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

Anonymous authors
Paper under double-blind review

Abstract

The Path-Dependent Neural Jump ODE (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.

1 Introduction

While the online prediction[1] 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 (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. For more related work we refer the interested reader to the respective sections in Herrera et al. (2021) and Krach et al. (2022).

1.1 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 B 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. Finally, in Section 5 we show empirically that the PD-NJ-ODE performs well in these generalised settings.

[1] 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.

2 Recall: the PD-NJ-ODE

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 $(\Omega, \mathcal{F}, \mathbb{F} := \{\mathcal{F}_t\}_{0 \le t \le T}, \mathbb{P})$ with an adapted càdlàg stochastic process $X := (X_t)_{t \in [0,T]}$ taking values in $\mathbb{R}^{d_X}$. We denote its running maximum process by $X^\star$ and the random set of its jump times by $\mathcal{J}$. The random observation framework is defined independently of X on another filtered probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{F}} := \{\tilde{\mathcal{F}}_t\}_{0 \le t \le T}, \tilde{\mathbb{P}})$ by

- $n : \tilde{\Omega} \to \mathbb{N}_{\ge 0}$, an $\tilde{\mathcal{F}}$-measurable random variable, the random number of observations,
- $K := \sup\left\{k \in \mathbb{N} \,|\, \tilde{\mathbb{P}}(n \ge k) > 0\right\} \in \mathbb{N} \cup \{\infty\}$, the maximal value of n,
- $t_i : \tilde{\Omega} \to [0,T] \cup \{\infty\}$ for $0 \le i \le K$, sorted stopping times, which are the random observation times, with $t_i(\tilde{\omega}) := \infty$ if $n(\tilde{\omega}) < i$,
- $\tau : [0,T] \times \tilde{\Omega} \to [0,T],\ (t, \tilde{\omega}) \mapsto \tau(t, \tilde{\omega}) := \max\{t_i(\tilde{\omega}) | 0 \le i \le n(\tilde{\omega}), t_i(\tilde{\omega}) \le t\}$, the last observation time before a certain time t, and
- $M = (M_k)_{0 \le k \le K}$, the observation mask, which is a sequence of random variables on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ taking values in $\{0,1\}^{d_X}$ such that M_k is $\tilde{\mathcal{F}}_{t_k}$-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 $M_{t_k} := M_k$.

In the following we consider the filtered product probability space $(\Omega \times \tilde{\Omega}, \mathcal{F} \otimes \tilde{\mathcal{F}}, \mathbb{F} \otimes \tilde{\mathbb{F}}, \mathbb{P} \times \tilde{\mathbb{P}})$ and the filtration of the currently available information $\mathbb{A} := (\mathcal{A}_t)_{t \in [0,T]}$ defined by

$$\mathcal{A}_t := \boldsymbol{\sigma}\left(X_{t_i,j}, t_i, M_{t_i} | t_i \le t,\, j \in \{1 \le l \le d_X | M_{t_i,l} = 1\}\right),$$

where $\boldsymbol{\sigma}(\cdot)$ denotes the generated σ-algebra. We note that $\mathcal{A}_t = \mathcal{A}_{\tau(t)}$ 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), is defined as $\hat{X} = (\hat{X}_t)_{0 \le t \le T}$, with $\hat{X}_t := \mathbb{E}_{\mathbb{P} \times \tilde{\mathbb{P}}}[X_t | \mathcal{A}_t]$. Moreover, for any $0 \le t \le T$ the j-th coordinate of the interpolated observation process $\tilde{X}^{\le t} \in \mathbb{R}^{2d_X}$ at time $0 \le s \le T$ is defined by

$$\tilde{X}^{\le t}_{s,j} := \begin{cases} X_{t_{l(s,t)},j} \frac{t_{\ell(s,t)} - s}{t_{\ell(s,t)} - t_{\ell(s,t)-1}} + X_{t_{\ell(s,t)},j} \frac{s - t_{\ell(s,t)-1}}{t_{\ell(s,t)} - t_{\ell(s,t)-1}}, & \text{if } t_{\ell(s,t)-1} < s \le t_{\ell(s,t)} \text{ and} \\ & 1 \le j \le d_X, \\ X_{t_{l(s,t)},j}, & \text{if } s \le t_{\ell(s,t)-1} \text{ and } 1 \le j \le d_X, \\ u_{t_{l(s,t)},j-d_X} + \frac{s - t_{\ell(s,t)-1}}{t_{\ell(s,t)} - t_{\ell(s,t)-1}}, & \text{if } t_{\ell(s,t)-1} < s \le t_{\ell(s,t)} \text{ and} \\ & d_X < j \le 2d_X, \\ u_{t_{l(s,t)},j-d_X}, & \text{if } s \le t_{\ell(s,t)-1} \text{ and} d_X < j \le 2d_X, \end{cases}$$

where $u_{t,j} := \sum_{k=0}^{K} M_{k,j} \mathbb{1}_{t_k \le t}$ is the jump process that counts the coordinate-wise observations and

$$l(s,t) := l(s,t,j) := \max\{0 \le l \le n | t_l \le \min(s,t), M_{t_l,j} = 1\},$$
$$\ell(s,t) := \ell(s,t,j) := \inf\{1 \le \ell \le n | s \le t_\ell \le t, M_{t_\ell,j} = 1\},$$

with $t_\infty := T$. The paths of $\tilde{X}^{\le t}$ belong to $BV^c([0,T])$, the set of continuous $\mathbb{R}^{d_X}$-valued paths of bounded variation on $[0,T]$. Since $\tilde{X}^{\le \tau(t)} \in \mathcal{A}_{\tau(t)}$ carries all available information, 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 $\hat{X}_{t,j} = F_j(t, \tau(t), \tilde{X}^{\le \tau(t)})$. In Krach et al. (2022) convergence of the PD-NJ-ODE model to the conditional expectation $\hat{X}$ was shown under the following assumptions.

Assumption 2.1. *We assume that*

1. *for every* $1 \leq k, l \leq K$, M_k *is independent of* t_l *and of* n, $\tilde{\mathbb{P}}(M_{k,j} = 1) > 0$ *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* $|M_k|_1 > 0$ *for every* $1 \leq k \leq K$ $\tilde{\mathbb{P}}$*-almost surely (at every observation time at least one coordinate is observed),*

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

3. *almost surely* X *is not observed at a jump, i.e.,* $(\mathbb{P} \times \tilde{\mathbb{P}})(t_j \in \mathcal{J} | j \leq n) = (\mathbb{P} \times \tilde{\mathbb{P}})(\Delta X_{t_j} \neq 0 | j \leq n) = 0$ *for all* $1 \leq j \leq K$,

4. 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* $X^\star$*, i.e.,*

$$|F_j(\tau(t), \tau(t), \tilde{X}^{\leq \tau(t)})| + |f_j(t, \tau(t), \tilde{X}^{\leq \tau(t)})| \leq B(X_t^\star + 1)^p,$$

5. $X^\star$ *is* L^{2p}*-integrable, i.e.,* $\mathbb{E}[(X_T^\star)^{2p}] < \infty$,

6. *the random number of observations* n *is integrable, i.e.,* $\mathbb{E}_{\tilde{\mathbb{P}}}[n] < \infty$.

Moreover, relaxations on the assumption of observing X_0 completely were discussed in Krach et al. (2022, Remark 2.3).

Definition 2.2. *Let* $c_0 := c_0(k) := (\tilde{\mathbb{P}}(n \geq k))^{-1}$. *A distance between càdlàg* $\mathbb{A}$*-measurable processes* $Z, \xi : [0, T] \times (\Omega \times \tilde{\Omega}) \to \mathbb{R}^r$ *is defined through the pseudo metrics*

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

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

The path-dependent generalisation of the Neural Jump ODE model (Herrera et al., 2021) uses the truncated signature transformation π_m (Krach et al., 2022, Definition 3.4) and bounded output neural networks $f_{(\tilde{\theta}, \gamma)}$, where $\tilde{\theta}$ 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 $\tilde{\mathcal{N}}$ of standard neural networks. In the following we assume that $\tilde{\mathcal{N}}$ is a set of standard feedforward neural networks with $\mathrm{id} \in \tilde{\mathcal{N}}$ that satisfies the standard universal approximation theorem with respect to the supremum-norm on compact sets, see for example Hornik (1991, Theorem 2).

Definition 2.3. *The Path-Dependent Neural Jump ODE (PD-NJ-ODE) model is given by*

$$\begin{aligned} H_0 &= \rho_{\theta_2}(0, 0, \pi_m(0), X_0), \\ dH_t &= f_{\theta_1}\left(H_{t-}, t, \tau(t), \pi_m(\tilde{X}^{\leq \tau(t)} - X_0), X_0\right) dt \\ &\quad + \left(\rho_{\theta_2}\left(H_{t-}, t, \pi_m(\tilde{X}^{\leq \tau(t)} - X_0), X_0\right) - H_{t-}\right) du_t, \\ Y_t &= \tilde{g}_{\tilde{\theta}_3}(H_t). \end{aligned} \tag{2}$$

The functions $f_{\theta_1}, \rho_{\theta_2} \in \mathcal{N}$ *are bounded output feedforward neural networks and* $\tilde{g}_{\tilde{\theta}_3} \in \tilde{\mathcal{N}}$ *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* Θ *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* $u_t := \sum_{k=1}^K \mathbb{1}_{t_k \leq t}$.

The existence and uniqueness of a solution of (2) is implied by Protter (2005, Thm. 7, Chap. V). To emphasize the dependence of the PD-NJ-ODE output Y on θ and X we write $Y^\theta(X)$.

The objective function (cf. *equivalent objective function* from Remark 4.7 & Appendix A.1.4 of Krach et al. (2022)) for the training of the PD-NJ-ODE is defined as

$$\Psi : \mathbb{D} \to \mathbb{R},$$

$$Z \mapsto \Psi(Z) := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(|M_i \odot (X_{t_i} - Z_{t_i})|_2 + |M_i \odot (X_{t_i} - Z_{t_i-})|_2\right)^2\right], \tag{3}$$

$$\Phi : \Theta \to \mathbb{R}, \theta \mapsto \Phi(\theta) := \Psi(Y^\theta(X)), \tag{4}$$

where $\odot$ is the element-wise multiplication (Hadamard product) and $\mathbb{D}$ the set of all càdlàg $\mathbb{R}^{d_X}$-valued $\mathbb{A}$-adapted processes on the product probability space $\Omega \times \tilde{\Omega}$. Moreover, for $N \in \mathbb{N}$ and $X^{(j)} \sim X$, $M^{(j)} \sim M$ and $(n^{(j)}, t_1^{(j)}, \cdots, t_{n^{(j)}}^{(j)}) \sim (n, t_1, \cdots, t_n)$ i.i.d. random processes (respectively variables) for $1 \leq j \leq N$ the Monte Carlo approximation of (4) is

$$\hat{\Phi}_N(\theta) := \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n^{(j)}}\sum_{i=1}^{n^{(j)}}\left(\left|M_i^{(j)} \odot \left(X_{t_i^{(j)}}^{(j)} - Y_{t_i^{(j)}}^{\theta,j}\right)\right|_2 + \left|M_i^{(j)} \odot \left(X_{t_i^{(j)}}^{(j)} - Y_{t_i^{(j)}-}^{\theta,j}\right)\right|_2\right)^2, \tag{5}$$

where $Y^{\theta,j} := Y^\theta(X^{(j)})$.

Based on these loss functions, the following convergence guarantees can be derived, where $\hat{\Theta}_m \subset \Theta$ is defined as the set of possible weights for the 3 (bounded output) neural networks, such that their widths are at most m and such that the truncated signature of level m or smaller is used. We use $\Theta_m := \{\theta = ((\tilde{\theta}_1, \gamma_1), (\tilde{\theta}_2, \gamma_2), \tilde{\theta}_3) \in \hat{\Theta}_m \,|\, |\tilde{\theta}_i|_2 \leq m, \gamma_i \leq m\} \subset \hat{\Theta}_m$, which is a compact subset of Θ_m.

Theorem 2.4. *Let $\theta_m^{\min} \in \Theta_m^{\min} := \operatorname{argmin}_{\theta\in\Theta_m}\{\Phi(\theta)\}$ for every $m \in \mathbb{N}$. If Assumption 2.1 is satisfied, then, for $m \to \infty$, the value of the loss function Φ (4) converges to the minimal value of Ψ (3) which is uniquely achieved by $\hat{X}$ up to indistinguishability, i.e.,*

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

Furthermore, for every $1 \leq k \leq K$ we have that $Y^{\theta_m^{\min}}$ converges to $\hat{X}$ in the metric d_k (1) as $m \to \infty$.

Let $\theta_{m,N}^{\min} \in \Theta_{m,N}^{\min} := \operatorname{argmin}_{\theta\in\Theta_m}\{\hat{\Phi}_N(\theta)\}$ for every $m, N \in \mathbb{N}$. Then, for every $m \in \mathbb{N}$, $(\mathbb{P}\times\tilde{\mathbb{P}})$-a.s.

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

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

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

In particular, one can define an increasing sequence $(N_m)_{m\in\mathbb{N}}$ in $\mathbb{N}$ such that for every $1 \leq k \leq K$ we have that $Y^{\theta_{m,N_m}^{\min}}$ converges to $\hat{X}$ in the metric d_k as $m \to \infty$.

3 PD-NJ-ODE with Noisy Observations

So far the PD-NJ-ODE model does not support measurement noise in the form of i.i.d. noise terms that are added to each observation. 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 this section, we introduce observation noise (e.g., measurement-noise), i.e., i.i.d. noise terms ϵ_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 $\mathbb{E}[X_t \,|\, O_{t_0}, \ldots, O_{\tau(t)}]$.

Inspecting the PD-NJ-ODE model and its loss function (replacing X by O), we notice two things. First that there is no reason why the model architecture should not be able to learn this modified objective, which should still have the same properties. And secondly 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_i}. However, this would not be the correct behaviour, since in general $\mathbb{E}[X_{t_i} \,|\, O_{t_0}, \ldots, O_{t_i}] \neq O_{t_i}$ as the conditional expectation filters out the noise as good as possible. Therefore we 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[2] if the observation noise ϵ_i 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 is enough to omit the first term of the loss function to recover the original results of Theorem 2.4 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 not optimizing the loss. In the following this is formalised.

3.1 Setting with Noisy Observations

The process X as well as the n, K, t_i, τ, 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_i}, we redefine the filtration of the currently available information via

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

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

Assumption 3.1. *We assume that assumptions 1, 2, 3 and 5 of Assumption 2.1 hold and additionally that*

4. *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 $X^\star$, i.e.,*

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

6. *n is square-integrable, i.e., $\mathbb{E}_{\tilde{\mathbb{P}}}[|n|^2] < \infty$.*
7. *the i.i.d. random noise variables ϵ_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 2.3) can equivalently be applied in this setting here.*

[2]Note that, in general we have $\mathbb{E}[X_{t_i} \,|\, O_{t_0}, \ldots, O_{t_i}] \neq \mathbb{E}[O_{t_i} \,|\, O_{t_0}, \ldots, O_{t_i}] = O_{t_i}$ at observation times, in contrast to their equality between observation times.

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 redefine the objective function as described before as

$$\Psi : \mathbb{D} \to \mathbb{R},\ Z \mapsto \Psi(Z) := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n} |M_i \odot (O_{t_i} - Z_{t_i-})|_2^2\right], \tag{6}$$

$$\Phi : \Theta \to \mathbb{R},\ \theta \mapsto \Phi(\theta) := \Psi(Y^\theta(X)), \tag{7}$$

and its Monte Carlo approximation accordingly.

3.2 Convergence Theorem with Noisy Observations

In the setting defined in Section 3.1, Theorem 2.4 holds equivalently as before. To prove this, we first need to adjust Krach et al. (2022, Lemma 4.2) for this setting.

Lemma 3.3. *For any $\mathbb{A}$-adapted process Z it holds that*

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^{n} |M_{t_i} \odot (O_{t_i} - Z_{t_i-})|_2^2\right] = \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (O_{t_i} - \hat{X}_{t_i-})\right|_2^2\right] + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (\hat{X}_{t_i-} - Z_{t_i-})\right|_2^2\right].$$

Proof. First note that by Assumption 3.1 point 3 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. Similarly as in Krach et al. (2022, Lemma 4.2), we can derive for $\hat{O}_{t_i-} := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[O_{t_i-} \,|\, \mathcal{A}_{t_i-}]$ that

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^{n} |M_{t_i} \odot (O_{t_i-} - Z_{t_i-})|_2^2\right] = \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (O_{t_i-} - \hat{O}_{t_i-})\right|_2^2\right] + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (\hat{O}_{t_i-} - Z_{t_i-})\right|_2^2\right].$$

To conclude the proof, it is enough to note that

$$\hat{O}_{t_i-} = \hat{X}_{t_i-} + \mathbb{E}[\epsilon_i|\mathcal{A}_{t_i-}] = \hat{X}_{t_i-} + \mathbb{E}[\epsilon_i] = \hat{X}_{t_i-}, \tag{8}$$

using that ϵ_i has expectation 0 and is independent of $\mathcal{A}_{t_i-}$. □

In the following we sketch the proof of Theorem 2.4 in the setting of noisy observations, by only outlining those parts of it that need to be changed in comparison with the original proof in Krach et al. (2022). A full proof is given in Appendix B.

Sketch of Proof of Theorem 2.4 for noisy observations. First, it follows directly from Lemma 3.3 that $\Psi(\hat{X}) = \min_{Z\in\mathbb{D}} \Psi(Z)$, i.e., that $\hat{X}$ is a minimizer of the redefined objective function Ψ. Secondly, again by Lemma 3.3, we have for any process $Z \in \mathbb{D}$ that

$$\Psi(Z) = \Psi(\hat{X}) + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\tfrac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (\hat{X}_{t_i-} - Z_{t_i-})\right|_2^2\right],$$

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 Ψ.

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

$$U := 3B\left((X_T^\star + 1)^p + \sum_{i=0}^{n} |\epsilon_i|\right).$$

Defining

$$c_m := c\,\varepsilon(T+1)d_X + c(T+1)d_X U\left(\mathbb{1}_{\{X_T^\star \geq 1/\varepsilon\}} + \mathbb{1}_{\{n \geq 1/\varepsilon\}} + \mathbb{1}_{\{\delta \leq \varepsilon\}}\right)$$

it follows that for all $t \in [0,T]$ we have $\left|Y_t^{\theta_m^*} - \hat{X}_t\right|_2 \leq c_m$. Convergence of $\Phi(\theta_m^\star)$ to $\Psi(\hat{X})$ then follows similarly as before, when noting that by Assumption 3.1

$$\mathbb{E}\left[(X_t^\star + 1)^{2p} + \left(\sum_{i=0}^{n} |\epsilon_i|\right)^2\right] \leq \mathbb{E}\left[(X_t^\star + 1)^{2p}\right] + \mathbb{E}[n^2]\mathbb{E}\left[|\epsilon_0|^2\right] < \infty, \tag{9}$$

using Cauchy–Schwarz and that the ϵ_i are i.i.d. and independent of n for the first step and the integrability of $X^\star$, ϵ_0 and n^2 for the upper bound. Moreover, the convergence of $d_k(\hat{X}, Y^{\theta_m^\star}) \to 0$ 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 $\sup_\theta h(\theta, \xi_j)$ follows from (9). □

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,

$$\beta_i(\tilde{O}^{\leq \tau(t)}) := \mathbb{E}[\epsilon_i | \mathcal{A}_{t_i-}],$$

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

$$\Psi : \mathbb{D} \to \mathbb{R},\ Z \mapsto \Psi(Z) := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left|M_i \odot \left(\left(O_{t_i} - \beta_i(\tilde{O}^{\leq t_{i-1}})\right) - Z_{t_i-}\right)\right|_2^2\right]. \tag{10}$$

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

$$\mathbb{E}\left[\left(O_{t_i-} - \beta_i(\tilde{O}^{\leq t_{i-1}})\right)|\mathcal{A}_{t_i-}\right] = \hat{X}_{t_i-} + \mathbb{E}[\epsilon_i \mid \mathcal{A}_{t_i-}] - \beta_i = \hat{X}_{t_i-}. \tag{11}$$

This implies that the statement of Lemma 3.3 holds equivalently under the reduced assumption of a known conditional bias function, when using the adjusted loss (10). Additionally assuming that $\mathbb{E}\left[\left(\sum_{i=0}^{n}|\epsilon_i|\right)^2\right] < \infty$, the following result follows as before.

Corollary 3.4. *In the setting described in this sub-section (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 2.4 holds equivalently as before when using the objective function* (11).

The following remark explains how Corollary 3.4 can be used to predict conditional higher moments (instead of only the conditional expectation) under certain assumptions.

Remark 3.5. *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 $O_{t_i-} = X_{t_i-} + \epsilon_i$, where we assume that*

- *ϵ_i is independent of $\mathcal{A}_{t_i-}$,*
- *ϵ_i is conditionally independent of X_{t_i-} given $\mathcal{A}_{t_i-}$*

- *and ϵ_i have known finite moments.*

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

$$O^q_{t_i-} = X^q_{t_i-} + \sum_{j=1}^{q} \binom{q}{j} X^{q-j}_{t_i-} \epsilon^j_i.^{3}$$

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

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

where we use the assumptions on ϵ_i together with Proposition A.7 for the second inequality.

Then an inductive argument shows that β^q_i is a known function of the observations, using the assumption that the moments of ϵ_i are known. Indeed, to compute β^q_i the conditional expectations of smaller moments $\mathbb{E}[X^{q-j}_{t_i-} | \mathcal{A}_{t_i-}]$ need to be computed, which can be done according to the induction hypothesis (note that the base case follows directly from Corollary 3.4 and the assumptions on ϵ_i). Therefore, Corollary 3.4 implies that we can compute $\mathbb{E}[X^q_{t_i-} | \mathcal{A}_{t_i-}]$ (assuming that we reach the limit where the PD-NJ-ODE output equals the conditional expectation). In case of an exponential moment assumption $\mathbb{E}[\exp(\lambda |X_{t_i-}|)] < \infty$ for some $\lambda > 0$ we can therefore infer the conditional law of X_{t_i-}.

3.4 Examples

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_i} instead of the original observations X_{t_i}. Therefore, we give one explicit example where we compute the conditional expectation in the noisy observation setting.

3.4.1 Brownian Motion with Gaussian Observation Noise

Let $X := W$ be a standard Brownian motion and let $\epsilon_0 = 0$, $\epsilon_i \sim N(0, \sigma^2)$ for $i \geq 1$ be the i.i.d. noise terms for some $\sigma > 0$. Then $O_{t_i} = X_{t_i} + \epsilon_i$ are the observations. Clearly, all integrability assumptions are satisfied by X and ϵ_i (cf. Krach et al. (2022, Section 7.6)). To compute the true conditional expectations we first note that the independent increments property of the Brownian motion imply for $t_k \leq t < t_{k+1}$

$$\mathbb{E}[X_t | \mathcal{A}_t] = \mathbb{E}[W_t - W_{t_k} | \mathcal{A}_{t_k}] + \mathbb{E}[W_{t_k} | \mathcal{A}_{t_k}] = \mathbb{E}[W_{t_k} | \mathcal{A}_{t_k}] = \mathbb{E}[W_{t_k} | O_{t_1}, \ldots, O_{t_k}],$$

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

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

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 $(\Sigma_{11})_{i,j} = \min(t_i, t_j) + \sigma^2 \mathbb{1}_{\{i=j\}}$, $\Sigma_{12}^\top = \Sigma_{21} = (t_1, \ldots, t_k) \in \mathbb{R}^{1 \times k}$ 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 $\hat{\mu} := \Sigma_{21}\Sigma_{11}^{-1}(O_{t_1}, \ldots, O_{t_k})^\top$ and variance $\hat{\Sigma} := \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ (Eaton, 2007, Proposition 3.13). In particular we have

$$\mathbb{E}[X_t | \mathcal{A}_t] = \mathbb{E}[W_{t_k} | O_{t_1}, \ldots, O_{t_k}] = \hat{\mu}.$$

[3] Note that q, j and $q - j$ denote exponents here rather than superscripts.

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 2.4.

4.1 Intuition on independence assumptions

In many real word applications the independence of the process X and the observation framework (i.e., $(t_i)_{i\in\{1,\dots,n\}}$ and M) is heavily violated. Think for example of irregular measurements of patients' health-parameters taken at a hospital. Usually (expensive) measurements are only taken if any information on the state X of the patient hints that this measurement might be relevant. 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 will precisely formulate in item 7 in Assumption 4.1). In theory this is a perfectly realistic assumption, *if* we assume that *every* piece of information the hospital gets from the patient is immediately logged as an (possibly incomplete and noisy) observation of X; in this case the hospital does not have any further information on X rather than the observations, which makes decisions whether a measurement should be taken conditionally independent of X given the past observations (thus fulfilling item 7).

However, in real-world practical applications one has to be careful, since it can for example happen that the hospital gets access to information on X that is not officially logged as an observation; e.g., imagine the doctor measures the body temperature of a patient who told her that he feels feverish. If the doctor often does not log the observation of feverish feelings of her patients, but only logs their temperature-measurements, our model (and most other classical forecasting methods too) would learn a quite biased forecast of the body temperature. For example, in the extreme case that the body temperature is only measured if patients feel very feverish, most measurements will give a high body temperature, which leads our model to always predicting 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. This problem can be mitigated in this example, if the hospital also logs whether the patient feels feverish in a further coordinate of X. Then one should first enter the feverish feeling of the patient as an observation and then enter the measurement of the body temperature as the next observation with a (slightly) later time stamp. In this case one would still have a problem with the conditional independence assumption if the patients mainly report feverish feelings rather than the absence of such a feeling. This could be fixed if the software automatically logs the absence of feverish feelings whenever they do not report a feverish feeling, if one can rely on the patients always reporting the intervals of feverish feelings. In such practical situations one should be aware that even our weakened, more realistic assumption of conditional independence is often not fully satisfied.

4.2 Setting with Dependence

To allow for dependence, we only consider the probability measure $\mathbb{P}$ and define $X, n, K, t_i, \tau, M, \mathcal{A}_t$ similar as before, but all on the same probability space 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 $\boldsymbol{\sigma}(Z, \mathcal{B}) := \boldsymbol{\sigma}(Z) \vee \boldsymbol{\sigma}(\mathcal{B})$. Then we need the following assumptions, where the main difference to the original Assumption 1 is in item 7 (and 8).

Assumption 4.1. *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. *the probability that any two observation times are closer than $\epsilon > 0$ converges to 0 when ϵ does, i.e., if $\delta(\omega) := \min_{0 \le i \le n(\omega)} |t_{i+1}(\omega) - t_i(\omega)|$ then $\lim_{\epsilon \to 0} \mathbb{P}(\delta < \epsilon) = 0$,*

3. *almost surely X is not observed at a jump, i.e., $\mathbb{P}(t_j \in \mathcal{J} \mid j \le n) = \mathbb{P}(\Delta X_{t_j} \neq 0 \mid j \le n) = 0$ for all $1 \le j \le K$,*

4. *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 $X^\star$, i.e.,*

$$\left|F_j(\tau(t), \tau(t), \tilde{X}^{\le \tau(t)})\right| + \left|f_j(t, \tau(t), \tilde{X}^{\le \tau(t)})\right| \le B(X_t^\star + 1)^p,$$

5. *$X^\star$ is L^{2p}-integrable, i.e., $\mathbb{E}[(X_T^\star)^{2p}] < \infty$,*

6. *the random number of observations n is integrable, i.e., $\mathbb{E}_{\tilde{\mathbb{P}}}[n] < \infty$,*

7. *for every $1 \le i \le n$, X_{t_i-} is conditionally independent of $\boldsymbol{\sigma}(n, M_{t_i})$ given $\mathcal{A}_{t_i-}$, and*

8. *for all $1 \le k \le K$, $1 \le j \le d_X$ there is some $\eta_{k,j} > 0$ such that $\mathbb{P}(M_{k,j} = 1 \mid \boldsymbol{\sigma}(n, \mathcal{A}_{t_k-})) > \eta_{k,j}$ (i.e., given the currently known information and n, for each coordinate the probability of observing it at the next observation time is positive).*

Remark 4.2. *The relaxations on the assumption of observing X_0 completely discussed in Krach et al. (2022, Remark 2.3) 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 2.4 holds equivalently. Before we can show this we need some additional results. First, we prove an extension of the L^2-orthogonality result Herrera et al. (2021, Proposition B.2).

Proposition 4.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a fixed probability space, and $\mathcal{A}, \mathcal{B}$ be sub-σ-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*

$$\mathbb{E}\left[|X - Z|_2^2 \mid \mathcal{B}\right] = \mathbb{E}\left[\left|X - \hat{X}\right|_2^2 \mid \mathcal{B}\right] + \mathbb{E}\left[\left|Z - \hat{X}\right|_2^2 \mid \mathcal{B}\right] \tag{12}$$

$$\ge \mathbb{E}\left[\left|X - \hat{X}\right|_2^2 \mid \mathcal{B}\right], \tag{13}$$

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.

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

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

$$\mathbb{E}\left[Z\hat{X} \mid \mathcal{B}\right] = \mathbb{E}\left[Z\mathbb{E}\left[X \mid \mathcal{A}\right] \mid \mathcal{B}\right] = \mathbb{E}\left[\mathbb{E}\left[ZX \mid \mathcal{A}\right] \mid \mathcal{B}\right] = \mathbb{E}\left[ZX \mid \mathcal{B}\right]$$

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 $\mathbb{E}[\hat{X}(X-\hat{X})\,|\,\mathcal{B}] = 0$ and therefore the cross term vanishes and the equality follows. The inequality holds since $\mathbb{E}[|Z-\hat{X}|_2^2\,|\,\mathcal{B}]$ 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 $\mathbb{E}[|Z-\hat{X}|_2^2\,|\,\mathcal{B}] = 0$ $\mathbb{P}$-a.s, which implies (by the tower property) that $\mathbb{E}[|Z-\hat{X}|_2^2] = 0$. This is only possible if $Z = \hat{X}$ $\mathbb{P}$-a.s., which contradicts the assumption that $\mathbb{P}(Z \neq \hat{X}) > 0$. □

Next we re-prove Krach et al. (2022, Lemma 4.2) under the relaxed assumptions.

Lemma 4.4. *For any $\mathbb{A}$-adapted process $Z \in L^2(\Omega, \mathbb{A}, \mathbb{P})$ it holds that*

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

Proof.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}|M_{t_i} \odot (X_{t_i-} - Z_{t_i-})|_2^2\right] &= \sum_{i=1}^{K}\mathbb{E}\left[\frac{1}{n}\mathbb{1}_{\{i\leq n\}}|M_{t_i} \odot (X_{t_i-} - Z_{t_i-})|_2^2\right] \\
&= \sum_{i=1}^{K}\mathbb{E}\left[\mathbb{E}\left[\frac{1}{n}\mathbb{1}_{\{i\leq n\}}|M_{t_i} \odot (X_{t_i-} - Z_{t_i-})|_2^2 \mid \boldsymbol{\sigma}(n, M_{t_i}, \mathcal{A}_{t_i-})\right]\right] \\
&= \sum_{i=1}^{K}\mathbb{E}\left[\frac{1}{n}\mathbb{1}_{\{i\leq n\}}\sum_{j=1}^{d_X}M_{t_i,j}\mathbb{E}\left[|X_{t_i-,} - Z_{t_i-,j}|^2 \mid \boldsymbol{\sigma}(n, M_{t_i}, \mathcal{A}_{t_i-})\right]\right] \\
&= \sum_{i=1}^{K}\mathbb{E}\left[\frac{1}{n}\mathbb{1}_{\{i\leq n\}}\sum_{j=1}^{d_X}M_{t_i,j}\mathbb{E}\left[|X_{t_i-,} - Z_{t_i-,j}|^2 \mid \mathcal{A}_{t_i-}\right]\right].
\end{aligned}$$

The first step follows by monotone convergence, the last by Lemma 4.5 below. Now we can conclude by first applying the equality from Proposition 4.3 with $\mathcal{A}, \mathcal{B} = \mathcal{A}_{t_i-}$ and then reversing the above steps to arrive at the desired form. □

Lemma 4.5. *Assume the context of Lemma 4.4. Then for all i, j it holds that*

$$\mathbb{E}\left[|X_{t_i-,j} - Z_{t_i-,j}|^2 \mid \boldsymbol{\sigma}(n, M_{t_i}, \mathcal{A}_{t_i-})\right] = \mathbb{E}\left[|X_{t_i-,j} - Z_{t_i-,j}|^2 \mid \mathcal{A}_{t_i-}\right].$$

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 $\mathcal{A}_{t_i-}$. 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 σ-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 $\mathbb{A}$-adapted, and so $Z_{t_i-,j}$ is $\mathcal{A}_{t_i-}$-measurable. Thus we have

$$\mathbb{E}\left[Z_{t_i-,j} \mid \boldsymbol{\sigma}(n, M_{t_i}, \mathcal{A}_{t_i-})\right] = Z_{t_i-,j} = \mathbb{E}\left[Z_{t_i-,j} \mid \mathcal{A}_{t_i-}\right]$$

as desired.

CASE $X_{t_i-,j}$: Assumption 4.1 point 7 implies that $X_{t_i-,j}$ is conditionally independent of $\boldsymbol{\sigma}(n, M_{t_i})$ given $\mathcal{A}_{t_i-}$. We therefore have by Proposition A.4

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

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 $\boldsymbol{\sigma}(n, M_{t_i})$ given $\mathcal{A}_{t_i-}$. Thus

$$\begin{aligned}\mathbb{E}\left[X_{t_i-,j}Z_{t_i-,j} \mid \boldsymbol{\sigma}(n, M_{t_i}, \mathcal{A}_{t_i-})\right] = \; & Z_{t_i-,j}\mathbb{E}\left[X_{t_i-,j} \mid \boldsymbol{\sigma}(n, M_{t_i}, \mathcal{A}_{t_i-})\right] \\ = \; & Z_{t_i-,j}\mathbb{E}\left[X_{t_i-,j} \mid \mathcal{A}_{t_i-}\right] = \mathbb{E}\left[Z_{t_i-,j}X_{t_i-,j} \mid \mathcal{A}_{t_i-}\right].\end{aligned}$$

Combining these 3 cases proves the claim. □

Finally, the following lemma shows that Assumption 4.1 point 8 can replace the independence assumption for M and X in the context of the following lemma.

Lemma 4.6. *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*

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

Proof. Assumption 4.1 point 8 states that $0 < \eta_{k,j} < \mathbb{P}(M_{k,j} = 1 \mid \sigma(n, A_{t_k-})) = \mathbb{E}[M_{t_k,j} \mid \boldsymbol{\sigma}(n, \mathcal{A}_{t_k-})]$ 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

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\{n\geq k\}}\left|M_{t_k} \odot (\hat{X}_{t_k-} - Z_{t_k-})\right|_1\right] &= \sum_{j=1}^{d_X}\mathbb{E}\left[\mathbb{1}_{\{n\geq k\}}M_{t_k,j}\left|\hat{X}_{t_k-,j} - Z_{t_k-,j}\right|\right] \\ &= \sum_{j=1}^{d_X}\mathbb{E}\left[\mathbb{1}_{\{n\geq k\}}\left|\hat{X}_{t_k-,j} - Z_{t_k-,j}\right|\mathbb{E}\left[M_{t_k,j} \mid \boldsymbol{\sigma}(n, \mathcal{A}_{t_k-})\right]\right] \\ &\geq c_2\,\mathbb{E}\left[\mathbb{1}_{\{n\geq k\}}\left|\hat{X}_{t_k-} - Z_{t_k-}\right|_1\right],\end{aligned}$$

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

Now we are ready to prove Theorem 2.4 in the setting with dependence. In the following we give again a sketch of the proof, by only outlining those parts of it that need to be changed in comparison with the original proof in Krach et al. (2022) and refer the interested reader to the full proof in Appendix B.

Sketch of Proof of Theorem 2.4 with dependence. As before it follows that $\hat{X}$ is a minimizer of Ψ. To show its uniqueness, we first note that

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\mathbb{1}_{\{n\geq k\}}\left|\hat{X}_{t_k-} - Z_{t_k-}\right|_2\right] \leq \frac{c_3}{c_2(k)}\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\mathbb{1}_{\{n\geq k\}}\left|M_{t_k} \odot (\hat{X}_{t_k-} - Z_{t_k-})\right|_2\right] \tag{14}$$

is implied as before when using Lemma 4.6 instead of the independence. Then

$$\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left|M_{t_i} \odot (\hat{X}_{t_i-} - Z_{t_i-})\right|_2^2\right] \geq \left(\frac{c_2(k)}{c_0c_1c_3}\right)^2 d_k(\hat{X}, Z)^2 > 0,$$

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

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

$$\min_{Z\in\mathbb{D}}\Psi(Z) \leq \Phi(\theta_m^{\min}) \leq \Phi(\theta_m^*)$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\left|M_{t_i}\odot(X_{t_i}-Y^{\theta^*_m}_{t_i})\right|_2+\left|M_{t_i}\odot(Y^{\theta^*_m}_{t_i}-Y^{\theta^*_m}_{t_i-})\right|_2\right)^2\right]\\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\left|M_{t_i}\odot(\hat{X}_{t_i}-Y^{\theta^*_m}_{t_i})\right|_2+\left|M_{t_i}\odot(Y^{\theta^*_m}_{t_i}-\hat{X}_{t_i})\right|_2\right.\right.\\
&\qquad\left.\left.+\left|M_{t_i}\odot(\hat{X}_{t_i}-\hat{X}_{t_i-})\right|_2+\left|M_{t_i}\odot(\hat{X}_{t_i-}-Y^{\theta^*_m}_{t_i-})\right|_2\right)^2\right]\\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\left|\hat{X}_{t_i}-Y^{\theta^*_m}_{t_i}\right|_2+\left|Y^{\theta^*_m}_{t_i}-\hat{X}_{t_i}\right|_2\right.\right.\\
&\qquad\left.\left.+\left|M_{t_i}\odot(\hat{X}_{t_i}-\hat{X}_{t_i-})\right|_2+\left|\hat{X}_{t_i-}-Y^{\theta^*_m}_{t_i-}\right|_2\right)^2\right]\\
&\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\left|M_{t_i}\odot(X_{t_i}-\hat{X}_{t_i-})\right|_2+3c_m\right)^2\right]\\
&= \Psi(\hat{X})+\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(6c_m\left|M_{t_i}\odot(X_{t_i}-\hat{X}_{t_i-})\right|_2+9c_m^2\right)\right]\\
&= \Psi(\hat{X})+\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}6c_m\left|M_{t_i}\odot(X_{t_i}-\hat{X}_{t_i-})\right|_2\right]+9\mathbb{E}\left[c_m^2\right]\\
&\leq \Psi(\hat{X})+6\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}c_m\left|X_{t_i}-\hat{X}_{t_i-}\right|_2\right]+9\mathbb{E}\left[c_m^2\right]\\
&\leq \Psi(\hat{X})+6\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}c_m\cdot 2TB(X^\star_T+1)^p\right]+9\mathbb{E}\left[c_m^2\right]\\
&= \Psi(\hat{X})+12TB\mathbb{E}\left[c_m(X^\star_T+1)^p\right]+9\mathbb{E}\left[c_m^2\right],
\end{aligned}
$$

where the triangle inequality was used in the third line, and we use Assumption 4.1 point 4 (which implies that $|\hat{X}_t|\leq TB(X^\star+1)^p$ for all t) to construct a crude bound in the second to last line. As in Krach et al. (2022), dominated convergence can be used to show

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

since $2p$ is again the largest power of $X^\star_T$ in both terms (remember c_m has an $(X^\star_T)^p$ term). The convergence of $d_k(\hat{X},Y^{\theta^\star_m})\to 0$ 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. □

4.4 Examples

Clearly, all the examples from Krach et al. (2022) are trivially valid for our generalised settings of Section 4.2 since points 6 & 7 of Assumption 4.1 are implied by independence. Furthermore, we provide a relatively general class of examples in Section 4.4.1 and extend two of the examples from Krach et al. (2022) to a setting where independence does not hold in Example 4.7 and Example 4.9.

4.4.1 Class of Examples Incorporating Dependence

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 7. One way to circumvent this problem is to define n conditionally independent of X_{t_i-} given $\mathcal{A}_{t_i-}$ for all i and to introduce a dummy variable, e.g. the time t itself, that is observed at every

t_i. Then we can control whether the process X is observed at an observation time t_i via the observation mask M_i^X (denoting the mask for the X-part of the observation, i.e., without the dummy variable). Since the dummy variable is always observed, the assumption that at least one coordinate of the mask has to be positive is trivially satisfied, such that $M_i^X = 0$ is admissible. In this way, the times at which X is observed can depend on X through the observation mask. 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 $\mathcal{A}_{t_i-}$ 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 7 and 8 are satisfied (assuming that a dummy variable is observed at every observation time), when defining the observation mask. A way to define M_i^X such that point 7 is satisfied is to define it as a function of random variables that are $\mathcal{A}_{t_i-}$-measurable and random variables that are independent of $\boldsymbol{\sigma}(\mathcal{A}_{t_i-}, X_{t_i-}, n)$. In particular, let $M_i^X := g_i\left((U_j^i)_{j\in J_1^i}, (V_j^i)_{j\in J_2^i}\right)$, where g_i is a measurable function (mapping to $\{0,1\}^{d_X}$), $J_1^i, J_2^i \subseteq \mathbb{N}$, U_j^i is a $\mathcal{A}_{t_i-}$-measurable random variable and V_j^i a random variable independent of $\boldsymbol{\sigma}(\mathcal{A}_{t_i-}, X_{t_i-}, n)$ 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 ϕ we have $\mathbb{E}[\phi(X_{t_i-}) \,|\, \boldsymbol{\sigma}(n, M_i, \mathcal{A}_{t_i-})] = \mathbb{E}[\phi(X_{t_i-}) \,|\, \mathcal{A}_{t_i-}]$. Indeed, M_i is $\boldsymbol{\sigma}(\mathcal{A}_{t_i-}, (V_j^i)_{j\in J_2^i})$-measurable by construction, therefore, we have for such a ϕ that

$$\mathbb{E}\left[\phi(X_{t_i-}) \mid \boldsymbol{\sigma}(n, M_i, \mathcal{A}_{t_i-})\right] = \mathbb{E}\left[\mathbb{E}\left[\phi(X_{t_i-}) \mid \boldsymbol{\sigma}(n, (V_j^i)_{j\in J_2^i}, \mathcal{A}_{t_i-})\right] \mid \boldsymbol{\sigma}(n, M_i, \mathcal{A}_{t_i-})\right],$$

by the tower property. On the other hand,

$$\mathbb{E}\left[\phi(X_{t_i-}) \mid \boldsymbol{\sigma}(n, (V_j^i)_{j\in J_2^i}, \mathcal{A}_{t_i-})\right] = \mathbb{E}\left[\phi(X_{t_i-}) \mid \boldsymbol{\sigma}(n, \mathcal{A}_{t_i-})\right] = \mathbb{E}\left[\phi(X_{t_i-}) \mid \mathcal{A}_{t_i-}\right],$$

using the independence of V_j^i 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 7.

For Assumption 4.1 point 8, we note that by Durrett (2010, Lemma 6.2.1) we have

$$\begin{aligned}
\mathbb{P}(M_k = 1 \mid \boldsymbol{\sigma}(n, A_{t_k-})) &= \mathbb{E}\left[M_{t_k} \mid \boldsymbol{\sigma}(n, \mathcal{A}_{t_k-})\right] \\
&= \mathbb{E}\left[g_k\left((U_j^k)_{j\in J_1^k}, (V_j^k)_{j\in J_2^k}\right) \mid \boldsymbol{\sigma}(n, \mathcal{A}_{t_k-})\right] \\
&= \tilde{g}_k((U_j^k)_{j\in J_1^k}),
\end{aligned}$$

for $\tilde{g}_k((u_j^k)_{j\in J_1^k}) := \mathbb{E}\left[g_k\left((u_j^k)_{j\in J_1^k}, (V_j^k)_{j\in J_2^k}\right)\right]$. 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 4.7 (Homogeneous Poisson Point Process with Dependent Observations)**.** *We use a homogeneous Poisson point process $X = N^\lambda$ (Krach et al., 2022, Section 7.4) 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 use time as dummy variable that is always observed and define the observation mask for X as*

$$M_i^X = \begin{cases} \mathbb{1}_{\{x_i \geq \lambda t_{i-1}\}},\ x_i \sim \mathcal{N}(N_{t_{i-1}}^\lambda, \sigma^2) & \text{if } M_{i-1}^X = 1, \\ u_i \sim \text{Bernoulli}(p) & \text{if } M_{i-1}^X = 0. \end{cases}$$

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 $\mathbb{E}[N_{t_j}^\lambda] = \lambda t_j$ for all j). Otherwise the mask value is sampled from a Bernoulli distribution.

To satisfy Assumption 4.1 point 1, we define $M_0 := 1$. *To see that Assumption 4.1 point 7 holds, let* $V_1^i \sim \text{Bernoulli}(p)$ *and* $V_2^i \sim \mathcal{N}(0, \sigma^2)$ *be independent random variables independent of* $\boldsymbol{\sigma}(\mathcal{A}_{t_i-}, N^\lambda_{t_i-})$. *Then*

$$M_i^X = \mathbb{1}_{\{M_{i-1}^X=1\}} \mathbb{1}_{\{N^\lambda_{t_{i-1}}+V_2^i \geq \lambda t_{i-1}\}} + V_1^i \mathbb{1}_{\{M_{i-1}^X=0\}} =: g_i(M_{i-1}^X, N^\lambda_{t_{i-1}}, t_{i-1}, V_1^i, V_2^i),$$

and the claim follows as explained above. Moreover, Assumption 4.1 point 8 is satisfied because

$$\begin{aligned}
\tilde{g}_i(M_{i-1}^X, N^\lambda_{t_{i-1}}, t_{i-1}) &= \mathbb{1}_{\{M_{i-1}^X=1\}} \mathbb{E}\left[\mathbb{1}_{\{a+V_2^i \geq \lambda b\}}\right]\Big|_{(a,b)=\left(N^\lambda_{t_{i-1}}, t_{i-1}\right)} + \mathbb{1}_{\{M_{i-1}^X=0\}} \mathbb{E}\left[V_1^i\right] \\
&\geq \min\left(\mathbb{P}\left[V_2^i \geq \lambda b - a\right]\Big|_{(a,b)=\left(N^\lambda_{t_{i-1}}, t_{i-1}\right)}, p\right) \\
&\geq \min\left(\mathbb{P}\left[V_2^i \geq \lambda T\right], p\right) =: \eta_i > 0,
\end{aligned}$$

using that $\lambda t_{i-1} - N^\lambda_{t_{i-1}} \leq \lambda T$, *since* $N^\lambda \geq 0$, $t_{i-1} \leq T$ *and* $\lambda > 0$, *in the last line.*

Remark 4.8. *We note that the choice* $X = N^\lambda$ *is only explicitly used for the fact that* N^λ *is lower bounded (by* 0*). Hence, the Poisson point process could be replaced by any other process that is lower bounded and satisfies Assumption 4.1 points 4 & 5.*

Example 4.9 (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* μ, σ *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^X. *We set* $M_0^X = 1$. *Moreover, we define the last time before* t *at which* X *was observed as* $\tau_X(t) = \max\{t_i \,|\, t_i < t, M_i^X = 1\}$. *Let* $V_1^i \sim \text{Bernoulli}\left(\frac{t_i - t_{i-1}}{t_i - \tau_X(t_i)}\right)$, $V_2^i \sim \mathcal{N}(0, \eta^2)$ *and* $V_3^i \sim \text{Bernoulli}\,(p)$ *be independent random variables for some* $\eta > 0$ *and* $p \in (0, 1)$. *Then we define*

$$M_i^X := V_1^i \, \mathbb{1}_{\{X_{\tau_X(t_i)} + V_2^i \geq x_0 e^{\mu t_i}\}} + (1 - V_1^i) \, V_3^i,$$

for all $i \geq 1$. *In particular, if* X *was observed at the previous observation time* $V_1^i = 1$ *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* $V_1^i = 0$ *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 point 7 and 8 are satisfied.*

5 Experiments

The code with all new experiments and those from Krach et al. (2022) is available at `https://github.com/???`. Further details about the experiments can be found in Appendix C. In particular, in Appendix C.1 we give details on the slightly 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.

$$\text{eval}(\hat{X}, Y^\theta) := \frac{1}{N_2} \sum_{j=1}^{N_2} \frac{1}{\kappa + 1} \sum_{i=0}^{\kappa} \left(\hat{X}^{(j)}_{\frac{iT}{\kappa}} - Y^{\theta,j}_{\frac{iT}{\kappa}}\right)^2,$$

where the outer sum runs over the test set of size N_2 and the inner sum runs over the equidistant grid points on the time interval $[0, T]$.

5.1 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.

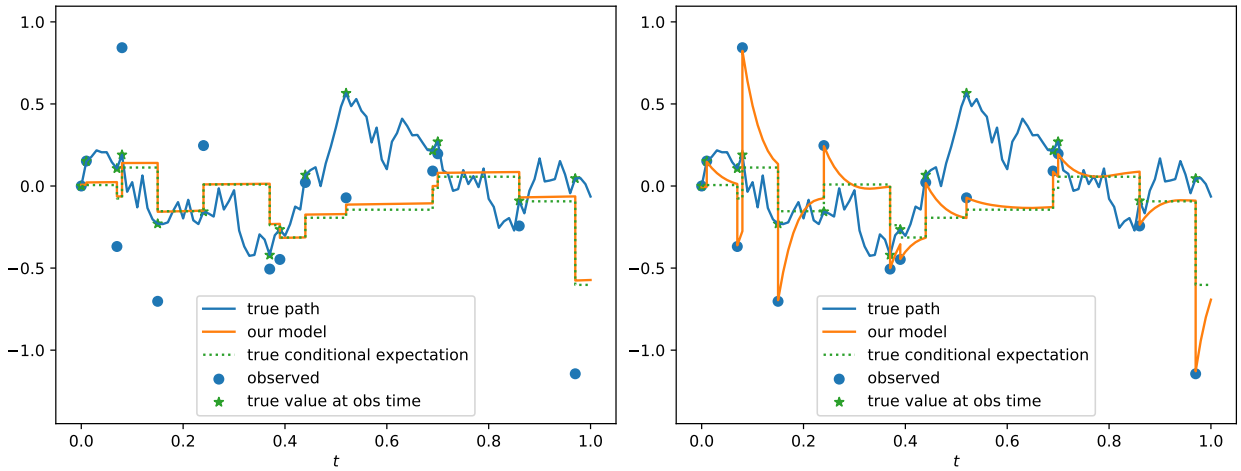

Figure 1: A test sample of a Brownian motion X with noisy observations $O_{t_i} = X_{t_i} + \epsilon_i$ together with the true and predicted conditional expectation. The PD-NJ-ODE is trained with the noise-adapted loss (left) and the original loss (right).

5.2 Dependent Observation Framework – a Black–Scholes Example

Based on Example 4.9 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 $V^i_{\{2,3\}}$ 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 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 noisy observations and for the observation framework (i.e., observation times and masks) to depend on previous information. 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.

References

Rick Durrett. Probability: Theory and Examples. *Cambridge University Press*, 2010. 10, 14, 19, 20

Morris L Eaton. *Multivariate Statistics: A Vector Space Approach.* Institute of Mathematical Statistics, 2007. 8

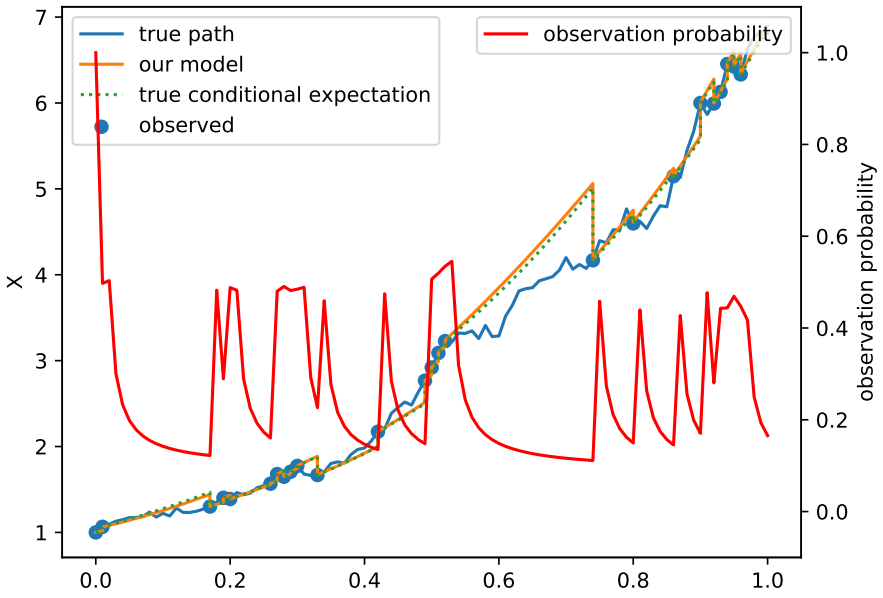

Figure 2: A test sample of a geometric Brownian motion X with observation probability depending on previously observed values of X. Plotted together with the true and predicted conditional expectation (scale on the left) and the observation probability over time (scale on the right).

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 2000. 1

Stefan Hansen. Conditional expectation on more than one sigma-algebra. Mathematics Stack Exchange, 2015. URL https://math.stackexchange.com/q/365319. URL:https://math.stackexchange.com/q/365319 (version: 2015-11-16). 18

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, 3, 10, 24, 25

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991. 3

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 1989. 22

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, 8, 10, 11, 12, 13, 14, 15, 16, 21, 22, 25, 26

Bernard Lapeyre and Jérôme Lelong. Neural network regression for bermudan option pricing. *Monte Carlo Methods and Applications*, 27(3):227–247, 2021. 24

Philip Protter. Stochastic integration and differential equations. *Springer-Verlag*, 2005. 4

Anton Maximilian Schäfer and Hans Georg Zimmermann. Recurrent neural networks are universal approximators. In *Artificial Neural Networks–ICANN 2006: 16th International Conference, Athens, Greece, September 10-14, 2006. Proceedings, Part I 16*, pp. 632–640. Springer, 2006. 1

Gunnar Taraldsen. Optimal learning from the doob-dynkin lemma. *arXiv*, 2018. 2, 7

Jisang Yoo. Conditional expectation on more than one sigma-algebra. Mathematics Stack Exchange, 2014. URL https://math.stackexchange.com/q/915935. URL:https://math.stackexchange.com/q/915935 (version: 2014-09-01). 18

Appendix

A Conditional independence

Let $(\Omega, \mathcal{F}, \mathbb{P})$ 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 $\mathbb{E}[X \,|\, \boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X \,|\, \mathcal{H}]$. 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 $\boldsymbol{\sigma}(X, \mathcal{H})$ is independent of $\mathcal{G}$ then $\mathbb{E}[X \,|\, \boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X \,|\, \mathcal{H}]$.*

Proof. We prove the desired equality in the context of basic measure theory. We first assume that X is integrable on $\boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})$, 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 $\boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})$-measurable,
2. Z is $\boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})$-integrable,
3. $\int_A X d\mathbb{P} = \int_A Z d\mathbb{P}$ for all $A \in \boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})$.

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

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

or equivalently

$$\mathbb{E}\left[X \cdot \mathbb{1}_{A \cap B}\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{H}\right] \cdot \mathbb{1}_{A \cap B}\right] \qquad \forall A \in \mathcal{G}, B \in \mathcal{H}.$$

Note that $\boldsymbol{\sigma}(X, \mathcal{H})$ being independent of $\mathcal{G}$ implies that $\mathcal{H}$ is independent of $\mathcal{G}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[X \cdot \mathbb{1}_{A \cap B}\right] &= \mathbb{E}\left[X \cdot \mathbb{1}_B \mathbb{1}_A\right] \\
&= \mathbb{E}\left[X \cdot \mathbb{1}_B\right] \mathbb{E}\left[\mathbb{1}_A\right] && (\boldsymbol{\sigma}(X, \mathcal{H}) \text{ indep. of } \mathcal{G}) \\
&= \mathbb{E}\left[\mathbb{E}\left[X \cdot \mathbb{1}_B \mid \mathcal{H}\right]\right] \cdot \mathbb{E}\left[\mathbb{1}_A\right] && (\text{tower property}) \\
&= \mathbb{E}\left[\mathbb{E}\left[X \cdot \mathbb{1}_B \mid \mathcal{H}\right] \cdot \mathbb{1}_A\right] && (\mathcal{H} \text{ indep. of } \mathcal{G}) \\
&= \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{H}\right] \cdot \mathbb{1}_A \mathbb{1}_B\right] && (\mathbb{1}_B \text{ is } \mathcal{H}\text{-measurable}) \\
&= \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{H}\right] \cdot \mathbb{1}_{A \cap B}\right],
\end{aligned}$$

completing the proof. □

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

Proof. This follows directly from Proposition A.1 upon replacing X by $\phi(X)$ and noting that $\boldsymbol{\sigma}(\phi(X), \mathcal{H}) \subseteq \boldsymbol{\sigma}(X, \mathcal{H})$, implying the needed condition. □

It should now be apparent that the assumption that X and $\mathcal{G}$ are independent is insufficient if we want to show $\mathbb{E}[X \,|\, \boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X \,|\, \mathcal{H}]$ (a counterexample is provided in Hansen (2015)). However, we can actually make a weaker assumption (though still stronger than the assumption that X and $\mathcal{G}$ are independent) to attain the same result, as shown in the next proposition.

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

$$\mathbb{E}\left[\mathbb{1}_x \mathbb{1}_A \mid \mathcal{H}\right] = \mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \mathbb{E}\left[\mathbb{1}_A \mid \mathcal{H}\right].$$

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

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

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

for all $A \in \mathcal{G}$ and $B \in \mathcal{H}$. 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., $X = \mathbb{1}_x$ for some $x \in \mathcal{F}$. Then

$$\begin{aligned}
\mathbb{E}\left[X \cdot \mathbb{1}_{A \cap B}\right] &= \mathbb{E}\left[\mathbb{1}_x \mathbb{1}_{A \cap B}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mathbb{1}_{A \cap B} \mid \mathcal{H}\right]\right] && \text{(tower property)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mathbb{1}_A \mid \mathcal{H}\right] \mathbb{1}_B\right] && (B \in \mathcal{H}) \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \mathbb{E}\left[\mathbb{1}_A \mid \mathcal{H}\right] \mathbb{1}_B\right] && \text{(cond. indep.)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \cdot \mathbb{1}_A \mid \mathcal{H}\right] \mathbb{1}_B\right] && (\mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \ \mathcal{H}\text{-mbl.}) \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \cdot \mathbb{1}_{A \cap B} \mid \mathcal{H}\right]\right] && (B \in \mathcal{H}) \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \cdot \mathbb{1}_{A \cap B}\right] && \text{(tower property)} \\
&= \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{H}\right] \cdot \mathbb{1}_{A \cap B}\right].
\end{aligned}$$

Thus in the most basic case we have the required property. In the second case we let $X = \sum_{i=1}^{n} c_i \mathbb{1}_{x_i}$ be a finite weighted sum of indicator random variables, where $c_i \in \mathbb{R}$ and $x_i \in \mathcal{F}$. 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 $X_k = \sum_{i=0}^{n2^n - 1} \frac{i}{2^n} \mathbb{1}\{\frac{i}{2^n} \leq X < \frac{i+1}{2^n}\} + n\mathbb{1}_{n \leq X}$, 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 $\mathcal{G}$ given $\mathcal{H}$ if and only if $\mathbb{E}[\phi(X) \mid \boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})] = \mathbb{E}[\phi(X) \mid \mathcal{H}]$ for all measurable and integrable functions ϕ.*

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 $\boldsymbol{\sigma}(X)$-measurable random variable.

The "$\Leftarrow$" direction is also easy to see. Let $x \in \boldsymbol{\sigma}(X)$, $A \in \mathcal{G}$ and define ϕ such that $\phi(X) = \mathbb{1}_x$. Then we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_x \mathbb{1}_A \mid \mathcal{H}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mathbb{1}_A \mid \boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})\right] \mid \mathcal{H}\right] && \text{(tower property)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mid \boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})\right] \mathbb{1}_A \mid \mathcal{H}\right] && (A \in \mathcal{G} \subseteq \boldsymbol{\sigma}(\mathcal{G}, \mathcal{H})) \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \mathbb{1}_A \mid \mathcal{H}\right] && \text{(RHS of claim with } \phi(X) = \mathbb{1}_x) \\
&= \mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \mathbb{E}\left[\mathbb{1}_A \mid \mathcal{H}\right] && (\mathbb{E}\left[\mathbb{1}_x \mid \mathcal{H}\right] \ \mathcal{H}\text{-mbl.})
\end{aligned}$$

This completes the proof. □

Proposition A.6. *If X is conditionally independent of $\mathcal{G}$ given $\mathcal{H}$, and Y is independent of $\boldsymbol{\sigma}(X, \mathcal{G}, \mathcal{H})$ then for any measurable and integrable function f also $f(X, Y)$ is conditionally independent of $\mathcal{G}$ given $\mathcal{H}$.*

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

$$\begin{aligned}
\mathbb{E}\left[f(X,Y) \mid \boldsymbol{\sigma}(\mathcal{G},\mathcal{H})\right] &= \mathbb{E}\left[\mathbb{E}\left[f(X,Y) \mid \boldsymbol{\sigma}(X,\mathcal{G},\mathcal{H})\right] \mid \boldsymbol{\sigma}(\mathcal{G},\mathcal{H})\right] && \text{(tower property)} \\
&= \mathbb{E}\left[g(X) \mid \boldsymbol{\sigma}(\mathcal{G},\mathcal{H})\right] && \text{(Durrett (2010, Lemma 6.2.1))} \\
&= \mathbb{E}\left[g(X) \mid \mathcal{H}\right], && \text{(Proposition A.5)} \\
&= \mathbb{E}\left[f(X,Y) \mid \mathcal{H}\right], && \text{(reversing step 1 \& 2)}
\end{aligned}$$

as wanted, proving the claim. □

Proposition A.7. *If X is conditionally independent of Y given $\mathcal{H}$ then*

$$\mathbb{E}[XY \mid \mathcal{H}] = \mathbb{E}[X \mid \mathcal{H}]\,\mathbb{E}[Y \mid \mathcal{H}].$$

Proof. From Proposition A.4 we know that $\mathbb{E}[X \mid \boldsymbol{\sigma}(Y,\mathcal{H})] = \mathbb{E}[X \mid \mathcal{H}]$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[XY \mid \mathcal{H}\right] &= \mathbb{E}\left[\mathbb{E}\left[XY \mid \boldsymbol{\sigma}(Y,\mathcal{H})\right] \mid \mathcal{H}\right] && \text{(tower property)} \\
&= \mathbb{E}\left[\mathbb{E}\left[X \mid \boldsymbol{\sigma}(Y,\mathcal{H})\right] Y \mid \mathcal{H}\right] && (Y \in \boldsymbol{\sigma}(Y,\mathcal{H})) \\
&= \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{H}\right] Y \mid \mathcal{H}\right] && \text{(Proposition A.4)} \\
&= \mathbb{E}\left[X \mid \mathcal{H}\right] \mathbb{E}\left[Y \mid \mathcal{H}\right], && (\mathbb{E}\left[X \mid \mathcal{H}\right] \mathcal{H}\text{-mbl.})
\end{aligned}$$

as wanted. □

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

Proof. Follows directly from Proposition A.5. □

B 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.

B.1 Setting

As in Section 4.2 we consider only the probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ on which we define X, n, K, t_i, τ, M. As in Section 3.1, we additionally define the observation noise $(\epsilon_k)_{0\le k\le K}$, the noisy observations $O_{t_k} := X_{t_k} + \epsilon_k$ for $0 \le k \le n$, the filtration of the currently available information via

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

$\tilde{O}^{\le t}$ and the functions F_j such that $\hat{X}_{t,j} = F_j(t, \tau(t), \tilde{O}^{\le \tau(t)})$. Then we make the following assumptions.

Assumption B.1. *We assume that*

1. *$M_{0,j} = 1$ for all $1 \le j \le d_X$ (X is completely observed at 0) and $|M_k|_1 > 0$ for every $1 \le k \le K$ $\mathbb{P}$-almost surely (at every observation time at least one coordinate is observed),*
2. *the probability that any two observation times are closer than $\epsilon > 0$ converges to 0 when ϵ does, i.e., if $\delta(\omega) := \min_{0\le i\le n(\omega)} |t_{i+1}(\omega) - t_i(\omega)|$ then $\lim_{\epsilon\to 0} \mathbb{P}(\delta < \epsilon) = 0$,*
3. *almost surely X is not observed at a jump, i.e., $\mathbb{P}(t_j \in \mathcal{J} \mid j \le n) = \mathbb{P}(\Delta X_{t_j} \ne 0 \mid j \le n) = 0$ for all $1 \le j \le K$,*

4. *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 $X^\star$, i.e.,*

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

5. *$X^\star$ is L^{2p}-integrable, i.e., $\mathbb{E}[(X_T^\star)^{2p}] < \infty$,*

6. *the i.i.d. ϵ_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 $\mathbb{E}[|\epsilon_k|^2] < \infty$*

7. *n is square-intgrable, i.e., $\mathbb{E}[n^2] < \infty$,*

8. *for every $1 \leq i \leq n$, X_{t_i-} is conditionally independent of $\boldsymbol{\sigma}(n, M_{t_i})$ given $\mathcal{A}_{t_i-}$, and*

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

Remark B.2. *The relaxations on the assumption of observing X_0 completely discussed in Krach et al. (2022, Remark 2.3) can equivalently be applied in this setting here.*

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 $\tilde{X}^{\leq \tau(t)}$ and is trained with the objective functions (6) respectively (7) and their Monte Carlo approximations.

B.2 Convergence Theorem

We start with the following result which is a combination of Lemma 3.3 and Lemma 4.4.

Lemma B.3. *For any $\mathbb{A}$-adapted process Z it holds that*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |M_{t_i} \odot (O_{t_i} - Z_{t_i-})|_2^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (O_{t_i} - \hat{X}_{t_i-})\right|_2^2\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (\hat{X}_{t_i-} - Z_{t_i-})\right|_2^2\right].$$

Proof. First note that by Assumption B.1 point 3 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 B.1 point 6 & 8 together with Proposition A.6 imply that O_{t_i-} is conditionally independent of $\boldsymbol{\sigma}(n, M_i)$ given $\mathcal{A}_{t_i-}$. Hence, for $\hat{O}_{t_i-} := \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[O_{t_i-} \mid \mathcal{A}_{t_i-}]$ it follows as in Lemma 4.4 that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} |M_{t_i} \odot (O_{t_i-} - Z_{t_i-})|_2^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (O_{t_i-} - \hat{O}_{t_i-})\right|_2^2\right] + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left|M_{t_i} \odot (\hat{O}_{t_i-} - Z_{t_i-})\right|_2^2\right].$$

Then we can conclude the proof as in Lemma 3.3, by noting that

$$\hat{O}_{t_i-} = \hat{X}_{t_i-} + \mathbb{E}[\epsilon_i | \mathcal{A}_{t_i-}] = \hat{X}_{t_i-} + \mathbb{E}[\epsilon_i] = \hat{X}_{t_i-},$$

using that ϵ_i has expectation 0 and is independent of $\mathcal{A}_{t_i-}$. □

Proof of Theorem 2.4 – Part 1. We start by showing that $\hat{X} \in \mathbb{D}$ is the unique minimizer of Ψ up to indistinguishability (as defined in Definition 2.2). Note that for every t_i we have $M_{t_i} \odot \hat{X}_{t_i} = M_{t_i} \odot X_{t_i}$ and that $X_{t_i} = X_{t_i-}$ if $t_i \notin \mathcal{J}$, hence with probability 1. It follows directly from Lemma B.3 that $\hat{X}$ is a minimizer of Ψ, since

$$\Psi(Z) = \Psi(\hat{X}) + \mathbb{E}\left[\tfrac{1}{n}\sum_{i=1}^{n}\left|M_{t_i} \odot (\hat{X}_{t_i-} - Z_{t_i-})\right|_2^2\right]. \tag{15}$$

Before we can show uniqueness of $\hat{X}$, we need some additional results. For those, let $Z \in \mathbb{D}$. Let $c_1 := \mathbb{E}\left[n\right]^{1/2} \in (0, \infty)$, then the Hölder inequality, together with the fact that $n \geq 1$, yields

$$\mathbb{E}\left[|Z|_2\right] = \mathbb{E}\left[\frac{\sqrt{n}}{\sqrt{n}}\,|Z|_2\right] \leq c_1\,\mathbb{E}\left[\frac{1}{n}\,|Z|_2^2\right]^{1/2}. \tag{16}$$

By Lemma 4.6 and by the equivalence of 1- and 2-norm we have for some constant $c_3 > 0$ and for any $1 \leq k \leq K$

$$\mathbb{E}\left[\mathbb{1}_{\{n\geq k\}}\left|\hat{X}_{t_k-} - Z_{t_k-}\right|_2\right] \leq \frac{c_3}{c_2(k)}\mathbb{E}\left[\mathbb{1}_{\{n\geq k\}}\left|M_{t_k} \odot (\hat{X}_{t_k-} - Z_{t_k-})\right|_2\right]. \tag{17}$$

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 $\hat{X}$. Hence, there exists some $1 \leq k \leq K$ such that $d_k(\hat{X}, Z) > 0$. We have

$$\begin{aligned}
\mathbb{E}\left[\tfrac{1}{n}\sum_{i=1}^{n}\left|M_{t_i} \odot (\hat{X}_{t_i-} - Z_{t_i-})\right|_2^2\right] &= \mathbb{E}\left[\tfrac{1}{n}\sum_{i=1}^{K}\mathbb{1}_{\{n\geq i\}}\left|M_{t_i} \odot (\hat{X}_{t_i-} - Z_{t_i-})\right|_2^2\right] \\
&\geq \mathbb{E}\left[\tfrac{1}{n}\mathbb{1}_{\{n\geq k\}}\left|M_{t_k} \odot (\hat{X}_{t_k-} - Z_{t_k-})\right|_2^2\right] \\
&\geq c_1^{-2}\mathbb{E}\left[\mathbb{1}_{\{n\geq k\}}\left|M_{t_k} \odot (\hat{X}_{t_k-} - Z_{t_k-})\right|_2\right]^2 \\
&\geq \left(\frac{c_2}{c_1 c_3}\right)^2 \mathbb{E}\left[\mathbb{1}_{\{n\geq k\}}\left|\hat{X}_{t_k-} - Z_{t_k-}\right|_2\right]^2 \\
&= \left(\frac{c_2}{c_0 c_1 c_3}\right)^2 d_k(\hat{X}, Z)^2 > 0,
\end{aligned} \tag{18}$$

where we used (16) for the 3rd, (17) for the 4th and (1) for the last line. Together with (15) 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 $\tilde{\theta}_3^\star$ such that $\tilde{g}_{\theta_3^\star} = \mathrm{id}$, which is possible since we assumed that $\mathrm{id} \in \tilde{\mathcal{N}}$. Let $\varepsilon > 0$, $N_\varepsilon := \lceil 2(T+1)\varepsilon^{-4}\rceil$ (implying that $\lim_{\varepsilon\to 0} N_\varepsilon = \infty$) and $\mathcal{P}_\varepsilon$ be the closure of the set A_{N_ε} 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 B.1 and can (by abuse of notation) equivalently be written as (continuous) function $f_j(t, \tau(t), \tilde{O}^{\leq\tau(t)} - O_0, O_0)$. 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 Π_ε, 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 $\tilde{\theta}_1^{\star,m_1} \in \tilde{\Theta}_{m_1}^1$ 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 $\tilde{f}_{\tilde{\theta}_1^{\star,m_1}} \in \tilde{\mathcal{N}}$ (denoted by $\tilde{f}_{\tilde{\theta}_1^{\star,m_1},j}$) on the compact set $[0,T]^2 \times \Pi_\varepsilon$. Hence, combining the two

approximations we get (by triangle inequality)

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

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 $\tilde{\theta}_2^{\star,m_2}\in\tilde{\Theta}_{m_2}^2$ such that for every $1\leq j\leq d_X$

$$\sup_{(t,X)\in[0,T]\times\mathcal{P}_\varepsilon}\left|F_j(t,t,X)-\tilde{\rho}_{\tilde{\theta}_2^{\star,m_2},j}(t,\pi_{m_1}(X-X_0),X_0)\right|\leq\varepsilon.$$

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

$$\gamma_1:=\max_{(t,\tau,X)\in[0,T]^2\times\mathcal{P}_\varepsilon}\left|\tilde{f}_{\tilde{\theta}_1^{\star,m_1}}(t,\tau,\pi_{m_0}(X-X_0),X_0)\right|$$

and γ_2 equivalently for $\tilde{\rho}_{\tilde{\theta}_2^{\star,m_2}}$. Since the neural networks are continuous functions they take a finite maximum on the compact sets, hence γ_1,γ_2 are finite. Then we define the bounded output neural networks $f_{\theta_1^{\star,m_1}},\rho_{\theta_2^{\star,m_2}}\in\mathcal{N}$ with $\theta_i^{\star,m_i}:=(\tilde{\theta}_i^{\star,m_i},\gamma_i)$. Clearly, these bounded output neural networks coincide with the neural networks on the compact sets. Therefore, they satisfy the same ε-approximation and since F_j,f_j are bounded by $U:=B\left((X_T^\star+1)^p+\sum_{i=0}^n|\epsilon_i|\right)$ (Assumption B.1 item 4), it follows that $f_{\theta_1^{\star,m_1}},\rho_{\theta_2^{\star,m_2}}$ are bounded by $U+\varepsilon$. In particular, we have for $\epsilon<B$ the global bounds $|f_j-f_{\theta_1^{\star,m_1},j}|_\infty\leq 3U$ and $|F_j-\rho_{\theta_2^{\star,m_2},j}|_\infty\leq 3U$. Setting $m:=\max(m_0,m_1,m_2,\gamma_1,\gamma_2,|\tilde{\theta}_i^{\star,m_2}|_2,|\tilde{\theta}_2^{\star,m_2}|_2)$, it follows that $\theta_m^\star:=(\theta_1^{\star,m_1},\theta_2^{\star,m_2},\tilde{\theta}_3^\star)\in\Theta_m$.

Now we can bound the distance between $Y_t^{\theta_m^\star}$ and $\hat{X}$. Whenever $X_T^\star<1/\varepsilon$, the number of observations satisfies $n<1/\varepsilon$, the minimal difference between any two consecutive observation times $\delta>\varepsilon$ and all noise terms satisfy $|\epsilon_i|_2<1/\varepsilon$ we know that the corresponding path $\tilde{X}^{\leq\tau(t)}-X_0$ is an element of A_{N_ε} and therefore the neural network approximations up to ε hold. Otherwise, one of those conditions is not satisfied and the global upper bound can be used. Hence, if $t\in\{t_1,\dots,t_n\}$, we have for $F=(F_j)_{1\leq j\leq d_X}$ and $f=(f_j)_{1\leq j\leq d_X}$

$$\begin{aligned}\left|Y_t^{\theta_m^*}-\hat{X}_t\right|_1&=\left|\rho_{\theta_2^{\star,m_2}}\left(H_{t-},t,\pi_m(\tilde{O}^{\leq t}-O_0),O_0\right)-F\left(t,t,\tilde{O}^{\leq t}\right)\right|_1\\&\leq\varepsilon d_X\mathbb{1}_{\{X_T^\star<1/\varepsilon\}}\mathbb{1}_{\{n<1/\varepsilon\}}\mathbb{1}_{\{\delta>\epsilon\}}\mathbb{1}_{\{\forall 0\leq i\leq n:|\epsilon_i|_2<1/\varepsilon\}}\\&\quad+d_X3U\left(\mathbb{1}_{\{X_T^\star\geq 1/\varepsilon\}}+\mathbb{1}_{\{n\geq 1/\varepsilon\}}+\mathbb{1}_{\{\delta\leq\epsilon\}}+\mathbb{1}_{\{\exists 0\leq i\leq n:|\epsilon_i|_2\geq 1/\varepsilon\}}\right),\end{aligned}$$

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

$$\begin{aligned}\left|Y_t^{\theta_m^*}-\hat{X}_t\right|_1&\leq\left|Y_{\tau(t)}^{\theta_m^*}-\hat{X}_{\tau(t)}\right|_1\\&\quad+\int_{\tau(t)}^t\left|f_{\theta_1^{\star,m_1}}\left(H_{s-},s,\tau(t),\pi_m(\tilde{O}^{\leq\tau(t)}-O_0),O_0\right)-f(s,\tau(t),\tilde{O}^{\leq\tau(t)})\right|_1ds\\&\leq\varepsilon(T+1)d_X\mathbb{1}_{\{X_T^\star<1/\varepsilon\}}\mathbb{1}_{\{n<1/\varepsilon\}}\mathbb{1}_{\{\delta>\varepsilon\}}\mathbb{1}_{\{\forall 0\leq i\leq n:|\epsilon_i|_2<1/\varepsilon\}}\\&\quad+(T+1)d_X3U\left(\mathbb{1}_{\{X_T^\star\geq 1/\varepsilon\}}+\mathbb{1}_{\{n\geq 1/\varepsilon\}}+\mathbb{1}_{\{\delta\leq\varepsilon\}}+\mathbb{1}_{\{\exists 0\leq i\leq n:|\epsilon_i|_2\geq 1/\varepsilon\}}\right).\end{aligned}$$

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

$$\begin{aligned}\left|Y_t^{\theta_m^*}-\hat{X}_t\right|_2&\leq c\,\varepsilon(T+1)d_X\\&\quad+c(T+1)d_X3U\left(\mathbb{1}_{\{X_T^\star\geq 1/\varepsilon\}}+\mathbb{1}_{\{n\geq 1/\varepsilon\}}+\mathbb{1}_{\{\delta\leq\varepsilon\}}+\mathbb{1}_{\{\exists 0\leq i\leq n:|\epsilon_i|_2\geq 1/\varepsilon\}}\right)\\&=:c_m.\end{aligned}$$

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 ε-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 ε_m when using an architecture with signature truncation level $m \in \mathbb{N}$ and weights in Θ_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 $\varepsilon_{m_1} \geq \varepsilon_{m_2}$ for any $m_1 \leq m_2$. In particular $(\varepsilon_m)_{m\geq 0}$ is a a decreasing sequence, hence, our derivations before proof that $\lim_{m\to\infty} \varepsilon_m = 0$. In the following we denote by $\theta_m^\star \in \Theta_m$ the best choice for the weights within the set Θ_m to approximate the functions F_j, f_j.

Since $\theta_m^{\min} \in \operatorname{argmin}_{\theta\in\Theta_m}\{\Phi(\theta)\}$ (note that at least one minimum exists in the compact set Θ_m since Φ is continuous), we get with Lemma B.3 that

$$\begin{aligned}\min_{Z\in\mathbb{D}} \Psi(Z) \leq \Phi(\theta_m^{\min}) \leq \Phi(\theta_m^*) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left|M_{t_i} \odot (O_{t_i} - Y_{t_i}^{\theta_m^*})\right|_2^2\right] \\ &= \Psi(\hat{X}) + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left|M_{t_i} \odot (\hat{X}_{t_i-} - Y_{t_i-}^{\theta_m^*})\right|_2^2\right] \leq \Psi(\hat{X}) + \mathbb{E}\left[c_m^2\right].\end{aligned} \tag{19}$$

Integrability of $|X_T^\star|_2$, $|n|$ and $|\epsilon_i|$ together with Assumption B.1 item 2 on δ imply that

$$\mathbb{1}_{\{X_T^\star \geq 1/\varepsilon_m\}} + \mathbb{1}_{\{n\geq 1/\varepsilon_m\}} + \mathbb{1}_{\{\delta\leq\varepsilon_m\}} + \mathbb{1}_{\{\exists 0\leq i\leq n: |\epsilon_i|_2 \geq 1/\varepsilon_m\}} \xrightarrow[m\to\infty]{\mathbb{P}-a.s.} 0.$$

Therefore, we have for a suitable constant $c > 0$ (not depending on ε_m and m),

$$\begin{aligned}\mathbb{E}\left[c_m^2\right] \leq c\varepsilon_m^2 + c\mathbb{E}\Big[U^2\Big(&\mathbb{1}_{\{X_T^\star\geq 1/\varepsilon_m\}} + \mathbb{1}_{\{n\geq 1/\varepsilon_m\}} \\ &+ \mathbb{1}_{\{\delta\leq\varepsilon_m\}} + \mathbb{1}_{\{\exists 0\leq i\leq n: |\epsilon_i|_2\geq 1/\varepsilon_m\}}\Big)\Big] \xrightarrow{m\to\infty} 0,\end{aligned}$$

by dominated convergence, since U is L^2-integrable. Indeed,

$$\begin{aligned}\mathbb{E}[U^2] \leq 8B^2\mathbb{E}\left[(X_t^\star+1)^{2p} + n\sum_{i=0}^n |\epsilon_i|^2\right] \\ = 8B^2\left(\mathbb{E}\left[(X_t^\star+1)^{2p}\right] + \mathbb{E}[n^2]\mathbb{E}\left[|\epsilon_0|^2\right]\right) < \infty,\end{aligned} \tag{20}$$

using Cauchy-Schwarz inequality for the first step, that the ϵ_i are i.i.d. and independent of n for the equality and the integrability of $X^\star$, ϵ_0 and n^2 for the upper bound. Using this and $\Psi(\hat{X}) = \min_{Z\in\mathbb{D}} \Psi(Z)$, we get from (19)

$$\min_{Z\in\mathbb{D}} \Psi(Z) \leq \Phi(\theta_m^{\min}) \leq \Phi(\theta_m^*) \xrightarrow{m\to\infty} \min_{Z\in\mathbb{D}} \Psi(Z).$$

Finally, we show that $\lim_{m\to\infty} d_k\left(\hat{X}, Y^{\theta_m^{\min}}\right) = 0$ for all $1 \leq k \leq K$. Applying (16), (17) and (1) in reverse order than it was done in (18) and finally Lemma B.3, yields

$$\begin{aligned}d_k\left(\hat{X}, Y^{\theta_m^*}\right) &\leq \frac{c_0\, c_1\, c_3}{c_2}\, \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left|M_{t_i} \odot (\hat{X}_{t_i-} - Y_{t_i-}^{\theta_m^*})\right|_2^2\right]^{1/2} \\ &= \frac{c_0\, c_1\, c_3}{c_2}\left(\Phi(\theta_m^*) - \Psi(\hat{X})\right)^{1/2} \xrightarrow{m\to\infty} 0,\end{aligned} \tag{21}$$

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 samples 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 $\mathcal{S} := \{x = (x_i)_{i\in\mathbb{N}} \in \ell^1(\mathbb{R}^d) \mid \|x\|_{\ell^1} < \infty\}$ for a suitable d (see below) with the norm $\|x\|_{\ell^1} := \sum_{i\in\mathbb{N}} |x_i|_2$, the function

$$F(x, y, m) := |m \odot (x - y)|_2$$

and $\xi_j := (\xi_{j,0}, \dots, \xi_{j,n^{(j)}}, 0, \dots)$, where $\xi_{j,k} := (t_k^{(j)}, O^{(j)}_{t_k^{(j)}}, M^{(j)}_{t_k^{(j)}}, \pi_m(\tilde{O}^{\leq t_k^{(j)},(j)})) \in \mathbb{R}^d$ and $t_k^{(j)}$, $M^{(j)}_{t_k^{(j)}}$ and $O^{(j)}_{t_i^{(j)}}$ (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 $n^j(\xi_j) := \max_{k\in\mathbb{N}}\{\xi_{j,k} \neq 0\}$, $t_k(\xi_j) := t_k^{(j)}$, $O_k(\xi_j) := O^{(j)}_{t_k^{(j)}}$ and $M_k(\xi_j) := M^{(j)}_{t_k^{(j)}}$. By this definition we have $n^{(j)} = n^j(\xi_j)$ $\mathbb{P}$-almost-surely. Moreover, we have that ξ_j are i.i.d. random variables taking values in $\mathcal{S}$. Let us write $Y_t^\theta(\xi)$ to make the dependence of Y on the input and the weight θ explicit. Then we define

$$h(\theta, \xi_j) := \frac{1}{n^j(\xi_j)} \sum_{i=1}^{n^j(\xi_j)} F\left(O_i(\xi_j), Y^\theta_{t_i(\xi_j)-}(\xi_j), M_i(\xi_j)\right)^2.$$

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

Lemma B.4. *Almost-surely the random function $\theta \in \Theta_M \mapsto Y_t^\theta$ is uniformly continuous for every $t \in [0, T]$.*

Now we are ready to prove the second part of our main theorem.

Proof of Theorem 2.4 – Part 2. First we note that, Y_t^θ 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 $|Y_t^\theta(\xi_j)| \leq \tilde{B}\left((X_T^{\star,(j)} + 1)^p + \sum_{i=0}^n |\epsilon_i^{(j)}|\right)$ for all $t \in [0, T]$ and $\theta \in \tilde{\Theta}_m$ for some constant $\tilde{B}$ (possibly depending on m), where $X^{\star,(j)}, \epsilon_i^{(j)}$ corresponds to the input ξ_j. Hence,

$$\begin{aligned} F\left(O_i(\xi_j), Y^\theta_{t_i(\xi_j)-}(\xi_j), M_i(\xi_j)\right)^2 &= \left|M_i(\xi_j) \odot (O_i(\xi_j) - Y^\theta_{t_i(\xi_j)-}(\xi_j))\right|_2^2 \\ &\leq \left((B + \tilde{B})\left((X_T^\star + 1)^p + \sum_{i=0}^n |\epsilon_i|\right)\right)^2 = \left(\frac{B + \tilde{B}}{B} U^{(j)}\right)^2, \end{aligned}$$

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

$$\mathbb{E}\left[\sup_{\theta\in\tilde{\Theta}_M} h(\theta, \xi_j)\right] \leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left(\frac{B + \tilde{B}}{B} U^{(j)}\right)^2\right] < \infty, \tag{22}$$

by Assumption 2.1 and (20). By Lemma B.4, the function $\theta \mapsto h(\theta, \xi_1)$ is continuous, hence, we can apply Herrera et al. (2021, Lemma E.15), yielding that almost-surely for $N \to \infty$ the function

$$\theta \mapsto \frac{1}{N}\sum_{j=1}^N h(\theta, \xi_j) = \hat{\Phi}_N(\theta) \tag{23}$$

converges uniformly on Θ_m to

$$\theta \mapsto \mathbb{E}[h(\theta, \xi_1)] = \Phi(\theta). \tag{24}$$

We deduce from Herrera et al. (2021, Lemma E.14) that $d(\theta^{\min}_{m,N}, \Theta^{\min}_m) \to 0$ a.s. when $N \to \infty$. Then there exists a sequence $(\hat{\theta}^{\min}_{m,N})_{N\in\mathbb{N}}$ in $\Theta^{\min}_m$ such that $|\theta^{\min}_{m,N} - \hat{\theta}^{\min}_{m,N}|_2 \to 0$ a.s. for $N \to \infty$. The uniform continuity of the random functions $\theta \mapsto Y_t^\theta$ on Θ_m implies that

$$|Y_t^{\theta^{\min}_{m,N}}(\xi_1) - Y_t^{\hat{\theta}^{\min}_{m,N}}(\xi_1)|_2 \to 0 \text{ a.s. for all } t \in [0, T] \text{ as } N \to \infty.$$

By continuity of F this yields $|h(\theta^{\min}_{m,N}, \xi_1) - h(\hat{\theta}^{\min}_{m,N}, \xi_1)| \to 0$ a.s. as $N \to \infty$. With (22) we can apply dominated convergence which yields

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

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

$$\lim_{N\to\infty} \Phi(\theta^{\min}_{m,N}) = \lim_{N\to\infty} \mathbb{E}\left[h(\theta^{\min}_{m,N}, \xi_1)\right] = \lim_{N\to\infty} \mathbb{E}\left[h(\hat{\theta}^{\min}_{m,N}, \xi_1)\right] = \Phi(\theta^{\min}_m). \tag{25}$$

Now by triangle inequality,

$$|\hat{\Phi}_N(\theta^{\min}_{m,N}) - \Phi(\theta^{\min}_m)| \leq |\hat{\Phi}_N(\theta^{\min}_{m,N}) - \Phi(\theta^{\min}_{m,N})| + |\Phi(\theta^{\min}_{m,N}) - \Phi(\theta^{\min}_m)|. \tag{26}$$

(23), (24) and (25) imply that both terms on the right hand side converge to 0 when $N \to \infty$, which 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 $m \in \mathbb{N}$

$$N_m := \min\left\{N \in \mathbb{N} \mid N > N_{m-1}, |\Phi(\theta^{\min}_{m,N}) - \Phi(\theta^{\min}_m)| \leq \tfrac{1}{m}\right\},$$

which is possible due to (25). Then Theorem 2.4 implies that

$$|\Phi(\theta^{\min}_{m,N_m}) - \Psi(\hat{X})| \leq \tfrac{1}{m} + |\Phi(\theta^{\min}_m) - \Psi(\hat{X})| \xrightarrow{m\to\infty} 0.$$

Therefore, we can apply the same arguments as in the first part of the proof (cf. (21)) to show that

$$d_k\left(\hat{X}, Y^{\theta^{\min}_{m,N_m}}\right) \leq \frac{c_0\, c_1\, c_3}{c_2} \left(\Phi(\theta^{\min}_{m,N_m}) - \Psi(\hat{X})\right)^{1/2} \xrightarrow{m\to\infty} 0,$$

for every $1 \leq k \leq K$. □

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

Corollary B.5. *In the setting of Theorem 2.4, we also have that $\mathbb{P}$-a.s.*

$$\Phi(\theta^{\min}_{m,N_m}) \xrightarrow{m\to\infty} \Psi(\hat{X}) \quad \textit{and} \quad \hat{\Phi}_{\tilde{N}_m}(\theta^{\min}_{m,\tilde{N}_m}) \xrightarrow{m\to\infty} \Psi(\hat{X}),$$

where $(\tilde{N}_m)_{m\in\mathbb{N}}$ is another increasing sequence in $\mathbb{N}$.

C 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 A for any details that are not provided here.

C.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 A.1.1) also apply here. In particular, we use standard neural networks for f_{θ_1} and ρ_{θ_2} 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 u-coordinates of $\tilde{X}$ in the implementation, since we anyways use the observation times and masks as additional inputs for ρ_{θ_2}.

C.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.

Training. 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).

C.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 $\mathbb{P}(M_i^X = 1 \,|\, \mathcal{A}_{t_i-}) = \mathbb{E}[M_i^X \,|\, \mathcal{A}_{t_i-}]$, where M_i^X is described in Example 4.9, using $\eta = 3$ and $p = 0.1$. We use the same discretisation grid and dataset sizes as in Section C.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 C.2 but only trained with original loss function (3).