Journal of Machine Learning Research 25 (2024) 1-33 Submitted 12/22; Published 01/24

Additive smoothing error in backward variational inference for general state-space models

Mathis Chagneux M.CHAGNEUX@GMAIL.COM
LTCI
Télécom Paris
Palaiseau, France

Élisabeth Gassiat ELISABETH.GASSIAT@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS
Laboratoire de mathématiques d'Orsay

Pierre Gloaguen PIERRE.GLOAGUEN@UNIV-UBS.FR
Universit Bretagne-Sud
Lorient, France

Sylvain Le Corff SYLVAIN.LE_CORFF@SORBONNE-UNIVERSITE.FR
LPSM
Sorbonne Université, UMR CNRS 8001
Paris, France

Editor: Anthony Lee

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 under mixing assumptions that 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. We illustrate our theoretical result with state-of-the art variational solutions based both on the backward parameterization and on alternatives using forward decompositions. This numerical study proposes guidelines for variational inference based on neural networks in state-space models.

Keywords: Variational inference, State-space models, Smoothing, Backward decomposition, State inference

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 dimensionality 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 contrast, sequential variational methods rely on tractable approximations of the smoothing distribution to compute these expectations. In the sequential context, a salient aspect of these approaches is the parameterization of the approximations which is mostly left as an implementation choice, despite the existing dependencies in the generative model. Some works introduce structured variational families by considering a variational smoothing distribution which is a product of approximate filtering distributions (e.g. Marino et al. (2018)). In this setting, the posterior dependencies are therefore misspecified. Recently, Bayer et al. (2021) highlighted the detrimental impact of misspecifyng dependencies on observations in the sequential setting. From a slightly different point of view, in the literature of message passing, some authors do not focus on an explicit form for the variational posterior, but rather propose a variational approximation of the distributions used in the classical forward-backward recursions for state space models (Johnson et al., 2016; Lin et al., 2018), leading to a fully conjugated framework giving promising results. In Krishnan et al. (2017), the authors specify a structured variational posterior satisfying the actual dependencies of the true smoothing distributions using a forward factorization, where a bi-directional recurrent neural network (RNN) is used to cover all temporal dependencies of the forward factorization. More recently, Campbell et al. (2021) proposed a new variational family which uses another factorization of the smoothing distributions, the *backward factorization*. This factorization has the appealing property of involving a product of distributions of latent variables that do not depend on future observations. We focus on the latter which we view as the most theoretically grounded given the SSM litterature, as well as the most

computationally appealing. Indeed, it is the only structured variational posterior satisfying the dependencies of the actual smoothing distribution, but which is also prone to online state estimation and parameter learning.

In this paper, we establish upper bounds for the error of the variational approximation of additive smoothing in state-space-models (see in particular Proposition 1 and Proposition 3), when the target expectations are approximated by expectations under a variational distribution satisfying the backward factorization of Campbell et al. (2021). The backward factorization of the variational posterior allows the decomposition of the global error into a sum of terms that can be controlled. To the best of our knowledge, these are the first theoretical results providing upper bounds on the state estimation error when using the latter, or in fact any variational posterior approximation (mean field or involving dependencies) in state-space models. This result is obtained in the context of a fixed sized sequence of observations, but leads to open questions in the context of online learning.

These theoretical results are empirically validated with various numerical experiments which also explore several choices of variational kernels. We consider linear and Gaussian state spaces to illustrate the linear growth as the ground truth can be computed in this case. We also use the backward variational approach in the case of nonlinear emission densities and compare it to sequential Monte Carlo smoothers and other state-of-the-art variational estimators. We finally explore the impact of the backward parametrization with nonlinear hidden dynamics and non-Gaussian observation noise in the framework proposed by Zhao et al. (2022).

In Section 2, we present the general background for SSMs and variational estimation using backward decompositions. 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. As a by-product, we also obtain an upper-bound which is uniform in time for the estimation error of the marginal smoothing expectations. In Section 4, we illustrate our theoretical results using a variety of numerical implementations for backward variational smoothing distributions, which we then illustrate for different generative models.

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* $\phi^\theta_{k_1:k_2}$ 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:

$$\begin{aligned}\phi^\theta_{0:n}h &= \mathbb{E}^\theta\left[h\left(X_{0:n}\right)|Y_{0:n}\right] \\ &= \mathsf{L}^\theta_n(Y_{0:n})^{-1}\int h(x_{0:n})\chi^\theta(x_0)g^\theta_0(x_0,Y_0)\prod_{k=0}^{n-1}\ell^\theta_k(x_k,x_{k+1})\mu(\mathrm{d}x_{0:n})\,,\end{aligned} \tag{1}$$

where[1]

$$\ell^\theta_k(x_k,x_{k+1}) = m^\theta_k(x_k,x_{k+1})g^\theta_{k+1}(x_{k+1},Y_{k+1})$$

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

$$\mathsf{L}^\theta_n(Y_{0:n}) = \int \chi^\theta(x_0)g^\theta_0(x_0,Y_0)\prod_{k=0}^{n-1}\ell^\theta_k(x_k,x_{k+1})\mu(\mathrm{d}x_{0:n})\,. \tag{2}$$

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 $\tilde{h}_k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{\mathsf{d}}$. Such functions are of great importance in many learning tasks. For instance, when θ is known, marginal state inference often boils down to estimate $\mathbb{E}^\theta[X_k|Y_{0:n}]$ which corresponds to $\tilde{h}_k(x_k,x_{k+1}) = x_k$ and $\tilde{h}_\ell(x_\ell,x_{\ell+1}) = 0$ for $\ell \neq k$. When θ is unknown, the MLE can be obtained through an Expectation Maximization (EM) algorithm. This algorithm relies on the computation of the intermediate quantity

$$\theta \mapsto Q(\theta,\theta') = \mathbb{E}^{\theta'}\left[\sum_{k=0}^{n-1}\log\ell^\theta_k(X_k,X_{k+1})\middle|Y_{0:n}\right],$$

which is another additive smoothing example where $\tilde{h}_k(x_k,x_{k+1}) = \log\ell^\theta_k(x_k,x_{k+1})$. As an alternative to EM, batch and recursive MLE (RMLE) methods express

$$\nabla_\theta\log\mathsf{L}^\theta_n = \mathbb{E}^\theta\left[\sum_{k=0}^{n-1}\nabla_\theta\log\ell^\theta_k(X_k,X_{k+1})\middle|Y_{0:n}\right]$$

via Fisher's identity under some regularity conditions (see Cappé et al. (2005), Chap. 10), in which case $\tilde{h}_k(x_k,x_{k+1}) = \nabla_\theta\log\ell^\theta_k(x_k,x_{k+1})$.

1. Note that the dependence of ℓ^θ_k on Y_{k+1} is omitted in the notation for better clarity.

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., 2022; Martin et al., 2023), 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 $\phi^\theta_{0:n}h$ (or of the conditional distribution of the states given the observations), the conditional law $\phi^\theta_{0:n}$ of $X_{0:n}$ given $Y_{0:n}$ is approximated by choosing a candidate in a parametric family $\{q^\lambda_{0:n}\}_{\lambda\in\Lambda}$, 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:

$$\mathcal{L}(\theta,\lambda) = \mathbb{E}_{q^\lambda_{0:n}}\left[\log \frac{p^\theta_{0:n}(X_{0:n},Y_{0:n})}{q^\lambda_{0:n}(X_{0:n})}\right] = \int \log \frac{p^\theta_{0:n}(x_{0:n},Y_{0:n})}{q^\lambda_{0:n}(x_{0:n})} q^\lambda_{0:n}(x_{0:n})\mu(\mathrm{d}x_{0:n}) \;, \quad (4)$$

where $p^\theta_{0:n}$ 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 $q^\lambda_{0:n}$ 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.

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 by ϕ^θ_k. One known factorization of $\phi^\theta_{0:n}$ exists by further introducing 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 $x_k \mapsto m^\theta_k(x_k,x_{k+1})\phi^\theta_k(x_k)$. 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*:

$$\phi^\theta_{0:n}(x_{0:n}) = \phi^\theta_n(x_n)\prod_{k=1}^{n}\frac{m^\theta_{k-1}(x_{k-1},x_k)\phi^\theta_{k-1}(x_{k-1})}{\int m^\theta_{k-1}(x,x_k)\phi^\theta_{k-1}(x)\mu(\mathrm{d}x)} \;.$$

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 $\phi^\theta_{0:k+1}$ only amounts

to computing ϕ^θ_{k+1} 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

$$q^\lambda_{0:n}(x_{0:n}) = q^\lambda_n(x_n) \prod_{k=1}^{n} q^\lambda_{k-1|k}(x_k, x_{k-1}) \,, \tag{5}$$

where q^λ_n (resp. $q^\lambda_{k-1|k}(x_k, \cdot)$) 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

3.1 Assumption and main result

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

$$\mathbf{L}^\theta_k(x_k, \mathrm{d}x_{k+1}) = m^\theta_k(x_k, x_{k+1}) g^\theta_{k+1}(x_{k+1}, Y_{k+1}) \mu(\mathrm{d}x_{k+1}) \,.$$

For additive functionals as in (3), the error between the target expectation $\phi^\theta_{0:n} h_{0:n}$ and its approximation $q^\lambda_{0:n} h_{0:n}$ can be upper bounded by controlling the bias in the estimation of $\mathbf{L}^\theta_k$ by the approximated model, see for instance Gloaguen et al., 2022. In the context of this paper, as the true model is defined by the forward distributions of X_k given X_{k-1}, and the variational approximation is defined by the backward distributions of X_{k-1} given X_k, we reformulate the discrepancy between the true model and the variational one as follows.

For all sequences of probability densities $\{\tilde{q}_k\}_{0 \leqslant k \leqslant n-1}$ with respect to μ, with the condition $\tilde{q}_n = q^\lambda_n$ with q^λ_n defined in (5), let $\tilde{\nu}^\lambda_{k-1:k}$ and $\tilde{\phi}^\theta_{k-1:k}$ be the distributions on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ defined, for all bounded measurable functions h on $\mathbb{R}^d \times \mathbb{R}^d$, by

$$\tilde{\nu}^\lambda_{k-1:k} h = \tilde{q}_k q^\lambda_{k-1|k} h = \int \tilde{q}_k(x_k) q^\lambda_{k-1|k}(x_k, x_{k-1}) h(x_{k-1}, x_k) \mu(\mathrm{d}x_{k-1}, \mathrm{d}x_k) \,,$$

$$\tilde{\phi}^\theta_{k-1:k} h = \frac{\tilde{q}_{k-1} \mathbf{L}^\theta_{k-1} h}{\tilde{q}_{k-1} \mathbf{L}^\theta_{k-1} \mathbb{1}} = \int \frac{\tilde{q}_{k-1}(x_{k-1}) \ell^\theta_{k-1}(x_{k-1}, x_k) h(x_{k-1}, x_k)}{\int \tilde{q}_{k-1}(u_{k-1}) \ell^\theta_{k-1}(u_{k-1}, u_k) \mu(\mathrm{d}u_{k-1}, \mathrm{d}u_k)} \mu(\mathrm{d}x_{k-1}, \mathrm{d}x_k) \,.$$

The discrepancy between these sequences of joint distributions is then defined with:

$$\tilde{c}_0(\theta) = \left\| \tilde{q}_0 - \phi^\theta_0 \right\|_{\mathrm{tv}} , \quad \text{and for all } k \geqslant 1 \quad \tilde{c}_k(\theta, \lambda) = \left\| \tilde{\phi}^\theta_{k-1:k} - \tilde{\nu}^\lambda_{k-1:k} \right\|_{\mathrm{tv}} \,, \tag{6}$$

where $\|\cdot\|_{\mathrm{tv}}$ is the total variation norm, and for all bounded measurable function h, $\phi^\theta_0 h = \chi^\theta g^\theta_0 h / \chi^\theta g^\theta_0 \mathbb{1}$. Note that for $k \geqslant 1$, $\tilde{c}_k(\theta, \lambda)$ depends on both $\tilde{q}_k$ and $\tilde{q}_{k+1}$.

H1 There exist constants $0 < \sigma_- < \sigma_+ < \infty$ such that for all $k \in \mathbb{N}$, $\theta \in \Theta$, $\lambda \in \Lambda$ and $(x_k, x_{k+1}) \in \mathbb{R}^d \times \mathbb{R}^d$,

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

and

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

Proposition 1 *Assume that H1 holds. Then, for all $n \in \mathbb{N}$, $\theta \in \Theta$, $\lambda \in \Lambda$, and all additive functionals $h_{0:n}$ as in* (3)*, and all probability densities $\tilde{q}_k$, $0 \leqslant k \leqslant n-1$, with the condition $\tilde{q}_n = q_n^\lambda$,*

$$\left|q_{0:n}^\lambda h_{0:n} - \phi_{0:n}^\theta h_{0:n}\right| \leq 2\frac{\sigma_+}{\sigma_-}\sum_{k=0}^{n-1}\left\|\tilde{h}_k\right\|_\infty \times \left(\tilde{c}_0(\theta) + \sum_{m=1}^{k}\rho^{k-m+1}\tilde{c}_m(\theta,\lambda) + \tilde{c}_{k+1}(\theta,\lambda) + \sum_{m=k+2}^{n}\rho^{m-k-1}\tilde{c}_m(\theta,\lambda)\right),$$

with $\rho = 1-\sigma_-/\sigma_+$, where σ_- and σ_+ are defined in H1, and $\tilde{c}_0(\theta)$ and $\tilde{c}_m(\theta,\lambda)$, $1 \leqslant m \leqslant n$ are defined in (6)*.*

Proof The proof is postponed to Appendix A. ∎

Marginal smoothing distributions are also of utmost importance as they appear in many applications for state estimation problems. These marginal smoothing expectations can be obtained as special cases of expectations of additive functionals, i.e. cases where $\tilde{h}_j = 0$ for all $j \neq k_\star$, for some $0 \leqslant k_\star \leqslant n-1$.

Corollary 2 *Assume that H1 holds. Then, for all $n \in \mathbb{N}$, $1 \leqslant k_\star \leqslant n-1$, $\theta \in \Theta$, $\lambda \in \Lambda$, all bounded measurable functions $\tilde{h}_{k_\star}$ on $\mathbb{R}^d \times \mathbb{R}^d$, and all probability densities $\tilde{q}_k$, $0 \leqslant k \leqslant n-1$, with the condition $\tilde{q}_n = q_n^\lambda$,*

$$\left|q_{0:n}^\lambda \bar{h}_{k_\star} - \phi_{0:n}^\theta \bar{h}_{k_\star}\right| \leq 2\frac{\sigma_+}{\sigma_-}\left\|\tilde{h}_{k_\star}\right\|_\infty \times \left(\tilde{c}_0(\theta) + \sum_{m=1}^{k}\rho^{k-m+1}\tilde{c}_m(\theta,\lambda) + \tilde{c}_{k+1}(\theta,\lambda) + \sum_{m=k+2}^{n}\rho^{m-k-1}\tilde{c}_m(\theta,\lambda)\right),$$

with $\bar{h}_{k_\star} : x_{0:n} \mapsto \tilde{h}_{k_\star}(x_{k_\star}, x_{k_\star+1})$, $\rho = 1-\sigma_-/\sigma_+$, where σ_- and σ_+ are defined in H1, and $\tilde{c}_0(\theta)$ and $\tilde{c}_m(\theta,\lambda)$, $1 \leqslant m \leqslant n$ are defined in (6)*.*

Note that if there exists c_+ such that for all $\theta \in \Theta$, $\lambda \in \Lambda$, $0 \leq m \leq n$, $\tilde{c}_m(\theta,\lambda) \leq c_+(\theta,\lambda)$, by Corollary 2

$$\left|q_{0:n}^\lambda \bar{h}_{k_\star} - \phi_{0:n}^\theta \bar{h}_{k_\star}\right| \leq 4\frac{\sigma_+}{\sigma_-}\left\|\tilde{h}_{k_\star}\right\|_\infty c_+(\theta,\lambda)\left(1+\frac{\rho}{1-\rho}\right),$$

so that the marginal smoothing errors are uniformly bounded in time.

Proof The proof is postponed to Appendix A. ∎

For all $1 \leqslant k \leqslant n$, let $b_{k-1|k}^\theta$ be the backward kernel at time k, defined for all bounded measurable functions h on $\mathbb{R}^d$ and all $x_k \in \mathbb{R}^d$, by

$$b_{k-1|k}^\theta h(x_k) = \frac{\int m_{k-1}^\theta(x_{k-1}, x_k)\phi_{k-1}^\theta(x_{k-1})h(x_{k-1})\mu(\mathrm{d}x_{k-1})}{\int m_{k-1}^\theta(x, x_k)\phi_{k-1}^\theta(x)\mu(\mathrm{d}x)}.$$

When the backward variational kernel is a sharp approximation of the true backward kernel, Proposition 3 provides an explicit control of the smoothing error.

Proposition 3 *Assume H1 holds. Let $n \in \mathbb{N}$, $\theta \in \Theta$, $\lambda \in \Lambda$. Assume that there exists $\varepsilon > 0$ such that $\left\|q_n^\lambda - \phi_n^\theta\right\|_{\mathrm{tv}} \leqslant \varepsilon$ and for all $1 \leqslant k \leqslant n$, $x_k \in \mathbb{R}^d$, $\left\|q_{k-1|k}^\lambda(x_k, \cdot) - b_{k-1|k}^\theta(x_k, \cdot)\right\|_{\mathrm{tv}} \leqslant \varepsilon$. Then, for all additive functionals $h_{0:n}$ as in (3),*

$$\left|q_{0:n}^\lambda h_{0:n} - \phi_{0:n}^\theta h_{0:n}\right| \leq 4\frac{\sigma_+}{\sigma_-}\left(1 + 2\frac{\rho}{1-\rho}\right)\sum_{k=0}^{n-1}\left\|\tilde{h}_k\right\|_\infty \varepsilon \ ,$$

where $\rho = 1 - \sigma_-/\sigma_+$, with σ_- and σ_+ defined in H1. Therefore, in the case where $\sup_{0\leqslant k\leqslant n-1}\|\tilde{h}_k\|_\infty \leqslant M$ for some $M \geqslant 0$, there exists $c \geqslant 0$ such that

$$\left|q_{0:n}^\lambda h_{0:n} - \phi_{0:n}^\theta h_{0:n}\right| \leq cn\varepsilon \ .$$

Proof The proof amounts to applying Proposition 1 with for all $0 \leqslant k \leqslant n-1$, $\tilde{q}_k = \phi_k^\theta$.

- $\tilde{c}_0(\theta) = \left\|\tilde{q}_0 - \phi_0^\theta\right\|_{\mathrm{tv}} = 0$, as $\tilde{q}_0 = \phi_0^\theta$.
- For all $1 \leqslant m \leqslant n-1$,

$$\begin{aligned}
\tilde{c}_m(\theta, \lambda) &= \left\|\tilde{\phi}_{m-1:m}^\theta - \tilde{\nu}_{m-1:m}^\lambda\right\|_{\mathrm{tv}} \ , \\
&= \left\|\frac{\tilde{q}_{m-1}\mathbf{L}_{m-1}^\theta}{\tilde{q}_{m-1}\mathbf{L}_{m-1}^\theta \mathbb{1}} - \tilde{q}_m q_{m-1|m}^\lambda\right\|_{\mathrm{tv}} \ , \\
&\leqslant \left\|\frac{\phi_{m-1}^\theta \mathbf{L}_{m-1}^\theta}{\phi_{m-1}^\theta \mathbf{L}_{m-1}^\theta \mathbb{1}} - \phi_m^\theta b_{m-1|m}^\theta\right\|_{\mathrm{tv}} + \left\|\phi_m^\theta b_{m-1|m}^\theta - \phi_m^\theta q_{m-1|m}^\lambda\right\|_{\mathrm{tv}} \leqslant \varepsilon \ ,
\end{aligned}$$

  where the first term in last inequality is zero as $\phi_{m-1}^\theta \mathbf{L}_{m-1}^\theta / \phi_{m-1}^\theta \mathbf{L}_{m-1}^\theta \mathbb{1}$ and $\phi_m^\theta b_{m-1|m}^\theta$ are both equal to the probability density of (X_{m-1}, X_m) given $Y_{0:m}$ under the law of the state-space model parameterized by θ.

- The last term is upper-bounded as follows:

$$\begin{aligned}
\tilde{c}_n(\theta, \lambda) &= \left\|\tilde{\phi}_{n-1:n}^\theta - \tilde{\nu}_{n-1:n}^\lambda\right\|_{\mathrm{tv}} \ , \\
&= \left\|\frac{\tilde{q}_{n-1}\mathbf{L}_{n-1}^\theta}{\tilde{q}_{n-1}\mathbf{L}_{n-1}^\theta \mathbb{1}} - q_n^\lambda q_{n-1|n}^\lambda\right\|_{\mathrm{tv}} \ , \\
&\leqslant \left\|\frac{\phi_{n-1}^\theta \mathbf{L}_{n-1}^\theta}{\phi_{n-1}^\theta \mathbf{L}_{n-1}^\theta \mathbb{1}} - \phi_n^\theta b_{n-1|n}^\theta\right\|_{\mathrm{tv}} + \left\|\phi_n^\theta b_{n-1|n}^\theta - \phi_n^\theta q_{n-1|n}^\lambda\right\|_{\mathrm{tv}} \\
&\qquad + \left\|\phi_n^\theta q_{n-1|n}^\lambda - q_n^\lambda q_{n-1|n}^\lambda\right\|_{\mathrm{tv}} \leqslant 2\varepsilon \ ,
\end{aligned}$$

  where the first term in last inequality is zero as $\phi_{n-1}^\theta \mathbf{L}_{n-1}^\theta / \phi_{n-1}^\theta \mathbf{L}_{n-1}^\theta \mathbb{1}$ and $\phi_n^\theta b_{n-1|n}^\theta$ are both equal to the probability density of (X_{n-1}, X_n) given $Y_{0:n}$ under the law of the state-space model parameterized by θ.

■

Remark 4 *By Proposition 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 $\theta \in \Theta$, $\lambda \in \Lambda$, $0 \leq m \leq n$, $\tilde{c}_m(\theta, \lambda) \leq c_+(\theta, \lambda)$ then*

$$\left|q^\lambda_{0:n} h_{0:n} - \phi^\theta_{0:n} h_{0:n}\right| \leq 4\frac{\sigma_+}{\sigma_-}\left(1 + \frac{\rho}{1-\rho}\right) c_+(\theta, \lambda) h_\infty n \,. \tag{7}$$

Remark 5 *Proposition 1 provides a criterion for assessing the sharpness of a variational approximation for $\phi^\theta_{0:n}$. Indeed, for such approximation, write*

$$c_{\inf}(\lambda, \theta) = \inf_{(\tilde{q}_k)_{0\leq k\leq n}} \sum_{k=0}^{n-1} \left(\tilde{c}_0(\theta) + \sum_{m=1}^{k} \rho^{k-m+1} \tilde{c}_m(\theta, \lambda) + \tilde{c}_{k+1}(\theta, \lambda) + \sum_{m=k+2}^{n} \rho^{m-k-1} \tilde{c}_m(\theta, \lambda) \right) .$$

Then, if there exists h_∞ such that for all $0 \leq k \leq n-1$, $\|\tilde{h}_k\|_\infty \leq h_\infty$, by Proposition 1, we have:

$$\left|q^\lambda_{0:n} h_{0:n} - \phi^\theta_{0:n} h_{0:n}\right| \leq 2\frac{\sigma_+}{\sigma_-} c_{\inf}(\theta, \lambda) h_\infty. \tag{8}$$

Although difficult to compute in practice, this criterion might be the focus of future research. An open question here is whether the optimal sequence $(\tilde{q}_k)_{0\leq k\leq n}$ is given by the sequence of true marginal smoothing distributions.

3.2 Comments on Proposition 1 and H1

Proposition 1 provides an upper-bound for the smoothing error for additive functionals which is linear in the number of observations. The sharpness of this bound depends on our ability to find a sequence of distributions $(\tilde{q}_k)_{0\leqslant k\leqslant n-1}$, so that each $c_k(\theta, \lambda)$, i.e., the total variation distance between $(x_{k-1}, x_k) \mapsto \tilde{q}_k(x_k) q^\lambda_{k-1|k}(x_k, x_{k-1})$ and the probability density proportional to $(x_{k-1}, x_k) \mapsto \tilde{q}_{k-1}(x_{k-1}) \ell^\theta_{k-1}(x_{k-1}, x_k)$, is small.
First, 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 $(\tilde{q}_k)_{k\geqslant 1}$ achieving $\tilde{c}_k(\theta, \lambda) = 0$ for all k is the sequence of true filtering distributions.
However, in generic cases (i.e. non linear gaussian cases), this joint minimization over this sequence of distributions appears to be an open challenge. In Section 4.2, we discuss empirically how the backward $q^\lambda_{k-1|k}(x_k, x_{k-1})$ can be parameterized by the user, depending on the form of $\ell^\theta_{k-1}(x_{k-1}, x_k)$ (see the experiments related to the results of Figure 2a).

Obtaining theoretical guarantees on the variational approximations remains of course an open problem but we believe that Proposition 1 provides a first result in this direction.

*About **H1**.* 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 in the context of smoothing, (see Douc et al., 2011; Dubarry and Le Corff, 2013; Olsson et al., 2017; Gloaguen et al., 2022. It is worth noting that in the context of approximating the filtering distributions, weaker assumptions exist (see Chigansky and Liptser, 2004; Douc et al., 2009), but the extension of these results to the smoothing context remains an open challenge.

Data set: We use "data set" rather than "dataset". d by the filtering recursions, from which the backward kernels are derived. In contrast, a recursion for $(q_k^\lambda)_{k\geq 0}$ is neither *a priori* defined by the factorization 5 nor easily derived from it. Suppose however that we want to recursively build $(q_{0:k}^\lambda)_{k\geq 0}$ with observations $(Y_k)_{k\geq 0}$.

4. Numerical experiments

We now present some practical examples of implementations of the backward variational factorization on which we validate our theoretical results.

4.1 Linear Gaussian SSMs

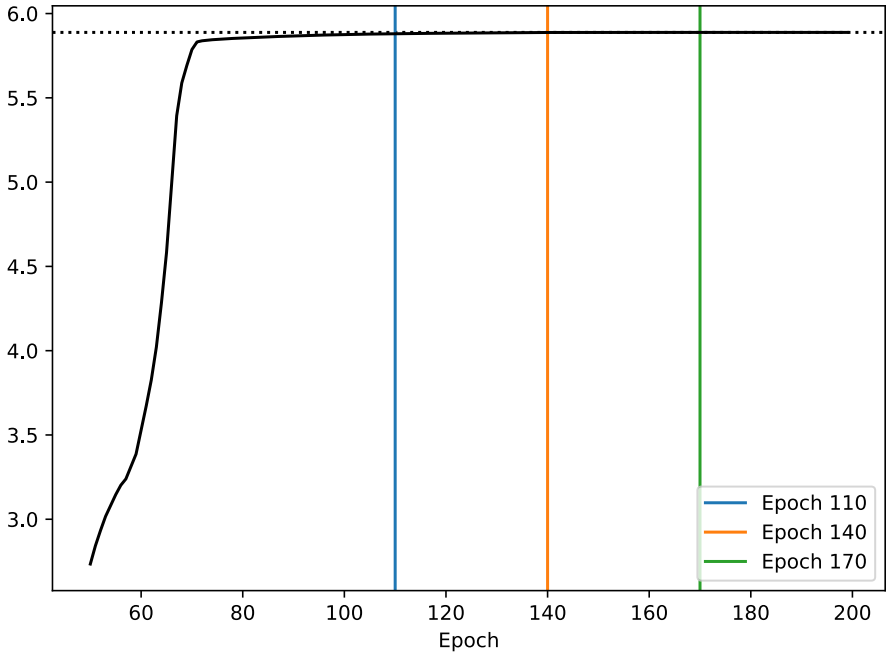
A first interesting case is when the variational family *contains* the true model. This is in particular possible when the latter is a linear and Gaussian SSM, i.e. when χ^θ (resp. $m_k^\theta(X_k,\cdot)$ and $g_k^\theta(X_k,\cdot)$) are densities of Gaussian distributions with mean A_0 (resp. AX_k and BX_k) and variance Q_0 (resp. Q and R), such that $\theta = (A_0, Q_0, A, Q, B, R)$. If we define a similar "mirror" model described with another set of parameters $\lambda = (\bar{A}_0, \bar{Q}_0, \bar{A}, \bar{Q}, \bar{B}, \bar{R})$, we can choose $q_n^\lambda \sim \mathcal{N}(\mu_n, \Sigma_n)$ where (μ_n, Σ_n) are provided by the Kalman filtering recursions, and $q_{k-1|k}^\lambda(x_k, x_{k-1}) \sim \mathcal{N}(A_{k-1|k}x_k + b_{k-1|k}, \Sigma_{k-1|k})$ where $(A_{k-1|k}, b_{k-1|k}, \Sigma_{k-1|k})$ are obtained through Kalman smoothing steps. In this case, $q_{0:n}^\lambda$ is of the same form as $\phi_{0:n}^\theta$ and $q_{0:n}^\lambda = \phi_{0:t}^\theta$ when $\lambda = \theta$.

When the latter case is reached, Section 3.2 shows that $c_k(\theta, \lambda) = 0$ for all k, suggesting that the additive error vanishes. In this section, we study the case where the parameter θ is known, $d = 5$ and λ is trained on a set of sequences of $n = 50$ observations. The evolution of the ELBO is given in Figure 1a. In Figure 1b, we depict the controlled term of Proposition 1 in the case of state estimation, i.e. for $h_{0:n} : x_{0:n} \mapsto \sum_{k=0}^n x_k$. This evaluation is performed on $J = 50$ evaluation sequences $(Y_{0:n}^j)_{1\leq j\leq J}$ of length $n = 500$ sampled from the generative model. Each plot 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. Additionally, for a given λ, slopes vary across sequences, which highlights the dependency of $(c_k(\theta, \lambda))_{0\leq k\leq n}$ on the observations.

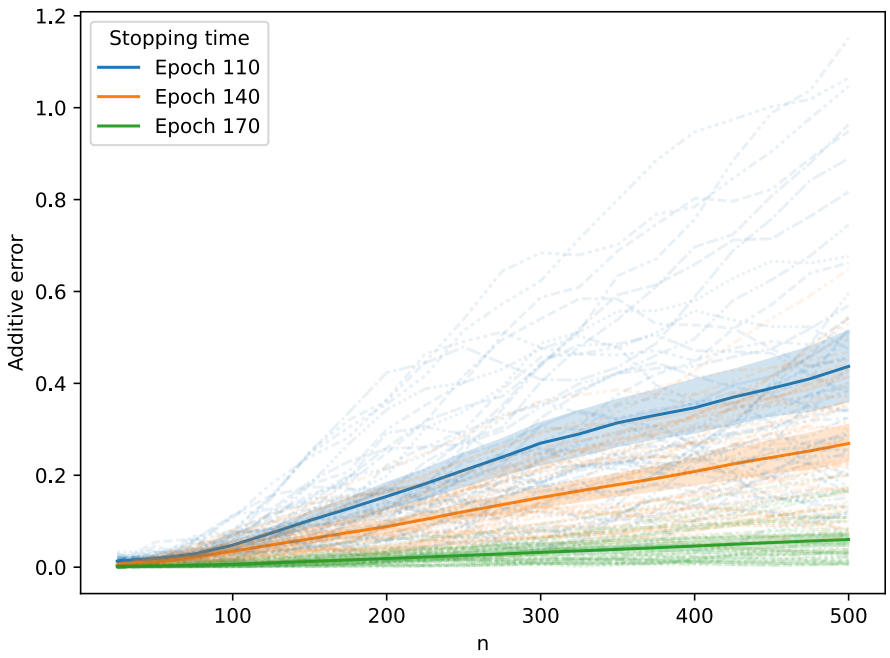
In the appendix, we provide more implementation details, as well as additional figures for the errors on the marginal distributions.

4.2 Nonlinear SSMs

The primary motivation to use variational inference is when $\phi_{0:n}^\theta$ cannot be computed analytically, which generally happens when the generative model contains nonlinearities and/or non-Gaussian noises. In this case - contrary to the previous section - there is no obvious choice for the form of the kernels in $q_{0:n}^\lambda$ and many options exist to balance the amount of approximation with the computational complexity. In the next subsections, we revisit some of the literature on sequential variational inference in the backward context to illustrate our theoretical result.



(a) L_n^θ (dotted line) and $\lambda \mapsto \mathcal{L}(\theta, \lambda)$ over epochs (full line).



(b) $\left|q_{0:n}^\lambda h_{0:n} - \phi_{0:n}^\theta h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$. The solid lines display the mean over the 50 independent replicates, the transparent filling is the standard deviation, shaded lines are the all sequences. Values are normalized by the state-space dimension.

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 on 50 different sequences (right)

4.2.1 Nonlinearity in the emission distribution

We first consider a generative model where the prior distribution and transition kernels are still linear, but $g_k^\theta(X_k, \cdot)$ is the Gaussian probability density with mean $d^\theta(X_k)$ and variance R, d^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 d^θ for identifiable state estimation. In particular d^θ need not to be an injective mapping and therefore we use an unconstrained and arbitrary multi layer perceptron (MLP).

In this context, Hälvä et al. (2021) obtained promising results via a parameterization of the factors in $q^\lambda_{0:n}$ which relies entirely on Gaussian conjugation and can be analytically marginalized, therefore allowing fast inference. A central element of their approximation is the idea from Johnson et al. (2016), which consists in mapping each observation y_k to a set of valid natural parameters (κ_k, Π_k) for some Gaussian distribution, using an *encoder* network r^λ such that $(\kappa_k, \Pi_k) = r^\lambda(y_k)$. By defining (as in Section 4.1) some additional parameters $(\bar{A}_0, \bar{Q}_0, \bar{A}, \bar{Q})$ for kernels $\chi_0^\lambda, m_k^\lambda$ (i.e. similar to the generative model but parameterized by λ) the authors design $q^\lambda_{0:n}$ using forward-backward recursions (see (Cappé et al., 2005, section 3.2.1)) where the forward and backward variables are updated analytically by Gaussian conjugation with the exponential factors $x_k \mapsto e^{\langle r^\lambda(y_k), t_{\mathcal{N}}(x_k)\rangle}$, $t_{\mathcal{N}}(x_k) = (x_k, x_k x_k^\top)$ being the set of sufficient statistics for a Gaussian distribution in x_k. This algorithm is a special form of two-filter smoothing, which is rather rooted in the alternate *forward* decomposition of the joint smoothing distribution, that is $q^\lambda_{0:n}(x_{0:n}) = q_0^\lambda(x_0) \prod_{k=1}^{n-1} q^\lambda_{k|k-1}(x_{k-1}, x_k)$ where each factor depends on the entire sequence of observations $y_{0:n}$ and is built using the so-called backward *variables* (which are non-normalized quantities distinct to the backward kernels). However, the core idea can be reframed under the backward factorisation very easily by defining a sequence of distributions $(q_k^\lambda)_{k\leq n}$ which are updated from q^λ_{k-1} to q_k^λ via:

- $\bar{q}_k^\lambda(x_k) = \mathbb{E}_{q^\lambda_{k-1}}\left[m_k^\lambda(\cdot, x_k)\right]$ similarly to a Kalman predict step
- $\eta_k = r^\lambda(y_k) + \bar{\eta}_k^\lambda$ where η_k and $\bar{\eta}_k$ are the natural parameters of q_k^λ and $\bar{q}_k^\lambda$, respectively.

and by defining the backward kernels with $q^\lambda_{k-1|k}(x_k, x_{k-1}) \propto q^\lambda_{k-1}(x_{k-1}) m_k^\lambda(x_{k-1}, x_k)$, such that their parameters are derived analytically at each time step from η_{k-1} and the parameters of m_k^λ. We refer to the models of Johnson et al. (2016) as the *Conjugate Forward* variational model and to the backward adaptation as the *Conjugate Backward* model.

These solutions are computationally very efficient because they allow closed-form updates of the factors with DNN-predicted encodings which are already Gaussian parameters. Under the backward factorization, more general implementations are possible that still allow analytical marginalisation by keeping the factors in (5) conjugated. For example, one may use a recurrent neural network which updates an internal state $(s_k)_{k\leq n}$ from which the backward kernels and the terminal distribution and built analytically via an intermediate linear-Gaussian kernel m_k^λ as before, e.g.

- $s_k = \mathsf{RNN}^\lambda(s_{k-1}, y_k)$ and $q_k^\lambda \sim \mathcal{N}(\mu_k, \Sigma_k)$ where $(\mu_k, \Sigma_k) = \mathsf{MLP}^\lambda(s_k)$
- $q^\lambda_{k-1|k}(x_k, x_{k-1}) \propto q^\lambda_{k-1}(x_{k-1}) m_k^\lambda(x_{k-1}, x_k)$ from which parameters of $q^\lambda_{k-1|k}$ are derived analytically.

We implement such version with a Gated Recurrent Unit (GRU) for the RNN, and refer to it as the *GRU Backward* implementation.

In the nonlinear setting, since the true smoothing distribution $\phi_{0:n}^{\theta}$ has no analytic form, we use the particle-based Forward Filtering Backward Simulation (FFBSi) algorithm as a surrogate for this ground truth. The FFBSi outputs trajectories 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 choose the case $d = 10$, where a high number of particles for the FFBSi (10000 for the bootstrap filtering, 2000 for the backward smoothing) to consider it as a proper ground truth.

We compare the additive error with respect to the FFBSi (i.e. the left hand term of equation (7)) for $h_{0:n} : x_{0:n} \mapsto \sum_{k=0}^{n} x_k$). In appendix, we report the quality of the FFBSi estimator in the form of the sample mean and variance of its error against the true states, which establishes the error made by the oracle reference estimator considered as ground truth.

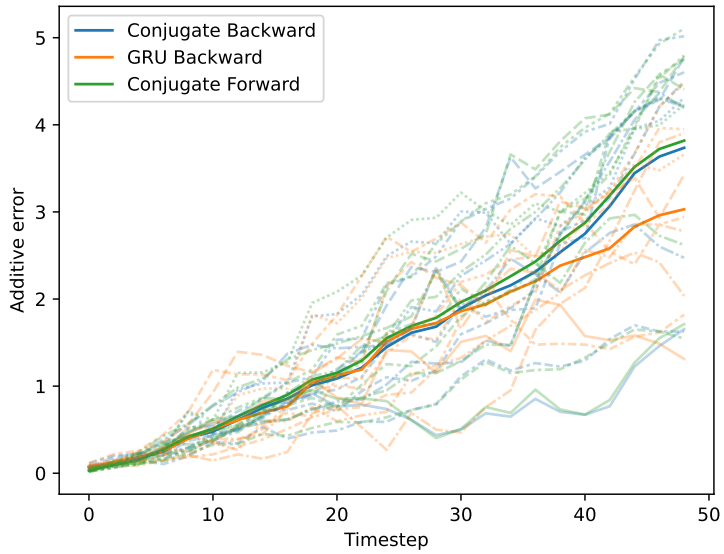
In Figure 2a, we plot the evolution of the additive error against this oracle. As predicted by our theoretical result, all backward methods have a linear dependency in the number of observations n. Interestingly, we observe that the *Conjugate Forward* model also shares this property, which suggests that our main theoretical result is also valid for other factorizations. However, while the two-filter formulation brings similar results using the same amount of parameters, it is much less convenient computationally because it requires to compute the entire sequence of backward variables for any new observation.

One hidden aspect of the fully conjugate models is that the natural parameters given by $r^{\lambda}(y_k)$ implicitly model the distribution of x_k given y_k (unconditionnally on the dynamics), yet this distribution is likely to admit several modes (especially if d^{θ} is strongly injective on some portions of the support). We observe a slight performance gain for the *GRU Backward* model in this context. In this model, the parameters of the intermediate distributions q_k^{λ} are updated without any intermediate Gaussian approximation which might explain the better performance.

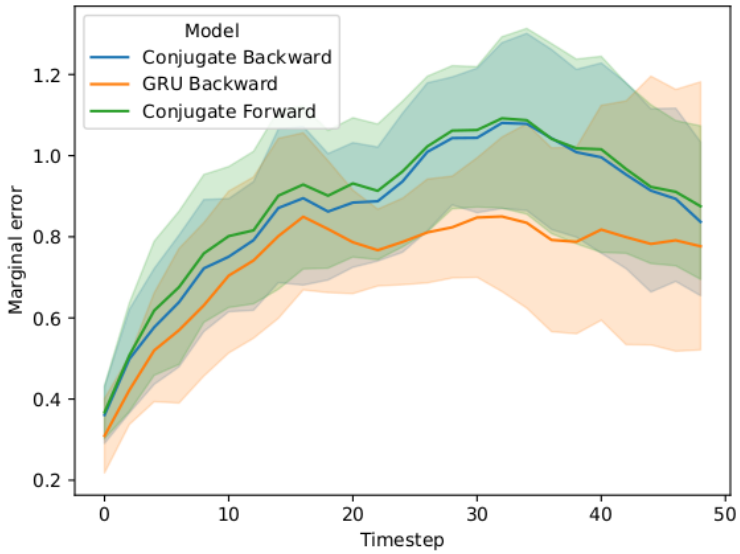
In Figure 2b, we provide the marginal errors over time in the same setting. The results coincide with the time-uniform bound presented in Corollary 2.

4.2.2 Nonlinear hidden dynamics with a non-Gaussian observation noise

We now consider a model introduced in Zhao et al. (2022), where $m_k^{\theta}(x_{k-1}, \cdot)$ is the density of $\mathcal{N}\left(x_{k-1} + \delta\left[\gamma W \tanh(x_{k-1}) - x_{k-1}\right]/\tau, Q\right)$ and g_k^{θ} is the density of a Student-t distribution with mean x_k, ν degrees of freedom and scale R. We start by reproducing this *chaotic recurrent neural network* setting as in Campbell et al. (2021), Section 5.2. That is, we fit the parameter λ on a given sequence $y_{0:n}$ and we evaluate the performance on the same sequence. To assess the variability of the performance, we train and evaluate on $J = 50$ sequences $(y_{0:n}^{(j)})_{1 \leq j \leq J}$, each drawn from a different model with parameter $\theta^{(j)}$, on which we learn a different variational parameter $\lambda^{(j)}$. In Figure 4, we plot the evolution of the error with $d = 5$ and $n = 500$ for both the Conjugate Forward and Conjugate Backward models together with the state-of-the-art online backward smoother of Campbell et al. (2021). Once again, all models show a linear dependency on the observations, which



(a) Smoothing errors $\left|q^{\lambda}_{0:n} h_{0:n} - \phi^{\theta}_{0:n} h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$. The thick solid lines display the mean over the 10 independent replicates for both approaches, shaded lines are single sequences.



(b) Marginal errors $\left(\left|q^{\lambda}_{0:n} h^m_{0:n} - \phi^{\theta}_{0:n} h^m_{0:n}\right|\right)_{m \leq n}$, i.e. for $\tilde{h}^m_k(x_k, x_{k+1}) = x_k \mathbb{1}_{k=m}$. The thick solid lines display the mean over the 10 independent replicates for both approaches, the filling is the standard deviation

Figure 2: Additive and marginal errors in the setting of section 4.2.1 where $\phi^{\theta}_{0:n}$ is obtained by the FFBSi algorithm. All values are normalized by the dimension of the state space. Experiments are produced on 10 independent sequences.

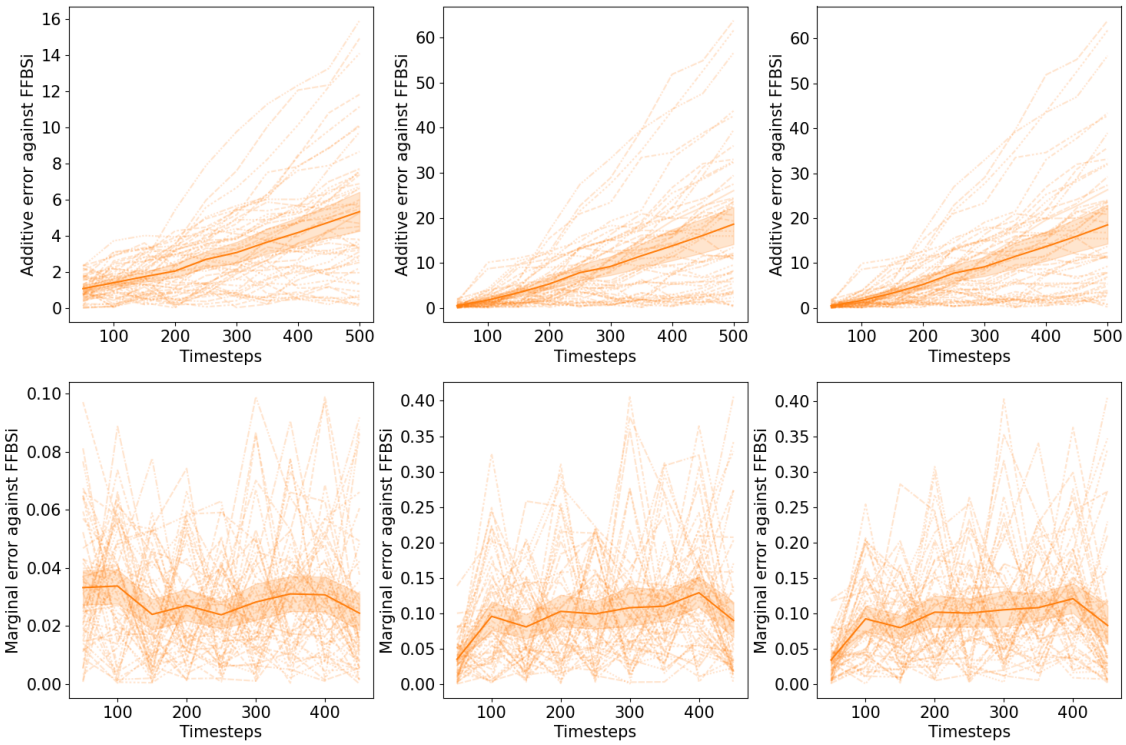


Figure 3: Additive (top) and marginal (bottom) errors against FFBSi estimates on the chaotic data with the *Conjugate Backward* model on three types of functionals, from left to right: (i) $\tilde{h}_k(x_{k-1}, x_k) = ||x_k||_1$ (ii) $\tilde{h}_k(x_{k-1}, x_k) = x_k^T x_k$ (iii) $\tilde{h}_k(x_{k-1}, x_k) = x_{k-1}^T x_k$.

supports our main theoretical claim. In Figure 3, we provide a more thorough analysis of the additive smoothing performance on other moments for the *Conjugate Backward* model by generating more sequences under a single θ and training for more epochs. Again, in this case, the estimates obtain using the FFBSi considered are considered as ground truth. For all moments, we observe the linearity of the additive smoothing error and the uniform bound on the marginal error. We also observe the dependency of $||h_k||_\infty$ through the increased slopes and higher error values for the additive and marginal errors, respectively.

This experiment also highlights an interesting aspect on the impact of the parameterization choices. In the previous sections, training was performed on multiple sequences of fixed length, therefore multiple learning signals are available to learn the terminal distribution q_n^λ (i.e. terminal observations of the sequences in the training set). In the setting of this section, on the contrary, only one data point is available at n. For the offline setting, we therefore do not expect the distributions q_k^λ to be good terminal laws of the subsequences $(y_{0:k})_{k<n}$ under (5). Indeed, except for $k = n$, the parameters of these distributions only appear indirectly during optimization (via their relationship with the backward kernels) when optimization of the joint ELBO is performed at a fixed length n. In contrast, the solution

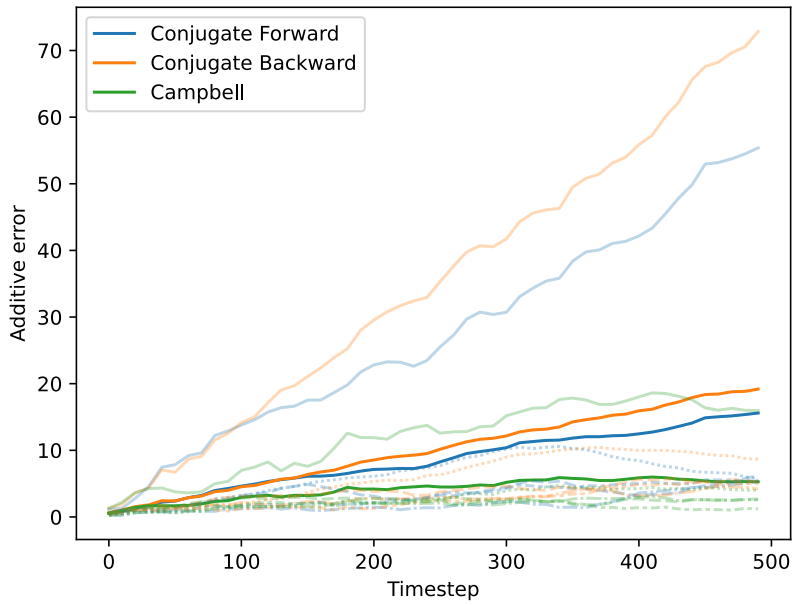


Figure 4: $\left|q^{\lambda}_{0:n}h_{0:n} - \phi^{\theta}_{0:n}h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$ in the setting of section 4.2.2. The solid lines display the mean over the 5 independent replicates which are shown in shaded lines.

of Campbell et al. (2021) explicitly performs gradient-descent on a new set of parameters λ_k at each timestep such that $q^{\lambda}_k = q^{\lambda_k}_k$ is always a good terminal law for $y_{0:k}$. Interestingly, the results for the *Conjugate Forward* and *Conjugate Backward* models - which do not have such regularisation - are only slightly worse than the state-of-the-art, albeit at a much lighter computational cost. Indeed, in practice, Figure 4 is obtained simply by using the distributions q^{λ}_k as terminal laws for $k \leq n$. This suggests that the associated parameterizations may provide good variational *filtering* distributions through the laws q^{λ}_k as a byproduct of the smoothing objective $q^{\lambda}_{0:n}$ with no additional regularisation. In section 5.3, we discuss more extensively the link between our theoretical results and the choice of parameterizations for the variational kernels.

On the contrary, the *GRU Backward* model has a different behaviour. In Figure 5, the dotted blue curve shows that a good approximation of $q^{\lambda}_{0:n}$ is obtained by fitting on $y_{0:n}$, however the associated parameter λ does not provide a good approximation of $(q^{\lambda}_{0:k})_{k<n}$. If we instead learn λ by computing the gradient of the ELBO for increasingly large subsequences $(y_{0:k})_{k\leq n}$ - i.e. mimicking the training scheme of Campbell et al. (2021) - we obtain a different type of approximation, which is suitable for $k < n$, even though this additional constraint results in slightly worse performance for $k = n$. In this case, the results are comparable with those of Campbell et al. (2021).

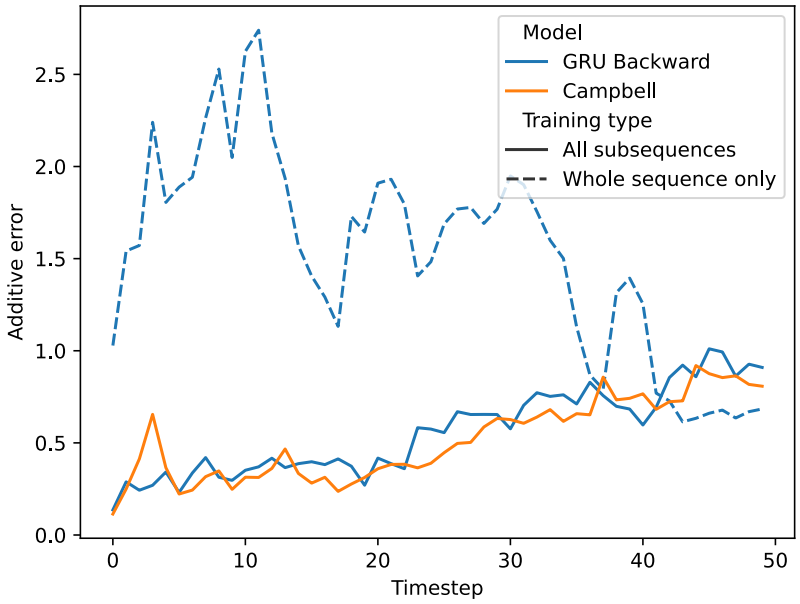


Figure 5: $\left|q^{\lambda}_{0:n}h_{0:n} - \phi^{\theta}_{0:n}h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$ when training the *GRU Backward* model in two different ways, alongside the solution of Campbell et al. (2021)

5. 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 cases to illustrate these results. We have also shown that some existing ideas from litterature on message passing or conjugate graphical models can be reframed to be used under the backward factorization. We believe that our theoretical result sheds light on important properties of sequential variational methods and provides perspectives for future research which we detail in this section.

5.1 Assumptions

The proposed strong mixing assumptions are classical to obtain theoretical guarantees in nonlinear smoothing problems. Weaker assumptions have been proposed in the literature to control filtering distributions. Although these results cannot be extended to smoothing distributions easily, obtaining similar upper bounds as in our contribution with weaker assumptions is an interesting perspective for future works. Our numerical experiments do not restrict to models satisfying these assumptions, suggesting that some relaxations of these classical hypothesis should be investigated.

5.2 Additional theoretical guarantees

- Recently, Tang and Yang (2021) proposed a general theoretical framework for analyzing the excess risk associated with empirical Bayes variational Auto Encoders, covering both parametric and nonparametric cases. The authors study the statistical properties of the VAE estimator using M-estimation theory. In our context of time series, extending the M-estimation theory requires to first analyze the asymptotic behavior of the ELBO. We believe this is another appealing property of the backward decomposition of the variational family, as in this case thee ELBO writes

$$\frac{1}{n}\mathcal{L}_n(\theta,\varphi) = \frac{1}{n}\ell_n(\theta) + \frac{1}{n}\mathbb{E}_{q^\lambda_{0:n}}\left[\log\frac{\phi_{\theta,n}(X_n)}{q^\lambda_n(X_n)}\right] + \frac{1}{n}\sum_{s=1}^{n}\mathbb{E}_{q^\lambda_{0:n}}\left[\log\frac{b_{\theta,s-1|s}(X_{s-1},X_s)}{q^\lambda_{s-1|s}(X_{s-1},X_s)}\right],$$

  where $\phi_{\theta,n}$ is the filtering distribution at time n, $(b_{\theta,s-1|s})_{1\le s\le n}$ are the backward kernels of the true model and $\ell_n(\theta)$ is the loglikelihood of the observations. Using this decomposition and additional assumptions, the limiting behavior of the ELBO can be derived to extend the results of Tang and Yang (2021) to state-space models. However, this requires to obtain the asymptotic behavior of various terms which relies on many technicalities and this is therefore left for future work.

  In addition, the backward factorization offers a suitable framework (combined with strong mixing assumptions and regularity conditions on the state-space model) to satisfy Condition A of Tang and Yang (2021). In an offline learning setting, with fixed n, this provides an interesting perspective to control the total variation distance between the true distribution of the observations and

$$y_{0:n} \mapsto \int \left(\frac{1}{N}\sum_{i=1}^{N} q^\lambda_{0:n}(z_{0:n}|y^i_{0:n})\right)\prod_{k=1}^{n} g^\theta_k(z_k,y_k)\mathrm{d}z_{0:n}\,,$$

where $y_{0:n}^i$, $1 \leq i \leq N$, are i.i.d. sequences with distribution parameterized by θ. These extensions are the focus on the ongoing work Gassiat and Le Corff (2023).

- The linear growth with the number of observations matches the results obtained when the true smoothing distributions are replaced by "skewed" or Monte Carlo estimators. Indeed, using for instance (Gloaguen et al., 2022, Theorem 4.10), we can show that even if the smoothing expectation is computed under the true model but not with the true parameter, the estimation error of the smoothing expectation grows linearly in the number of observations:

$$\left|\phi_{0:n}^{\theta'} h_{0:n} - \phi_{0:n}^{\theta} h_{0:n}\right| \leq c(\theta', \theta) n \,.$$

  Therefore, even if the variational family contains the true model, if the minimization of the ELBO does not recover the true parameter, we recover the upper bound linear in the number of observations.

- Obtaining lower bounds for the estimation error of joint smoothing expectation is an open problem, especially in a variational inference framework. This is also an open problem in variational inference for state-space models. We believe that it also relies on important theoretical results which have not been developed yet for the analysis of variational inference of state space models.

5.3 Variational kernels parameterization

We do not provide *constructive* assumptions on the variational model, i.e. further works may provide more explicitly the form of the optimal variational factors when the variational kernels belong to a parametric family. Obtaining specific conditions on the variational kernels to optimize the upper bound in Proposition 3 is also an open problem. This leaves a lot of room for implementation choices, even when restricted to the backward factorization. As we did however explore several implementations, we now discuss qualitatively their possible impact on performance and the link with our theoretical results.

Amortization. In Section 3, we deliberately do not specify explicitly what λ is. In the offline setting with sequences of fixed length n, our results hold in these two cases.

- $\lambda = (\lambda_0, \ldots, \lambda_n)$ is directly the set of all parameters of the kernels, where λ_k denotes the parameters of the variational terms involved at k (e.g the parameters for the k-th backward kernel, and for $k = n$, the parameters of the terminal distribution q_n^λ). This corresponds to *non-amortized* inference.

- λ is the global (temporally-shared) parameter of a function f^λ which itself outputs the (local) parameters of the variational kernels from observations, i.e. $f^\lambda(y_{0:n}) = (\lambda_1, \ldots, \lambda_n)$. This is usually referred to as *amortized* inference.

One example of non-amortized setting is the implementation of Campbell et al. (2021), while both the *Conjugate Backward, Conjugate Forward, GRU Backward* are amortized implementations. While experiments all show the linear behaviour of the additive error, some elements may be discussed with respect to the assumptions involved in the theoretical

results. In particular, in Proposition 3, the sharpness of the bound on the additive error is linear in ε, where ε is an upper bound for the error between the variational kernels and their counterparts under the true model. As such, minimizing these local distances with a small ε is key to obtaining a low additive error. In the non-amortized scenario, the parameters of the kernels can be individually tuned during minimization of the joint ELBO and independently of each other. Intuitively, this leaves the highest flexibility to minimize local distances $\|q_n^{\lambda_n} - \phi_n^\theta\|_{\mathrm{tv}}$ and $\|q_{k-1|k}^{\lambda_k}(x_k, \cdot) - b_{k-1|k}^\theta(x_k, \cdot)\|_{\mathrm{tv}}$ for all $k \leq n$, $x_k \in \mathbb{R}^d$, under chosen parameteric families for these kernels. One perspective of this work that remains is to analyse quantitatively how these two types of implementation differ in terms of the local distances recalled above, which is not direct since such distances are not readily available explicitely.

Recursions for parameters of the variational kernels. Under the true model, recursions exist that relate the filtering distributions and the backward kernels explicitly, and approximating these recursions is at the core of sequential Bayesian inference algorithms. One question that remains in our study of backward variational methods is whether reproducing similar recursions to build the variational kernels leads to better practical solutions. Again, our results hold irrespective of the dependencies between the parameters of the variational kernels, but experimentally we explored many scenarios. In that respect, experiments of Section 4.2.2 are somehow informative. Indeed, we observe, for example, that the *Conjugate Backward* exhibits the linear additive behaviour for any $k \leq n$ when using the $(q_k^\lambda)_{k \leq n}$ to build the terminal distributions, even when trained on a sequence of fixed length n. Contrarily, the *GRU Backward* does not. In the former implementation, denoting $\psi_k^\lambda : x_k \mapsto e^{\langle r^\lambda(y_k), t_{\mathcal{N}}(x_k) \rangle}$, one has, for all $k \leq n$, $q_k^\lambda \propto \psi_k^\lambda \int m_k^\lambda q_{k-1}^\lambda$ and $q_{k-1|k}^\lambda \propto q_{k-1}^\lambda m_k^\lambda$, which is similar to the true model where $\phi_k^\theta(\cdot) \propto g_k^\theta(\cdot) \int m_k^\theta(x_{k-1}, \cdot) \phi_{k-1}^\theta(\mathrm{d}x_{k-1})$ and $b_{k-1|k}^\theta(x_k, \cdot) \propto \phi_{k-1}^\theta(\cdot) m_k^\theta(\cdot, x_k)$. On the contrary, for the *GRU Backward*, no such link can be made.

Relating the distributions $\{\tilde{q}_k^\lambda\}_{k<n}$ with implementation choices. The previous discussion is tightly linked to the practical existence and meaning of distributions q_k^λ for $k < n$ in the offline setting that we studied. Indeed, the theoretical study only prescribes implementing explicitly a term for $k = n$. The proof of Proposition 3 suggests that when this terminal distribution q_n^λ is the last term of a sequence $(q_k^\lambda)_{k \leq n}$ where $q_k^\lambda = \phi_k^\theta$ for $k < n$, then it only remains to have the variational backward kernels closest to the true ones to reduce the additive smoothing error. However it is unclear whether this is the optimal scenario in the sense of Proposition 1, i.e. the discrepancies $\tilde{c}_k$ may be lower for some sequence $(\tilde{q}_k)_{k \leq n}$ which is not an approximation of the sequence of true filtering distributions, and an implementation of this optimum might not yield - as is the case for some of our models - good approximations of the latter as a byproduct.

Nonetheless, coincidentally, standard implementations of the backward factorization using DNNs involve recursions on some running quantity (i.e. the RNN states $(s_k)_{k \geq 0}$ in the *GRU Backward* case). In the offline setting, the parameters of q_n^λ are obtained as a transformation of such quantity applied at $k = n$ only. In practice, the latter transformation can always be applied for $k < n$ to yield a sequence of distributions $(q_k^\lambda)_{k \leq n}$ from which q_n^λ is the terminal term. Therefore, intermediate distributions appear naturally both when deriving the theoretical bound of Proposition 1 and in efficient implementations. An interesting

perspective of this work would be to more clearly relate these aspects, which are currently distinct. In the general case, this falls back on the mathematical problem of computing the infimum over $\{\tilde{q}_k^\lambda\}_{k<n}$ to gain insight on which sequence should be implemented in practice. Another approach would be to derive a result under the additional assumption that $q_{k-1|k}^\lambda \propto q_{k-1}^\lambda m_{parvec}^k$ for all $k \leq n$, given some sequence $(q_k^\lambda)_{k\leq n}$ which is implemented in practice - and which explicitly affects the overally joint distribution $q_{0:t}^\lambda$ through the backward kernels - as is the case in the *Conjugate Backward* implementation. While this would yield less general theoretical results, it is more approachable mathematically, as one may adapt the proof of Proposition 1 to this setting.

Appendix A. Proofs of the main results

A.1 Proof of Proposition 1

Following Gloaguen et al. (2022), write

$$q^{\lambda}_{0:n} h_n - \phi^{\theta}_{0:n} h_n = \sum_{k=0}^{n-1} \left(q^{\lambda}_{0:n} \bar{h}_{k|n} - \phi^{\theta}_{0:n} \bar{h}_{k|n} \right), \tag{9}$$

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

$$\bar{h}_{k|n} : x_{0:n} \mapsto \tilde{h}_k(x_k, x_{k+1}) \,. \tag{10}$$

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

$$\mathbf{L}^{\theta}_{m,n}(x'_{0:m}, \mathrm{d}x_{0:n}) := \delta_{x'_{0:m}}(\mathrm{d}x_{0:m}) \prod_{\ell=m}^{n-1} \mathbf{L}^{\theta}_{\ell}(x_\ell, \mathrm{d}x_{\ell+1}) \tag{11}$$

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

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

where for all $1 \leq m \leq n$, $\tilde{q}_{0:m} = \tilde{q}_m \prod_{k=1}^{m} q^{\lambda}_{k-1|k}$, $\tilde{q}_{0:0} = \tilde{q}_0$, 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 $m \in \{1, n\}$,

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

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

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

Therefore,

$$q^{\lambda}_{0:n} \bar{h}_{k|n} - \phi^{\theta}_{0:n} \bar{h}_{k|n} = \sum_{m=1}^{n} \left(\frac{\tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \bar{h}_{k|n}}{\tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} - \frac{\tilde{q}_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n} \bar{h}_{k|n}}{\tilde{q}_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n} \mathbb{1}} \right) + \frac{\tilde{q}_0 \mathbf{L}^{\theta}_{0,n} \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}^{\theta}_{0,n} \mathbb{1}} - \frac{\chi^{\theta} g^{\theta}_0 \mathbf{L}^{\theta}_{0,n} \bar{h}_{k|n}}{\chi^{\theta} g^{\theta}_0 \mathbf{L}^{\theta}_{0,n} \mathbb{1}}. \tag{13}$$

By Lemma 6,

$$\left| \frac{\tilde{q}_0 \mathbf{L}^{\theta}_{0,n} \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}^{\theta}_{0,n} \mathbb{1}} - \frac{\phi^{\theta}_0 \mathbf{L}^{\theta}_{0,n} \bar{h}_{k|n}}{\phi^{\theta}_0 \mathbf{L}^{\theta}_{0,n} \mathbb{1}} \right| \leqslant 2 \left\| \tilde{q}_0 - \phi^{\theta}_0 \right\|_{\mathrm{tv}} \frac{\sigma_+}{\sigma_-} \|\tilde{h}_k\|_{\infty} \, .$$

Consider now the error term at time $m > 0$ in (13). Define the kernel

$$\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n}(x'_{m-1}, x'_m, \mathrm{d}x_{0:n}) := \delta_{x'_{m-1}}(\mathrm{d}x_{m-1}) \prod_{\ell=0}^{m-2} q^{\lambda}_{\ell|\ell+1}(x_{\ell+1}, \mathrm{d}x_{\ell}) \delta_{x'_m}(\mathrm{d}x_m) \prod_{\ell=m}^{n-1} \mathbf{L}^{\theta}_{\ell}(x_{\ell}, \mathrm{d}x_{\ell+1}), \tag{14}$$

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

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

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

$$\begin{aligned} \mathcal{L}^{*,\lambda,\theta}_{m,n} \bar{h}_{k|n}(x_{m-1}, x_m) &= \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \bar{h}_{k|n}(x_{m-1}, x_m)}{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}(x_{m-1}, x_m)} - \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \bar{h}_{k|n}(x^*_{m-1}, x^*_m)}{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}(x^*_{m-1}, x^*_m)} \, , \\ &= \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \bar{h}_{k|n}(x_{m-1}, x_m)}{\mathbf{L}^{\theta}_{m,n} \mathbb{1}(x_m)} - \frac{\tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \bar{h}_{k|n}(x^*_{m-1}, x^*_m)}{\mathbf{L}^{\theta}_{m,n} \mathbb{1}(x^*_m)} \end{aligned} \tag{15}$$

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 7, $\left\| \mathcal{L}^{*,\lambda,\theta}_{m,n} \bar{h}_{k|n} \right\|_{\infty}$ can be upper bounded and note that

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

By definition of the normalized measure $\tilde{\phi}^{\theta}_{m-1:m}$,

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

Then, using that

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

and the fact that $\tilde{\nu}^{\lambda}_{m-1:m} = \tilde{q}_m q^{\lambda}_{m-1|m}$,

$$\left| \frac{\tilde{\phi}^{\theta}_{m-1:m} \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1} - \tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}}{\tilde{\phi}^{\theta}_{m-1:m} \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} \right| \leqslant \left\| \tilde{\phi}^{\theta}_{m-1:m} - \tilde{\nu}^{\lambda}_{m-1:m} \right\|_{\mathrm{tv}} \frac{\left\| \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1} \right\|_{\infty}}{\tilde{\phi}^{\theta}_{m-1:m} \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} ,$$

and

$$\begin{aligned}
&\left| \frac{\tilde{q}_m q^{\lambda}_{m-1|m} \left\{ \mathcal{L}^{*,\lambda,\theta}_{m,n} \bar{h}_{k|n} \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1} \right\} - \tilde{\phi}^{\theta}_{m-1:m} \left\{ \mathcal{L}^{*,\lambda,\theta}_{m,n} \bar{h}_{k|n} \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1} \right\}}{\tilde{\phi}^{\theta}_{m-1:m} \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} \right| \\
&\qquad\qquad \leqslant \left\| \tilde{\phi}^{\theta}_{m-1:m} - \tilde{\nu}^{\lambda}_{m-1:m} \right\|_{\mathrm{tv}} \frac{\left\| \mathcal{L}^{*,\lambda,\theta}_{m,n} \bar{h}_{k|n} \right\|_{\infty} \left\| \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1} \right\|_{\infty}}{\tilde{\phi}^{\theta}_{m-1:m} \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} ,
\end{aligned}$$

yields

$$\left| \frac{\tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \bar{h}_{k|n}}{\tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} - \frac{\tilde{q}_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n} \bar{h}_{k|n}}{\tilde{q}_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n} \mathbb{1}} \right| \leqslant 2 \left\| \tilde{\phi}^{\theta}_{m-1:m} - \tilde{\nu}^{\lambda}_{m-1:m} \right\|_{\mathrm{tv}} \frac{\left\| \mathcal{L}^{*,\lambda,\theta}_{m,n} \bar{h}_{k|n} \right\|_{\infty} \left\| \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1} \right\|_{\infty}}{\tilde{\phi}^{\lambda}_{m} \tilde{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} .$$

Note also that by H1,

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

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,

$$\left| \frac{\tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \bar{h}_{k|n}}{\tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} - \frac{\tilde{q}_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n} \bar{h}_{k|n}}{\tilde{q}_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n} \mathbb{1}} \right| \leqslant 2 \frac{\sigma_+}{\sigma_-} \left\| \tilde{\phi}^\theta_{m-1:m} - \tilde{\nu}^\lambda_{m-1:m} \right\|_{\mathrm{tv}} \left\| \mathcal{L}^{*,\lambda,\theta}_{m,n} \bar{h}_{k|n} \right\|_\infty .$$

The proof is completed using Lemma 7.

A.2 Proof of Corollary 2

It is enough to introduce the same decomposition as the one used in Proposition 1:

$$q^\lambda_{0:n} \bar{h}_{k_\star|n} - \phi^\theta_{0:n} \bar{h}_{k_\star|n} = \sum_{m=1}^n \left(\frac{\tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \bar{h}_{k_\star|n}}{\tilde{q}_m \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m,n} \mathbb{1}} - \frac{\tilde{q}_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n} \bar{h}_{k_\star|n}}{\tilde{q}_{m-1} \boldsymbol{\mathcal{L}}^{\lambda,\theta}_{m-1,n} \mathbb{1}} \right) + \frac{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k_\star|n}}{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} - \frac{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k_\star|n}}{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n} \mathbb{1}}.$$

Each term is then controlled similarly as in the proof of Proposition 1. By Lemma 6,

$$\left| \frac{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k_\star|n}}{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} - \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k_\star|n}}{\phi^\theta_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} \right| \leqslant 2 c_0(\theta) \frac{\sigma_+}{\sigma_-} \|\tilde{h}_k\|_\infty .$$

On the other hand, the error term at time $m > 0$ is upper bounded by

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

The proof is completed using Lemma 7.

Appendix B. Technical results

Lemma 6 *Assume that H1 holds. Then for all, $\theta \in \Theta$, $\lambda \in \Lambda$, $n \geq 1$, $k \in \{0, n-1\}$, bounded and measurable function $\tilde{h}_k$,*

$$\left| \frac{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} - \frac{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n}}{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} \right| \leq 2 \left\| \tilde{q}_0 - \phi^\theta_0 \right\|_{\mathrm{tv}} \frac{\sigma_+}{\sigma_-} \|\tilde{h}_k\|_\infty ,$$

where $\bar{h}_{k|n}$ is defined in (10).

Proof Consider the following decomposition of the first term:

$$\begin{aligned} \frac{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} - \frac{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n}}{\chi^\theta g^\theta_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} &= \frac{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} - \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n}}{\phi^\theta_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} , \\ &= \frac{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n} - \phi^\theta_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} \\ &\quad + \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n} \bar{h}_{k|n}}{\phi^\theta_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} \frac{\phi^\theta_0 \mathbf{L}^\theta_{0,n} \mathbb{1} - \tilde{q}_0 \mathbf{L}^\theta_{0,n} \mathbb{1}}{\tilde{q}_0 \mathbf{L}^\theta_{0,n} \mathbb{1}} , \end{aligned}$$

where ϕ_0^θ the filtering distribution at time 0, i.e the law defined as $\phi_0^\theta h = \chi^\theta g_0^\theta h / \chi^\theta g_0^\theta$. Note that

$$\left| \frac{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \bar{h}_{k|n} - \phi_0^\theta \mathbf{L}_{0,n}^\theta \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}} \right| \leq \left\| \tilde{q}_0 - \phi_0^\theta \right\|_{\mathrm{tv}} \frac{\|\mathbf{L}_{0,n}^\theta \bar{h}_{k|n}\|_\infty}{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}} \leq \left\| \tilde{q}_0 - \phi_0^\theta \right\|_{\mathrm{tv}} \frac{\|\mathbf{L}_{0,n}^\theta \mathbb{1}\|_\infty \|\bar{h}_{k|n}\|_\infty}{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}}$$

and, using that $\phi_0^\theta \mathbf{L}_{0,n}^\theta \bar{h}_{k|n} / \phi_0^\theta \mathbf{L}_{0,n}^\theta \mathbb{1} \leq \|\bar{h}_{k|n}\|_\infty$,

$$\left| \frac{\phi_0^\theta \mathbf{L}_{0,n}^\theta \bar{h}_{k|n}}{\phi_0^\theta \mathbf{L}_{0,n}^\theta \mathbb{1}} \frac{\phi_0^\theta \mathbf{L}_{0,n}^\theta \mathbb{1} - \tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}}{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}} \right| \leq \left\| \tilde{q}_0 - \phi_0^\theta \right\|_{\mathrm{tv}} \frac{\|\mathbf{L}_{0,n}^\theta \mathbb{1}\|_\infty \|\bar{h}_{k|n}\|_\infty}{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}}$$

Then,

$$\left| \frac{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}} - \frac{\phi_0^\theta \mathbf{L}_{0,n}^\theta \bar{h}_{k|n}}{\phi_0^\theta \mathbf{L}_{0,n}^\theta \mathbb{1}} \right| \leqslant 2 \left\| \tilde{q}_0 - \phi_0^\theta \right\|_{\mathrm{tv}} \frac{\|\mathbf{L}_{0,n}^\theta \mathbb{1}\|_\infty \|\bar{h}_{k|n}\|_\infty}{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}} \, .$$

By H1, for all $x_0 \in \mathbb{R}^d$,

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

and

$$\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1} = \int \tilde{q}_0(\mathrm{d}x_0) \ell_{0,\theta}(x_0, x_1) \mu(\mathrm{d}x_1) \mathbf{L}_{1,n}^\theta \mathbb{1}(x_1) \geqslant \sigma_- \int \mu(\mathrm{d}x_1) \mathbf{L}_{1,n}^\theta \mathbb{1}(x_1) \, ,$$

which yields

$$\left| \frac{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \bar{h}_{k|n}}{\tilde{q}_0 \mathbf{L}_{0,n}^\theta \mathbb{1}} - \frac{\phi_0^\theta \mathbf{L}_{0,n}^\theta \bar{h}_{k|n}}{\phi_0^\theta \mathbf{L}_{0,n}^\theta \mathbb{1}} \right| \leqslant 2 \left\| \tilde{q}_0 - \phi_0^\theta \right\|_{\mathrm{tv}} \frac{\sigma_+}{\sigma_-} \|\tilde{h}_k\|_\infty \, .$$

■

Lemma 7 *Assume that H1 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 $\tilde{h}_k$,*

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

where $\rho = 1 - \sigma_-/\sigma_+$ and $\bar{h}_{k|n}$ is defined in (10) *and $\mathcal{L}_{m,n}^{*,\lambda,\theta} \bar{h}_{k|n}$ is defined in* (15)*.*

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,

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

Therefore,

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

By H1, 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

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

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 H1,

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

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,

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

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.

■

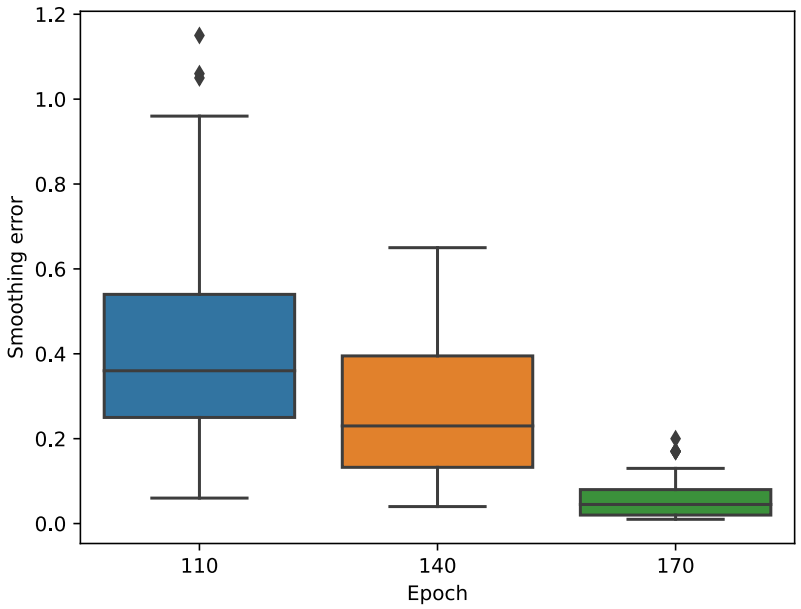


Figure 6: Smoothing errors $\left|q^{\lambda}_{0:n}h_{0:n} - \phi^{\theta}_{0:n}h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$ at $n = 500$, when $\phi^{\theta}_{0:n}$ is given via Kalman smoothing with the true parameters θ and $q^{\lambda}_{0:n}$ is given via Kalman smoothing with parameters λ selected at epochs 110,140 and 170. Each plot is generated from the $J = 50$ sequences $(Y^j_{0:n})_{1\leq j\leq J}$ drawn from p^{θ}

Appendix C. Experimental details

C.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.

C.2 Linear Gaussian models

We provide here additional figures for the experiments of Section 4.1. Figure 6 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 fully fitted models against the ground truth Kalman smoother which uses the true parameters θ.

C.3 Nonlinear models

Here we provide additional details on the experiments of section 4.2.

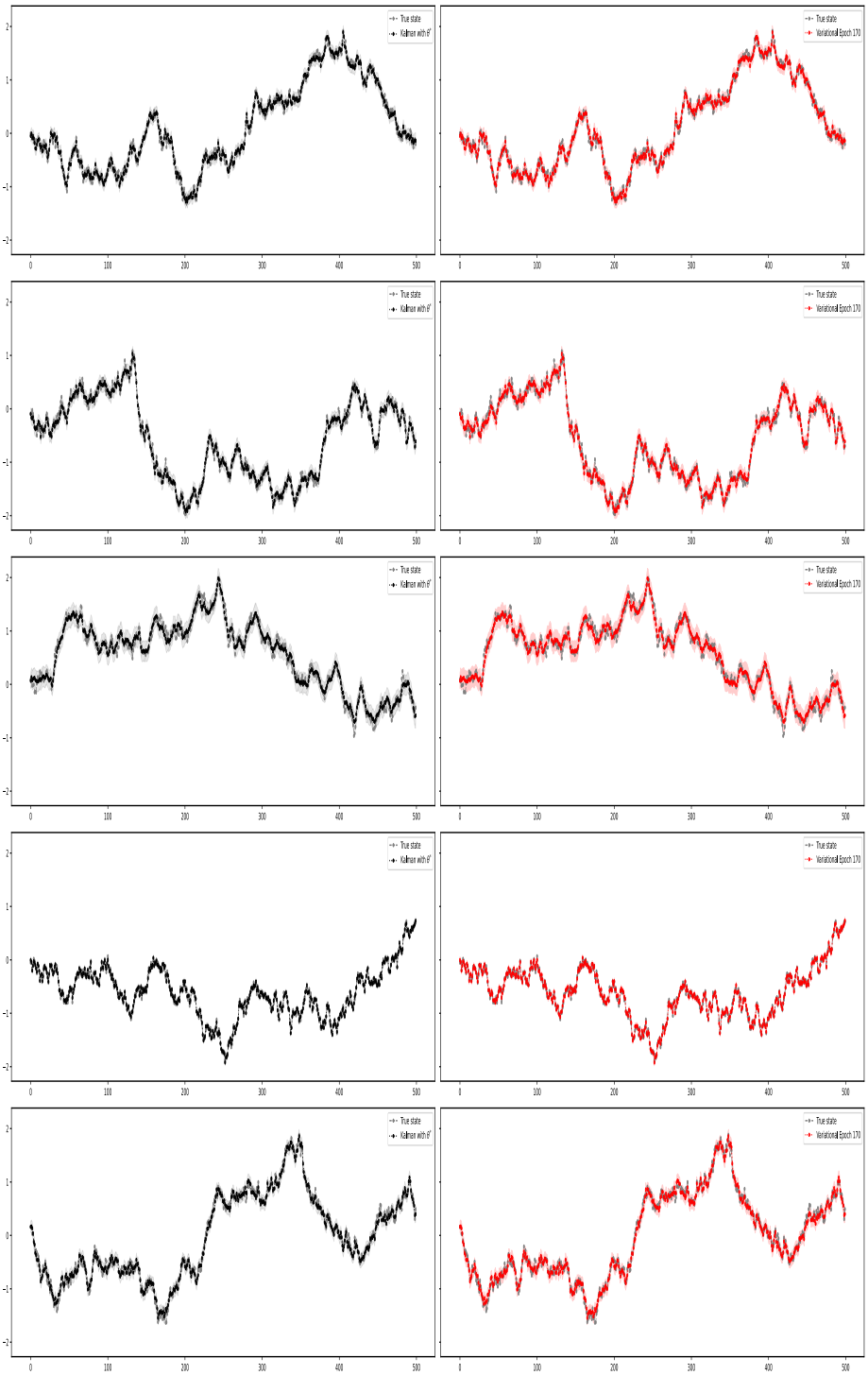


Figure 7: Example of smoothed states when the dimension of the state space is 5 and the observations is 5. Left column: component-wise (from top to bottom) smoothed states with true parameters θ. Right column: same thing with learnt parameters λ. The dashed fillings are the 95% confidence intervals. The horizontal axis is the time axis.

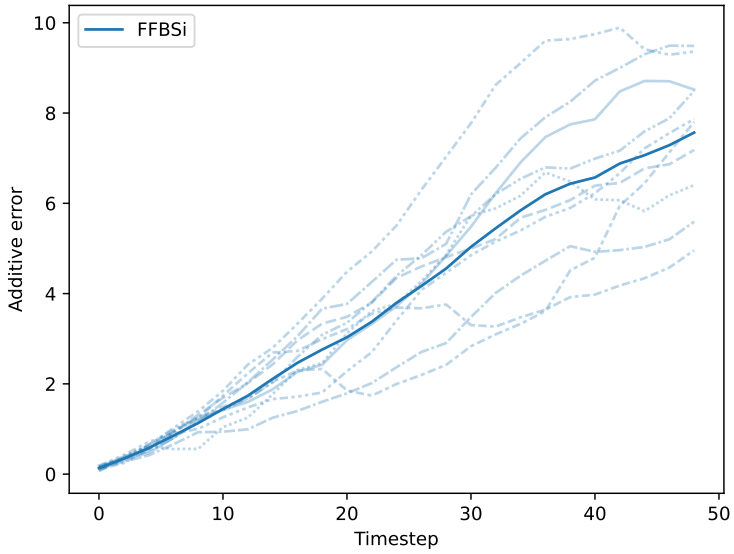


Figure 8: Smoothing errors $\left|h_{0:n}(x^*_{0:n}) - \phi^\theta_{0:n} h_{0:n}\right|$ for $\tilde{h}_k(x_k, x_{k+1}) = x_k$, where $x^*_{0:n}$ is the true sequence of hidden states and $\phi^\theta_{0:n}$ is obtained by the FFBsi algorithm. All values are normalized by the dimension of the state space. Experiments are produced on 10 independent sequences. The thick solid lines display the mean over the 10 independent replicates for both approaches, shaded lines are single sequences.

- For the nonlinear emission function d^θ of the data model, we used a single-layer perceptron with a ReLU activation function (which induces non-injectivty on some portions of the support).
- For the *Conjugate Forward* and *Conjugate Backward* methods, the encoder r^λ is a multi-layer perceptron (MLP) 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 GRU Backward model, H^λ is a Deep GRU as implemented in the Haiku library from the JAX ecosystem.

For the experiments of section 4.2, we use small networks with two hidden layers of size 8 (both for r^λ and the GRU in the corresponding models). For the experiments of section 4.2.2, we use configurations similar to that of Campbell et al. (2021) for fair comparison, i.e. neural networks with a single hidden layer of size 100.

In Figure 8, we plot the evolution of the additive error of the FFBSi oracle against the true states.

References

Justin Bayer, Maximilian Soelch, Atanas Mirchev, Baris Kayalibay, and Patrick van der Smagt. Mind the gap when conditioning amortised inference in sequential latent-variable models. In *International Conference on Learning Representations*, 2021.

Andrew Campbell, Yuyang Shi, Thomas Rainforth, and Arnaud Doucet. Online variational filtering and parameter learning. *Advances in Neural Information Processing Systems*, 34, 2021.

O. Cappé, É. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

Pavel Chigansky and Robert Liptser. Stability of nonlinear filters in nonmixing case. *The Annals of Applied Probability*, 14(4):2038 – 2056, 2004. doi: 10.1214/105051604000000873.

Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*. Springer, 2020.

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

R. Douc, É. Moulines, and D. Stoffer. *Nonlinear time series: theory, methods and applications with R examples*. CRC Press, 2014.

Randal Douc, Gersende Fort, Eric Moulines, and Pierre Priouret. Forgetting the initial distribution for hidden markov models. *Stochastic processes and their applications*, 119 (4):1235–1256, 2009.

Randal Douc, Aurélien Garivier, Eric Moulines, and Jimmy Olsson. Sequential monte carlo smoothing for general state space hidden markov models. *The Annals of Applied Probability*, 21(6):2109–2145, 2011.

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

E. Gassiat and S. Le Corff. Variational autoencoder excess risk bound for state space models. *Work in progress*, 2023.

Élisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. Identifiability and consistent estimation of nonparametric translation hidden Markov models with general state space. *Journal of Machine Learning Research*, 21, 04 2020.

Pierre Gloaguen, Sylvain Le Corff, and Jimmy Olsson. A pseudo-marginal sequential monte carlo online smoothing algorithm. *Bernoulli*, 28(4):2606–2633, 2022.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Élisabeth Gassiat, and Aapo Hyvärinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems (NeurIPS)*, 29, 2016.

Ilyes Khemakhem, Diederik P. Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2207–2217, 2020.

Diederik Kingma and Max Welling. Auto-encoding variational bayes. 12 2014.

Rahul Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Wu Lin, Mohammad Emtiyaz Khan, and Nicolas Hubacher. Variational message passing with structured inference networks. In *International Conference on Learning Representations*, 2018.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Scholkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *International Conference on Machine Learning (ICML)*, 119:6348–6359, 2020.

Joseph Marino, Milan Cvitkovic, and Yisong Yue. A general method for amortizing variational filtering. In *Advances in neural information processing systems (NeurIPS)*, volume 31, 2018.

Alice Martin, Marie-Pierre Etienne, Pierre Gloaguen, Sylvain Le Corff, and Jimmy Olsson. Backward importance sampling for online estimation of state space models. *Journal of Computational and Graphical Statistics*, pages 1–14, 2023.

Alessandro Mastrototaro, Jimmy Olsson, and Johan Alenlöv. Fast and numerically stable particle-based online additive smoothing: the adasmooth algorithm. *Journal of the American Statistical Association*, pages 1–12, 2022.

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

Rong Tang and Yun Yang. On empirical bayes variational autoencoder: An excess risk bound. In *Conference on Learning Theory*, 2021.

Yuan Zhao, Josue Nassar, Ian Jordan, Mónica Bugallo, and Il Memming Park. Streaming variational monte carlo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1150–1161, 2022.