Amortized backward variational inference in nonlinear state-space models

Mathis Chagneux, Élisabeth Gassiat, Pierre Gloaguen, Sylvain Le Corff

Amortized backward variational inference in nonlinear state-space models

Mathis Chagneux[†], Élisabeth Gassiat[‡], Pierre Gloaguen[⋆], and Sylvain Le Corff[⊤]

[†]Télécom Paris, Institut Polytechnique de Paris, Palaiseau.

[‡]Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay.

[⋆]AgroParisTech, UMR MIA 518, Palaiseau.

[⊤]Samovar, Télécom SudParis, département CITI, TIPIC, Institut Polytechnique de Paris, Palaiseau.

Abstract

We consider the problem of state estimation in general state-space models using variational inference. For a generic variational family defined using the same backward decomposition as the actual joint smoothing distribution, we establish for the first time that, under mixing assumptions, the variational approximation of expectations of additive state functionals induces an error which grows at most linearly in the number of observations. This guarantee is consistent with the known upper bounds for the approximation of smoothing distributions using standard Monte Carlo methods. Moreover, we propose an amortized inference framework where a neural network shared over all times steps outputs the parameters of the variational kernels. We also study empirically parametrizations which allow analytical marginalization of the variational distributions, and therefore lead to efficient smoothing algorithms. Significant improvements are made over state-of-the art variational solutions, especially when the generative model depends on a strongly nonlinear and noninjective mixing function.

1 Introduction

When generative data models involve so-called hidden or *latent* states, providing statistical estimates of the latter given observed data - also known as *state inference* - is the cornerstone of many machine learning algorithms [Dempster et al., 1977, Kingma and Welling, 2014]. Traditional models usually introduce low-dimensional states having directly interpretable meaning, while benefiting from accurate inference via exact or consistent Monte Carlo methods. In contrast, modern latent-data machine learning models are rooted in the so-called manifold hypothesis which views high dimensional data as originating from hidden representations in an unknown space and via a complex nonlinear mapping. In the context of unsupervised representation learning, state inference is a goal in itself. Due to the intricacy and dimentionality of the inverse problems involved, most of these works resort to a combination of deep neural networks (DNNs) and variational approximations which allow tractable inference and serve as a principled proxy for maximum likelihood estimation (MLE) [Higgins et al., 2017, Locatello et al., 2020].

The particular case of dependent data is of special importance as it guarantees identifiability results [Khemakhem et al., 2020], especially in the *sequential* setting [Gassiat et al., 2020, Hälvä et al., 2021]. This in turn renews interest in a more solid theoretical understanding of the behaviour of sequential variational methods. In this work, we focus on the case where the true generative model is assumed to be a *state-space model* (SSM). In the general SSM litterature, theoretical analysis of the conditional distribution of the states given the observations - commonly referred to as the *smoothing* distribution - has been extensively conducted to derive efficient estimation algorithms with good convergence properties. Among these works, a keystone in sequential inference is the computation of expected values of additive state functionals under the smoothing distribution, known as additive

smoothing ([Cappé et al., 2005], Chap. 4), and more precisely the control of the additive smoothing error when the target expectations are approximated. Theoretical guarantees have been provided when the approximation is performed using a surrogate of the true smoothing distribution provided by Sequential Monte Carlo (SMC) methods [Douc et al., 2011, Dubarry and Le Corff, 2013, Olsson et al., 2017, Gloaguen et al., 2022]. In addition, in [Gloaguen et al., 2022], a control has also been derived when the smoothed expectations are computed under a biased joint distribution of the hidden states and the observations.

In parallel to these works, sequential variational methods rely on a tractable approximation of the smoothing distribution to compute these expectations. However, this variational approximation has to account for the dependencies implied by the data model [Bayer et al., 2021], and typically does not recover the true distribution in the limit of infinite data when using mean-field variational families. This is why introducing dependency in the variational family has been recently explored in the literature. In [Johnson et al., 2016], the authors obtained promising results by combining conjugate graphical models with variational inference, see also [Lin et al., 2018] for variational methods based on graphical models in the inference network fostering fast amortized inference. In [Krishnan et al., 2017], the variational approximation uses a forward decomposition, parameterized by recurrent neural networks, which allows to mimic the forward decomposition of the true posterior distribution. More recently, [Campbell et al., 2021] proposed a variational family using the so-called *backward factorization*. Such a choice has very appealing properties as it is prone to online state estimation and parameter learning in SSMs.

However, the question of whether these variational families suited to SSMs lead to good variational approximations for additive smoothing remains open. Indeed, to the best of our knowledge, there are no theoretical results providing upper bounds on the state estimation error when using any (mean field or involving dependencies) variational posterior in state-space models. In this paper, we establish the first theoretical guarantees for the variational approximation of additive smoothing in state-space-models, see Proposition 3.1.

In Section 3, we prove that, in the case of strongly mixing state hidden Markov models, the variational estimation error of smoothed additive functional grows at most linearly with the number of observations. In Section 4, we build a backward variational inference algorithm involving fully amortized networks and amenable to recursive learning. In Section 5, we illustrate the theoretical results numerically, and additionally show that a linear Gaussian parametrization of the backward variational kernels can achieve good performance at a small computational cost, even in the case of a strongly nonlinear and noninjective observation model.

2 Background

Notations. Let $\Theta \subset \mathbb{R}^q$ be a parameter space and consider a *state-space model* depending on $\theta \in \Theta$ where the hidden Markov chain in $\mathbb{R}^d$ is denoted by $(X_k)_{k \geqslant 0}$. The distribution of X_0 has density χ^θ with respect to the Lebesgue measure μ and for all $k \geqslant 0$, the conditional distribution of X_{k+1} given $X_{0:k}$ has density $m_k^\theta(X_k, \cdot)$, where $a_{u:v}$ is a short-hand notation for $(a_u, \ldots, a_v)$ for $0 \leqslant u \leqslant v$ and any sequence $(a_\ell)_{\ell \geqslant 0}$. In SSMs, it is assumed that this state is partially observed through an observation process $(Y_k)_{0 \leqslant k \leqslant n}$ taking values in $\mathbb{R}^m$. The observations $Y_{0:n}$ are assumed to be independent conditionally on $X_{0:n}$ and, for all $0 \leqslant k \leqslant n$, the distribution of Y_k given $X_{0:n}$ depends on X_k only and has density $g_k^\theta(X_k, \cdot)$ with respect to the Lebesgue measure.

In the following, for any measure ν on a measurable space $(\mathsf{X}, \mathcal{X})$ and any measurable function h on X, write $\nu h = \int h(x)\nu(\mathrm{d}x)$. In addition, for any measurable spaces $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$, any measure ν on $(\mathsf{X}, \mathcal{X})$, any kernel $K : (\mathsf{X}, \mathcal{Y}) \to \mathbb{R}_+$ and any measurable function h on $\mathsf{X} \times \mathsf{Y}$, write $Kh : x \mapsto \int h(x, y)K(x, \mathrm{d}y)$ and $\nu Kh = \int h(x, y)\nu(\mathrm{d}x)K(x, \mathrm{d}y)$. For simplicity, if for all $x \in \mathsf{X}$, $K(x, \cdot)$ has a density $k(x, \cdot)$ with respect to a reference measure ν, we write $kh : x \mapsto \int h(x, y)K(x, \mathrm{d}y) = \int h(x, y)k(x, y)\nu(\mathrm{d}y)$. Let also $\mathbb{1}$ be the constant function which equals 1 on $\mathbb{R}^d$.

2.1 Latent data models and additive state functionals

In this context, for any $0 \leqslant k_1 \leqslant k_2 \leqslant n$ the *joint smoothing distribution* $\phi_{k_1:k_2}^\theta$ is the conditional law of $X_{k_1:k_2}$ given $Y_{0:n}$. For any function h from $\mathbb{R}^{d \times (n+1)}$ to $\mathbb{R}^{\mathsf{d}}$, we define its *smoothed expectation* when the model is

parameterized by θ as:

$$\begin{aligned}\phi_{0:n}^{\theta} h &= \mathbb{E}^{\theta}\left[h\left(X_{0:n}\right)|Y_{0:n}\right] \\ &= \mathsf{L}_n^{\theta}(Y_{0:n})^{-1} \int h(x_{0:n}) \chi^{\theta}(x_0) g_0^{\theta}(x_0, Y_0) \prod_{k=0}^{n-1} \ell_k^{\theta}(x_k, x_{k+1}) \mu(\mathrm{d}x_{0:n}) \ ,\end{aligned} \tag{1}$$

where[1]

$$\ell_k^{\theta}(x_k, x_{k+1}) = m_k^{\theta}(x_k, x_{k+1}) g_{k+1}^{\theta}(x_{k+1}, Y_{k+1})$$

and $\mathsf{L}_n^{\theta}(Y_{0:n})$ is the likelihood of the observations:

$$\mathsf{L}_n^{\theta}(Y_{0:n}) = \int \chi^{\theta}(x_0) g_0^{\theta}(x_0, Y_0) \prod_{k=0}^{n-1} \ell_k^{\theta}(x_k, x_{k+1}) \mu(\mathrm{d}x_{0:n}) \ . \tag{2}$$

In the context of state-space models, *additive state functionals* are functions $h_{0:n}$ from $\mathbb{R}^{d\times(n+1)}$ to $\mathbb{R}^{\mathsf{d}}$ satisfying:

$$h_{0:n} : x_{0:n} \mapsto \sum_{k=0}^{n-1} \tilde{h}_k(x_k, x_{k+1}) \ , \tag{3}$$

where $\tilde{h}_k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{\mathsf{d}}$. In Bayesian inference, point estimates of most quantities of interest are naturally expressed as posterior means of random functionals belonging to this class. For state inference at a fixed θ, i.e. the recovery of X_k for $0 \leqslant k \leqslant n$ given the observations $Y_{0:n}$, a standard estimator is $\mathbb{E}^{\theta}[X_k|Y_{0:n}]$ which corresponds to $\tilde{h}_k(x_k, x_{k+1}) = x_k$. In Expectation Maximization-based MLE estimation, the intermediate quantity $\theta \mapsto Q(\theta, \theta') = \mathbb{E}^{\theta'}[\sum_{k=0}^{n-1} \log \ell_k^{\theta}(X_k, X_{k+1})|Y_{0:n}]$ is another example where $\tilde{h}_k(x_k, x_{k+1}) = \log \ell_k^{\theta}(x_k, x_{k+1})$. Recursive MLE (RMLE) methods express $\nabla_{\theta} \log \mathsf{L}_n^{\theta} = \mathbb{E}^{\theta}[\sum_{k=0}^{n-1} \nabla_{\theta} \log \ell_k^{\theta}(X_k, X_{k+1})|Y_{0:n}]$ via Fisher's identity under some regularity conditions (see [Cappé et al., 2005], Chap. 10), in which case $\tilde{h}_k(x_k, x_{k+1}) = \nabla_{\theta} \log \ell_k^{\theta}(x_k, x_{k+1})$.

The challenge of computing (1) is twofold, i) the smoothing distribution is generally intractable, ii) under this distribution, expectations are also intractable. A classical approach is to learn both the distribution and expectations using Markov chain or sequential Monte Carlo methods, (see [Chopin et al., 2020], Chapter 12, for a recent review of SMC methods). In the case of additive functionals, more recent generic estimators based on SMC have been designed [Mastrototaro et al., 2021, Martin et al., 2022], and their theoretical properties (consistency, asymptotic variance and normality) have been studied [Gloaguen et al., 2022]. However, Monte Carlo methods show limitations when the dimension d of the latent space is large, and alternatives using variational inference are appealing and computationally efficient solutions.

2.2 Variational inference for sequential data

In variational approaches, instead of designing Monte Carlo estimators of $\phi_{0:n}^{\theta} h$ (or of the conditional distribution of the states given the observations), the conditional law $\phi_{0:n}^{\theta}$ of $X_{0:n}$ given $Y_{0:n}$ is approximated by choosing a candidate in a parametric family $\{q_{0:n}^{\lambda}\}_{\lambda\in\Lambda}$, referred to as the *variational* family, where Λ is a parameter set. Parameters are then learned by maximizing the *evidence lower bound* (ELBO) defined as:

$$\mathcal{L}(\theta, \lambda) = \mathbb{E}_{q_{0:n}^{\lambda}}\left[\log \frac{p_{0:n}^{\theta}(X_{0:n}, Y_{0:n})}{q_{0:n}^{\lambda}(X_{0:n})}\right] = \int \log \frac{p_{0:n}^{\theta}(x_{0:n}, Y_{0:n})}{q_{0:n}^{\lambda}(x_{0:n})} q_{0:n}^{\lambda}(x_{0:n}) \mu(\mathrm{d}x_{0:n}) \ , \tag{4}$$

where $p_{0:n}^{\theta}$ is the joint probability density function of $(X_{0:n}, Y_{0:n})$ when the model is parametrized by θ. A critical point therefore lies in the form of the variational family. Motivated by the sequential nature of the data, most works impose further structure on the variational family via a factorized decomposition of $q_{0:n}^{\lambda}$ over $x_{0:n}$ [Johnson et al., 2016, Krishnan et al., 2017, Lin et al., 2018, Marino et al., 2018]. Here, the natural strategy is to reintroduce part or all of the conditional independence properties of the true generative model.

[1] Note that the dependence of ℓ_k^{θ} on Y_{k+1} is omitted in the notation for better clarity.

2.3 Backward factorization of the smoothing distribution

Under the true model, the *filtering* distribution at time k is defined as the distribution of X_k given $Y_{0:k}$, with density w.r.t the Lebesgue measure denoted as ϕ_k^θ. One known factorization of $\phi_{0:n}^\theta$ - albeit not used in the aforementioned works - exists by further introducing the distribution of the so-called *backward kernels*, that is, for each $0 \leqslant k \leqslant n-1$, the conditional distribution of X_k given $(X_{k+1}, Y_{0:k})$ whose density is proportional to $x_k \mapsto m_k^\theta(x_k, x_{k+1})\phi_k^\theta(x_k)$. A key result for SSMs is that, conditionally on the observations, the reverse-time process $(X_{n-k})_{0\leqslant k\leqslant n}$ is an *inhomogeneous* Markov chain whose initial distribution is the filtering distribution at n, and whose transition kernels are precisely the backward kernels. This allows the following *backward factorization*:

$$\phi_{0:n}^\theta(x_{0:n}) = \phi_n^\theta(x_n) \prod_{k=1}^{n} \frac{m_{k-1}^\theta(x_{k-1}, x_k)\phi_{k-1}^\theta(x_k)}{\int m_{k-1}^\theta(x, x_k)\phi_{k-1}^\theta(x)\mu(\mathrm{d}x)} \ .$$

Since each backward kernel at time k only depends on observations up to time k, a major practical advantage of this decomposition is to allow *recursive* estimation of the smoothing distributions: when a new observation Y_{k+1} is processed, obtaining $\phi_{0:k+1}^\theta$ only amounts to computing ϕ_{k+1}^θ and the associated backward kernel, while previous terms in the product stay fixed. Recently, [Campbell et al., 2021] proposed a related variational family by introducing

$$q_{0:n}^\lambda(x_{0:n}) = q_n^\lambda(x_n) \prod_{k=1}^{n} q_{k-1|k}^\lambda(x_k, x_{k-1}) \ , \tag{5}$$

where q_n^λ (resp. $q_{k-1|k}^\lambda(x_k, \cdot)$) are user-chosen p.d.f. whose parameters typically would depend on $Y_{0:n}$ (resp. $Y_{0:k}$). Under (5), the ELBO (4) becomes an expectation of an additive functional.

3 A control on backward variational additive smoothing

In the context where the variational factorization follows 5, we now present our main theoretical result.

For all $x_k \in \mathbb{R}^d$ and $\theta \in \Theta$, define $\mathbf{L}_k^\theta(x_k, \cdot)$ the kernel with density $\ell_k^\theta(x_k, \cdot)$ with respect to the Lebesgue measure:

$$\mathbf{L}_k^\theta(x_k, \mathrm{d}x_{k+1}) = m_k^\theta(x_k, x_{k+1}) g_{k+1}^\theta(x_{k+1}, Y_{k+1})\mu(\mathrm{d}x_{k+1}) \ .$$

H1 There exist distributions $\tilde{q}_k^\lambda$, $\lambda \in \Lambda$, and functions c_k, $0 \leqslant k \leqslant n$, such that $\tilde{q}_n^\lambda = q_n^\lambda$ and for all $1 \leqslant k \leqslant n$, $\theta \in \Theta$, $\lambda \in \Lambda$, all bounded measurable functions h on $\mathbb{R}^d \times \mathbb{R}^d$,

$$\left| \tilde{q}_k^\lambda q_{k-1|k}^\lambda h - \frac{\tilde{q}_{k-1}^\lambda \mathbf{L}_{k-1}^\theta h}{\tilde{q}_{k-1}^\lambda \mathbf{L}_{k-1}^\theta \mathbb{1}} \right| \leqslant c_k(\theta, \lambda)\|h\|_\infty,$$

and for all bounded measurable functions h on $\mathbb{R}^d$,

$$\left| \tilde{q}_0^\lambda h - \phi_0^\theta h \right| \leqslant c_0(\theta, \lambda)\|h\|_\infty \ ,$$

where ϕ_0^θ is the filtering distribution at time 0, i.e. $\phi_0^\theta h = \chi^\theta g_0^\theta h / \chi^\theta g_0^\theta \mathbb{1}$.

Note that under H1, choosing h such that there exists $\tilde{h}$ satisfying $h : (x_{k-1}, x_k) \mapsto \tilde{h}(x_k)$, yields for all $\theta \in \Theta$, $\lambda \in \Lambda$,

$$\left| \tilde{q}_k^\lambda \tilde{h} - \frac{\tilde{q}_{k-1}^\lambda \mathbf{L}_{k-1}^\theta \tilde{h}}{\tilde{q}_{k-1}^\lambda \mathbf{L}_{k-1}^\theta \mathbb{1}} \right| \leqslant c_k(\theta, \lambda)\|\tilde{h}\|_\infty.$$

H2 There exist constants $0 < \sigma_- < \sigma_+ < \infty$ such that for all $k \in \mathbb{N}$, $\theta \in \Theta$, $\lambda \in \Lambda$ and $(x_k, x_{k+1}) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\sigma_- \leq \ell_k^\theta(x_k, x_{k+1}) \leq \sigma_+$$

and

$$\sigma_- \leq q_{k|k+1}^\lambda(x_{k+1}, x_k) \leq \sigma_+.$$

Proposition 3.1. *Assume that H1 and H2 hold. Then, for all $n \in \mathbb{N}$, $\theta \in \Theta$, $\lambda \in \Lambda$, and all additive functionals $h_{0:n}$ as in* (3)*,*

$$\left|q^{\lambda}_{0:n} h_{0:n} - \phi^{\theta}_{0:n} h_{0:n}\right| \leq 2\frac{\sigma_+}{\sigma_-} \sum_{k=0}^{n-1} \left\|\tilde{h}_k\right\|_\infty \times \left(c_0(\theta,\lambda) + \sum_{m=1}^{k} \rho^{k-m+1} c_m(\theta,\lambda) + c_{k+1}(\theta,\lambda) + \sum_{m=k+2}^{n} \rho^{m-k-1} c_m(\theta,\lambda) \right) ,$$

where $\rho = 1 - \sigma_-/\sigma_+$ and where σ_- and σ_+ are defined in H2.

Proof. The proof is postponed to Appendix A. □

By Proposition 3.1, if there exist h_∞ and c_+ such that for all $0 \leq k \leq n-1$, $\|\tilde{h}_k\|_\infty \leq h_\infty$ and for all $\theta \in \Theta$, $\lambda \in \Lambda$, $0 \leq m \leq n$, $c_m(\theta,\lambda) \leq c_+(\theta,\lambda)$ then

$$\left|q^{\lambda}_{0:n} h_{0:n} - \phi^{\theta}_{0:n} h_{0:n}\right| \leq 4\frac{\sigma_+}{\sigma_-} \left(1 + \frac{\rho}{1-\rho}\right) c_+(\theta,\lambda) h_\infty n \, . \tag{6}$$

On the other hand, if we are interested in marginal smoothing distributions, i.e. cases where $\tilde{h}_j = 0$ for all $j \neq k$, Proposition 3.1 yields a uniform control in time:

$$\left|q^{\lambda}_{0:n} \tilde{h}_k - \phi^{\theta}_{0:n} \tilde{h}_k\right| \leq 4\frac{\sigma_+}{\sigma_-} \left(1 + \frac{\rho}{1-\rho}\right) c_+(\theta,\lambda) h_\infty \, .$$

3.1 Comments on assumptions H1 and H 2

Assumption **H1** is a pivotal technical tool to prove Proposition 3.1. Nonetheless, it is not a strong assumption as for any sequence of distributions $\left(\tilde{q}^{\lambda}_k\right)_{1 \leqslant k \leqslant n}$, the sequence $c_k(\theta,\lambda)$ can be chosen to be the total variation between $(x_{k-1}, x_k) \mapsto \tilde{q}^{\lambda}_k(x_k) q^{\lambda}_{k-1|k}(x_k, x_{k-1})$ and the probability density proportional to $(x_{k-1}, x_k) \mapsto \tilde{q}^{\lambda}_{k-1}(x_{k-1}) \ell^{\theta}_{k-1}(x_{k-1}, x_k)$. However, a challenging task for future research would be to find the best sequence of $\tilde{q}^{\lambda}_k$ in terms of $c_k(\theta,\lambda)$. We now show that in some specific examples, given a sequence of $\tilde{q}^{\lambda}_k$, an explicit sequence of $c_k(\theta,\lambda)$ can be given.

Exact inference. It is worth noting that if q^{λ}_n is the true filtering distribution at time n and $(q^{\lambda}_{k-1|k})_{k \geqslant 1}$ are the true backward distributions, then the unique sequence $(\tilde{q}^{\lambda}_k)_{k \geqslant 1}$ that achieves $c_k(\theta,\lambda) = 0$ in **H1** for all k is the sequence of true filtering distributions.

Linear and Gaussian case. In the linear and Gaussian case, we assume that for all x_{k-1}, $m^{\theta}_{k-1}(x_{k-1}, \cdot)$ is the Gaussian p.d.f with mean $A^{\theta}_k x_{k-1}$ and variance R^{θ}_k and that $g^{\theta}_k(x_k, \cdot)$ is the Gaussian p.d.f with mean $B^{\theta}_k x_k$ and variance S^{θ}_k.

In this setting, assume that $\tilde{q}^{\lambda}_k$ is the Gaussian p.d.f. with mean μ^{λ}_k and variance Σ^{λ}_k and that for all x_k, $q^{\lambda}_{k-1|k}(x_k, \cdot)$ is the Gaussian p.d.f with mean $A^{\lambda}_k x_k$ and variance R^{λ}_k. Therefore, we assume (i) that the variational backward kernels are linear as are the backward kernels of the true model and (ii) that the instrumental intermediate distributions $\tilde{q}^{\lambda}_k$, $0 \leq k \leq n-1$, are Gaussian as the filtering distributions of the true model. As described below, this choice allows to obtain an explicit upper bound for c_k, $0 \leq k \leq n$. This highlights that assumption **H1** can be made usable in practice. This also emphasizes the versatility of **H1** as other instrumental densities could be tuned, since this specific choice is not proved to be optimal.

Choosing $q^{\lambda}_{k-1:k}$ (resp. $q^{\lambda,\theta}_{k-1:k}$) as a short-hand notation for the joint distribution $\tilde{q}^{\lambda}_k q^{\lambda}_{k-1|k}$ (resp. $\tilde{q}^{\lambda}_{k-1} \mathbf{L}^{\theta}_{k-1} h / \tilde{q}^{\lambda}_{k-1} \mathbf{L}^{\theta}_{k-1} \mathbb{1}$), standard computations show that $q^{\lambda}_{k-1:k}$ (resp. $q^{\lambda,\theta}_{k-1:k}$) is a multivariate Gaussian distributions with known mean

M_k^λ (resp. $\mathsf{M}_k^{\lambda,\theta}$) and variance V_k^λ (resp. $\mathsf{V}_k^{\lambda,\theta}$). In this case, for all bounded and measurable function h,

$$\left| \tilde{q}_k^\lambda q_{k-1|k}^\lambda h - \frac{\tilde{q}_{k-1}^\lambda \mathbf{L}_{k-1}^\theta h}{\tilde{q}_{k-1}^\lambda \mathbf{L}_{k-1}^\theta \mathbb{1}} \right| \leqslant 2 \left\| q_{k-1:k}^\lambda - q_{k-1:k}^{\lambda,\theta} \right\|_{\mathrm{tv}} \|h\|_\infty,$$

where $\|\cdot\|_{\mathrm{tv}}$ is the total variation distance. Therefore, we can choose $c_k(\theta,\lambda) = 2\|q_{k-1:k}^\lambda - q_{k-1:k}^{\lambda,\theta}\|_{\mathrm{tv}}$. It remains to use the fact that $q_{k-1:k}^\lambda$ and $q_{k-1:k}^{\lambda,\theta}$ are Gaussian distributions, that $\|q_{k-1:k}^\lambda - q_{k-1:k}^{\lambda,\theta}\|_{\mathrm{tv}} \leqslant (\mathrm{KL}(q_{k-1:k}^\lambda \| q_{k-1:k}^{\lambda,\theta})/2)^{1/2}$ and that we have an explicit expression of the KL divergence between Gaussian distributions which yields

$$c_k(\theta,\lambda) \quad \propto \quad \left(\log \left| \mathsf{V}_k^{\lambda,\theta} \right| / \left| \mathsf{V}_k^\lambda \right| + \left(\Delta_k^{\lambda,\theta} \right)^\top (\mathsf{V}_k^{\lambda,\theta})^{-1} \left(\Delta_k^{\lambda,\theta} \right) + \mathrm{Tr}\left((\mathsf{V}_k^{\lambda,\theta})^{-1} \mathsf{V}_k^\lambda \right) - d \right)^{-1/2} ,$$

where $\Delta_k^{\lambda,\theta} = \mathsf{M}_k^\lambda - \mathsf{M}_k^{\lambda,\theta}$, Tr is the Trace operator and $\propto$ means up to a multiplicative constant independent of θ and λ.

About H2 This assumption is rather strong, but typically satisfied in models where the state space is compact. This assumption is classic in the SMC literature in order to obtain quantitative bounds for errors or variance of estimators.

4 Recursive backward variational learning with amortizing networks

Written as in (5), the backward factorization of the variational family only imposes dependencies between the latent states. This minimal setup, sufficient to derive the theoretical results above, leaves a lot of freedom for implementation.

4.1 Amortized parametrization of the variational distribution

First, suppose that we want to learn the variational parameters by computing ELBO gradients on sequences of fixed length n. Implementing (5) requires to define $n+1$ distributions $(q_{k-1|k}^\lambda)_{0\leq k\leq n}$ and q_n^λ. A direct approach would be to freely parameterize these distributions. In this case, the number of parameters to learn would grow linearly with n, which is prohibitive for long sequences. To reduce the computational burden, a popular alternative is amortized inference, which in this context amounts to output the parameters of each kernel via a common highly expressive mapping, – typically, a DNN which itself holds a fixed number of parameters.

For this purpose, an appealing property of the backward kernels of the true data model is the incremental dependency on the observations (see Section 2.3). Indeed, the backward distribution of x_{k-1} depends on the observations up to time $Y_{0:k-1}$ (through the filtering distribution at time $k-1$) and on the state x_k. Our first step is then to encode sequentially the dependencies on the observations through a recurrent neural network $f_{k-1}^\lambda(y_{0:k-1})$, such that parameters of $q_{k-1|k}^\lambda$ are given by a non linear function $g_k^\lambda\left(f_{k-1}^\lambda(y_{0:k-1}), x_k\right)$. Finally, parameters of q_n^λ are given by a last non linear function $f_n^\lambda(y_{0:n})$, that typically would depend on $f_{n-1}^\lambda(y_{0:n-1})$.

4.2 Variational recursions and online computation of the ELBO

In the setting presented above an interesting implementation choice is when the RNN f_k^λ outputs the parameters of a p.d.f. The RNN therefore indirectly outputs a sequence of distributions $\left(q_k^\lambda\right)_{1\leqslant k\leqslant n}$. These distributions can be used at each time k to define online variational distributions that factorize as in (5) (replacing q_n^λ by q_k^λ). From there, the ELBO can be computed online. Indeed, note that at time n, using the tower property of expectations, $\mathcal{L}(\theta,\lambda) = \mathbb{E}_{q_n^\lambda}[T_n(X_n)]$ where $T_n(X_n) = \mathbb{E}_{q_{0:n}^\lambda}[\log p_{0:n}^\theta(X_{0:n}, Y_{0:n})/q_{0:n}^\lambda(X_{0:n})|X_n]$. This statistic can be

computed recursively, since, for all $k \geq 0$,

$$T_k(X_k) = \mathbb{E}_{q^\lambda_{k-1|k}}\left[T_{k-1}(X_{k-1}) + \log \frac{\ell^\theta_k(X_{k-1}, X_k) q^\lambda_{k-1}(X_{k-1})}{q^\lambda_{k-1|k}(X_k, X_{k-1}) q^\lambda_k(X_k)} \middle| X_k\right] . \tag{7}$$

A more detailed derivation is provided in the appendix. An important point is that contrary to [Campbell et al., 2021], we only assume that each of the *joint* distributions $(q^\lambda_{0:k})_{k\geq 0}$ is an approximation of $(\phi^\theta_{0:k})_{k\geq 0}$. Interestingly, **H1** hints that best results may be obtained by actually enforcing that $(q^\lambda_k)_{k\geq 0}$ and $(q^\lambda_{k-1|k})_{k\geq 0}$ approximate the densities of the true filtering and backward distributions. Still, we find that competitve results are obtained without further regularization, even in the amortized setting where the global set of parameters is optimized jointly over time.

5 Numerical experiments

5.1 Linear Gaussian SSMs and equality in H1

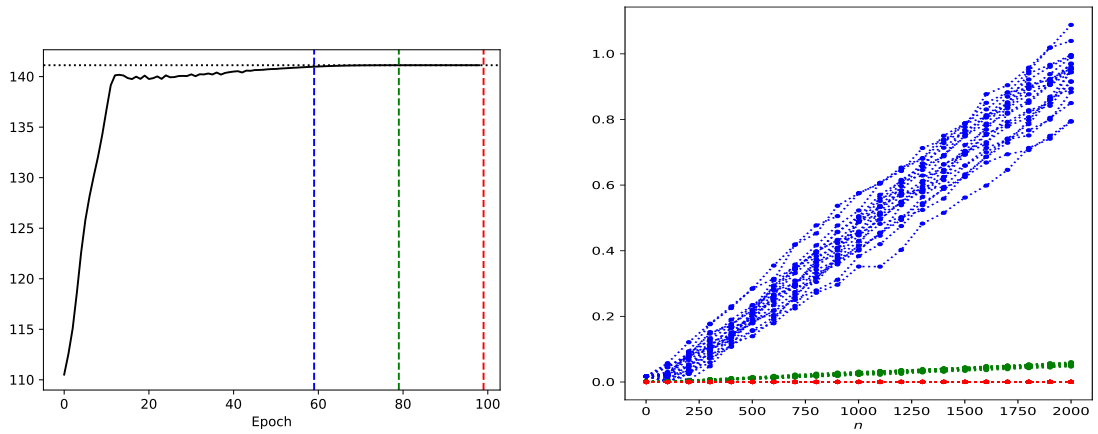
First, we want to study empirically the special case where the variational family *contains* the true model. This can be achieved when the true state-space model is a linear and Gaussian SSM, i.e. when χ^θ (resp. $m^\theta_k(X_k, \cdot)$ and $g^\theta_k(X_k, \cdot)$) are densities of Gaussian distributions with mean A_0 (resp. AX_k and BX_k) and variance Q_0 (resp. Q and R), such that $\theta = (A_0, Q_0, A, Q, B, R)$. If we define $(q^\lambda_{k-1|k})_{k\leq n}$ and q^λ_n as the backward and filtering densities of a similar model with parameters $\lambda = (\bar{A}_0, \bar{Q}_0, \bar{A}, \bar{Q}, \bar{B}, \bar{R})$, then $q^\lambda_{0:n} = \phi^\theta_{0:n}$ for $\lambda = \theta$. When this is achieved, Section 3.1 shows that $c_k(\theta, \lambda) = 0$ for all k, suggesting that the additive error vanishes. In this setting, the form and parameters of the variational backward and filtering kernel is given analytically via the Kalman filtering and smoothing recursions, thus the computations of $\phi^\theta_{0:n}$, $q^\lambda_{0:n}$ and all expectations in (7) are fully tractable. In this example, the parameter θ is known and λ is trained in the case $d = 1$ and using samples of $n = 64$ observations. The training curve is given in Figure 1a.

In Figure 1b, we depict the controlled term of Proposition 3.1 in the case of state estimation, i.e. for $h_{0:n} : x_{0:n} \mapsto \sum_{k=0}^n x_k$. This is done by sampling $J = 20$ observation sequences $(Y^j_{0:n})_{1\leq j\leq J}$ of length $n = 2000$ using the true model with parameter θ. This clearly illustrates the linear dependency on the number of observations. We also find that the error rates can vary greatly between parameters $\lambda_1 \neq \lambda_2$, even when $|\mathcal{L}(\theta, \lambda_1) - \mathcal{L}(\theta, \lambda_2)|$ is small. This is observed by computing the errors for different stopping points of the optimization. Sampling distinct sequences $(Y^j_{0:n})_{1\leq j\leq J}$ highlights the dependency of $(c_k(\theta, \lambda))_{0\leq k\leq n}$ on the observations. In the appendix, we provide more implementation details, as well as additional figures for the errors on the marginal distributions.

5.2 Expressive capabilities of backward variational families in nonlinear Gaussian SSMs

We now consider a generative model where the prior distribution and transition kernels are still linear, but $g^\theta_k(X_k, \cdot)$ is the Gaussian probability density with mean $h^\theta(X_k)$ and variance R, h^θ being a nonlinear mapping commonly referred to as the *decoder*. In this setting, [Hälvä et al., 2021] showed for the first time that no assumptions are required on h^θ for identifiable state estimation. The authors obtained promising results via a variational approximation $q^\lambda_{0:n}$ which can be analytically marginalized and therefore allows fast inference. We briefly explain how this variational approximation can be generalized in our context. For all $k \geq 0$, q^λ_k (resp. $q^\lambda_{k-1|k}(X_k, \cdot)$) is a Gaussian probability density with mean μ_k (resp. $\overleftarrow{A}_k X_k + \overleftarrow{a}_k$) and variance Σ_k (resp. $\overleftarrow{\Sigma}_k$). Moreover, a variational prior $\bar{\chi}^\lambda$ and variational transition kernels $\bar{m}^\lambda_k(X_k, \cdot)$ are introduced as Gaussian densities with mean $\bar{A}_0$ (resp. $\bar{A}X_k$) and variance $\bar{Q}_0$ (resp. $\bar{Q}$) which enforces hidden dynamics of the variational model to have the same form as the data model. We then suppose that:

- $(\mu_k, \Sigma_k) = r^\lambda(u_k, y_k)$, where $u_k = (\bar{A}\mu_{k-1}, \bar{A}\Sigma_{k-1}\bar{A}^T + \bar{Q})$ and r^λ is a mapping to be specified below.
- $q^\lambda_{k-1|k}(X_k, X_{k-1}) \propto \bar{m}^\lambda_k(X_{k-1}, X_k) q^\lambda_k(X_k)$.



(a) L_n^θ (dotted line) and $\lambda \mapsto \mathcal{L}(\theta, \lambda)$ over epochs (full line). (b) Smoothing errors between the variational model and the true model, i.e. $\left|q_{0:n}^\lambda h_{0:n} - \phi_{0:n}^\theta h_{0:n}\right|$ for $\bar{h}_k(x_k, x_{k+1}) = x_k$;

Figure 1: ELBO during the training of λ (left). Additive smoothing error for a linear Gaussian variational model at successive stopping points of the optimization (blue, green and red), on $J = 20$ different observation sequences (right).

The linear dynamics of $\bar{m}_k^\lambda(X_k, \cdot)$ prescribe Kalman-type *predict* and *backward* updates, while u_k are the parameters of an intermediate *predictive* Gaussian distribution. The mapping r^λ then performs the Bayesian *update* step and can be of any form. In [Hälvä et al., 2021], the authors do not use of a generic form for this update step but follow [Johnson et al., 2016] and impose that μ_k, Σ_k is the result of the conjugation of two Gaussian distributions: the predictive whose parameters are u_k, and a variational approximation of $x_k \mapsto g_k^\theta(x_k, y_k)$ whose parameters are given by a DNN f_{enc}^λ (referred to as the *encoder*) which takes only y_k as input. While this form is required for tractable inference in their framework (as they build $q_{0:n}^\lambda$ from it with the sum-product algorithm for SSMs) our backward formulation does not require this, and we show that higher performance can be obtained by letting a DNN r^λ learn a more realistic conjugation of new observations with the running variational filtering estimates.

In this context, the true smoothing distribution $\phi_{0:n}^\theta$ has no analytic form. As a surrogate for this ground truth, we use the particle-based Forward Filtering Backward Simulation (FFBSi) algorithm. The FFBSi outputs trajectories (here, 1000 samples) approximately sampled from the true target smoothing distributions using sequential importance sampling and resampling steps. This algorithm is also based on a forward-backward decomposition of the smoothing distributions (see [Douc et al., 2014], Chapter 11, for details). We remain in the case $d = 1$ to ensure that this approximation is good. We provide additional implementation details and figures in the appendix.

In the case where h^θ is a non-injective mapping, we compare the additive error with respect to the FFBSi (i.e. the left hand term of equation (6)) obtained for our parametrization and the one of [Hälvä et al., 2021] for $h_{0:n} : x_{0:n} \mapsto \sum_{k=0}^n x_k$. Figure 2 shows that our method reduces significantly this error. In Figure 3, we report the quality of the FFBSi estimator in the form of the sample mean and variance of its error against the true states. We then report the final additive smoothing errors of the variational methods after processing all of the $n = 500$ observations of the evaluation sequences. The results confirm our intuition that our framework leads to more expressive variational distributions, especially when the distribution of X_k given Y_k admits several modes. Indeed, the framework of [Hälvä et al., 2021] approximates $x_k \mapsto p^\theta(x_k|y_k)$ by an encoder $f_{enc}^\lambda(y_k)$ that outputs a Gaussian density. In contrast, our parametrization only assumes Gaussianity for the variational filtering distribution and does not attempt to solve the inverse problem of modeling the distribution of X_k given Y_k without the dynamics. Therefore, here, the backwards formulation allows to conserve analytical marginalisation of $q_{0:n}^\lambda$ without modeling the previous distribution as an intermediate step, which increases performance.

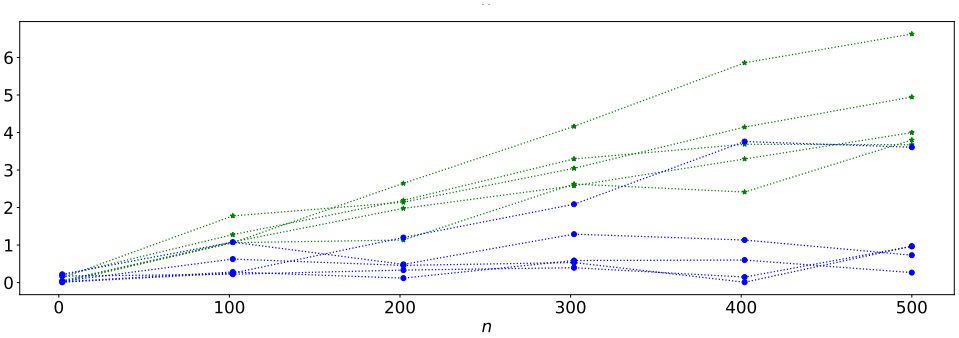


Figure 2: Smoothing errors $\left|q^{\lambda}_{0:n}h_{0:n} - \phi^{\theta}_{0:n}h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$, with our variational approach (blue dots) and that of [Hälvä et al., 2021] (green stars). Experiments were produced on 5 independent (simulated) data set, hence the 5 replicates.

Seq.	Mean err.$_{\text{FFBSi}}$	Var err.$_{\text{FFBSi}}$	Smooth err.$_{\text{FFBSi/Johnson}}$	Smooth err.$_{\text{FFBSi/Ours}}$
0	0.05	0.01	4.95	0.73
1	0.04	0.00	3.80	0.97
2	0.05	0.01	4.00	0.27
3	0.03	0.00	3.67	0.97
4	0.07	0.02	6.63	3.61

Figure 3: First column: empirical mean of $\{(\hat{x}_{k,FFBSi} - x^*_k)^2\}_{0\leqslant k\leqslant n}$ where x^*_k is the true state and $\hat{x}_{k,FFBSi}$ is the marginal mean of $\phi^{\theta}_{0:n}$ at time k provided by the FFBSi algorithm. Second column: empirical variance of the same quantity. Third and fourth column: smoothing errors $\left|q^{\lambda}_{0:n}h_{0:n} - \phi^{\theta}_{0:n}h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$ of the two compared methods at time $n = 500$ when $\phi^{\theta}_{0:n}$ is given by the FFBSi algorithm.

6 Discussion

We have provided the first bound on the additive smoothing error in the context of sequential variational inference using a backward factorization. We have empirically presented clear cases to highlight the practical consequences of this theoretical result. We have also shown that existing methods can be reframed into filtering and backward recursions: in this case, we found that more flexible updates are available without increasing the computational workload. Some limitations of our work and challenges for further research are the following.

- Our theoretical result sheds light on important properties of sequential variational methods, but the assumptions involved are not fully *constructive*, i.e. we believe that further works may provide more explicitly the form of the optimal variational factors under given parametric families of the variational kernels.
- Empirically, we have restricted to the case where analytical computations are available to marginalize the joint variational smoothing distribution. More computationally heavy approaches requiring Monte Carlo sampling for marginalisation are possible, and may further improve the state estimation results shown in Section 5.
- Since the DNNs involved in our implementation take the estimations of the current dynamics as input, we find that training in our context suffers more easily from the drawbacks of gradient descent in recurrent models, e.g. it is more amenable to vanishing / exploding gradients.

As a novel variational approach for sequential data, this work has potential applications in many areas. This work does not present any foreseeable societal consequence.

References

[Bayer et al., 2021] Bayer, J., Soelch, M., Mirchev, A., Kayalibay, B., and van der Smagt, P. (2021). Mind the gap when conditioning amortised inference in sequential latent-variable models. In *International Conference on Learning Representations*.

[Campbell et al., 2021] Campbell, A., Shi, Y., Rainforth, T., and Doucet, A. (2021). Online variational filtering and parameter learning. *Advances in Neural Information Processing Systems*, 34.

[Cappé et al., 2005] Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer.

[Chopin et al., 2020] Chopin, N., Papaspiliopoulos, O., et al. (2020). *An introduction to sequential Monte Carlo*. Springer.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–38.

[Douc et al., 2011] Douc, R., Garivier, A., Moulines, E., and Olsson, J. (2011). Sequential monte carlo smoothing for general state space hidden markov models. *The Annals of Applied Probability*, 21(6):2109–2145.

[Douc et al., 2014] Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear time series: theory, methods and applications with R examples*. CRC Press.

[Dubarry and Le Corff, 2013] Dubarry, C. and Le Corff, S. (2013). Non-asymptotic deviation inequalities for smoothed additive functionals in nonlinear state-space models. *Bernoulli*, 19(5B):2222–2249.

[Gassiat et al., 2020] Gassiat, E., Le Corff, S., and Lehéricy, L. (2020). Identifiability and consistent estimation of nonparametric translation hidden Markov models with general state space. *Journal of Machine Learning Research*, 21.

[Gloaguen et al., 2022] Gloaguen, P., Le Corff, S., and Olsson, J. (2022). A pseudo-marginal sequential Monte Carlo online smoothing algorithm. *Bernoulli*, To appear(-):–.

[Hälvä et al., 2021] Hälvä, H., Le Corff, S., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., and Hyvärinen, A. (2021). Disentangling identifiable features from noisy data with structured nonlinear ICA. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34.

[Higgins et al., 2017] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

[Johnson et al., 2016] Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., and Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems (NeurIPS)*, 29.

[Khemakhem et al., 2020] Khemakhem, I., Kingma, D. P., and Hyvärinen, A. (2020). Variational autoencoders and nonlinear ICA: A unifying framework. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2207–2217.

[Kingma and Welling, 2014] Kingma, D. and Welling, M. (2014). Auto-encoding variational bayes.

[Krishnan et al., 2017] Krishnan, R., Shalit, U., and Sontag, D. (2017). Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

[Lin et al., 2018] Lin, W., Khan, M. E., and Hubacher, N. (2018). Variational message passing with structured inference networks. In *International Conference on Learning Representations*.

[Locatello et al., 2020] Locatello, F., Poole, B., Rätsch, G., Scholkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 119:6348–6359.

[Marino et al., 2018] Marino, J., Cvitkovic, M., and Yue, Y. (2018). A general method for amortizing variational filtering. In *Advances in neural information processing systems (NeurIPS)*, volume 31.

[Martin et al., 2022] Martin, A., Étienne, M.-P., Gloaguen, P., Le Corff, S., and Olsson, J. (2022). Backward importance sampling for online estimation of state space models. *ArXiv:2002.05438*.

[Mastrototaro et al., 2021] Mastrototaro, A., Olsson, J., and Alenlöv, J. (2021). Fast and numerically stable particle-based online additive smoothing: the adasmooth algorithm.

[Olsson et al., 2017] Olsson, J., Westerborn, J., et al. (2017). Efficient particle-based online smoothing in general hidden markov models: the PaRIS algorithm. *Bernoulli*, 23(3):1951–1996.

A Proof of Proposition 3.1

Following [Gloaguen et al., 2022], write

$$q_{0:n}^{\lambda} h_n - \phi_{0:n}^{\theta} h_n = \sum_{k=0}^{n-1} \left(q_{0:n}^{\lambda} \bar{h}_{k|n} - \phi_{0:n}^{\theta} \bar{h}_{k|n} \right) ,$$

where, for each $k \in \{0, n-1\}$, $\bar{h}_{k|n}$ is defined on $(\mathbb{R}^d)^{n+1}$ by

$$\bar{h}_{k|n} : x_{0:n} \mapsto \tilde{h}_k(x_k, x_{k+1}) . \tag{8}$$

Define, for each $n \in \mathbb{N}$ and $m \in \{0, n\}$, the kernel

$$\mathbf{L}^{\theta}_{m,n}(x'_{0:m}, \mathrm{d}x_{0:n}) := \delta_{x'_{0:m}}(\mathrm{d}x_{0:m}) \prod_{\ell=m}^{n-1} \mathbf{L}^{\theta}_{\ell}(x_\ell, \mathrm{d}x_{\ell+1}) \tag{9}$$

on $(\mathbb{R}^d)^{n+1} \times \mathcal{B}((\mathbb{R}^d)^{n+1})$, with the convention $\prod_{\ell=n}^{n-1} f(\ell) = 1$. This yields the following decomposition:

$$q^{\lambda}_{0:n}\bar{h}_{k|n} - \phi^{\theta}_{0:n}\bar{h}_{k|n} = \sum_{m=1}^{n} \left(\frac{\tilde{q}^{\lambda}_{0:m}\mathbf{L}^{\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{0:m}\mathbf{L}^{\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^{\lambda}_{0:m-1}\mathbf{L}^{\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{0:m-1}\mathbf{L}^{\theta}_{m-1,n}\mathbb{1}} \right) + \frac{\tilde{q}^{\lambda}_{0}\mathbf{L}^{\theta}_{0,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{0}\mathbf{L}^{\theta}_{0,n}\mathbb{1}} - \frac{\chi^{\theta} g^{\theta}_0 \mathbf{L}^{\theta}_{0,n}\bar{h}_{k|n}}{\chi^{\theta} g^{\theta}_0 \mathbf{L}^{\theta}_{0,n}\mathbb{1}},$$

where $\tilde{q}^{\lambda}_{0:m} = \tilde{q}^{\lambda}_{m} \prod_{k=1}^{m} q^{\lambda}_{k-1|k}$, $(1 \leq m \leq n)$, $\tilde{q}^{\lambda}_{0:0} = \tilde{q}^{\lambda}_{0}$, and since $\chi^{\theta} g^{\theta}_0 \mathbf{L}^{\theta}_{0,n}\bar{h}_{k|n} / \chi^{\theta} g^{\theta}_0 \mathbf{L}^{\theta}_{0,n}\mathbb{1} = \phi^{\theta}_{0:n}\bar{h}_{k|n}$. For each $n \in \mathbb{N}$, define $\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{0,n}(x'_0, \mathrm{d}x_{0:n}) := \delta_{x'_0}(\mathrm{d}x_0) \prod_{\ell=0}^{n-1} \mathbf{L}^{\theta}_{\ell}(x_\ell, \mathrm{d}x_{\ell+1})$ and for $m \in \{1, n\}$,

$$\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}(x'_m, \mathrm{d}x_{0:n}) := \delta_{x'_m}(\mathrm{d}x_m) \prod_{\ell=0}^{m-1} q^{\lambda}_{k|k+1}(x_{\ell+1}, \mathrm{d}x_\ell) \prod_{\ell=m}^{n-1} \mathbf{L}^{\theta}_{\ell}(x_\ell, \mathrm{d}x_{\ell+1}), \tag{10}$$

on $\mathbb{R}^d \times \mathcal{B}((\mathbb{R}^d)^{n+1})$. As for all $m \in \{1, n\}$ and measurable function h, $\tilde{q}^{\lambda}_{0:m}\mathbf{L}^{\theta}_{m,n}h = \tilde{q}^{\lambda}_{m}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}h$,

$$\frac{\tilde{q}^{\lambda}_{0:m}\mathbf{L}^{\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{0:m}\mathbf{L}^{\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^{\lambda}_{0:m-1}\mathbf{L}^{\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{0:m-1}\mathbf{L}^{\theta}_{m-1,n}\mathbb{1}} = \frac{\tilde{q}^{\lambda}_{m}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{m}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^{\lambda}_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}}.$$

Therefore,

$$q^{\lambda}_{0:n}\bar{h}_{k|n} - \phi^{\theta}_{0:n}\bar{h}_{k|n} = \sum_{m=1}^{n} \left(\frac{\tilde{q}^{\lambda}_{m}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{m}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^{\lambda}_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}} \right) + \frac{\tilde{q}^{\lambda}_{0}\mathbf{L}^{\theta}_{0,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{0}\mathbf{L}^{\theta}_{0,n}\mathbb{1}} - \frac{\chi^{\theta} g^{\theta}_0 \mathbf{L}^{\theta}_{0,n}\bar{h}_{k|n}}{\chi^{\theta} g^{\theta}_0 \mathbf{L}^{\theta}_{0,n}\mathbb{1}}. \tag{11}$$

By Lemma B.1,

$$\left| \frac{\tilde{q}^{\lambda}_{0}\mathbf{L}^{\theta}_{0,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{0}\mathbf{L}^{\theta}_{0,n}\mathbb{1}} - \frac{\phi^{\theta}_{0}\mathbf{L}^{\theta}_{0,n}\bar{h}_{k|n}}{\phi^{\theta}_{0}\mathbf{L}^{\theta}_{0,n}\mathbb{1}} \right| \leqslant 2c_0(\theta, \lambda)\frac{\sigma_+}{\sigma_-}\|\tilde{h}_k\|_\infty \, .$$

Consider now the error term at time $m > 0$ in (11). Define the kernel

$$\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}(x'_{m-1}, x'_m, \mathrm{d}x_{0:n}) := \delta_{x'_{m-1}}(\mathrm{d}x_{m-1}) \prod_{\ell=0}^{m-2} q^{\lambda}_{\ell|\ell+1}(x_{\ell+1}, \mathrm{d}x_\ell)\delta_{x'_m}(\mathrm{d}x_m) \prod_{\ell=m}^{n-1} \mathbf{L}^{\theta}_{\ell}(x_\ell, \mathrm{d}x_{\ell+1}), \tag{12}$$

on $(\mathbb{R}^d)^2 \times \mathcal{B}((\mathbb{R}^d)^{n+1})$ so that for all $x_{m-1}, x_m \in \mathbb{R}^d$,

$$\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1}, x_m) = \begin{cases} q^{\lambda}_{m-2|m-1} \cdots q^{\lambda}_{k|k+1}\tilde{h}_k(x_{m-1})\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m) & \text{if } k \leq m-2\,, \\ \tilde{h}_k(x_{m-1}, x_m)\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m) & \text{if } k = m-1\,, \\ \mathbf{L}^{\theta}_{m,n}\tilde{h}_k(x_m) & \text{if } k \geq m\,. \end{cases}$$

Then, write

$$\frac{\tilde{q}^{\lambda}_{m}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{m}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^{\lambda}_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}} = \frac{\tilde{q}^{\lambda}_{m}q^{\lambda}_{m-1|m}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{m}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^{\lambda}_{m-1}\mathbf{L}^{\theta}_{m-1}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^{\lambda}_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}} \, .$$

Let $1 \leq m \leq n$ and x^*_{m-1} and x^*_m be arbitrary elements in $\mathbb{R}^d$. For $k \neq m-1$, define

$$\begin{aligned}\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m) &= \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}(x_{m-1},x_m)} - \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x^*_{m-1},x^*_m)}{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}(x^*_{m-1},x^*_m)}\,, \\ &= \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} - \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x^*_{m-1},x^*_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x^*_m)}\end{aligned} \tag{13}$$

and for $k = m-1$, $\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m) = \tilde{h}_k(x_{m-1},x_m)$. By Lemma B.2, $\left\|\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\right\|_\infty$ can be upper bounded and note that

$$\begin{aligned}&\frac{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^\lambda_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}} \\ &\qquad = \frac{\tilde{q}^\lambda_m q^\lambda_{m-1|m}\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\}}{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^\lambda_{m-1}\mathbf{L}^{\theta}_{m-1}\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\}}{\tilde{q}^\lambda_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}}\,.\end{aligned}$$

Define now the normalized measure $\tilde{\phi}^\lambda_m h$ by $\tilde{q}^\lambda_{m-1}\mathbf{L}^{\theta}_{m-1}h / \tilde{q}^\lambda_{m-1}\mathbf{L}^{\theta}_{m-1}\mathbb{1}$, so that

$$\begin{aligned}&\frac{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^\lambda_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}} \\ &\qquad = \frac{\tilde{q}^\lambda_m q^\lambda_{m-1|m}\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\}}{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{\phi}^\lambda_m\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\}}{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} \\ &\qquad = \frac{\tilde{q}^\lambda_m q^\lambda_{m-1|m}\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\} - \tilde{\phi}^\lambda_m\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\}}{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} \\ &\qquad\qquad + \frac{\tilde{q}^\lambda_m q^\lambda_{m-1|m}\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\}}{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}\left(\frac{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1} - \tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}\right)\,.\end{aligned}$$

Then, using that

$$\left|\frac{\tilde{q}^\lambda_m q^\lambda_{m-1|m}\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\}}{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}\right| \leqslant \left\|\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\right\|_\infty\,,$$

and by H1,

$$\left|\frac{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1} - \tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}\right| \leqslant c_m(\theta,\lambda)\frac{\left\|\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\|_\infty}{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}\,,$$

$$\left|\frac{\tilde{q}^\lambda_m q^\lambda_{m-1|m}\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\} - \tilde{\phi}^\lambda_m\left\{\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\}}{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}\right| \leqslant c_m(\theta,\lambda)\frac{\left\|\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\right\|_\infty\left\|\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\|_\infty}{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}\,,$$

yields

$$\left|\frac{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_m\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^\lambda_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_{m-1}\boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}}\right| \leqslant 2c_m(\theta,\lambda)\frac{\left\|\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}\right\|_\infty\left\|\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}\right\|_\infty}{\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}}\,.$$

Note also that by H2,

$$\tilde{\phi}^\lambda_m\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1} \geqslant \sigma_-\mu\mathbf{L}^{\theta}_{m+1,n-1}\mathbb{1}\,,$$

and for all $x_m \in \mathbb{R}^d$,

$$\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}(x_m) \leqslant \sigma_+ \mu \mathbf{L}^\theta_{m+1,n-1}\mathbb{1} \ .$$

Therefore,

$$\left| \frac{\tilde{q}^\lambda_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n}\mathbb{1}} - \frac{\tilde{q}^\lambda_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n}\mathbb{1}} \right| \leqslant 2\frac{\sigma_+}{\sigma_-} c_m(\theta,\lambda) \left\| \mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n} \right\|_\infty \ .$$

The proof is completed using Lemma B.2.

B Technical results

Lemma B.1. *Assume that H1 and H2 hold. Then for all, $\theta \in \Theta$, $\lambda \in \Lambda$, $n \geq 1$, $k \in \{0, n-1\}$, bounded and measurable function $\tilde{h}_k$,*

$$\left| \frac{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} - \frac{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} \right| \leq 2c_0(\theta,\lambda)\frac{\sigma_+}{\sigma_-}\|\tilde{h}_k\|_\infty \ ,$$

where $\bar{h}_{k|n}$ is defined in (8)*.*

Proof. Consider the following decomposition of the first term:

$$\begin{aligned}
\frac{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} - \frac{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} &= \frac{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} - \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} \ , \\
&= \frac{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n} - \phi^\theta_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} \\
&\qquad + \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\mathbb{1} - \tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}}{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} \ ,
\end{aligned}$$

where ϕ^θ_0 the filtering distribution at time 0, i.e the law defined as $\phi^\theta_0 h = \chi^\theta g^\theta_0 h / \chi^\theta g^\theta_0$. Then, by H1,

$$\left| \frac{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} - \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} \right| \leqslant 2c_0(\theta,\lambda)\frac{\|\mathbf{L}^\theta_{0,n}\mathbb{1}\|_\infty \|\bar{h}_{k|n}\|_\infty}{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} \ .$$

By H2, for all $x_0 \in \mathbb{R}^d$,

$$\mathbf{L}^\theta_{0,n}\mathbb{1}(x_0) = \int \ell_{0,\theta}(x_0,x_1)\mu(\mathrm{d}x_1)\mathbf{L}^\theta_{1,n}\mathbb{1}(x_1) \leqslant \sigma_+ \int \mu(\mathrm{d}x_1)\mathbf{L}^\theta_{1,n}\mathbb{1}(x_1)$$

and

$$\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1} = \int \tilde{q}^\lambda_0(\mathrm{d}x_0)\ell_{0,\theta}(x_0,x_1)\mu(\mathrm{d}x_1)\mathbf{L}^\theta_{1,n}\mathbb{1}(x_1) \geqslant \sigma_- \int \mu(\mathrm{d}x_1)\mathbf{L}^\theta_{1,n}\mathbb{1}(x_1) \ ,$$

which yields

$$\left| \frac{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\tilde{q}^\lambda_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} - \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\bar{h}_{k|n}}{\phi^\theta_0 \mathbf{L}^\theta_{0,n}\mathbb{1}} \right| \leqslant 2c_0(\theta,\lambda)\frac{\sigma_+}{\sigma_-}\|\tilde{h}_k\|_\infty \ .$$

□

Lemma B.2. *Assume that H2 holds. Then for all $n \in \mathbb{N}$, $\theta \in \Theta$, $\lambda \in \Lambda$, $m \in \{1, n\}$, $k \in \{0, n-1\}$, $x_{m-1}, x_m, x^*_{m-1}, x^*_m$ in $\mathbb{R}^d$, bounded and measurable function $\tilde{h}_k$,*

$$\left|\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)\right| \leq \begin{cases} \|\tilde{h}_k\|_\infty \rho^{m-k-1} & \text{if } k \leq m-2 \ , \\ \|\tilde{h}_k\|_\infty & \text{if } k = m-1 \ , \\ \|\tilde{h}_k\|_\infty \rho^{k-m+1} & \text{if } k \geq m \ . \end{cases}$$

where $\rho = 1 - \sigma_-/\sigma_+$ and $\bar{h}_{k|n}$ is defined in (8) *and $\mathcal{L}^{*,\lambda,\theta}_{m,n}\bar{h}_{k|n}$ is defined in* (13)*.*

Proof. The proof is adapted from [Gloaguen et al., 2022, Lemma D.3] and given here for completeness. Assume first that $k \leq m-2$. Then,

$$\frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} = q^{\lambda}_{m-2|m-1}\dots q^{\lambda}_{k|k+1}\tilde{h}_k(x_{m-1})$$

Therefore,

$$\frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} - \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x^*_{m-1},x^*_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x^*_m)} = (\delta_{x_{m-1}} - \delta_{x^*_{m-1}})q^{\lambda}_{m-2|m-1}\dots q^{\lambda}_{k|k+1}\tilde{h}_k \ .$$

By H2, the Dobrushin coefficient of the variational backward kernels is upper-bounded by $1-\sigma_-/\sigma_+$ so that

$$\left|\frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} - \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x^*_{m-1},x^*_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x^*_m)}\right| \leqslant \left(1-\frac{\sigma_-}{\sigma_+}\right)^{m-k-1}\left\|\tilde{h}_k\right\|_\infty \ .$$

In the case where $k = m-1$,

$$\frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} = \tilde{h}_k(x_k,x_{k+1}) \ ,$$

so that the result is straightforward. Assume now first that $k \geq m$. Note that

$$\frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} = \frac{\mathbf{L}^{\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} = \frac{F^{\theta}_{m|n}\dots F^{\theta}_{k|n}\bar{h}_{k|n}(x_m)\cdot\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} \ ,$$

where the forward kernel $\mathsf{F}^{\theta}_{\ell|n}$ is given by

$$\mathsf{F}^{\theta}_{\ell|n}h(x_\ell) = \frac{\mathbf{L}^{\theta}_{\ell}(h\mathbf{L}^{\theta}_{\ell+1,n-1}\mathbb{1})(x_\ell)}{\mathbf{L}^{\theta}_{\ell,n-1}\mathbb{1}(x_\ell)} \ .$$

By H2,

$$\mathsf{F}^{\theta}_{\ell|n}h(x_\ell) \geqslant \frac{\sigma_-}{\sigma_+}\mu_{\ell|n}h \ ,$$

with $\mu_{\ell|n}h = \mu(h\mathbf{L}^{\theta}_{\ell+1,n-1}\mathbb{1})(x_\ell)/\mu\mathbf{L}^{\theta}_{\ell+1,n-1}\mathbb{1}$. Therefore, the Dobrushin coefficients of the kernels $F^{\theta}_{\ell|n}$ are also upper-bounded by $1-\sigma_-/\sigma_+$. On the other hand,

$$\frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} - \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x^*_{m-1},x^*_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x^*_m)} = (\lambda_{m|n}-\lambda'_{m|n})\mathsf{F}^{\theta}_{m|n}\dots\mathsf{F}^{\theta}_{k|n}\bar{h}_{k|n},$$

where $\lambda_{m|n}h = \delta_{x_m}h\mathbf{L}^{\theta}_{m,n}\mathbb{1}/\delta_{x_m}\mathbf{L}^{\theta}_{m,n}\mathbb{1}$ and $\lambda'_{m|n}h = \delta_{x'_m}h\mathbf{L}^{\theta}_{m,n}\mathbb{1}/\delta_{x'_m}\mathbf{L}^{\theta}_{m,n}\mathbb{1}$. This yields

$$\left|\frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x_{m-1},x_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x_m)} - \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}\bar{h}_{k|n}(x^*_{m-1},x^*_m)}{\mathbf{L}^{\theta}_{m,n}\mathbb{1}(x^*_m)}\right| \leqslant \left(1-\frac{\sigma_-}{\sigma_+}\right)^{k-m+1}\left\|\tilde{h}_k\right\|_\infty \ ,$$

which concludes the proof.

□

C Deriving the recursive form of the ELBO

To obtain a recursion on $T_n(X_n) = \mathbb{E}_{q^{\lambda}_{0:n}}[\log p^{\theta}_{0:n}(X_{0:n},Y_{0:n})/q^{\lambda}_{0:n}(X_{0:n})|X_n]$, we notice, as in [Campbell et al., 2021], that

$$q^{\lambda}_{0:k}(x_{0:k}) = q^{\lambda}_{0:k-1}(x_{0:k-1})\bar{q}^{\lambda}_{k|k-1}(x_{k-1},x_k) \ ,$$

where $\bar{q}^{\lambda}_{k|k-1}(x_{k-1}, x_k) = q^{\lambda}_{k-1|k}(x_k, x_{k-1}) q^{\lambda}_k(x_k) / q^{\lambda}_k(x_{k-1})$. The function $x_k \mapsto \bar{q}_{k|k-1}(x_{k-1}, x_k)$ is not the density of a Markov kernel but allows an alternate decomposition of the variational family forward in time. Since the density of the complete data model $x_{0:n} \mapsto p^{\theta}_{0:n}(x_{0:n}, Y_{0:n})$ also factorizes via the densities $x_k \mapsto \ell^{\theta}_k(x_{k-1}, x_k)$ of the forward kernels, the statistic $T_n(X_n)$ writes:

$$T_n(X_n) = \mathbb{E}_{q^{\lambda}_{0:n}}\left[\log \frac{\chi^{\theta}(X_0) g^{\theta}_0(X_0) \prod_{k=1}^n \ell^{\theta}_k(X_{k-1}, X_k)}{q^{\lambda}_0(X_0) \prod_{k=1}^n \bar{q}^{\lambda}_{k|k-1}(X_{k-1}, X_k)} \middle| X_n\right] .$$

By applying again the tower property of expectations, this yields:

$$\begin{aligned} T_n(X_n) &= \mathbb{E}_{q^{\lambda}_{0:n}}\left[\mathbb{E}_{q^{\lambda}_{0:n}}\left[\log \frac{\chi^{\theta}(X_0) g^{\theta}_0(X_0) \prod_{k=1}^n \ell^{\theta}_k(X_{k-1}, X_k)}{q^{\lambda}_0(X_0) \prod_{k=1}^n \bar{q}^{\lambda}_{k|k-1}(X_{k-1}, X_k)} \middle| X_{n-1}, X_n\right] \middle| X_n\right] \\ &= \mathbb{E}_{q^{\lambda}_{0:n}}\left[\mathbb{E}_{q^{\lambda}_{0:n-1}}\left[\log \frac{\chi^{\theta}(X_0) g^{\theta}_0(X_0) \prod_{k=1}^{n-1} \ell^{\theta}_k(X_{k-1}, X_k)}{q^{\lambda}_0(X_0) \prod_{k=1}^{n-1} \bar{q}^{\lambda}_{k|k-1}(X_{k-1}, X_k)} \middle| X_{n-1}\right]\right. \\ &\qquad \left. + \log \frac{\ell^{\theta}_n(X_{n-1}, X_n)}{\bar{q}^{\lambda}_{n|n-1}(X_{n-1}, X_n)} \middle| X_n\right] . \end{aligned}$$

The inner expectation is $T_{k-1}(X_{k-1})$ by definition. Since all terms in the outer expectation are only functions of X_{n-1}, the expectation under $q^{\lambda}_{0:n}$ reduces to an expectation under the backward kernel $q^{\lambda}_{n-1|n}$, i.e.

$$T_n(X_n) = \mathbb{E}_{q^{\lambda}_{n-1|n}}\left[T_{n-1}(X_{n-1}) + \log \frac{\ell^{\theta}_n(X_{n-1}, X_n)}{\bar{q}^{\lambda}_{n|n-1}(X_{n-1}, X_n)} \middle| X_n\right] ,$$

which is the recursion proposed in (7).

D Experiment details

D.1 Hardware configuration

We ran all experiments on a machine with the following specifications.

- CPUs: 4x Intel(R) Xeon(R) Gold 6154 (total 72 cores, 144 threads).
- RAM: 260 Go.

No GPU was used.

D.2 Linear Gaussian models

We provide here additional figures for the experiments of Section 5.1. Figure 4 shows the marginal errors across time for the variational models for the different stopping points of Figure 1. Table 5 shows the accuracy of the optimal Kalman smoothing (with true parameters θ) w.r.t the true states, as well as the numerical values for the smoothing errors at the three stopping points of the optimization.

We also provide examples of smoothed states for the multivariate case. In Figure 6, we plot the paths of an evaluation sequence where the state space is of dimension 3 and the observation space is of dimension 4. We visualise the results by marginalizing $\phi^{\theta}_{0:n}\ q^{\lambda}_{0:n}$ on each dimension of the state space (and for each timestep). Note that here we do not learn the emission matrix and the variances to avoid having to fix indeterminacies of the multidimensional case (scaling and permutations).

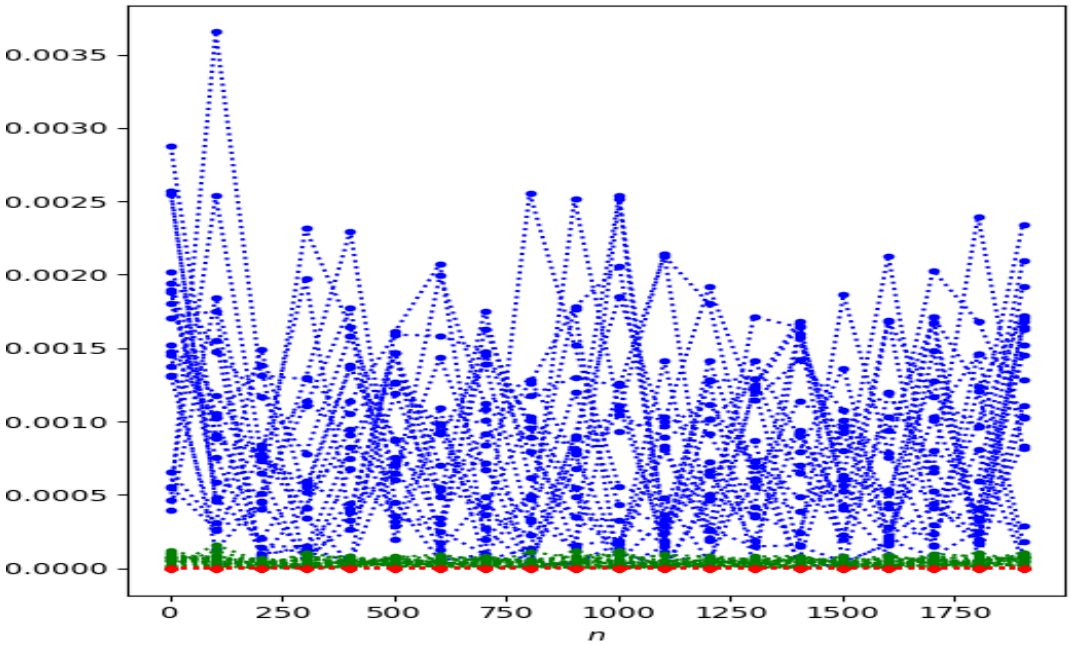


Figure 4: Marginal errors between the variational model and the true model at three different points of the optimization: 60 epochs (blue), 80 epochs (green) and 100 epochs (red).

D.3 Nonlinear models

Here we provide additional details on the experiments of section 5.2.

- For the nonlinear emission function h^θ of the data model, we used a single-layer perceptron with a $\tanh$ activation function. We found that this mapping is sufficiently nonlinear to evaluate and compare the models, but we further apply the $\cos$ function to the output to ensure noninjectivity.
- For the parameterization of the method based on [Johnson et al., 2016], the mapping f^λ_{enc} is a multi-layer perceptron (MLP) with two hidden layers of 16 neurons and a $\tanh$ activation function. The activation function is not applied to the output layer to ensure that the values can exceed values outside the range $[-1, 1]$, being natural parameters of Gaussian distributions. The output of the network is split into two natural parameters η_1 and η_2, the latter being constrained to strictly negative values by applying the softplus function $x \mapsto -\log(1 + e^x)$. We use Xavier initialization for the matrix parameters, and random normal initialisation for the bias parameters.
- For the parameterization of r^λ in our method, we use the exact same MLP as f^λ_{enc} described above but take the predictive parameters u_k as additinal input (we denote it f'^λ_{enc}). We add a forget gate to mitigate vanishing / exploding gradient issues, where the forget state is computed by a single-layer perceptron. If we denote by s this forget layer, then

$$r^\lambda(u_k, y_k) = s(u_k, y_k) * u_k + [1 - s(u_k, y_k)] * f'^\lambda_{enc}(u_k, y_k) \ ,$$

  where $*$ is the element-wise product.

Seq nb.	Mean err.$_\theta$	Var err.$_\theta$	Smoothing err.$_{\theta/\lambda_{60}}$	Smoothing err.$_{\theta/\lambda_{80}}$	Smoothing err.$_{\theta/\lambda_{100}}$
0	0.001432	0.000004	0.954874	0.054593	0.000232
1	0.001432	0.000004	0.914786	0.052843	0.000227
2	0.001466	0.000004	1.039261	0.058327	0.000243
3	0.001420	0.000004	0.953855	0.054565	0.000232
4	0.001394	0.000004	0.943506	0.054108	0.000231
5	0.001473	0.000004	1.088147	0.060461	0.000249
6	0.001445	0.000004	0.944956	0.054160	0.000231
7	0.001415	0.000004	0.994196	0.056330	0.000237
8	0.001390	0.000004	0.849939	0.050003	0.000219
9	0.001404	0.000004	0.990084	0.056150	0.000237
10	0.001408	0.000004	0.959414	0.054805	0.000233
11	0.001378	0.000004	0.883776	0.051488	0.000223
12	0.001412	0.000004	0.952297	0.054494	0.000232
13	0.001405	0.000004	0.793487	0.047527	0.000212
14	0.001497	0.000005	0.996393	0.056432	0.000237
15	0.001495	0.000004	0.893889	0.051933	0.000225
16	0.001564	0.000005	0.916225	0.052907	0.000227
17	0.001406	0.000004	0.794842	0.047594	0.000212
18	0.001386	0.000004	0.968786	0.055223	0.000234
19	0.001373	0.000004	0.970476	0.055282	0.000234

Figure 5: First column: empirical mean of $\{(\hat{x}_{k,\theta} - x_k^*)^2\}_{0\leqslant k\leqslant n}$ where x_k^* is the true state and $\hat{x}_{k,\theta}$ is the marginal mean of $\phi_{0:n}^\theta$ at time k provided by Kalman smoothing with true parameters θ. Second column: empirical variance of the same quantity. Third, fourth and fifth columns: smoothing errors $\left|q_{0:n}^\lambda h_{0:n} - \phi_{0:n}^\theta h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$ at $n = 2000$, when $\phi_{0:n}^\theta$ is given via Kalman smoothing with the true parameters θ and $q_{0:n}^\lambda$ is given via Kalman smoothing with parameters λ selected at epochs 60,80 and 100. Each line is corresponds to one observation sequence in $(Y_{0:n}^j)_{1\leq j\leq J}$, $J = 20$.

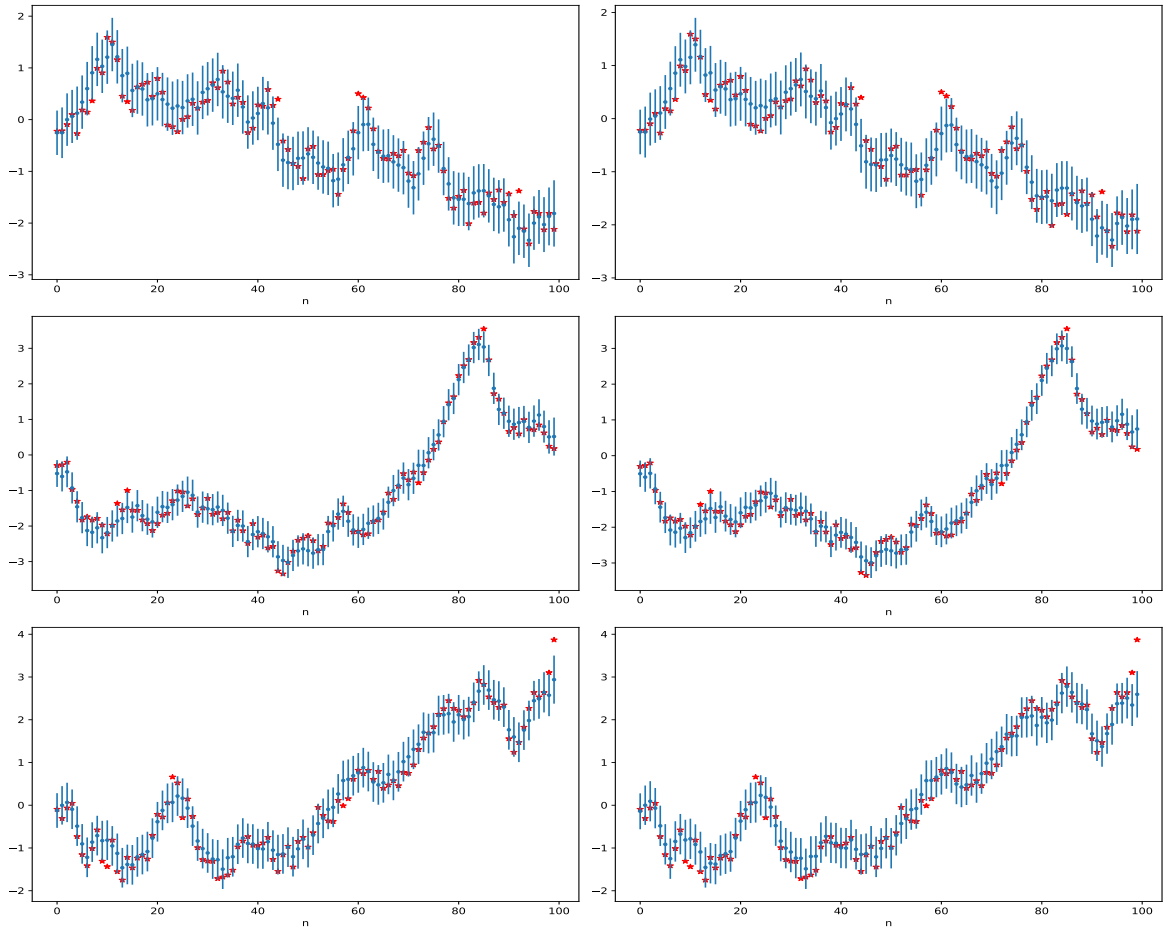


Figure 6: Example of smoothed states when the dimension of the state space is 3 and the dimension of the observations is 4. Left column: component-wise (from top to bottom) smoothed states with true parameters θ. Right column: same thing with learnt parameters λ. Red stars: true state components. Blue dots: smoothed marginal means of each component. Blue vertical lines: 95% confidence regions built from the smoothed marginal variances of each component. The horizontal axis is the time axis.