Learning Partially Known Stochastic Dynamics with Empirical PAC Bayes —APPENDIX—

1 CONTINUOUS TIME SDES

Solving the SDE system in (1) for a time interval $[0, T]$ and fixed θ_f requires computing integrals of the form

$$\int_0^T d\mathbf{h}_t = \int_0^T f_{\theta_f}(\mathbf{h}_t, t)dt + \int_0^T G(\mathbf{h}_t, t)dW_t.$$

This operation is intractable for almost any practical choice of $f_{\theta_f}(\cdot, \cdot)$ and $G(\cdot, \cdot)$ for two reasons. First, the integral around the drift term $f_{\theta_f}(\cdot, \cdot)$ does not have an analytical solution, due both to potential nonlinearities of the drift and to the fact that $\mathbf{h}_t \sim p(\mathbf{h}_t, t)$ is a stochastic variable following an implicitly defined distribution. Second, the diffusion term involves the Itô integral (Oksendal, 1992) about W_t which multiplies the non-linear function $G(\cdot, \cdot)$.

For each of the SDEs in (6) and (7), we could alternatively to the Euler-Maruyama integration theme use the Fokker-Planck-Kolmogorov equation to derive a partial differential equation (PDE) system

$$\begin{aligned}\partial p_{\text{hyb}}(\mathbf{h}_t, t|\theta_f)/\partial t &= -\nabla \cdot \big[\big(f_{\theta_f}(\mathbf{h}_t, t) + \boldsymbol{\gamma} \circ r_\xi(\mathbf{h}_t, t)\big)p_{\text{hyb}}(\mathbf{h}_t, t|\theta_f)\big] + \nabla \cdot \big(\mathbf{1}\nabla \cdot G(\mathbf{h}_t, t)p_{\text{hyb}}(\mathbf{h}_t, t|\theta_f)\big),\\ \partial p_{\text{pri}}(\mathbf{h}_t, t)/\partial t &= -\nabla \cdot \big[\big(\boldsymbol{\gamma} \circ r_\xi(\mathbf{h}_t, t)\big)p_{\text{pri}}(\mathbf{h}_t, t)\big] + \nabla \cdot \big(\mathbf{1}\nabla \cdot G(\mathbf{h}_t, t)p_{\text{pri}}(\mathbf{h}_t, t)\big),\end{aligned}$$

where $\nabla\cdot$ is the divergence operator and $\mathbf{1} = (1, \ldots, 1)^\top$. Theoretically, these distributions can be obtained by solving the Fokker-Planck PDE. As this requires solving a PDE which is not analytically tractable, we instead resort to the discrete time Euler-Maruyama integration.

2 PROOFS

This section gives a more detailed derivation of the individual results stated in the main paper.

Lemma 1. *For the process distributions*[1] $Q_{0\to T}$ *and* $P_{0\to T}$ *the following property holds*

$$D_{KL}\big(Q_{0\to T}||P_{0\to T}\big) = \frac{1}{2}\int_0^T \mathbb{E}_{Q_{0\to T}}\Big[f_{\theta_f}(\mathbf{h}_t, t)^\top \mathbf{J}_t^{-1} f_{\theta_f}(\mathbf{h}_t, t)\Big]dt + D_{KL}\big(p_\phi(\theta_f)||p_{\text{pri}}(\theta_f)\big)$$

for some $T > 0$, *where* $\mathbf{J}_t = G(\mathbf{h}_t, t)G(\mathbf{h}_t, t)^\top$.

[1] See the main paper for their definitions.

Proof. Assume Euler-Maruyama discretization for the process $Q_{0\to T}$ on arbitrarily chosen K time points within the interval $[0, T]$. Then we have $D_{KL}(Q||P)$ denoting the Kullback-Leibler divergence between processes $Q_{0\to T}$ and $P_{0\to T}$ up to discretization into T time points as:

$$\begin{aligned}
D_{KL}(Q||P) &= \iint \log \frac{\prod_{t=0}^{K-1}\Big(\mathcal{N}\big(\mathbf{h}_{t+1}|\left(f_{\theta_f}(\mathbf{h}_t,t)+\boldsymbol{\gamma}\circ r_\xi(\mathbf{h}_t,t)\right)\Delta t, \mathbf{J}_t\Delta t\big)\Big)}{\prod_{t=0}^{K-1}\Big(\mathcal{N}\big(\mathbf{h}_{t+1}|\boldsymbol{\gamma}\circ r_\xi(\mathbf{h}_t,t)\Delta t, \mathbf{J}_t\Delta t\big)\Big)} \cdot \frac{\cancel{p(\mathbf{h}_0)}p_\phi(\theta_f)}{\cancel{p(\mathbf{h}_0)}p_{\text{pri}}(\theta_f)} Q_{0\to T} d\mathbf{H} d\theta_f \\
&= \sum_{t=0}^{K-1}\iint \log \mathcal{N}\big(\mathbf{h}_{t+1}|\left(f_{\theta_f}(\mathbf{h}_t,t)+\boldsymbol{\gamma}\circ r_\xi(\mathbf{h}_t,t)\right)\Delta t, \mathbf{J}_t\Delta t\big) \\
&\qquad - \log \mathcal{N}\big(\mathbf{h}_{t+1}|\boldsymbol{\gamma}\circ r_\xi(\mathbf{h}_t,t)\Delta t, \mathbf{J}_t\Delta t\big) Q_{0\to T} d\mathbf{H} d\theta_f \\
&\quad + D_{KL}\big(p_\phi(\theta_f)||p_{\text{pri}}(\theta_f)\big).
\end{aligned}$$

For simplicity, let us modify notation and adopt $\mathbf{f}_t := f_{\theta_f}(\mathbf{h}_t,t)+\boldsymbol{\gamma}\circ r_\xi(\mathbf{h}_t,t)$, $\mathbf{g}_t := \boldsymbol{\gamma}\circ r_\xi(\mathbf{h}_t,t)$, and $\Delta\mathbf{h}_{t+1} := \mathbf{h}_{t+1}-\mathbf{h}_t$. Now writing down the $\log(\cdot)$ terms explicitly, we get

$$\begin{aligned}
D_{KL}(Q||P) = \frac{1}{2}\sum_{t=0}^{K-1}\iiint \Big[&- (\Delta\mathbf{h}_{t+1}-\mathbf{f}_t\Delta t)^\top(\mathbf{J}_t\Delta t)^{-1}(\Delta\mathbf{h}_{t+1}-\mathbf{f}_t\Delta t) \\
&+ (\Delta\mathbf{h}_{t+1}-\mathbf{g}_t\Delta t)^\top(\mathbf{J}_t\Delta t)^{-1}(\Delta\mathbf{h}_{t+1}-\mathbf{g}_t\Delta t)\Big] \\
&\cdot p_{\text{hyb}}(\mathbf{h}_{0\to T}|\theta_f)p_\phi(\theta_f)d\mathbf{H}d\theta_f \\
&+ D_{KL}\big(p_\phi(\theta_f)||p_{\text{pri}}(\theta_f)\big).
\end{aligned}$$

Expanding the products, removing the terms that cancel out, and rearranging the rest, we get

$$\begin{aligned}
D_{KL}(Q||P) = \frac{1}{2}\sum_{t=0}^{K-1}\iint \Big[&- \mathbf{f}_t^\top\mathbf{J}_t^{-1}\mathbf{f}_t\Delta t + 2\Delta\mathbf{h}_{t+1}\mathbf{J}_t^{-1}\mathbf{f}_t + \mathbf{g}_t^\top\mathbf{J}_t^{-1}\mathbf{g}_t\Delta t - 2\Delta\mathbf{h}_{t+1}\mathbf{J}_t^{-1}\mathbf{g}_t\Big] \\
&\cdot p_{\text{hyb}}(\mathbf{h}_{0\to T}|\theta_f,)p_\phi(\theta_f)d\mathbf{H}d\theta_f \\
&+ D_{KL}\big(p_\phi(\theta_f)||p(\theta_f)\big).
\end{aligned}$$

Note that from the definition of the process it follows that

$$\int \Delta\mathbf{h}_{t+1}\ \ p_{\text{hyb}}(\mathbf{h}_{0\to T}|\theta_f)d\Delta\mathbf{h}_{t+1} = \mathbf{f}_t\Delta t.$$

Plugging this fact into the KL term, we have

$$D_{KL}(Q||P) = \frac{1}{2}\sum_{t=0}^{K-1}\int \Big[\mathbf{f}_t^\top\mathbf{J}_t^{-1}\mathbf{f}_t\Delta t + \mathbf{g}_t^\top\mathbf{J}_t^{-1}\mathbf{g}_t\Delta t - 2\mathbf{f}_t\mathbf{J}_t^{-1}\mathbf{g}_t\Delta t\Big]p_{\phi_f}(\theta_f))d\theta_f + D_{KL}\big(p_\phi(\theta_f)||p_{\text{pri}}(\theta_f)\big).$$

For any pair of vectors $\mathbf{a}, \mathbf{b}\in\mathbb{R}^P$ and symmetric matrix $\mathbf{C}\in\mathbb{R}^{P\times P}$, the following identity holds:

$$\mathbf{a}^\top\mathbf{C}\mathbf{a} - \mathbf{b}^\top\mathbf{C}\mathbf{b} = (\mathbf{a}-\mathbf{b})^\top\mathbf{C}(\mathbf{a}-\mathbf{b}) + 2\mathbf{a}^\top\mathbf{C}\mathbf{b}.$$

Applying this identity to the above, we attain

$$D_{KL}(Q||P) = \frac{1}{2}\sum_{t=0}^{K-1}\int \Big[(\mathbf{f}_t-\mathbf{g}_t)^\top\mathbf{J}_t^{-1}(\mathbf{f}_t-\mathbf{g}_t)\Delta t\Big]p_{\phi_f}(\theta_f)d\theta_f + D_{KL}\big(q(\theta_f)||p(\theta_f)\big).$$

Plugging back the original terms and setting K to the limit, we arrive at the desired outcome

$$\begin{aligned}
&\lim_{K\to+\infty}\left\{\frac{1}{2}\sum_{t=0}^{K-1}\int \Big[(f_{\theta_f}(\mathbf{h}_t,t))^\top\mathbf{J}_t^{-1}f_{\theta_f}(\mathbf{h}_t,t)\Delta t\Big]p_{\phi_f}(\theta_f)d\theta_f + D_{KL}\big(p_\phi(\theta_f)||p_{\text{pri}}(\theta_f)\big)\right\} \\
&= \frac{1}{2}\int\Big[\int f_{\theta_f}(\mathbf{h}_t,t)^\top\mathbf{J}_t^{-1}f_{\theta_f}(\mathbf{h}_t,t)p_{\phi_f}(\theta_f)d\theta_f\Big]dt + D_{KL}\big(p_\phi(\theta_f)||p_{\text{pri}}(\theta_f)\big) \\
&= \frac{1}{2}\int_0^T \mathbb{E}_{Q_{0\to T}}\Big[f_{\theta_f}(\mathbf{h}_t,t)^\top\mathbf{J}_t^{-1}f_{\theta_f}(\mathbf{h}_t,t)\Big]dt + D_{KL}\big(p_\phi(\theta_f)||p_{\text{pri}}(\theta_f)\big).
\end{aligned}$$

□

Theorem 1. *Let $p(\mathbf{y}_t|\mathbf{h}_t)$ be uniformly bounded likelihood function with density $p(\mathbf{y}_t|\mathbf{h}_t)$ everywhere and $Q_{0\to T}$ and $P_{0\to T}$ be the joints stochastic processes defined on the hypothesis class of the learning task, respectively. Define the true risk of a draw from $Q_{0\to T}$ on an i.i.d. sample $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_K\}$ at discrete and potentially irregular time points $t_1, \ldots, t_K$ drawn from an unknown ground-truth stochastic process $\mathfrak{G}(t)$ as the expected model misfit as on the sample as defined via the following risk over hypotheses $H = (\mathbf{h}_{0\to T}, \theta_f)$*

$$R(H) = 1 - \mathbb{E}_{\mathbf{Y}\sim\mathfrak{G}(t)}\left[\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{h}_k)/\overline{B}\right], \tag{1}$$

for time horizon $T > 0$ and the corresponding empirical risk on a data set $\mathcal{D} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_N\}$ as

$$R_{\mathcal{D}}(H) = 1 - \frac{1}{N}\sum_{n=1}^{N}\left[\prod_{k=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k)/\overline{B}\right]. \tag{2}$$

Then the expected true risk is bounded above by the marginal negative log-likelihood of the predictor and a complexity functional as

$$\mathbb{E}_{H\sim Q_{0\to T}}\left[R(H)\right] \leq \mathbb{E}_{H\sim Q_{0\to T}}\left[R_{\mathcal{D}}(H)\right] + \mathcal{C}_{\delta}(Q_{0\to T}, P_{0\to T}), \tag{3}$$

$$\leq -\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{S}\sum_{s=1}^{S}\prod_{k=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k^{n,s})\right) + \mathcal{C}_{\delta/2}(Q_{0\to T}, P_{0\to T}) + \sqrt{\frac{\log(2N/\delta)}{2S}} + K\log\overline{B} \tag{4}$$

$$\leq -\frac{1}{SN}\sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K}\log\left(p(\mathbf{y}_k^n|\mathbf{h}_k^{s,n})\right) + \mathcal{C}_{\delta/2}(Q_{0\to T}, P_{0\to T}) + \sqrt{\frac{\log(2N/\delta)}{2S}} + K\log\overline{B}, \tag{5}$$

where $\overline{B} := \max_{\mathbf{y}_k,\mathbf{h}_k} p(\mathbf{y}_k|\mathbf{h}_k)$ is the uniform bound, S is the sample count taken independently for each observed sequence, and the complexity functional is given as

$$\mathcal{C}_{\delta}(Q_{0\to T}, P_{0\to T}) := \sqrt{\frac{D_{KL}\big(Q_{0\to T}||P_{0\to T}\big) + \log(2\sqrt{N}) - \log(\delta/2)}{2N}}$$

with $D_{KL}\big(Q_{0\to T}||P_{0\to T}\big)$ as in Lemma 1 for some $\delta > 0$.

Proof. To be able to apply known PAC bounds, we first define the hypothesis class $H \in \mathcal{H}_K$ that contain latent states $\mathbf{h}_k, \theta_f$ that explain the observations $\mathbf{y}_k$. Then, we define the true risk as

$$R(H) = \mathbb{E}_{\mathbf{Y}_k\sim\mathfrak{G}(t)}\left[1 - \frac{1}{\overline{B}_K}\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{h}_k)\right]$$

and the empirical risk as

$$R_{\mathcal{D}}(H) = \frac{1}{N}\sum_{n=1}^{N}\left\{1 - \frac{1}{\overline{B}_K}\prod_{kn=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k^n)\right\},$$

where we defined

$$\overline{B}_K := \max_{\mathbf{y},\mathbf{h}_k}\prod_{k=1}^{K} p\left(\mathbf{y}_k|\mathbf{h}_k\right) \leq \left(\max_{\mathbf{y},\mathbf{h}_k} p\big(\mathbf{y}_k|\mathbf{h}_k\big)\right)^K.$$

The data set $\mathcal{D} = \{\mathbf{Y}_k^n\}_{k,n}$ was generated by an unknown stochastic process $\mathfrak{G}(t)$. Note that we normalize the risks $R(H)$ and $R_D(H)$ by the maximum of the likelihood and thereby obtaining a possible range of these risk of $[0, 1]$. The likelihood can be bounded, as the term $p(\mathbf{y}_k|\mathbf{h}_k)$ can be bounded from above, as we model this by a Gaussian.Therefore, it is bounded, if we assume a minimal allowed variance.

To obtain a tractable bound, it is common practice is to upper bound its analytically intractable inverse (Germain et al., 2016) using Pinsker's inequality (Catoni, 2007; Dziugaite and Roy, 2017). Indeed, by applying Pinsker's inquality to the PAC-Theorem from Maurer (2004), we obtain the following theorem.

PAC-theorem For any $[0,1]$-valued loss function giving rise to empirical and true risk $R_{\mathcal{D}}(H), R(H)$, for any distribution Δ, for any $N \in \mathbb{N}$, $N > 8$, for any distribution $P_{0\to T}$ on a hypothesis set $\mathcal{Q}_K$, and for any $\delta \in (0,1]$, the following holds with probability at least $1-\delta$ over the training set $\mathcal{D} \sim \Delta^N$:

$$\forall Q_{0\to T}: \quad \mathbb{E}_{H\sim Q_{0\to T}}\left[R(H)\right] \leq \mathbb{E}_{H\sim Q_{0\to T}}\left[R_{\mathcal{D}}(H)\right] + \sqrt{\frac{\mathrm{KL}\left(Q_{0\to T} \parallel P_{0\to T}\right) + \log\left(\frac{2\sqrt{N}}{\delta}\right)}{2N}}$$

Here, $\mathrm{KL}\left(Q_{0\to T} \parallel P_{0\to T}\right)$ acts as a complexity measure that measures, how much the posterior predictive governing the SDE $Q_{0\to T}$ needed to be adapted to the data when compared to an a priori chosen SDE that could alternatively have generated data $P_{0\to T}$. In our situation, $Q_{0\to T}$ is obtained by our approximation scheme, resulting in a bounded likelihood of observations $\mathbf{y}_k$ which factorizes over different observations n. The $P_{0\to T}$ can be arbitrarily chosen as long as it does not depend on the observations. As mentioned in the main paper, we chose an SDE with the same diffusion term which also factorizes over observations. Using this setting, we can analytically compute the KL-distance (as shown in Lemma 1).

On the right hand side of this PAC-bound, we need to evaluate $\mathbb{E}_{H\sim Q_{0\to T}}\left[R_{\mathcal{D}}(H)\right]$. To this end, we note

$$\begin{aligned}
\mathbb{E}_{H\sim Q_{0\to T}}\left[R_{\mathcal{D}}(H)\right] &= \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{H\sim Q_{0\to T}}\left[1 - \frac{1}{\overline{B}_K}\left(\prod_{k=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k^n)\right)\right] \\
&= 1 - \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{H\sim Q_{0\to T}}\left[\frac{1}{\overline{B}_K}\prod_{k=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k^n)\right] \\
&\overset{\text{Hoeffding}}{\leq} 1 - \frac{1}{SN}\sum_{n=1}^{N}\sum_{s=1}^{S}\left[\frac{1}{\overline{B}_K}\prod_{k=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k^{n,s})\right] + \sqrt{\frac{\log(2N/\delta)}{2S}} \\
&= \frac{1}{N}\sum_{n=1}^{N}\left\{1 - \frac{1}{S}\sum_{s=1}^{S}\left[\frac{1}{\overline{B}_K}\prod_{k=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k^{n,s})\right]\right\} + \sqrt{\frac{\log(2N/\delta)}{2S}} \\
&\overset{-\log(z)\geq 1-z}{\leq} -\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{S}\sum_{s=1}^{S}\prod_{k=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k^{n,s})\right) + \log\overline{B}_K + \sqrt{\frac{\log(2N/\delta)}{2S}} \\
&\overset{\text{Jensen's ineq.}}{\leq} -\frac{1}{SN}\sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K}\left[\log p(\mathbf{y}_k^n|\mathbf{h}_k^{n,s})\right] + \log\overline{B}_K + \sqrt{\frac{\log(2N/\delta)}{2S}},
\end{aligned}$$

where we have used Hoeffding's inequality for estimating the true expectation over hypotheses with a K samples trace $\mathbf{h}_k^{n,s}, k = 1,\ldots,K, s = 1,\ldots,S$ for each observation n. As we approximate the integral for each time-series n separately via sampling, we require Hoeffding to hold simultaneously for all n. Using a union bound, we have to scale δ for each n by N. Splitting confidences between the PAC-bound and the sampling based approximation results an additional factor of 2. With $\delta/(2N)$, the corresponding inequality holds with a probability of $\mathbb{P} > \delta/2$. Also using $\delta/2$ in PAC-theorem, we obtain that with $\mathbb{P} \geq 1-\delta$ we have for all $Q_{0\to T}$ that

$$\begin{aligned}
&\mathbb{E}_{H\sim Q_{0\to T}}\left[R(H)\right] \leq \mathbb{E}_{H\sim Q_{0\to T}}\left[R_{\mathcal{D}}(H)\right] + \sqrt{\frac{\mathrm{KL}\left(Q_{0\to T} \parallel P_{0\to T}\right) + \log\left(\frac{2\sqrt{N}}{\delta/2}\right)}{2N}} \\
&\leq -\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{S}\sum_{s=1}^{S}\prod_{k=1}^{K} p(\mathbf{y}_k^n|\mathbf{h}_k^{n,s})\right) + \sqrt{\frac{\mathrm{KL}\left(Q_{0\to T} \parallel P_{0\to T}\right) + \log\left(\frac{2\sqrt{N}}{\delta/2}\right)}{2N}} + \log\overline{B}_K + \sqrt{\frac{\log(2N/\delta)}{2S}} \\
&\leq -\frac{1}{SN}\sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{k=1}^{K}\left[\log p(\mathbf{y}_k^n|\mathbf{h}_k^{n,s})\right] + \sqrt{\frac{\mathrm{KL}\left(Q_{0\to T} \parallel P_{0\to T}\right) + \log\left(\frac{2\sqrt{N}}{\delta/2}\right)}{2N}} + \log\overline{B}_K + \sqrt{\frac{\log(2N/\delta)}{2S}}
\end{aligned}$$

□

Corollary 1. *Given a L-Lipschitz continuous function set*

$$\Big\{f_\theta^n(x):\mathbb{R}\to[0,1]\Big|n=1,\cdots,N\Big\}\bigcup\Big\{g_\theta(x):\mathbb{R}\to[0,+\infty]\Big\},$$

for the two losses:

$$l_1(\theta)=-\sum_{n=1}^N f_\theta^n(x)+g_\theta(x)\quad\textit{and}\quad l_2(\theta)=-\sum_{n=1}^N\log f_\theta^n(x)+g_\theta(x),$$

the sequential updates ($\theta^0:=\theta$)

$$\begin{aligned}\theta^{(n)}&\leftarrow\theta^{(n-1)}+\alpha_n\nabla\big(\log f^n_{\theta^{(n-1)}}(x)\big),\ n=1,\dots,N,\\ \theta^{(N+1)}&\leftarrow\theta^{(N)}-\alpha_{N+1}\nabla g_{\theta^{(N)}}(x),\end{aligned}$$

where $\alpha_n\in(0,f^n_{\theta^{(n-1)}}(x)/L)\ \forall n$ and $\alpha_{N+1}\in(0,1/L)$, satisfy both $l_1(\theta^{(N+1)})\le l_1(\theta)$ and $l_2(\theta^{(N+1)})\le l_2(\theta)$.

Proof. As we only consider updates in θ for constant x, we simplify the notation for this proof to $f^n(\theta):=f^n_\theta(x)$, $g(\theta)=g_\theta(x)$. I.e. we have as the two loss terms

$$l_1(\theta)=-\sum_{n=1}^N f^n(\theta)+g(\theta)\quad\text{and}\quad l_2(\theta)=-\sum_{n=1}^N\log f^n(\theta)+g(\theta).$$

In general we have with $\log f(\theta)<f(\theta)$ that $l_1(\theta)<l_2(\theta)$. Similarly we have

$$\nabla l_2(\theta)=-\sum_n\underbrace{\frac{1}{f^n(\theta)}}_{\ge1}\nabla f^n(\theta)+\nabla g(\theta)\le-\sum_n\nabla f^n(\theta)+\nabla g(\theta)=\nabla l_1(\theta).$$

Due to the sequential updates we can consider each term separately. For an L-Lipschitz function $f^n(\theta)$, we have that for arbitrary x,y

$$f(y)\le f(x)+\nabla f(x)^\top(y-x)+\frac{L}{2}||y-x||_2^2.$$

Choosing $y=\theta^{(n-1)}$ and $x=\theta^{(n)}=\theta^{(n-1)}+\alpha_n\nabla\log f^n$ this gives us

$$\begin{aligned}f(\theta^{(n-1)})&\le f(\theta^{(n)})-\frac{\alpha_n}{f^n(\theta^{(n)})}||\nabla f^n(\theta^{(n)})||_2^2+\frac{L\alpha_n^2}{2f^n(\theta^{(n)})^2}||\nabla f^n(\theta^{(n)})||_2^2\\&=f^n(\theta^{(n)})-\underbrace{\frac{\alpha_n}{f^n(\theta^{(n)})}}_{\ge0}\underbrace{\left(1-\frac{L\alpha_n}{2f^n(\theta^{(n)})}\right)}_{>0}||\nabla f^n(\theta^{(n)})||_2^2\le f^n(\theta^{(n)}),\end{aligned}$$

and hence chaining the update steps gives the desired result. □

That is, updating the terms in $l_2(\theta)$ sequentially, one can ensure concurrent optimization of $l_1(\theta)$. Note that $l_1(\theta)$ and $l_2(\theta)$ are not necessarily dual objectives, hence may have different extrema. Nevertheless, a gradient step that decreases one loss also decreases the other with potentially a different magnitude. In practice, we observe this behavior to also hold empirically for joint gradient update steps with shared learning rates. Applying Lemma 2 to the setup in Theorem 2, we establish a useful link between Empirical Bayes and PAC learning.

Theorem 2 (strong convergence). *Let $\mathbf{h}_t^\theta$ be an Itô process as in (4) with drift parameters θ and its Euler-Maruyama approximation $\widetilde{\mathbf{h}}_t^\theta$ for some regular step size $\Delta t>0$. For some coefficient $R>0$ and any $T>0$, the following inequality holds*

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\mathbb{E}_\theta[\mathbf{h}_t^\theta]-\frac{1}{S}\sum_{s=1}^S\widetilde{\mathbf{h}}_t^{\theta^{(s)}}\right|\right]\le R\Delta t^{1/2},$$

as $S\to\infty$, where $\{\theta^{(s)}\sim p_\phi(\theta_f)|s=1,\dots,S\}$ are i.i.d. draws from a prior $p_\phi(\theta_f)$.

Proof: The Euler-Maruyama (EM) approximation converges strongly as

$$\mathbb{E}\left[\left|\mathbf{h}_T^\theta - \widetilde{\mathbf{h}}_T^\theta\right|\right] \leq R\Delta t^{1/2},$$

for a positive constant R and a suitably small step size Δt as discussed e.g. by Kloeden and Platen (2011). To simplify the mathematical notation we follow their approach of comparing the absolute error of the end of the trajectory throughout the proof. As our sampling scheme is unbiased it is a consistent estimator and we have that asymptotically for $S \to \infty$

$$\frac{1}{S}\sum_{s=1}^{S} \widetilde{h}_T^{\theta(s)} = \mathbb{E}_\theta[\widetilde{h}_T^\theta].$$

We then have for the marginal $\mathbf{h}_T$, $\tilde{\mathbf{h}}_T$ that

$$\begin{aligned}
\mathbb{E}\left[\left|\mathbf{h}_T - \widetilde{\mathbf{h}}_T\right|\right] &= \mathbb{E}\left[\left|\mathbb{E}_\theta \mathbf{h}_T^\theta - \mathbb{E}_\theta \widetilde{\mathbf{h}}_T^\theta\right|\right] \\
&= \mathbb{E}\left[\left|\mathbb{E}_\theta\left[\mathbf{h}_T^\theta - \widetilde{\mathbf{h}}_T^\theta\right]\right|\right] \\
&\leq \mathbb{E}\left[\mathbb{E}_\theta\left[\left|\mathbf{h}_T^\theta - \widetilde{\mathbf{h}}_T^\theta\right|\right]\right] \\
&\leq \mathbb{E}_\theta\left[R\Delta t^{1/2}\right] = R\Delta t^{1/2},
\end{aligned}$$

where the first inequality is due to Jensen and the second due to the strong convergence result for a fixed set of parameters. □

3 COMPUTATIONAL COST

We present the runtimes of the different approaches in Table 1. D-BNN samples the weights of the neural network directly leading to the runtime term $\mathcal{O}(MTF)$. All other approches do not sample the weights but the linear activations of the each data points leading to $\mathcal{O}(2MTF)$. When we apply empirical Bayes, we dot not use any regularization term on the weights, while all other approaches contain a penalty term with cost $\mathcal{O}(W)$. Using the PAC-framework, we employ a second regularization term that leads to an additional runtime cost of $\mathcal{O}(TMD^3)$. However, the cubic cost in D is invoked by inverting the diffusion matrix $G(h_t, t)$ and can be further reduced by choosing a simpler form for $G(h_t, t)$ (e.g. diagonal). In case that prior knowledge is available in ODE form, we need to compute the corresponding drift term for each time point and each MC sample leading to the term $\mathcal{O}(MTP)$.

Table 1: Computational cost analysis in FLOPs for time series of length **T**. **M:** Number of Monte Carlo Samples. **W:** Number of weights in the neural net. **F:** Forward pass cost of a neural net. **L:** Cost for computing the likelihood term. **D:** Number of dimensions. **P:** Cost of a prior SDE integration.

Model	**Training per Iteration**
D-BNN (SGLD)	$\mathcal{O}(MTF + MTDL + W)$
Variational Bayes	$\mathcal{O}(2MTF + MTDL + W)$
E-Bayes	$\mathcal{O}(2MTF + MTDL)$
E-PAC-Bayes	$\mathcal{O}(2MTF + MTDL + W + TMD^3)$
E-Bayes-Hybrid	$\mathcal{O}(2MTF + MTDL + MTP)$
E-PAC-Bayes-Hybrid	$\mathcal{O}(2MTF + MTDL + W + TMD^3 + MTP)$

4 FURTHER DETAILS ON THE EXPERIMENTS

Here we provide the details of the experiment setup we used in obtaining our results reported in the main paper. We observed our results to be robust against most of the design choices. We provide a reference implementation at https://github.com/manuelhaussmann/bnsde.

4.1 Lorenz Attractor

We took 200000 Euler-Maruyama steps ahead with a time step size of 10^{-4} and downsampling by factor 0.01, which gives a sequence of 2000 observations with frequency 0.01. We split the first half of this data set into 20 sequences of length 50 and use them for training, and the second half to 10 sequences of length 100 and use for test. For all model variants, we used an Adam optimizer learning rate 0.001, minibatch size of two, a drift net with two hidden layers of 100 neurons and softplus activation function.We trained all models for 100 epochs and observed this training period to be sufficient for convergence.

4.2 CMU Motion Capture

In this experiment, we tightly follow the design choices reported by Yildiz et al. (2019) to maintain commensurateness. This setup assumes the stochastic dynamics are determined in a six-dimensional latent space. Yildiz et al. (2019) use an auto-encoder to map this latent space to the 50−dimensional observation space back and forth. We adopt their exact encoder-decoder architecture and incorporate it into our BNSDE, arriving at the data generating process

$$\begin{aligned}
\theta_f &\sim p_{\phi_f}(\theta_f),\\
d\mathbf{h}_t|\theta_f &\sim f_{\theta_f}\big(b_\lambda(\mathbf{h}_t), t\big)dt + G\big(b_\lambda(\mathbf{h}_t), t\big)d\boldsymbol{\beta}_t,\\
\mathbf{z}_t|\mathbf{h}_t &\sim \mathcal{N}(\mathbf{z}_t|a_\psi(\mathbf{h}_t), 0.5\cdot 10^{-6}\mathbb{1}),\\
\mathbf{y}_t|\mathbf{z}_t &\sim \mathcal{N}(\mathbf{y}_t|\mathbf{z}_t, 0.5\cdot 10^{-6}\mathbb{1}), \qquad \forall t \in \mathbf{t}.
\end{aligned}$$

Above, $b_\lambda(\cdot,\cdot)$ is the encoder which takes the observations of the last three time points as input, passes them through two dense layers with 30 neurons and softplus activation function, and then linearly projects them to a six-dimensional latent space, where the dynamics are modeled. The decoder $a_\psi(\mathbf{h}_t)$ follows the same chain of mapping operations in reverse order. The only difference is that the output layer of the decoder emits only one observation point, as opposed to the encoder admitting three points at once.

The drift function $f_{\theta_f}(\cdot,\cdot)$ is governed by another separate Bayesian neural net, again with one hidden layer of 30 neurons and softplus activation function on the hidden layer. The diffusion function is fixed to be a constant.

We train all models except SGLD with the Adam optimizer for 3000 epochs on seven randomly chosen snippets at a time with a learning rate of 10^{-3}. We use snippet length 30 for the first 1000 epochs, 50 until epoch 2500, and 100 afterwards. SGLD demonstrates significant training instability for this learning rate, hence for it we drop its learning rate to the largest possible stable value 10^{-5} and increase the epoch count to 5000.

5 FURTHER EXPERIMENTS

5.1 Lotka Volterra

We demonstrate the benefits of incorporating prior knowledge although it is a coarse approximation to the true system. We consider the Lotka-Volterra system specified as:

$$\begin{aligned}
dx_t &= (\theta_1 x_t - \theta_2 x_t y_t)dt + 0.2\ d\beta_t,\\
dy_t &= (-\theta_3 y_t + \theta_4 x_t y_t)dt + 0.3\ d\beta_t.
\end{aligned}$$

with $\boldsymbol{\theta} = (2.0, 1.0, 4.0, 1.0)$. Assuming that the trajectory is observed on the interval $t = [0, 1]$ with a resolution of $dt = 0.01$, we compare the following three methods: *i)* the black-box BNSDE without prior knowledge, *ii)* the white-box SDE in (7) representing partial prior knowledge (parameters are sampled from a normal distribution centered on the true values with a standard deviation of 0.5), and finally *iii)* combining them in our proposed hybrid method. The outcome is summarized in Figure 1. While the plain black-box model delivers a poor fit to data, our hybrid BNSDE brings significant improvement from relevant but inaccurate prior knowledge.

5.1.1 Experimental details

We took 10^5 Euler-Maruyama steps on the interval $[0, 10]$ with a time step size of 10^{-4}, downsampling them by a factor of 100 giving us 1000 observations with a frequency of 0.01. We take the first 500 observations on

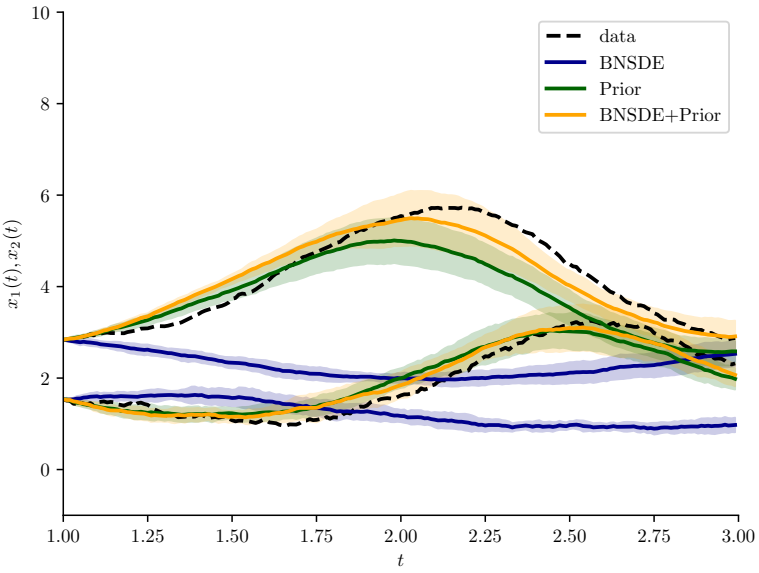


Figure 1: Lotka-Volterra visualization. Error bars indicate three standard deviations over 10 trajectories starting from the true value at $t = 1$. The predictions over 200 time steps ($dt = 0.01$) are for: *i)* a BNSDE trained without prior knowledge, *ii)* an SDE with known prior parameters, *iii)* the joint hybrid BNSDE. The dashed lines are the observed trajectories for x_t and y_t.

the interval [0,5] to be the training data and the observations in $(5, 10]$ to be the test data. Each sequence is split into ten sequences of length 50. Assuming the diffusion parameters to be known and fixed, both BNSDEs (i.e. with and without prior knowledge) get a 4 layer net as the drift function with 50 neurons per layer and ReLU activation functions. The BNSDE with prior knowledge as well as the raw SDE estimate each get an initial sample of $\tilde{\boldsymbol{\theta}}$ parameters as the prior information by sampling from a normal distribution centered around the true parameters ($\tilde{\boldsymbol{\theta}} \sim \mathcal{N}(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}, \sigma^2 \mathbb{1}_4)$). The models are each trained for 50 epochs with the Adam optimizer and a learning rate of $1e-3$. Since both the latent and observed spaces are only two dimensional, we did not need an observation model in this experiment. We directly linked the BNSDE to the likelihood.

5.2 Lorenz Attractor

As discussed in the main paper, the model is trained solely on the first 1000 observations of a trajectory consisting of 2000 observations, leaving the second half for the test evaluation. Figure 2 visualizes the qualitative difference between the two. Note also the single loop the trajectory performs which we will see again in the 1d projections below. To visualize explore the qualitative difference of our proposed model with weak prior knowledge compared to one lacking this knowledge we consider the situation where we we have structural prior knowledge only about the third SDE (i.e. the penultimate case in Table 1 with $\boldsymbol{\rho} = [0, 0, 1]$.

In order to properly visualize it we switch from the 3d plot to 1d plots showing always one of the three dimensions vs the time component. We always start at $T = 10$, forcasting either 100 steps (as in the numerical evaluation), 200 or 1000 steps. All the following figures show the mean trajectory averaged over 21 trajectories, as well as an envelope of $\pm$ 2 standard deviations. Figure 3 visualizes that at that time scale the qualitative behavior is similar without clear differences. Doubling the predicted time interval as shown in Figure 4 the baseline starts to diverge from the true test sequence, while our proposed model still tracks it closely be it at an increased variance. Finally predicting for 1000 time steps (Figure 5) the chaotic behavior of the Lorenz attractor becomes visible as the mean in both setups no longer tracks the true trajectory. Note however that the baseline keeps has rather small variance and a strong tendency in its predictions that do not replicate the qualitative behavior of the Lorenz attractor. While the proposed model also shows an unreliable average, the large variance, which nearly always includes the true trajectory shows that the qualitative behavior is still replicated properly by individual trajectories of the model. See Figure 6 for seven individual trajectories of each of the two models. All trajectories of *E-PAC-Bayes-Hybrid* show the qualitatively correct behavior, including even the characteristic loop.

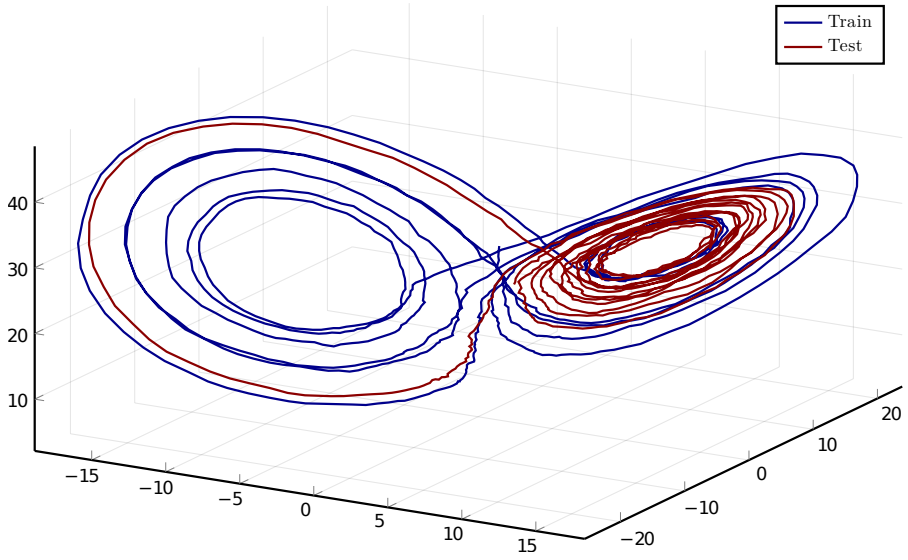


Figure 2: Visualization of the stochastic Lorenz attractor. Of the 2000 observations, the first 1000 constitute the training data (marked in blue), while the second 1000 are the test observations (marked in red). Note the qualitative difference of the two sets.

References

O. Catoni. PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. *IMS Lecture Notes Monograph Series*, 56, 2007.

G. Dziugaite and D.M. Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. In *UAI*, 2017.

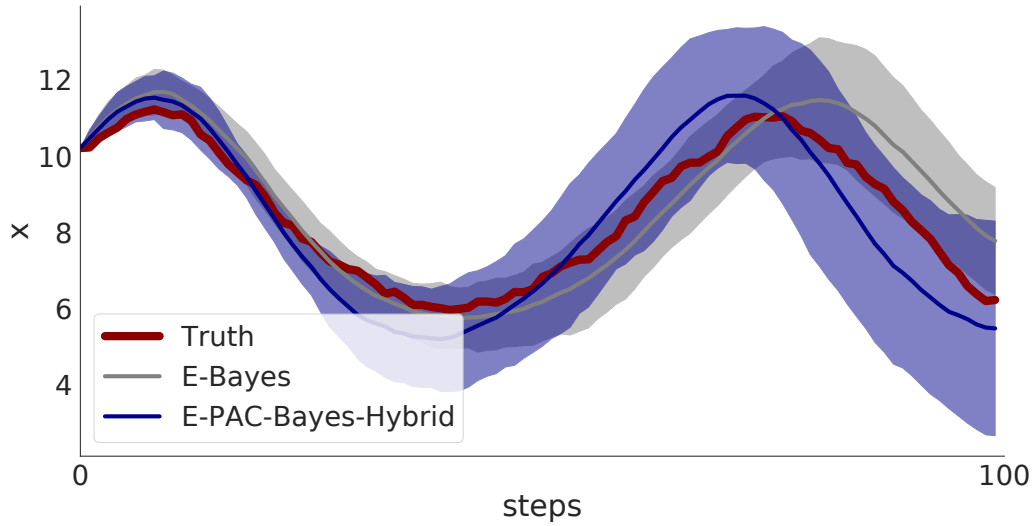
P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. PAC-Bayesian Theory Meets Bayesian Inference. In *NIPS*, 2016.

P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag, 2011.

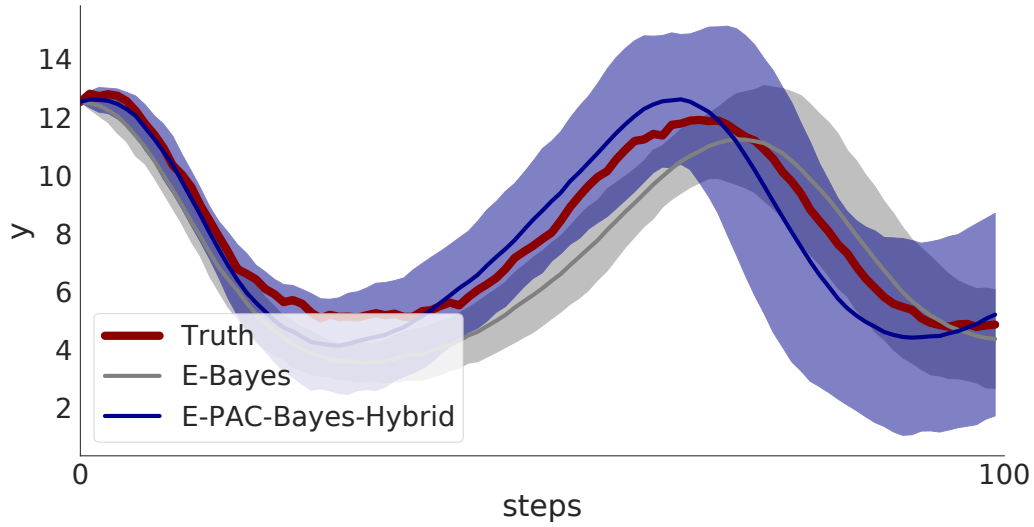
A. Maurer. A note on the PAC Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

B. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer-Verlag, 1992.

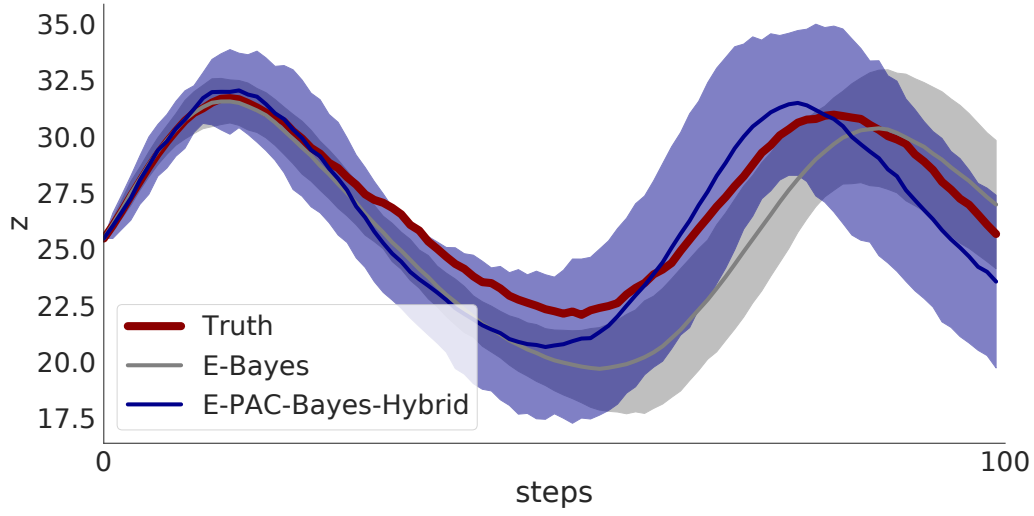
C. Yildiz, M. Heinonen, and H. Lahdesmaki. ODE2VAE: Deep Generative Second Order ODEs with Bayesian Neural Networks. In *NeurIPS*. 2019.



(a) x coordinate over time

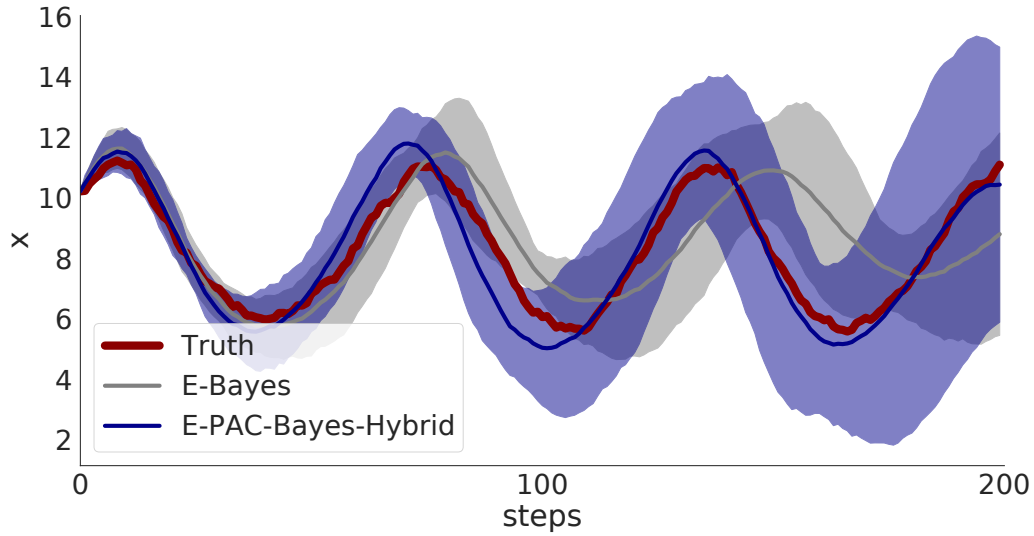


(b) y coordinate over time

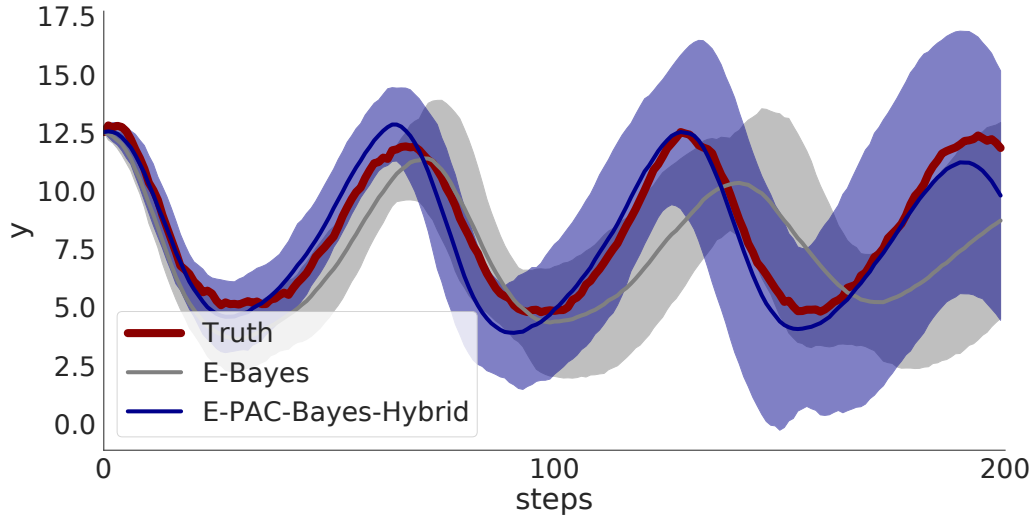


(c) z coordinate over time

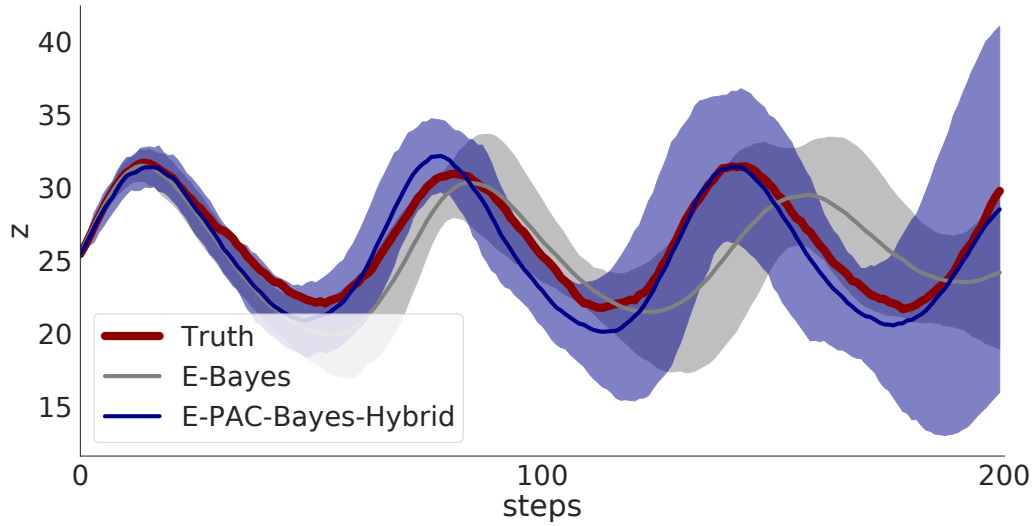
Figure 3: Predicting 100 time steps ahead.



(a) x coordinate over time

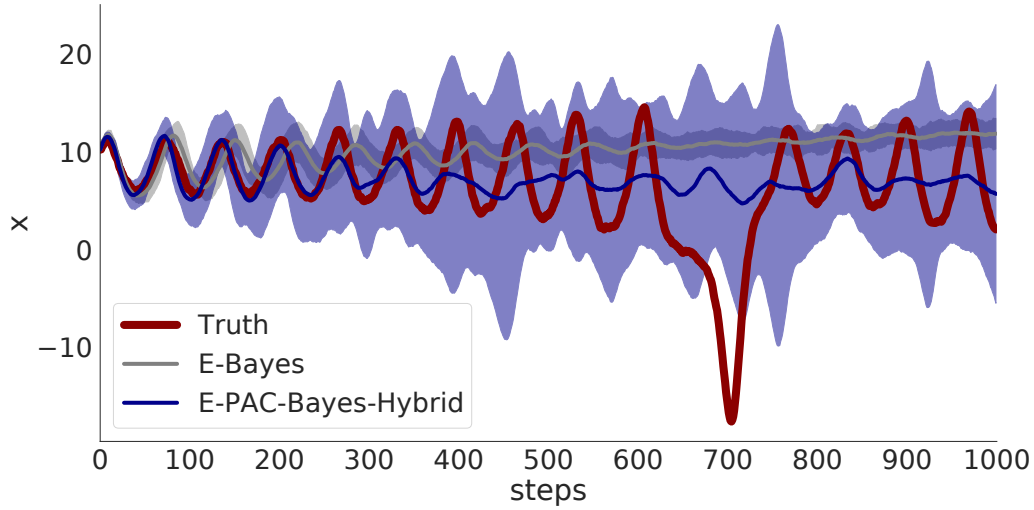


(b) y coordinate over time

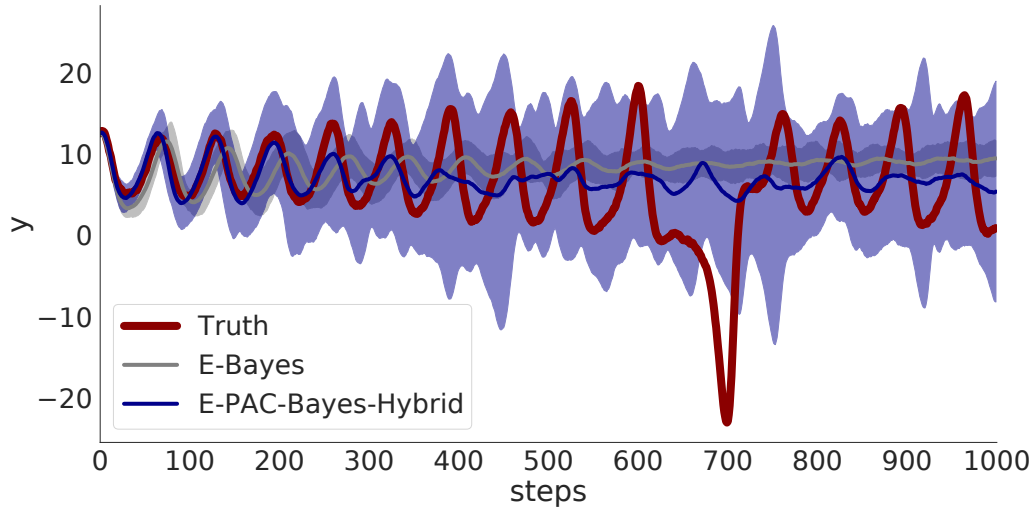


(c) z coordinate over time

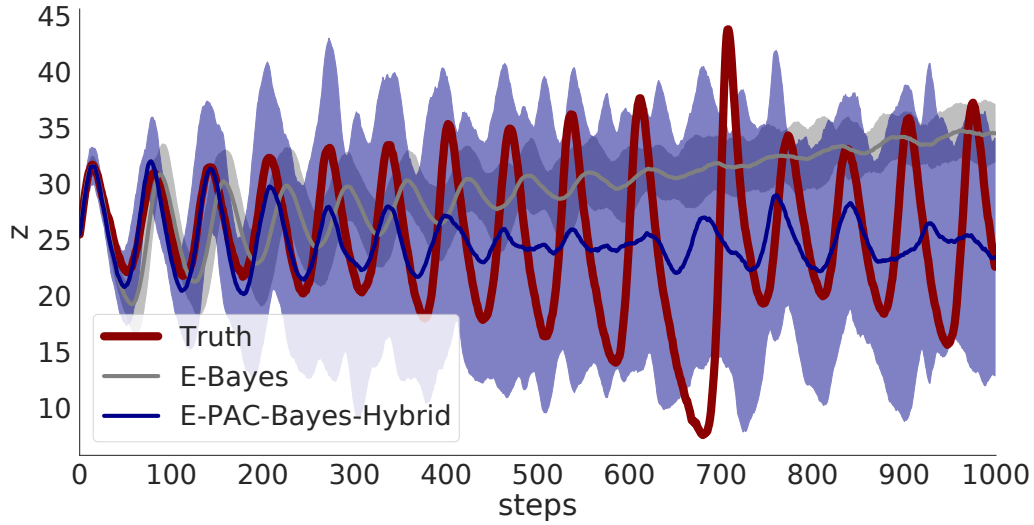
Figure 4: Predicting 200 time steps ahead.



(a) x coordinate over time

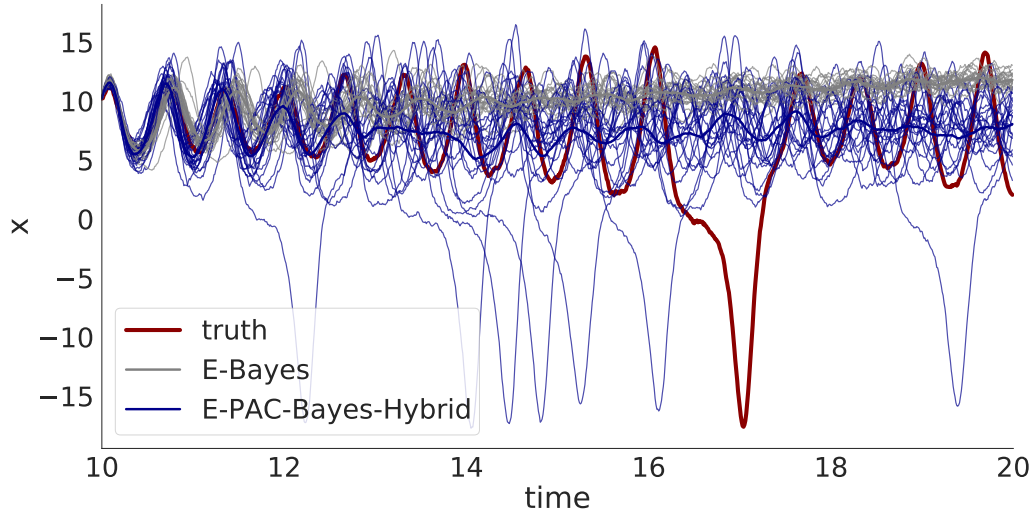


(b) y coordinate over time

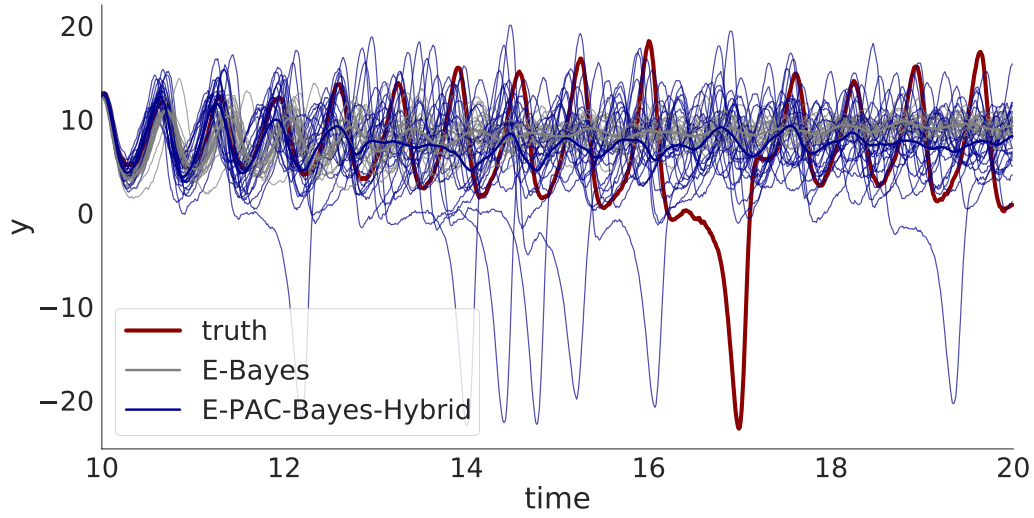


(c) z coordinate over time

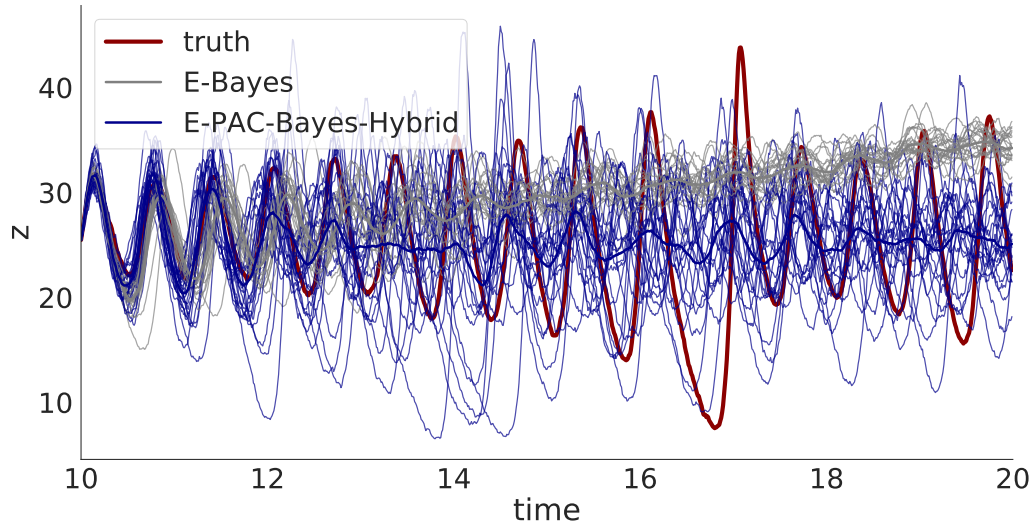
Figure 5: Predicting 1000 time steps ahead.



(a) x coordinate over time



(b) y coordinate over time



(c) z coordinate over time

Figure 6: Predicting 1000 time steps ahead. Shows individual trajectories.