) and added to the observation. The model never sees the observation of the original process, but only the noisy observation. We sample 20'000 paths of which 80% are used as training set and the remaining 20% as test set.

**Architecture.** We use the PD-NJ-ODE with the following architecture. The latent dimension is \( d_H = 100 \), the readout network is a linear map and the other 2 neural networks have the same structure of 1 hidden layer with ReLU activation function and 100 nodes. The signature is used up to truncation level 3, the encoder is recurrent and the decoder uses a residual connection.
We use the Adam optimizer with the standard choices $\beta = (0.9, 0.999)$, weight decay of 0.0005 and learning rate 0.001. Moreover, a dropout rate of 0.1 is used for every layer and training is performed with a mini-batch size of 200 for 200 epochs. The model is trained once with the noise-adapted objective function (6) and once with the original one (3).

### D.3 Details for Dependent Observation Framework

**Dataset.** We use the Euler scheme to sample paths from a 1-dimensional Black-Scholes model (geometric Brownian motion) with drift $\mu = 2$, volatility $\sigma = 0.3$, and starting value $X_0 = 1$. At each time point we observe the process with the probability $P(M_t = 1 | A_{t-}) = E[M_t | A_{t-}]$, where $M_t$ is described in Example 4.13, using $\eta = 3$ and $p = 0.1$. We use the same discretisation grid and dataset sizes as in Section D.2.

**Architecture.** We use the PD-NJ-ODE with the following architecture. The latent dimension is $d_H = 50$ and all 3 neural networks have the same structure of 2 hidden layers with tanh activation function and 50 nodes. The signature is used up to truncation level 3, the encoder is recurrent and the decoder does not use a residual connection.

**Training.** Same as in Section D.2 but only trained with original loss function (3).

### D.4 Details for Physionet with Observation Noise

**Dataset.** Details on the standard Physionet dataset are given in Herrera et al. (2021, Appendix F.5.3). In particular, we use a train-test split of 80% to 20% (i.e., $N = 6400$ training paths, and $N_{test} = 1600$ test paths). On the train set, the entire path amounting to 48 hours is used as input to the model, while on the test set only the first 24 hours are used as input and the second 24 hours are used to compute the evaluate MSE. In particular, the model predicts starting from 24 hours up to 48 hours and at every time point where there is an observation, the squared error is computed between the observation and the prediction. Importantly, the observation is not used as an input to the model afterwards. I.e., we evaluate $Y_{t_i}^{\text{min},N}(\bar{X} \leq 24 \text{ hours})$, while in theory we could get even better results by evaluating $Y_{t_i}^{\text{min},N}(\bar{X} \leq 24 \text{ hours})$. For the datasets with synthetically added noise, the variance of the noise is chosen relative to the variance of the data. Since the $d_X = 41$ different coordinates of the Physionet dataset describe very different health parameters on different scales, we first compute each coordinate’s standard deviation $\sigma_{\text{data},j}$ on the training set. Then we add i.i.d. noise samples $\epsilon_{i,j} \sim N(0, \zeta^2 \sigma_{\text{data},j}^2)$, for $\zeta \geq 0$, to each observed coordinate $X_{t_i,j}$ of input samples for the model, i.e., to each observation of the training set as well as to each observation on the first 24 hours of the test set (but not to any of the observations on the second 24 hours of the test set, which are exclusively used for evaluating the model). This is done for $\zeta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, where the same seed is used for each of them, to make the results more comparable.

**Architecture.** We use the PD-NJ-ODE with the following architecture. The latent dimension is $d_H = 50$ and all 3 neural networks have the same structure of 1 hidden layer with tanh activation function and 50 nodes. The signature is used up to truncation level 2, the encoder is recurrent and the decoder does a residual connection.

**Training.** For each dataset the model is trained with mini-batch size 50 for 100 epochs, once with the noise-adapted objective function (6) and once with the original one (3). Otherwise the same as in Section D.2.

---

23In practice both predictions $Y_{t_i}^{\text{min},N}(\bar{X} \leq 24 \text{ hours})$ and $Y_{t_i}^{\text{min},N}(\bar{X} \leq t_i-)$ can be useful depending on the situation. One has to use $Y_{t_i}^{\text{min},N}(\bar{X} \leq 24 \text{ hours})$ if one is at the situation that one has observed the patient for exactly 24 hours and wants to forecast the next 24 hours. On the other hand, at any time $s$ one should use all the currently available information to make the optimal prediction $Y_{t_i}^{\text{min},N}(\bar{X} \leq s) = Y_{t_i}^{\text{min},N}(\bar{X} \leq t_i)$ (s). While our model is flexible enough to make predictions $Y_{t_i}^{\text{min},N}(\bar{X} \leq 24 \text{ hours})$ for every combination $t > s$, we only evaluate $Y_{t_i}^{\text{min},N}(\bar{X} \leq 24 \text{ hours})$, since some competing methods compared to in Herrera et al. (2021) are limited to this restricted setting of one fixed value of $s = 24$ hours. However, especially if one is interested in now-casting, the performance of $Y_{t_i}^{\text{min},N}(\bar{X} \leq t_i-)$ would be the more relevant metric (in which we expect our model to outperform its competitors even more clearly).
We always report the minimal evaluation MSE among all trained epochs, as was done in prior works. Due to the shorter training compared to the 175 epochs in Krach et al. (2022), the evaluation MSE of the original PD-NJ-ODE on the dataset without synthetic noise is slightly larger (2 · 10^{-3} compared to 1.93 · 10^{-3}). We could imagine that for our new loss, one should actually use even more epochs rather than less epochs, since the additional term that was only present in the old loss pushed the jump-network $\rho_{t_k}$ faster into the right direction directly from the start of the training, while the new loss trains the jump-network only indirectly and thus probably slower.

### E Pseudocode

In addition to our code repository (https://github.com/FlorianKrach/PD-NJODE), we provide pseudocode for the key parts of our method. We remark that multiple python packages are available for computing the signature of a process numerically, as for example ESig, iisignature and signatory. Therefore, we do not provide pseudocode for its computation. For our implementation, we use the great iisignature package implemented by Reizenstein & Graham (2018). We note that our definition of $\tilde{X}^{\leq t}$, with its time-consistency, allows for efficient online updates of the signature (i.e., we do not have to recompute the signature for the whole path at every new observation time $t_k$).

In Algorithm 1 we present the forward pass of the PD-NJ-ODE, which can be used for training as well as for evaluating the model. Algorithms 2 and 3 present the standard loss function and the loss function for noisy observations, respectively. The training for the model follows the standard neural network training approach via stochastic gradient descent (SGD) and is presented in Algorithm 4. We assume that a library is used for the neural network implementation which allows for automatic differentiation, as for example PyTorch (Paszke et al., 2019), which we used in our implementation. For simplicity, we present the simplest type of SGD, with a fixed step size and no momentum or weight decay. In our implementation we actually use the more sophisticated method Adam (Kingma & Ba, 2014).

### Notation

$F_j$ The Doob-Dynkin Lemma (Taraldsen, 2018, Lemma 2) implies the existence of measurable functions $F_j : [0, T] \times [0, T] \times BV^c([0, T]) \to \mathbb{R}$ such that $\tilde{X}_{t,j} = F_j (t, \tau(t), \tilde{X}^{\leq \tau(t)})$, since $\tilde{X}_{t,j} := \mathbb{E}[X_{t,j} | A_t] = \mathbb{E}[X_{t,j} | \tilde{X}^{\leq \tau(t)}]$ (see Section 2). The goal of our model $Y^\theta$ is to learn this function $F = (F_j)_{1 \leq j \leq dx}$, i.e., $\lim_{m \to \infty} Y^\theta_{m,n} (X^{\leq \tau(t)}) = F (t, \tau(t), \tilde{X}^{\leq \tau(t)}) = \tilde{X}_t$, while we are usually not able to write down $F$ explicitly. In settings with noise (e.g., in Section 3 and in the appendix) we replace $\tilde{X}^{\leq \tau(t)}$ by $\tilde{O}^{\leq \tau(t)}$. 4, 8, 9, 16, 34, 36, 38, 39

$K$ The “maximal” value of $n$, i.e., the essential supremum $K := \sup \{ k \in \mathbb{N} | \tilde{P}(n \geq k) > 0 \} \in \mathbb{N} \cup \{ \infty \}$ of $n$ (see Section 2). In Sections 2 and 3, $K$ is defined with respect to $\tilde{P}$, but in Section 4 (and all sections thereafter including the appendix), $K$ is defined with respect to $\mathbb{P}$, 3–8, 13–15, 18, 20–24, 36–38, 40, 42, 44, 46, 49

$M$ The observation mask $M = (M_k)_{0 \leq k \leq K}$, which is a sequence of random variables on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ taking values in $\{0, 1\}^d$, such that $M_k$ is $\mathcal{F}_{t_k}$-measurable. The $j$-th coordinate of the $k$-th element of the sequence $M$, i.e., $M_{k,j}$, signals whether $X_{t_k,j}$, denoting the $j$-th coordinate of the stochastic process at observation time $t_k$, is observed. By abuse of notation we also write $M_{t_k} := M_k$ (see Section 2). In Sections 2 and 3, $M$ is defined on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, but in Section 4 (and all sections thereafter including the appendix), $M$ is defined on $(\Omega, \mathcal{F}, \mathbb{P})$. 3–6, 8–10, 12–20, 33, 36–41, 33, 45–47, 49

$O$ The noisy observations $O_{t_k} := X_{t_k} + \epsilon_k$ for $0 \leq k \leq n$, where $\epsilon_k$ is the noise with known mean (see Section 3.1). For the majority of the paper we assume that the noise has zero mean (whenever we use (the MC-approximation of) the loss (6) and (7), except Section 3.3 where we assume any known mean $\beta_j (O^{\leq \tau(t)}) := \mathbb{E}[\epsilon_j | A_{t_\tau}]$ and thus use (the MC-approximation of) the loss (10). We define $O_{t_-} := X_{t_-} + \epsilon_i$ and therefore also have that $O_{t_k} = O_{t_-}$ almost surely. In Section 3, $\epsilon_k$ are i.i.d.
**Algorithm 1** Forward pass of the PD-NJ-ODE. A small step size $\Delta t$ is fixed and we denote $t_{n+1} := T$. $\text{ODESolve}(f, x, (a, b))$ numerically solves the 1st-order ODE defined by $f$, taking inputs $x$, for $t \in (a, b)$.

**Input:** Data points with timestamps and masks $\{(X_i, t_i, M_i)\}_{i=0...n}$

**Output:** Prediction process $Y$

set $H_{0i} = 0$

for $i = 0$ to $n$ do
- construct $\hat{X}^i \leq t_i$ from data
- $S_i = \pi_m(\hat{X}^i - X_0)$
- $H_{t_i} = \rho_{\theta_3}(H_{t_i-}, t_i, S_i, X_0)$
- $Y_{t_i} = \bar{g}_{\theta_3}(H_{t_i})$
- $s \leftarrow t_i$
  while $s + \Delta t < t_{i+1}$ do
    $H_{s+\Delta t} = \text{ODESolve}(f_{\theta_1}, (H, s, t_i, S_i, X_0), (s, s + \Delta t))$
    $Y_{s+\Delta t} = \bar{g}_{\theta_3}(H_{s+\Delta t})$
    $s \leftarrow s + \Delta t$
  end while
- $H_{t_{i+1}-} = \text{ODESolve}(f_{\theta_1}, (H_{t_i-}, s, t_i, S_i, X_0), (s, t_{i+1}))$
- $Y_{t_{i+1}-} = \bar{g}_{\theta_3}(H_{t_{i+1}-})$
end for

**Algorithm 2** Computation of the standard loss for PD-NJ-ODE. A small $\epsilon > 0$ is used for numerical stability.

**Input:** $N$ samples of data points with timestamps and masks $\{(X_i^{(j)}, t_i^{(j)}, M_i^{(j)})\}_{i=0...n^{(j)}}$ and the respective prediction process $Y_i^{(j)}$ of PD-NJ-ODE for $1 \leq j \leq N$

**Output:** Loss $L$

$L = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \left| M_i^{(j)} \odot (X_i^{(j)} - Y_i^{(j)}) + \epsilon \right|_2 + \left| M_i^{(j)} \odot (X_i^{(j)} - Y_{i-}^{(j)}) + \epsilon \right|_2^2$

**Algorithm 3** Computation of the noisy observation loss for PD-NJ-ODE.

**Input:** $N$ samples of noisy data points with timestamps, masks and conditional means of the noise $\{(O_i^{(j)}, t_i^{(j)}, M_i^{(j)}, \beta_i^{(j)})\}_{i=0...n^{(j)}}$ and the respective prediction process $Y_i^{(j)}$ of PD-NJ-ODE for $1 \leq j \leq N$

**Output:** Loss $L$

$L = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} \left| M_i^{(j)} \odot \left( (O_i^{(j)} - \beta_i^{(j)}) - Y_i^{(j)} \right) \right|_2$

**Algorithm 4** Training of PD-NJ-ODE.

**Input:** step size $\alpha$; number of epochs $E$; $N$ samples of noisy data points with timestamps, masks and conditional means of the noise $\{(O_i^{(j)}, t_i^{(j)}, M_i^{(j)}, \beta_i^{(j)})\}_{i=0...n^{(j)}}$ for $1 \leq j \leq N$

**Output:** trained network parameters $\theta = (\theta_1, \theta_2, \theta_3)$ randomly initialize $\theta$

for $e = 0$ to $E$ do
- split the dataset into random batches $B$
  for $b$ in $B$ do
    for each sample $j$ in $b$ do
      compute prediction $Y_i^{(j)}$ with Algorithm 1
    end for
  compute loss $L$ over all samples in the batch $b$ with Algorithm 2 or 3
  update NN parameters $\theta \leftarrow \theta - \alpha \nabla_\theta L$
end for
random variables on \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})\), but in Section 4 (and all sections thereafter including the appendix), \(\epsilon_k\) are i.i.d. random variables on \((\Omega, \mathcal{F}, \mathbb{P})\). 7–12, 25, 33, 34, 36–41, 45–48

\[ T \] The largest time \(T \in \mathbb{R}_{>0}\) we consider (see Section 2). 3–5, 8–10, 17–19, 21–24, 36, 38–42, 44–49

\[ X^* \] The running maximum process of \(X\), i.e., \(X^*_t(\omega) := \sup_{s \in [0,t]} |X_s(\omega)| \) \(\forall t \in [0,T] \forall \omega \in \Omega\) with \(X^* : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}_{\geq 0}[0,T]\) (see Section 2). 3–5, 8–10, 17, 36, 39–41

\[ X \] The stochastic process \(X := (X_t)_{t \in [0,T]}\) we want to study (which is an adapted càdlàg stochastic process \(X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}^{d_{\mathcal{F}}}[0,T]\)). In our notation \(X^{(j)}_{k}\) refers to the \(k\)-th coordinate of the \(j\)-th path at time \(t\) for \(1 \leq k \leq d_X, 1 \leq j \leq N\) and \(t \in [0,T]\) (see Section 2). E.g., for a medical data-set, the index \(j\) could correspond to a patient and the first coordinate could correspond to the body-temperature and the second coordinate could correspond to the blood-pressure, then \(X^{(7)}_t(\omega) = 36.9\) would mean that the 7-th patient had a body temperature of 36.9°C at time (or age) \(t\). 2–27, 33–37, 43–49

\[ Y \] The output \(Y^\theta(X^{\leq \tau(\cdot)}(\cdot))\) of our PD-NJ-ODE (2) which should approximate \(\hat{X}\) for properly trained parameters \(\theta\) (see Definition 2.6). We use the short notation \(Y^\theta \hat{\beta} := Y^\theta(X^{\leq \tau(\cdot)}(\cdot))\). 6–10, 14, 17, 21–24, 33–35, 37, 39–45

\(\hat{\theta}\) The conditional expectation of \(O\), which is its \(L^2\)-optimal prediction (Krach et al., 2022, Proposition 2.5) given the currently available information, is defined as \(\hat{\theta}_{t,-} := \mathbb{E}_{\mathbb{P}}[O_t \mid A_{t,-}] = \hat{X}_{t,-}\) (see the proof of Lemma 3.4). Only directly at observation times, \(\hat{\theta}_{t_k}\) and \(\hat{X}_{t_k}\) deviate from each other: i.e., in general \(\hat{X}_{t_k} = \mathbb{E}[X_{t_k} \mid A_t] \neq \hat{\theta}_{t_k} = \mathbb{E}[O_t \mid A_t] = \hat{X}_{t_k}\) if \(M_{t_k} = 1\) (see Section 3). In Section 3, the expectation is taken with respect to \(\mathbb{P} \times \mathbb{P}\), but in Section 4 (and all sections thereafter including the appendix), the expectation is taken with respect to \(\mathbb{P}\). 3, 4, 6–11, 14–17, 20–24, 34, 36–40, 42, 44, 46

\(\hat{X} := (\hat{X}_t)_{0 \leq t \leq T}\) The conditional expectation process of \(X\), which is its \(L^2\)-optimal prediction (Krach et al., 2022, Proposition 2.5) given the currently available information, is defined as \(\hat{X} := (\hat{X}_t)_{0 \leq t \leq T}\), with \(\hat{X}_t := \mathbb{E}_{\mathbb{P}}[X_t \mid A_t]\) (see Section 2). In Sections 2 and 3, the expectation is taken with respect to \(\mathbb{P} \times \mathbb{P}\), but in Section 4 (and all sections thereafter including the appendix), the expectation is taken with respect to \(\mathbb{P}\). 3, 4, 6–11, 14–17, 20–24, 34, 36–40, 42, 44, 46

\(\hat{\Phi}_N\) The objective function \(\hat{\Phi}_N : \Theta \to \mathbb{R}, \theta \mapsto \hat{\Phi}_N(\theta) := \Psi(Y^\theta(X))\) on the parameter-space \(\Theta\) for a finite number of \(N \in \mathbb{N}\) training paths (i.e., the Monte Carlo approximation of \(\Phi\)). In Sections 2, 4 and 6.2, \(\hat{\Phi}_N\) is defined in Equation (5) (this old objective function only leads to the correct results for noiseless observations, but in Section 3 and in the appendix, \(\hat{\Phi}_N\) is defined analogously via (6) and (7) (this new objective function is different from the old one and even leads to the right result for noisy observations). In Sections 6.1 and 6.3 we compare both losses against each other. Note that in Section 3.3 we use another modification of the loss (10) for the case of noise with known non-zero mean. Note that these 3 variants of \(\hat{\Phi}_N\) are the loss functions on which we actually train our parameters in practice with finitely many training data. 6–8, 34, 36, 37, 41, 42, 47, 48

\((\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})\) The filtered probability space \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}}) := (\mathcal{F}_t)_{0 \leq t \leq T, \hat{\mathbb{P}}})\) on which \(X\) is defined (see Section 2). In Section 4 (and all sections thereafter including the appendix) all stochastic objects \(X, n, K, t_i, \tau, M, A_t\) are defined on this probability space. However, in Sections 2 and 3 only \(X\) is defined on this space, while the observation framework \(n, K, t_i, \tau, M\) is defined on a different filtered probability space \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})\). 3–10, 13–16, 19, 21–23, 36, 37, 40–44, 46–49

\((\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}})\) The filtered probability space \((\hat{\Omega}, \hat{\mathcal{F}}, \hat{\mathbb{P}}) := (\mathcal{F}_t)_{0 \leq t \leq T, \hat{\mathbb{P}}})\) on which \(n, K, t_i, \tau, M\) are defined describing the observation framework, but \(X\) is not defined on this space (see Section 2). In Section 4 (and all sections thereafter including the appendix) this space is not needed since all stochastic objects \(X, n, K, t_i, \tau, M, A_t\) are defined on \((\Omega, \mathcal{F}, \mathbb{P})\). However, in Sections 2 and 3 this space is used for \(n, K, t_i, \tau, M\) related to the events of observations. 3–10, 13, 44, 46–49

\(A := (A_t)_{t \in [0,T]}\) The filtration of the currently available information \(A := (A_t)_{t \in [0,T]}\) defined by \(A_t := \sigma(X_{t_i,j}, t_i, M_{t_i} \mid t_i \leq t, j \in \{1 \leq l \leq d_X \mid M_{t_i,l} = 1\})\),
in all sections, where we have noiseless observations \(X_t\), and analogously defined as

\[ \mathcal{A}_t := \sigma \left( O_{t,j}, t_i, M_{t_i} | t_i \leq t, j \in \{1 \leq l \leq d_X | M_{t_i,t} = 1 \} \right), \]

in all other sections, where we have noisy observations \(O_t\). In both cases this corresponds to the information obtained from seeing the observations until time \(t\) where \(\sigma(\cdot)\) denotes the generated \(\sigma\)-algebra (see Sections 2 and 3.1) We use the notion of stopped \(\sigma\)-algebras

\[ \mathcal{A}_{t_k} := \sigma \left( X_{t_i,j}, t_i, M_{t_i} | i \leq k, j \in \{1 \leq l \leq d_X | M_{t_i,t} = 1 \} \right), \]

and pre-stopped \(\sigma\)-algebras

\[ \mathcal{A}_{t_k}^- := \sigma \left( X_{t_i,j}, t_i, M_{t_i}, t_k | i < k, j \in \{1 \leq l \leq d_X | M_{t_i,t} = 1 \} \right) \]

from Karandikar & Rao (2018, Definitions 2.37 and 8.1). In the sections with noise, we replace \(X\) by \(O\). 3–6, 8–23, 33, 36, 37, 43, 44, 46, 47, 49

\(\mathbb{D}\) The set of all càdlàg \(\mathbb{R}^d\)-valued \(\mathcal{A}\)-adapted processes (see Section 2). In Sections 2 and 3, the stochastic processes in \(\mathbb{D}\) live on \((\Omega \otimes \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{P} \otimes \tilde{\mathbb{P}}, \mathbb{P} \times \tilde{\mathbb{P}})\), but in Section 4 (and all sections thereafter including the appendix), they live on \((\Omega, \mathcal{F}, \mathbb{P})\). 6–10, 16, 17, 37–40, 47

\(\mathcal{J}\) The random set of the jump times \(\mathcal{J} : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow \mathcal{P}([0, T])\) of \(X\) (see Section 2). 3, 4, 36, 37

\(\mathcal{N}\) The set \(\mathcal{N}\) of bounded output neural networks consists of all feed-forward neural networks that have bounded outputs. In particular we assume that the final activation function applied to the output of the neural network is the bounded output activation function \(\Gamma_{\gamma} = (\cdot) \min \left(1, \frac{\gamma}{1+|\cdot|} \right)\) (see Section 2).

Throughout the paper we assume that the functions \(f_{\theta_1}, \rho_{\theta_2} \in \mathcal{N}\) in the PD-NJ-ODE (2) are bounded output feedforward neural networks. Note that this assumption is not really important in practice but facilitates our theoretical proof. 5, 6, 39, 48, 49

\(\mathcal{N}(\mu, \sigma^2)\) The normal distribution \(\mathcal{N}(\mu, \sigma^2)\) (also known as Gaussian distribution) with mean \(\mu\) and standard deviation \(\sigma\). 12, 19, 20, 26, 43

\(\mathcal{P}\) The powerset \(\mathcal{P}(S)\) of a set \(S\) is the set of all subsets of \(S\). 47

\(\Phi\) The objective function \(\Phi : \Theta \rightarrow \mathbb{R}, \theta \mapsto \Phi(\theta) := \Psi(Y^x(\theta))\) on the parameter-space \(\Theta\) in expectation (i.e., “for an infinite amount of training paths”). In Sections 2, 4 and 6.2, the old objective function given in (3) and (4) that only works for noiseless observations is used, but in Section 3 and in the appendix, the new objective function given in (6) and (7) that also works for noisy observations is used. In Sections 6.1 and 6.3 we compare both losses against each other. Note that in Section 3.3 we use another modification of the loss (10) for the case of noise with known non-zero mean. \(\Phi(\theta)\) is the expected value of \(\hat{\Phi}_N(\theta)\), where \(\hat{\Phi}_N\) is the loss we actually train on based on \(N\) training paths. 6–8, 10, 16, 17, 24, 39–42, 46, 48

\(\Psi\) The objective function \(\Psi : \mathbb{D} \rightarrow \mathbb{R}\) (cf. equivalent objective function from Remark 4.7 & Appendix A.1.4 of Krach et al. (2022)) on the path-space \(\mathbb{D}\) in expectation (i.e., “for an infinite amount of training paths”). In Sections 2, 4 and 6.2, the old objective function (3) that only works for noiseless observations is used, but in Section 3 and in the appendix, the new objective function (6) that also works for noisy observations is used. In Sections 6.1 and 6.3 we compare both losses against each other. Note that in Section 3.3 we use another modification of the loss (10) for the case of noise with known non-zero mean. 6–10, 16, 17, 20, 24, 34, 36–40, 42, 46, 47

\(\Theta_m\) The compact subset \(\Theta_m \subset \Theta\) consists of all \(\theta = (\theta_1, \theta_2, \theta_3) = (\tilde{\theta}_1, \gamma_1), (\tilde{\theta}_2, \gamma_2), (\tilde{\theta}_3) \in \Theta\) that correspond to (bounded output) neural networks with widths and depths that are at most \(m\) and such that the truncated signature of level \(m\) or smaller is used and such that the norms of the weights \(\tilde{\theta}_i\) and the bounds \(\gamma_i\) are bounded by \(m\). I.e., \(\Theta_m := \{\theta = (\tilde{\theta}_1, \gamma_1), (\tilde{\theta}_2, \gamma_2), (\tilde{\theta}_3) \in \Theta | \|\tilde{\theta}_i\|_2 \leq m, \gamma_i \leq m \} \subset \Theta_m \subset \Theta\), where \(\Theta_m \subset \Theta\) is defined as the set of possible parameters for the \(3\) (bounded output) neural networks, such that their widths and depths are at most \(m\) and such that the truncated signature
of level $m$ or smaller is used (see Section 2). We use the notation $\Theta_m := \{\theta \mid (\theta_1, \theta_2, \theta_3) \in \Theta_m\}$ and $\Theta^i_m := \{\theta_i \mid ((\theta_1, \gamma_1), (\theta_2, \gamma_2), (\theta_3) \in \Theta_m\}$ for the projections of the sets on the weights $\theta_i$ and $\bar{\theta}_i$ respectively. 6, 7, 10, 35, 37–41, 48

$\Theta$ The set of all possible trainable parameters $\theta = (\theta_1, \theta_2, \bar{\theta}_3) \in \Theta$ for our our PD-NJ-ODE (2) (see Definition 2.6). 6–8, 24, 34, 35, 46–49

$\gamma$ The parameters $\gamma$ are contained in $\theta = ((\bar{\theta}_1, \gamma_1), (\bar{\theta}_2, \gamma_2), (\bar{\theta}_3)$ and bound the outputs of bounded output neural networks. I.e., for every bounded output neural network $f_{\theta_1}, \rho_{\theta_2} \in \mathcal{N}$ it holds that $|f_{\theta_1}(x)|_2 \leq \gamma_1$ and $|\rho_{\theta_2}(x)|_2 \leq \gamma_2$, because of the bounded output activation function $\Gamma_{\gamma_1} : \mathbb{R}^d \to \mathbb{R}^d$, $x \mapsto \Gamma_{\gamma_1}(x) = x \cdot \min\left(1, \frac{\gamma_1}{|x|_2}\right)$ (see Section 2). I.e., $f_{\bar{\theta}_1}(x) = f_{(\bar{\theta}_1, \gamma_1)}(x) = \Gamma_{\gamma_1}(f_{\bar{\theta}_1}(x))$ and $\rho_{\bar{\theta}_2}(x) = \rho_{(\bar{\theta}_2, \gamma_2)}(x) = \Gamma_{\gamma_2}(\rho_{\bar{\theta}_2}(x))$. 5–7, 34, 39, 47–49

$\tau$ The last observation time $\tau(t)$ before a certain time $t$, i.e., $\tau : [0, T] \times \bar{\Omega} \to [0, T]$, $(t, \bar{\omega}) \mapsto \tau(t, \bar{\omega}) := \max\{t_i(\bar{\omega}) \mid 0 \leq i \leq n(\bar{\omega}), t_i(\bar{\omega}) \leq t\}$. In Sections 2 and 3, $\tau$ is defined on $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$, but in Section 4 (and all sections thereafter including the appendix), $\tau$ is defined on $(\Omega, \mathcal{F}, \mathbb{P})$, i.e., $\tau : [0, T] \times \Omega \to [0, T]$. 3, 4, 6–8, 10, 12, 13, 20, 23, 34, 36–39, 43, 44, 46, 48

$\theta_m^{\min} \in \Theta_m^{\min}$ The set of all minimizers $\theta_m^{\min} := \arg\min_{\theta \in \Theta_m} \{\Phi(\theta)\}$ of the objective function $\Phi$ (corresponding to infinitely many training data) under the constraints of $\Theta_m$ for any given $m \in \mathbb{N}$ (see Theorem 2.7). 7, 9, 14, 16, 17, 21–23, 37, 39–43

$\theta_{m,N}^{\min} \in \Theta_{m,N}^{\min}$ The set of all minimizers $\theta_{m,N}^{\min} := \arg\min_{\theta \in \Theta_m} \{\hat{\Phi}_N(\theta)\}$ of the objective function $\hat{\Phi}_N$ (corresponding to N training paths) under the constraints of $\Theta_m$ for any given $m, N \in \mathbb{N}$ (see Theorem 2.7). 7, 21–23, 33, 35, 41, 42, 44

$\theta$ The trainable parameters $\theta = (\theta_1, \theta_2, \bar{\theta}_3) \in \Theta$ contain all trainable parameters. I.e., $\theta$ contains $\theta_i = (\bar{\theta}_i, \gamma_i)$ for $i \in \{1, 2\}$ which parameterise the bounded output feedforward neural networks $f_{\theta_1}, \rho_{\theta_2} \in \mathcal{N}$ and $\theta_3$ parameterize the feedforward neural network $\rho_{\bar{\theta}_3} \in \bar{\mathcal{N}}$. Thus, $\theta = (\theta_1, \theta_2, \bar{\theta}_3) \in \Theta$ parameterize all 3 parameterized functions $f_{\theta_1}, \rho_{\theta_2}, \rho_{\bar{\theta}_3}$ in our PD-NJ-ODE (2) (see Definition 2.6). 6–8, 10, 16, 17, 24, 33–37, 39–42, 44–49

$\pi_m$ The truncated signature $\pi_m(X)$ of order $m \in \mathbb{N}$ of a continuous path with finite variation $X$ is defined in Definition 2.5. In simple words, $\pi_m(X)$ is a finite dimensional feature-vector representing a continuous path $X$. For every finite truncation level $m$, $\pi_m(X)$ does not capture all the information about the infinite dimensional object $X$, but in our proof we use that there always exists a $m \in \mathbb{N}$ such that $\pi_m(X)$ describes $X$ sufficiently well. 5, 6, 34, 38–40, 45

$\bar{O}^{\leq \tau}$ The interpolated observation process $\bar{O}^{\leq \tau}$ continuously interpolates the noisy observations $O_t$ that where observed before time $t$ (see Section 3.1). To be precise the first $d_x$ coordinates of $\bar{O}^{\leq \tau}$ interpolate the noisy observations $O_t$, while the next $d_x$ coordinates of $\bar{O}^{\leq \tau}$ captures explicit information on when which coordinate was observed and the last coordinate is just the time. At time $s \in [0, T]$, the interpolated observation process $\bar{O}_s^{\leq \tau}(\omega) \in \mathbb{R}^{2d_x+1}$ is defined analogously to $\bar{X}_s^{\leq \tau}(\omega) \in \mathbb{R}^{2d_x+1}$ from Section 2, by replacing $X$ by $O$. In Section 3, $\bar{O}^{\leq \tau}$ is an adapted stochastic process on $(\bar{\Omega} \times \bar{\mathcal{F}} \otimes \mathcal{F} \otimes \bar{\mathcal{F}} \otimes \bar{\mathcal{F}} \otimes \bar{\mathcal{F}} \otimes \bar{\mathcal{F}})$, but in Section 4 (and all sections thereafter including the appendix), $\bar{O}^{\leq \tau}$ is an adapted stochastic process on $(\Omega \times \mathcal{F} \otimes \mathcal{F} \otimes \bar{\mathcal{F}} \otimes \bar{\mathcal{F}} \otimes \bar{\mathcal{F}} \otimes \bar{\mathcal{F}} \otimes \bar{\mathcal{F}} \otimes \bar{\mathcal{F}})$, 8, 10, 12, 36–40, 44

$\bar{X}^{\leq \tau}$ The interpolated observation process $\bar{X}^{\leq \tau}$ continuously interpolates the observations of $X$ that where observed before time $t$. To be precise the first $d_x$ coordinates of $\bar{X}^{\leq \tau}$ interpolate the observations of $X$, while the next $d_x$ coordinates of $\bar{X}^{\leq \tau}$ captures explicit information on when which coordinate was observed and the last coordinate is just the time. At time $s \in [0, T]$, the $j$-th coordinate $\bar{X}^{\leq \tau}_j(\omega)$ of $\bar{X}^{\leq \tau}(\omega) \in \mathbb{R}^{2d_x+1}$ is defined in Section 2. In Sections 2 and 3, $\bar{X}^{\leq \tau}$ is an adapted stochastic process

---

24 The definition is less ambiguous, if we write more precisely: $\Theta^i_m := \{\theta_i \mid \exists (\theta'_1, \theta'_2, \theta'_3) \in \Theta_m : \theta_i = \theta'_i\}$.

25 Interpolation also includes extrapolation within this paper. $\bar{X}^{\leq \tau}$ interpolates the observations before time $t$ without leaking any information from observations after time $t$. Furthermore, its time-consistency allows for efficient online updates of its signature instead of recomputing its signature for the whole path at every new observation time $t_k$. 48
The jump process $u$ on $(\Omega \times \bar{\Omega}, \mathcal{F} \otimes \bar{\mathcal{F}}, \bar{\mathcal{P}} \otimes \bar{\mathcal{P}} \times \bar{\mathcal{P}})$, but in Section 4 (and all sections thereafter including the appendix), $\tilde{X}^{\leq t}$ is an adapted stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$. 4, 6, 8, 21–24, 33, 34, 37, 42–46

$\tilde{N}$ The set $\tilde{N}$ of feedforward neural networks consists of all (classical) feed-forward neural networks. We assume that $\tilde{N}$ is a set of standard feedforward neural networks with $id \in \tilde{N}$ that satisfies the standard universal approximation theorem with respect to the supremum-norm on compact sets, see for example Hornik (1991, Theorem 2) Throughout the paper we assume that the functions $\tilde{g}_{\theta_1} \in \tilde{N}$ in the PD-NJ-ODE (2) is a feedforward neural networks. 5, 6, 38, 48, 49

$\theta$ The trainable parameters $\theta = (\theta_1, \theta_2, \theta_3) \in \Theta$ but not the bounds $\gamma_1$ and $\gamma_2$. I.e., the classical feedforward neural network $\tilde{g}_{\tilde{\theta}_3} \in \tilde{N}$ is fully parametrized by $\tilde{\theta}_3$, while the bounded output feedforward neural networks $\tilde{f}_{\theta_1}, \tilde{\rho}_{\theta_2} \in \tilde{N}$ are parametrized by $\theta_i = (\tilde{\theta}_i, \gamma_i)$ for $i \in \{1, 2\}$ which also includes $\gamma_1$ and $\gamma_2$ additionally to $\tilde{\theta}_1$ and $\tilde{\theta}_2$ (see Definition 2.6). 5–7, 34, 35, 37–39, 45, 47–49

$\tilde{u}$ The jump process $\tilde{u}_{i,j} := \sum_{k=0}^{K} M_{k,j} \mathbb{I}_{I_k \leq t}$ counts the coordinate-wise observations (see Section 2). 4, 42

$d_X$ The dimension $d_X \in \mathbb{N}$ of $X$ (i.e., $X_t(\omega) \in \mathbb{R}^{d_X}$) (see Section 2). 3, 4, 6, 8, 10, 11, 13–16, 18, 20, 34, 36–39, 42–44, 46–48

$d_k$ A pseudo metric between two càdlàg $\mathbb{A}$-adapted processes defined in Definition 2.4 in Section 2 to measure the distance between processes. 5, 7, 9, 10, 14, 16, 17, 20, 21, 24, 37, 38, 40, 42

$n$ The random number of observations $n : (\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathcal{P}}) \to \mathbb{N}_{\geq 0}$ up to time $T$ (see Section 2). Every observation time $t_i \in [0, T]$ counts as 1 observation for this count (also for incomplete observations). In Sections 2 and 3, $n$ is defined on $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathcal{P}})$, but in Section 4 (and all sections thereafter including the appendix), $n$ is defined on $(\Omega, \mathcal{F}, \mathbb{P})$. In our medical example, $n^{(j)}$ denotes the number of observation times for the $j$-th patient. 3–6, 8–10, 12–23, 36–40, 44–46, 48, 49

$t_i$ The random observation times $t_i : (\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathcal{P}}) \to [0, T] \cup \{\infty\}$ for $0 \leq i \leq K$ are sorted stopping times, with $t_i(\omega) := \infty$ if $n(\omega) < i$ (see Section 2). In our practical implementation we replace “$\infty$” by “$T$” (see Algorithm 1), since we are not interested in times after $T$ anyway. In Sections 2 and 3, $t_i$ are defined on $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathcal{P}})$, but in Section 4 (and all sections thereafter including the appendix), $t_i$ are defined on $(\Omega, \mathcal{F}, \mathbb{P})$. 3–26, 33, 34, 36–41, 43–49

$u$ The jump process $u_{i,k} := \sum_{k=1}^{K} \mathbb{I}_{I_k \leq t}$ counts the observations without considering which coordinates where observed (see Section 2). 6