Structured Variational Inference in Bayesian State-Space Models

Honggang Wang
Texas A&M University

Yun Yang
University of Illinois
at Urbana-Champaign

Debdeep Pati
Texas A&M University

Anirban Bhattacharya
Texas A&M University

Abstract

Variational inference is routinely deployed in Bayesian state-space models as an efficient computational technique. Motivated by the inconsistency issue observed by Wang and Titterington (Wang & Titterington, 2004) for the mean-field approximation in linear state-space models, we consider a more expressive variational family for approximating the joint posterior of the latent variables to retain their dependence, while maintaining the mean-field (i.e. independence) structure between latent variables and parameters. In state-space models, such a latent structure adapted mean-field approximation can be efficiently computed using the belief propagation algorithm. Theoretically, we show that this adapted mean-field approximation leads to consistent variational estimates. Furthermore, we derive a non-asymptotic risk bound for an averaged α-divergence from the true data generating model, suggesting that the posterior mean of the best variational approximation for the static parameters shows optimal concentration. From a broader perspective, we add to the growing literature on statistical accuracy of variational approximations by allowing dependence between the latent variables, and the techniques developed here should be useful in related contexts.

1 Introduction

Variational Inference (VI) has gained major prominence as an approximate Bayesian computational tool in the past two decades. VI aims to find the closest member to an otherwise intractable posterior distribution from a structured family of distributions, commonly referred to as the variational family. The closeness is most commonly defined in terms of the Kullback–Leibler (KL) divergence, and various iterative procedures have been developed to optimize the resulting objective function.

The choice of the variational family is a crucial component of VI, as it trades-off between computational ease and approximation capability. Perhaps the most commonly used variational family is the mean-field family (henceforth MFVI), which assumes mutual independence across (blocks of) parameters. General purpose algorithms, such as the coordinate ascent variational inference (CAVI), have been developed to obtain the MFVI solution. Motivated by its practical success, there has been a flurry of recent activity studying theoretical properties of MFVI (Blei *et al.*, 2017; Pati *et al.*, 2018; Yang *et al.*, 2020; Zhang & Gao, 2020; Alquier & Ridgway, 2020; Ghosh *et al.*, 2020; Plummer *et al.*, 2021; Ray & Szabó, 2021), which establish optimality properties of MFVI in various statistical models. Yang *et al.* (2020) in particular considered latent (or hidden) variable models, albeit under the assumption that the observation-latent variable pairs are mutually independent. A more general treatment of latent variable models has thus far been missing.

An important class of models where the above assumption of independence is violated is state-space models (Turner & Sahani, 2011). Moreover, Wang & Titterington (2004) noted a lack of consistency of the emission parameter in linear Gaussian state-space models (LGSSM), suggesting the mean-field approximation may not be adequate for these types of models. Motivated by this observation, Barber & Chiappa (2007) proposed a more expressive decomposition, where the dependence between the latent state variables is retained, and a mean-field decomposition is otherwise employed between the static vectors and the collection of latent state variables. Such structured decompositions have been subsequently termed generalized mean-field (GMF) in the literature (Xing *et al.*,

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

2012). A more flexible approach is the belief propagation (BP) (Pearl, 2014), which tries to find local approximations, which are exactly or approximately the marginals of the target density. This is also done in an iterative way called the message passing, where messages are passed along the edges of a factor graph. Belief propagation routinely extends beyond the factor graph as loopy belief propagation (Weiss, 2000) which is known to converge to the first order stationary points of the variational objective.

In this article, we offer a systematic theoretical treatment of such structured VI in SSMs to complement the algorithmic development. We operate in the α-VB framework of Yang *et al.* (2020), who introduce a temperature parameter $\alpha \in (0,1]$ and modify the usual variational objective in latent variable models to target a particular tempered posterior distribution. Yang *et al.* (2020) showed that frequentist risk bounds for $\alpha < 1$ require fewer assumptions compared to $\alpha = 1$, the usual VI. We shall restrict our attention to $\alpha < 1$ throughout this article. Operating in this framework, we first extend the inconsistency result of Wang & Titterington (2004) to show an inconsistency result for the full mean-field approximation in scalar LGSSM for an integrated α-Rényi Bayes risk. Next, we establish a general risk bound for the gMF decomposition for general state-space models. In deriving the risk bounds, while we follow the general technique of obtaining Kullback-Leibler concentration for non-i.i.d models (Ghosal & Van Der Vaart, 2007), a key innovation is in the introduction of the blow-up functions $f_\lambda(\cdot)$ and $f_\mu(\cdot)$ which offers additional flexibility in choosing a neighborhood of the latent variables and in obtaining a theoretically optimal variational guess. Applying our general risk bound to scalar and multivariate LGSSMs, we obtain a near-optimal (up to logarithmic terms commonly appearing in related Bayesian calculations) rate of convergence for the static parameters. We also delineate sufficient conditions on the model structure for our risk bound to be applicable beyond LGSSMs.

1.1 Notation

Given probability measures p and q with a common dominating measure μ, we denote the KL divergence as $D(p \,\|\, q) := \int p \log(p/q) d\mu$. Also, define the V-divergence $V(p \,\|\, q) = \int p \log^2(p/q) d\mu$. For $\alpha \in (0,1)$, the α-Rényi divergence between p and q is given by $D_\alpha(p \,\|\, q) = \frac{1}{\alpha-1} \log \int p^\alpha q^{1-\alpha} d\mu$. Given a probability density q and a measurable function H, we use $\langle H \rangle_q$ to denote $\mathbb{E}_{\theta\sim q}[H(\theta)]$. For any positive integer m, $[m] := \{1, \dots, m\}$.

2 Background

We begin this section with a brief introduction on general state-space models, with a special attention to the linear Gaussian state-space model as our main illustrative example. After that, we discuss the computational aspect of implementing inference for state-space models. In particular, we focus on variational inference and review some recent developments.

2.1 State-Space Models

State-space models (SSM; (Koller & Friedman, 2009; Zeng & Wu, 2013)) are extremely popular in time-series applications across diverse disciplines. Given a time-series consisting of observations $\{Y_t\}$ in $\mathbb{R}^{d_H}$, the general formulation of a state-space model involves (a) specifying a model for $Y_t \mid X_t$, where $X_t \in \mathbb{R}^{d_V}$ denotes the collection of all hidden variables at time t, and (b) specifying a Markovian evolution for the hidden variables $\{X_t\}$ over time. Specifically, for $t \in [n]$, assume

$$Y_t \mid X_t \overset{\text{ind.}}{\sim} p(\cdot \mid X_t, \mu), \quad X_t \mid X_{t-1} \overset{\text{ind.}}{\sim} \tilde{p}(\cdot \mid X_{t-1}, \lambda). \tag{1}$$

Here, p and $\tilde{p}$ respectively denote the distribution families for $Y_t \mid X_t$ and $X_t \mid X_{t-1}$, and μ and λ respectively denote unknown static parameters specifying these distributions. Throughout this article, we shall use $\theta = (\mu, \lambda)$ to denote the collection of static parameters that are the primary quantities of interest.

An important sub-class of state-space models is linear Gaussian state-space models (LGSSM) where p and $\tilde{p}$ are both multivariate Gaussian distributions whose conditional expectations $E(Y_t \mid X_t)$ and $E(X_t \mid X_{t-1})$ are linear in X_t and X_{t-1} respectively:

$$\begin{aligned} Y_t \mid X_t &\overset{\text{ind.}}{\sim} \mathcal{N}_d(BX_t, \Sigma_H), \\ X_t \mid X_{t-1} &\overset{\text{ind.}}{\sim} \mathcal{N}_d(AX_{t-1}, \Sigma_V). \end{aligned} \tag{2}$$

$A \in \mathbb{R}^{d_V \times d_V}$ is called the *transmission* matrix and B is the $d_H \times d_V$ *emission* matrix. Σ_H and Σ_V are separately $d_H \times d_H$ and $d_V \times d_V$ positive definite matrices describing innovation covariance structures, which we shall assume to be diagonal for technical ease, that is, $\Sigma_V = \text{Diag}(\sigma^2_{V,1}, \dots, \sigma^2_{V,d_V})$ and $\Sigma_H = \text{Diag}(\sigma^2_{H,1}, \dots, \sigma^2_{H,d_H})$, although our theory can be extended to non-diagonal covariances. In this paper, we focus on the time-homogeneous setting where the collection of parameters $\theta = (A, B, \Sigma_H, \Sigma_V)$ does not involve over time.

A Bayesian specification of the state-space model is completed by endowing the static parameters in θ with a prior distribution $\pi(\cdot)$, and performing inference based on the joint posterior distribution

$p(\theta, X^n|Y^n) \propto p(Y^n|X^n, \mu)\, p(X^n|\lambda)\, \pi(\theta)$. In particular, parameter estimation can be carried out based on the marginal posterior distribution $p(\theta\,|Y^n) = \int p(\theta, X^n|Y^n)\, dX^n$. In the LGSSM, it is common to endow the transition and emission matrices with entrywise independent Gaussian priors, and diagonal entries $\{\sigma^2_{V,i},\, \sigma^2_{H,j} : i \in [d_V],\, j \in [d_H]\}$ of the innovation covariance matrices with independent inverse-gamma priors. More generally, one may consider matrix-Gaussian priors for A and B, and inverse-Wishart priors for Σ_H and Σ_V (without imposing the diagonal structure) to deliver matrix-level prior structures. Although the latent variable distribution $p(X^n|Y^n, \theta)$ may fall into some common distribution families in conditionally conjugate models, the joint posterior of (θ, X^n) is generally intractable. For example, in the scalar LGSSM where $(A, B) = (a, b)$ and $\Sigma_H = \Sigma_V = \sigma_0^2$, the latent variable distribution of $[X^n \mid Y^n, \theta]$ is $\mathcal{N}_n(\Lambda, \sigma_0^2\, \Omega^{-1})$, where

$$\Omega = \begin{bmatrix} a^2 + b^2 + 1 & -a & \dots & 0 \\ -a & a^2 + b^2 + 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -a \\ 0 & 0 & \dots & b^2 + 1 \end{bmatrix}$$

and $\Lambda = b\,\Omega^{-1} Y^n$. Thus, the conditional posterior distribution of latent variables given static parameters lies in the multivariate Gaussian family with a tri-diagonally structured precision matrix.

The analytic intractability of the joint posterior distribution necessitates numerical techniques for approximate Bayesian computation. Sampling-based approaches such as Markov chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) have received a lot of attention (Kantas *et al.*, 2009) in dealing with state space models. However, the relative high accuracy of these approaches comes at the price of expensive computations, making them only suitable for small to moderate data sets. More recently, variational approximations have gained popularity due to their scalability to massive data by exploiting the computational advantages of optimizations over samplings. However, unlike MCMC and SMC, which carry asymptotic guarantees of convergence to the exact marginal posterior of θ given Y^n, variational inference approximates this marginal posterior from a restrictive class which may induce non-negligible systematic bias. Hence it is important to study the quality of estimation provided by VI, which we undertake in this article.

2.2 Variational Inference in State-Space Models

Variational inference aims to find the best approximation (in terms of the KL divergence) to the joint posterior of parameters and latent variables from a tractable family of probability distributions Γ, called the variational family. A common structural constraint on the variational family is the mean-field structure, which posits that all constituent variable components are mutually independent. In the context of SSMs, letting $W^n = (\theta, X^n)$ collect latent variables and static parameters, the mean-field family consists of all distributions of the form

$$q_{\boldsymbol{W}^n}(W^n) = \left[\prod_{t=1}^n q_{\boldsymbol{X}_t}(X_t)\right] q_{\boldsymbol{\theta}}(\theta). \tag{3}$$

Here and elsewhere, we use bold subscripts to distinguish the various variational densities.

For a class of latent variable models where the observation-latent variable pairs are mutually independent, Yang *et al.* (2020) established that the mean of the optimal mean-field variational solution $\widehat{q}_{\boldsymbol{\theta}}$ attains the minimax rate of parameter estimation up to logarithmic factors. However, in SSMs, this mutually independence assumption is clearly violated due to the dependence among the X_t's across time, and the theory of Yang *et al.* (2020) cannot be directly applied. In fact, Wang & Titterington (2004) provided a counter-example that illustrated the lack of consistency of mean-field VI (MFVI) in the context of LGSSMs. By inspecting their derivation, it turns out that the inconsistency is due to the ignorance of the dependence structure among the latent variables $X^n = (X_1, \dots, X_n)$. In light of this, a natural idea is to retain this dependence within the variational family by considering a block mean-field decomposition:

$$q_{\boldsymbol{W}^n}(W^n) = q_{\boldsymbol{X}^n}(X^n)\, q_{\boldsymbol{\theta}}(\theta). \tag{4}$$

Such a variational family was employed in the influential article (Barber & Chiappa, 2007). More generally, in situations where retaining such a partial dependence within the variational family is warranted, a decomposition as in (4) is called generalized mean-field VI (GMFVI; (Xing *et al.*, 2012)).

Due to a natural trade-off between the flexibility of the variational family and the complexity of numerically computing the variational solution using iterative procedures, the GMFVI solution in general is computationally more expensive than MFVI. However, in LGSSMs and other conjugate SSMs, carefully designed algorithms can mitigate this computational gap. A commonly used algorithm to compute the variational solution is the coordinate ascent VI (CAVI) algorithm, which iteratively updates each component in the mean-field factorization by minimizing the KL divergence to the target while keeping the other components fixed at their current states. In conditionally conjugate models, these updates naturally lie in an exponential family without any additional assumptions on the constituent components. Specifically, in the context of block mean-field approximation (4), the CAVI alternatively updates the two blocks with formulas

$$\begin{aligned} q_{\boldsymbol{X}^n}(X^n) &\propto \exp\left[\langle \log p(\theta, X^n \mid Y^n)\rangle_{q_{\boldsymbol{\theta}}}\right], \\ q_{\boldsymbol{\theta}}(\theta) &\propto \exp\left[\langle \log p(\theta, X^n \mid Y^n)\rangle_{q_{\boldsymbol{X}^n}}\right]. \end{aligned} \tag{5}$$

The main challenge lies in computing and manipulating $q_{\boldsymbol{X}^n}$. In LGSSMs, $q_{\boldsymbol{X}^n}(X^n)$ lies in the Gaussian family with a non-diagonal covariance matrix, and a naive manipulation with this multivariate Gaussian density renders an $O(n^3)$ complexity due to matrix inversion. However, the precision matrix of this Gaussian distribution has a banded structure (e.g. the Ω in the scalar case), which can be exploited in designing more efficient algorithms. In particular, the computation of univariate and bivariate marginals $q^{(s)}_{\boldsymbol{X}^n}(X_k)$ and $q^{(s)}_{\boldsymbol{X}^n}(X_k, X_{k+1})$ at any iteration s can be efficiently carried out using Belief Propagation (BP; (Yedidia *et al.*, 2000, 2003)) and its many related variants For example, Barber & Chiappa (2007) used the related Rauch–Tung–Striebe (RTS) Smoother (Welch *et al.*, 1995; Koller & Friedman, 2009). It is worth noting here that $q_{\boldsymbol{X}^n}(X^n)$ cannot be computed exactly in general when the graph structure is not a tree. In such circumstances, loopy belief propagation (LBP) is commonly employed. LBP can be recognized as a first-order stationary point of the variational objective; however, global convergence is not generally guaranteed. Further, in non-conjugate models, one may require a hybrid algorithm which combines MCMC to approximate certain analytically intractable steps in the iterative procedure.

Although the block mean-field approximation (4) that preserves latent variable dependence does not impose significantly more computational burdens thanks to the aforementioned algorithmic developments using BP, the theoretical properties of GMFVI, such as consistency and the rate of convergence, remain an open question. In this article, we close this gap and offer for the first time a theoretical treatment of GMFVI in SSMs. A key upshot of our analysis is that the variational Bayes estimator $\hat{\theta}_B$ defined as the mean of the optimal GMFVI posterior $\widehat{q}_{\boldsymbol{\theta}}$ is estimation consistent and achieves a near-optimal rate of convergence. We conduct our analysis under the α-VB framework of Yang *et al.* (2020), where the joint likelihood of the observed and hidden variables is raised to a fractional power $\alpha \in (0, 1)$, and the corresponding α-fractional posterior (Bhattacharya *et al.*, 2019)

$$p_\alpha(\theta, X^n \mid Y^n) \propto \left[p(Y^n \mid X^n, \theta)\, p(X^n \mid \theta)\right]^\alpha p(\theta) \tag{6}$$

replaces the usual posterior distribution as the target. With this setup, the variational solution is defined as

$$\begin{aligned} \left(\widehat{q}_{\boldsymbol{\theta},\alpha}, \widehat{q}_{\boldsymbol{X^n},\alpha}\right) = \underset{(q_{\boldsymbol{\theta}}, q_{\boldsymbol{X}^n})}{\operatorname{argmin}} \Big\{ & D[q_{\boldsymbol{W}^n}(\cdot) \,\|\, p_\alpha(\cdot \mid Y^n)] \\ & + (1-\alpha)\mathcal{H}(q_{\boldsymbol{X}^n}) \Big\}, \end{aligned} \tag{7}$$

where $\mathcal{H}(q_{\boldsymbol{X}^n}) = -\int q_{\boldsymbol{X}^n} \log q_{\boldsymbol{X}^n} dX^n$ is the differential entropy of $q_{\boldsymbol{X}^n}$. When $\alpha = 1$, the above criterion reduces to the usual variational objective function. From a computational perspective, the α-VB objective can be trivially incorporated into existing algorithms, while certifying theoretical optimality for α-VB requires fewer assumptions than usual VB.

3 Variational Risk Bounds for State-Space Models

We now provide finite-sample risk bounds for GMFVI in SSMs. We operate in a frequentist framework by assuming the observations Y^n to be generated from model (1) with true parameters μ^* and λ^*. Define $\theta^* = (\mu^*, \lambda^*)$ as the true static parameter. We use a sample size rescaled α-Rényi divergence

$$D^{(n)}_\alpha(\theta, \theta^*) = n^{-1}\, D_\alpha(p^{(n)}_\theta \,||\, p^{(n)}_{\theta*})$$

as a measure of discrepancy between θ and θ^*, where recall that D_α is the α-Rényi divergence and $p^{(n)}_\theta$ denotes the marginal density of Y^n under parameter θ obtained by integrating out the latent variables:

$$p^{(n)}_\theta(Y^n) = \int p(Y^n | \, X^n, \theta)\, p(X^n | \, \theta)\, dX^n. \tag{8}$$

Here, we rescale the α-Rényi divergence by a factor of n^{-1} since $p^{(n)}_\theta$ is the joint distribution of n (dependent) observations. In the i.i.d. setting, $D^{(n)}_\alpha(\theta, \theta^*)$ reduces to $D_\alpha(p_\theta || p_{\theta*})$ due to the additivity, or tensorization, of D_α.

In the sequel, we aim to provide an upper bound to the variational risk $\int D^{(n)}_\alpha(\theta, \theta^*)\, \widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta)$, which is the expected discrepancy to the true data generating model with respect to the α-variational posterior $\widehat{q}_{\boldsymbol{\theta},\alpha}$ of the static parameter. The variational risk tends to be small as $\widehat{q}_{\boldsymbol{\theta},\alpha}$ becomes concentrated around θ^*.

Now, let us briefly discuss why the considered variational risk is meaningful. In fact, the commonly used risk criterion for characterizing distribution estimation with non i.i.d. data, the averaged square-Hellinger distance $h^2(\theta, \theta^*) := n^{-1} h^2(p^{(n)}_\theta, p^{(n)}_{\theta*})$ (Ghosal & Van Der Vaart, 2007), is bounded above by $D^{(n)}_{1/2}(\theta, \theta^*)$. In addition, the D_α family indexed by $\alpha \in (0, 1)$ satisfies $\frac{\alpha}{\beta}\frac{1-\beta}{1-\alpha} D_\beta \leqslant D_\alpha \leqslant D_\beta$ for any $0 < \beta \leqslant \alpha < 1$

(Van Erven & Harremos, 2014). Consequently, a bound on $\int D_\alpha^{(n)}(\theta,\theta^*)\,\widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta)$ for any fixed $\alpha \in (0,1)$ leads to the same bound up to a constant multiple on $\int h^2(\theta,\theta^*)\,\widehat{q}_{\theta,\alpha}(d\theta)$. In particular, a desirable variational risk bound should scale like n^{-1} in sample size n, since it implies the $n^{-1/2}$ parametric convergence rate of some point estimator based on $\widehat{q}_{\boldsymbol{\theta},\alpha}$ (Yang *et al.*, 2020).

We introduce some notation next. Let

$$\pi_{X^n} := p(X^n \mid Y^n, \theta)$$

be a shorthand notation for the conditional posterior of the latent variables given data and static parameter θ. We also use $\pi^*_{X^n}$ to denote the corresponding quantity conditioned on the true parameter θ^*. Given two non-decreasing functions $f_\lambda(n)$ and $f_\mu(n)$, define the following KL neighborhoods around $\pi^*_{X^n}$ and μ^* with respective radius ε_λ and ε_μ as

$$\begin{aligned}\mathcal{B}_n\left(\pi^*_{X^n},\varepsilon_\lambda\right) = \Big\{ & D\left(\pi^*_{X^n} \,\|\, \pi_{X^n}\right) \leqslant f_\lambda(n)\,\varepsilon_\lambda^2, \\ & V\left(\pi^*_{X^n} \,\|\, \pi_{X^n}\right) \leqslant f_\lambda(n)\,\varepsilon_\lambda^2\Big\},\end{aligned} \tag{9}$$

$$\begin{aligned}\mathcal{B}_n\left(\mu^*,\varepsilon_\mu\right) = \Big\{ & \max_{1\leqslant i\leqslant n} \mathbb{E}_{X^n\mid\theta^*}\, D_i(\mu^*,\mu) \leqslant f_\mu(n)\,\varepsilon_\mu^2, \\ & \max_{1\leqslant i\leqslant n} \mathbb{E}_{X^n\mid\theta^*}\, V_i(\mu^*,\mu) \leqslant f_\mu(n)\,\varepsilon_\mu^2\Big\},\end{aligned} \tag{10}$$

$$\begin{aligned}\text{where }\; D_i(\mu^*,\mu) &:= D\left[p\left(\cdot \mid \mu^*, X_i\right) \,\|\, p(\cdot \mid \mu, X_i)\right], \\ V_i(\mu^*,\mu) &:= V\left[p\left(\cdot \mid \mu^*, X_i\right) \,\|\, p(\cdot \mid \mu, X_i)\right].\end{aligned}$$

Here, recall that $V(p||q)$ denotes the V-divergence which will be useful in controlling the variance of the log-likelihood function when applying Chebyshev's inequality to derive a high probability bound. We are now prepared to state our main theoretical result; a proof sketch is provided in Section 6.

Theorem 1. *For any fixed* $(\varepsilon_\lambda,\varepsilon_\mu) \in (0,1)^2$ *and* $D > 1$, *with* $\mathbb{P}^{(n)}_{\theta^*}$ *probability at least* $1 - 5/\{(D-1)^2(\varepsilon_\lambda^2 f_\lambda(n) + \varepsilon_\mu^2 n f_\mu(n))\}$, *where the* $f_\lambda(n)$ *and* $f_\mu(n)$ *are defined in equation* (9) *and* (10), *it holds that*

$$\begin{aligned}\int D_\alpha^{(n)}\left(\theta,\theta^*\right)\,\widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta) \leqslant{} & \frac{D\alpha}{1-\alpha}\bigg(\frac{f_\lambda(n)}{n}\varepsilon_\lambda^2 \\ & + f_\mu(n)\varepsilon_\mu^2\bigg) + \left\{-\frac{1}{n(1-\alpha)}\log P_\lambda\left[\mathcal{B}_n\left(\pi^*_{X^n},\varepsilon_\lambda\right)\right]\right\} \\ & + \left\{-\frac{1}{n(1-\alpha)}\log P_\mu\left[B_n\left(\mu^*,\varepsilon_\mu\right)\right]\right\}.\end{aligned} \tag{11}$$

In the above display, P_π and P_μ respectively denote the prior probabilities defined for λ and μ respectively. As in Yang *et al.* (2020), the variational risk bound in Theorem 1 offers a trade-off between the sizes of the neighborhood and the prior probability assigned to these neighborhoods. A crucial difference from Yang *et al.* (2020) is the introduction of the blow-up factor $f_\lambda(n)$ and $f_\mu(n)$ inside the neighborhoods. Since Yang *et al.* (2020) only considered the case where the observation-latent variable pairs are mutually independent, such a quantity did not appear in their analysis. However, in the present case, our variational family preserves the dependence among latent variables via $q_{\boldsymbol{X}^n}$. As a consequence, the tensorization property (for i.i.d. variables) $D\big(\pi^*_{X^n} \,\|\, \pi_{X^n}\big) = nD\big(\pi^*_{X_1} \,\|\, \pi_{X_1}\big) \approx Cn\|\lambda - \lambda^*\|^2$ is no longer true, and we need to figure out the leading factor $f(n)$ so that $D\big(\pi^*_{X^n} \,\|\, \pi_{X^n}\big) \leqslant f(n)\|\lambda-\lambda^*\|^2$ holds for all λ's that are sufficiently close to λ^*, by dealing with the n-dimensional joint distribution of X^n. Introduction of $f_\lambda(n)$ and $f_\mu(n)$ allows added flexibility in calibrating the prior probability of the neighborhood. In particular, as we will show in Lemma 2, for weakly dependent latent variables, for example, those form an α-mixing Markov chain with $\alpha \in (0,1)$ (see, e.g. Bradley (2005) for a precise definition), we can take $f_\lambda(n)$ to be a multiple of n and $f_\mu(n) = O(1)$, which corresponds to the scalar LGSSM in Section 2.2 with $a^* \in (-1,1)$. Therefore, the folklore in time series analysis that weakly dependent data are not too much different from independent data also applies to a carefully designed variational inference.

The next theorem shows the necessity of preserving the latent variable dependence in $q_{\boldsymbol{X}^n}$ by showing that the full mean-field approximation (3) that factorizes everything will lead to estimation inconsistency.

Theorem 2. *Consider the full mean-field decomposition* (3) *for a scalar* LGSSM *model* (12). *Assume the true* $a^* \in (0,1)$. *Then we have* $\lim_{n\to+\infty}\int D_\alpha^{(n)}(\theta,\theta^*)\widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta) > c$ *for some constant* $c > 0$.

Theorem 2 extends the point estimation inconsistency result of Wang & Titterington (2004) (which is for the usual posterior, i.e., $\alpha-$VB with $\alpha = 1$) by first adapting their proof for the inconsistency of the variational posterior mean of a to the current α-VB framework, based on which a lack of consistency for the integrated variational Bayes risk can be proved

4 Applications to Concrete State-Space Models

We now illustrate how to apply Theorem 1 to specialized contexts. We begin with the simplest setup of a scalar LGSSM, and then extend the result to general LGSSMs. We finally consider a more general class of

SSMs under appropriate control over the KL and V divergences.

4.1 Risk Bound in Scalar LGSSMs

Recall the following scalar version of the LGSSM described in Section 2.2:

$$Y_t|X_t \overset{\text{ind.}}{\sim} \mathcal{N}(bX_t, \sigma_H^2),\ X_t|X_{t-1} \overset{\text{ind.}}{\sim} \mathcal{N}(aX_{t-1}, \sigma_V^2), \tag{12}$$

where static parameter $\theta = (a, b, \sigma_H^2, \sigma_V^2)$. We assume independent priors are imposed on the components of θ as $a \sim \mathcal{N}(0, \sigma_A^2)$, $b \sim \mathcal{N}(0, \sigma_B^2)$, $\sigma_H^2 \sim \mathrm{IG}(d_{H_1}, d_{H_2})$, $\sigma_V^2 \sim \mathrm{IG}(d_{V_1}, d_{V_2})$, where $\mathrm{IG}(a, b)$ denotes an inverse-Gamma distribution with shape parameter a and rate parameter b.

To apply Theorem 1 in this setup, let us analyze how to choose some theoretical tuning parameters ε_λ and ε_μ to make the variational risk upper bound in Theorem 1 small. Recall that this upper bound is the sum of two terms:

$$\frac{D\alpha\varepsilon_\lambda^2 f_\lambda(n)}{(1-\alpha)n} + \left\{-\frac{1}{n(1-\alpha)} \log P_\lambda \left[\mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_\lambda\right)\right]\right\}$$

$$\text{and } \frac{D\alpha\varepsilon_\mu^2 f_\mu(n)}{1-\alpha} + \left\{-\frac{1}{n(1-\alpha)} \log P_\mu \left[B_n\left(\mu^*, \varepsilon_\mu\right)\right]\right\}.$$

First, we need to determine the scaling of functions $f_\lambda(n)$ and $f_\mu(n)$ in the context of scalar LGSSM, which is summarized in the following lemma.

Lemma 1. *For LGSSM model* (12), *we have the following bounds on $f_\lambda(n)$ and $f_\mu(n)$ depending on the absoulte value of a:*

1. *If $|a| < 1$, then $f_\lambda(n) \leqslant Cn$ and $f_\mu(n) \leqslant C$;*
2. *If $|a| = 1$, then $f_\lambda(n) \leqslant Cn^3$ and $f_\mu(n) \leqslant Cn^2$;*
3. *If $|a| > 1$, then $f_\lambda(n) \geqslant e^{cn}$ and $f_\mu(n) \geqslant e^{cn}$.*

Here C, c are positive constants independent of n.

In this lemma, the cubic $O(n^3)$ lower bound for $f_\lambda(n)$ at $|a| = 1$ is due to the growth rate of the V divergence. Note that the time series $(X_i : i \in [n])$ forms an $|a|$-mixing Markov chain when $|a| < 1$, corresponding to the good regime where $f_\lambda(n) = O(n)$ and the overall variational risk bound behaves as the $O(n^{-1})$ parametric risk bound.

Based on the above lemma, $\big(f_\lambda(n), f_\mu(n)\big)$ can be chosen as (Cn, C) when $|a^*| < 1$ and (Cn^3, Cn^2) when $|a^*| = 1$ respectively. Then, we set $\frac{\varepsilon_\lambda^2 f_\lambda(n)}{n} = \varepsilon_\mu^2 f_\mu(n) = \frac{(\log n)^\beta}{n}$ to ensure that $(\varepsilon_\lambda^2 f_\lambda(n) + \varepsilon_\mu^2 n f_\mu(n)) \to \infty$, and $(\varepsilon_\lambda^2 f_\lambda(n) + \varepsilon_\mu^2 n f_\mu(n))/n \to 0$ no slower than $\frac{\log n}{n}$. With these choices, by carefully estimating $P_\lambda\left[\mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_\lambda\right)\right]$ and $P_\mu\left[B_n\left(\mu^*, \varepsilon_\mu\right)\right]$ as a function of n, we obtain the following corollary.

Corollary 1. *Suppose the true $|a^*| \leqslant 1$. Then, there exists $\beta > 0$, $C > 0$ and $D > 0$ s.t. with $\mathbb{P}_{\theta^*}^{(n)}$ probability at least $1 - D^{-2}(\log n)^{-\beta}$, it holds that*

$$\int D_\alpha^{(n)}\left(\theta, \theta^*\right) \widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta) \leqslant \frac{CD(\log n)^{\beta \vee 1}}{n}.$$

From this Corollary, it is evident that the GM-FVI achieves the optimal theoretical convergence rate $\mathcal{O}(1/n)$ under the condition $|a^*| \leqslant 1$. Minimax rates for SSMs in the literatures (De Castro *et al.*, 2016; Lehéricy, 2018) involve the $\log n/n$ scaling in n, so the presence of the logarithmic factor in the rate is not entirely an artifact of our proof.

4.2 Risk Bound in Multivariate LGSSMs

Next, we consider the general LGSSM defined in (2). We again assume that the following independent priors are imposed on all components of static parameter $\theta = (A, B, \Sigma_H, \Sigma_V)$ where Σ_H and Σ_V are diagonal matrices,

$$A_{i,j} \overset{ind.}{\sim} \mathcal{N}(0, \sigma_A^2),\ B_{k,l} \overset{ind.}{\sim} \mathcal{N}(0, \sigma_B^2),$$
$${\sigma_H}_k^2 \overset{ind.}{\sim} \mathrm{IG}(\tau_{H1}, \tau_{H2}),\ {\sigma_V}_j^2 \overset{ind.}{\sim} \mathrm{IG}(\tau_{V1}, \tau_{V2}),$$
$$\text{where } i, j, l \in [d_V],\ k \in [d_H].$$

Let $\rho_{\max}(A)$ denote the spectral radius of a square matrix matrix A, i.e., its maximum absolute eigenvalue. Also, denote $d^2 := d_V^2 \vee (d_V d_H)$ and use the same $f_\lambda(n)$ and $f_\mu(n)$ from Lemma 1. We then have the following result in the multivariate setup.

Corollary 2. *If $\rho_{\max}(A^*) \leqslant 1$, then there exists $\beta, C, D \geqslant 0$, such that with $\mathbb{P}_{\theta^*}^{(n)}$ probability at least $1 - D^{-2}(\log n)^{-\beta}$, it holds that*

$$\int D_\alpha^{(n)}\left(\theta, \theta^*\right) \widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta) \leqslant CD\left(\frac{(\log n)^\beta}{n} \vee \frac{d^2 \log n}{n}\right)$$

The assumption $\rho_{\max}(A^*) \leqslant 1$ can be viewed as a generalization of the assumption $|a^*| \leqslant 1$ in Corollary 1 in the scalar case. Since the number of parameters now is of order d^2, the optimal rate is d^2/n modulo logarithmic terms.

4.3 Risk Bound in General State-Space Models

Finally, consider the general SSM as in (1). Still writing the parameters as $\theta = (\lambda, \mu)$, and assuming that they have dimension (d_λ, d_μ), we have the following convergence bound.

Corollary 3. *Assume there exists $C, C_1, C_2, D, \beta > 0$*

such that

$$\begin{aligned}&\max\left\{D\left(\pi^*_{X^n}\parallel\pi_{X^n}\right), V\left(\pi^*_{X^n}\parallel\pi_{X^n}\right)\right\}\\&\leqslant C_1 f_\lambda(n)\|\lambda-\lambda^*\|^2,\\&\max\max_{1\leqslant k\leqslant n}\left\{\mathbb{E}_{X^n|\theta*}\,D_k(\mu^*,\mu),\mathbb{E}_{X^n|\theta*}\,V_k(\mu^*,\mu)\right\}\\&\leqslant C_2 f_\mu(n)\|\mu-\mu^*\|^2.\end{aligned}$$

where $\|\cdot\|$ is the Euclidean metric. Also assume that the multivariate prior densities p_λ and p_μ are Riemann integrable over their domain. Then, we have the following bound with $\mathbb{P}^{(n)}_{\theta*}$ probability at least $1-D^{-2}(\log n)^{-\beta}$,

$$\int D^{(n)}_\alpha\left(\theta,\theta^*\right)\widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta)\leqslant CD\left(\frac{(\log n)^\beta}{n}\vee\frac{d_\lambda\log f_\lambda(n)}{n}\vee\frac{d_\mu\log(nf_\mu(n))}{n}\right). \tag{13}$$

Corollary can be applied to general state-space models subject to the quadratic bounds above on appropriate KL and V divergences. Note that a similar Lipschitz assumption on the densities also appears in Example 7.1 of Ghosal *et al.* (2000) and is typically satisfied for a wide range of parametric densities. Since in our case, the assumption is directly on the joint densities as opposed to on the independent marginals as in Ghosal *et al.* (2000), it is important to characterize the role of the data (in particular through the sample size) in the Lipschitz constant. From this result, one can always achieve the parametric variational risk $O(1/n)$ (modulo poly-$\log n$ terms) as long as $f_\lambda(n)$ and $f_\mu(n)$ grow at most polynomially in n. This explains why Corollary 1 (as a special case of Corollary 3) requires condition $|a^*|\leqslant 1$ (cf. Lemma 2). The same comment applies to Corollary 2.

Remark. *The choice of the Euclidean metric $\|\cdot\|$ in Corollary 3 is not a necessary assumption. To accommodate a more complex dependence structure of θ on p_θ, usage of other metrics may be considered leading to a possible alteration of the convergence rate.*

5 Comments on Implementation

Although the primary focus of this article is theoretical, we offer some discussion about implementation in general state-space models. We also provide numerical illustration for a scalar LGSSM in the supplemental document, where we compute the Bayes risk in a small-scale simulation study. The CAVI algorithm for LGSSMs in (5) can be adapted to general conditionally conjugate SSMs as long as the belief propagation can be analytically carried out, and the variational distribution for the static parameters can be updated in a conjugate fashion. The pseudo-code in Algorithm 1 offers a general template. Beyond the linear Gaussian setup, there is subclass of exponential family named exponential dispersion family (EDM) (Jørgensen, 1992) whose probability density functions or probability mass functions taking the form of

$$f(y;\theta,\phi)=a(y,\phi)\exp\left[\frac{1}{\phi}\{y\theta-\kappa(\theta)\}\right],\quad y\in\mathbb{R}$$

for suitable known functions $a(\cdot,\cdot)$ and $\kappa(\cdot)$, where θ the canonical parameter taking values in an open interval such that the log-partition function $\kappa(\theta)<\infty$, and $\phi>0$ is the dispersion parameter. This family retains the conjugacy of $p(X_i|\,Y^n,\theta)$ so that the belief propagation can be performed with analytically closed forms. This family has been used as the generative model for the latent process; see Vidoni (1999) for example. In particular, further restricting to the Tweedie EDM (Dunn & Smyth, 2005), whose variance function takes polynomial form, an iterative algorithm very similar to Barber & Chiappa (2007) can be developed. Alternatively, a hybrid algorithm combining Monte Carlo approximation to the expectations in the updating formula (5) can be employed in the absence of conjugacy.

Algorithm 1: Pseudo Algorithm for CAVI in a General Conjugate SSM

```
Data: $Y^n$;
Result: $\widetilde{q}(X^n,\theta) := q_{\boldsymbol{X}^n}(X^n)q_{\boldsymbol{\theta}}(\theta)$;
while $q_{\boldsymbol{X}^n}(X^n)$ and $q_{\boldsymbol{\theta}}(\theta)$ haven't converged do
    for $i = 1$ to $i = n$ do
        Calculate $q^{(s)}_{\boldsymbol{X}^n}(X_i)$ using the belief
         propagation in the subroutine
         Algorithm 2 ;   /* Belief Propagation
         */
    end
    Calculate $q^{(s+1)}_{\boldsymbol{\theta}}$ using conditional conjugacy;
     /* Conjugacy */
end
```

6 Proof Sketch of Theorem 1

Here we provide a sketch of the proof of Theorem 1. It follows from Yang *et al.* (2020) that the α-VB objective function in the right hand side of (7) can be equivalently expressed as

$$\begin{aligned}\Psi_{n,\alpha}\left(q_{\boldsymbol{\theta}},q_{\boldsymbol{X}^n}\right):=&\int_\Theta\left(\ell_n\left(\theta^*\right)-\widehat{\ell}_n(\theta)\right)q_{\boldsymbol{\theta}}(d\theta)\\&+\alpha^{-1}D\left(q_{\boldsymbol{\theta}}\|p_{\boldsymbol{\theta}}\right),\end{aligned}$$

where $\ell_n\left(\theta^*\right)=\log p(Y^n\mid\theta^*)$ is the (marginal) log-likelihood function evaluated at the true parameter θ^*,

Algorithm 2: Forward-Backward Algorithm for the Belief Propagation in Algorithm 1

```
Data: $Y^n, q_\theta(\theta)$;
Result: $q_{X^n}(X_i)$;
FORWARD:
for $i = 1$ to $i = n$ do
    Calculate $q(X_i \mid Y_{1:i}, \theta) \propto \int q(X_{i-1} \mid Y_{1:i-1}, \theta) q(X_{i-1} \mid X_i) q(Y_i \mid X_i) dX_{i-1}$ ;
end
BACKWARD:
for $j = n-1$ to $j = 1$ do
    Calculate $q(X_j \mid X_{j+1}, Y^n, \theta) \propto q(X_{j+1} \mid X_j, \theta) q(X_j \mid Y_{1:j}, \theta)$;
    Then, calculate $q(X_j \mid Y^n, \theta) = \int q(X_j \mid X_{j+1}, Y^n, \theta) q(X_{j+1} \mid Y^n, \theta) dX_{j+1}$ ;
    Then, $q_{X^n}(X_i) = \int q(X_j \mid Y^n, \theta) q_\theta(\theta) d\theta$ ;
end
```

$p_{\boldsymbol{\theta}}$ is the prior distribution on θ, and $\widehat{\ell}_n(\theta)$ is defined as

$$\widehat{\ell}_n(\theta) = \int_{X^n} q_{\boldsymbol{X}^n}(X^n) \log \frac{p(Y^n \mid \mu, X^n) \pi_{X^n}}{q_{\boldsymbol{X}^n}(X^n)}.$$

In their Theorem 3.1, Yang *et al.* (2020) showed that for any $\zeta \in (0,1)$, it holds with $\mathbb{P}_{\theta*}^n$ probability at least $(1-\zeta)$ that for any probability measure $q_{\boldsymbol{\theta}} \in \Gamma_{\boldsymbol{\theta}}$ with $q_{\boldsymbol{\theta}} \ll p_{\boldsymbol{\theta}}$ and any probability measure $q_{\boldsymbol{X}^n} \in \Gamma_{\boldsymbol{X}^n}$ on X^n,

$$\int D_\alpha^{(n)}(\theta, \theta^*) \widehat{q}_{\boldsymbol{\theta},\alpha}(\theta) d\theta \leqslant \frac{\alpha}{n(1-\alpha)} \Psi_{n,\alpha}(q_{\boldsymbol{\theta}}, q_{\boldsymbol{X}^n}) + \frac{1}{n(1-\alpha)} \log(1/\zeta).$$

This result relates the variational risk bound to the variational objective $\Psi_{n,\alpha}$. The general strategy then is to make judicious choices of $q_{\boldsymbol{\theta}}$ and $q_{\boldsymbol{X}^n}$ in the right hand side of the above display to appropriately control $\Psi_{n,\alpha}(q_{\boldsymbol{\theta}}, q_{\boldsymbol{X}^n})$. However, in all their statistical examples, q_{X^n} had a further mean-field decomposition, and additional care needs to be exercised to make these choices in the presence of dependence as in our context.

In our context, we first set

$$q_{\boldsymbol{X}^n} := \widetilde{q}_{\boldsymbol{X}^n}(X^n) \propto P(Y^n \mid \mu, X^n) P(X^n \mid \theta^*)$$

to get the following bound

$$\int D_\alpha^{(n)}(\theta, \theta^*) \widehat{q}_{\theta,\alpha}(\theta) d\theta \leqslant \frac{\alpha}{n(1-\alpha)} \Bigg[- \int_\Theta \int_{X^n} \widetilde{q}_{\boldsymbol{X}^n}(X^n) \log \frac{P(Y^n \mid \mu, X^n) \pi_{X^n}}{P(Y^n \mid \mu^*, X^n) \pi_{X^n}^*} q_{\boldsymbol{\theta}}(d\theta) + \frac{D(q_{\boldsymbol{\theta}} \| p_\theta)}{\alpha} + \frac{\log(1/\zeta)}{\alpha} \Bigg].$$

The above inequality is true for any $q_{\boldsymbol{\theta}} \ll p_{\boldsymbol{\theta}}$. We specifically choose $q_{\boldsymbol{\theta}} := \widetilde{q}_{\boldsymbol{\theta}}$ as the probability density function of the probability measure $\widetilde{Q}_{\boldsymbol{\theta}}$ given by

$$\widetilde{Q}_{\boldsymbol{\theta}}(\cdot) = \frac{P_\pi[\cdot \cap \mathcal{B}_n(\pi_{X^n}^*, \varepsilon_\lambda)] \otimes P_\mu[\cdot \cap \mathcal{B}_n(\mu^*, \varepsilon_\mu)]}{P_\pi[\mathcal{B}_n(\pi_{X^n}^*, \varepsilon_\lambda)] \cdot P_\mu[\mathcal{B}_n(\mu^*, \varepsilon_\mu)]}.$$

With this choice, we proceed to further bound the right hand side of the previous display. By applying Fubini's theorem, we obtain

$$\begin{aligned} &\mathbb{E}_{\theta*}\Bigg[\int_\Theta \widetilde{q}_{\boldsymbol{\theta}}(\theta) \Bigg[\int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) p(X^n \mid \theta)}{p(Y^n \mid \mu^*, X^n) p(X^n \mid \theta^*)} dX^n \Bigg] d\theta \Bigg] \\ &= \int_\Theta \mathbb{E}_{\theta*}\Bigg[\int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) p(X^n \mid \theta)}{p(Y^n \mid \mu^*, X^n) p(X^n \mid \theta^*)} dX^n \Bigg] \widetilde{q}_{\boldsymbol{\theta}}(\theta) d\theta \end{aligned}$$

Simplifying the the inner integral over a number of steps, we obtain:

$$\mathbb{E}_{\theta*}\Bigg[\int_{X^n} -\widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) p(X^n \mid \theta)}{p(Y^n \mid \mu^*, X^n) p(X^n \mid \theta^*)} dX^n \Bigg] \leqslant f_\lambda(n) \varepsilon_\lambda^2 + n f_\mu(n) \varepsilon_\mu^2$$

A similar analysis leads to

$$\mathrm{Var}_{\theta*}\Bigg[\int_\Theta \widetilde{q}_{\boldsymbol{\theta}}(\theta) \Bigg[\int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n)}{p(Y^n \mid \mu^*, X^n)} \frac{p(X^n \mid \theta)}{p(X^n \mid \theta^*)} dX^n \Bigg] d\theta \Bigg] \leqslant 2 f_\lambda(n) \varepsilon_\lambda^2 + 2n f_\mu(n) \varepsilon_\mu^2$$

Putting pieces together, we obtain by applying Chebyshev's inequality that

$$\begin{aligned} &P_{\theta*}\Bigg\{\int_\Theta \widetilde{q}_{\boldsymbol{\theta}}(\theta) \int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) \pi_{X^n}}{p(Y^n \mid \mu^*, X^n) \pi_{X^n}^*} dX^n d\theta \leqslant -D(f_\lambda(n) \varepsilon_\lambda^2 + n f_\mu(n) \varepsilon_\mu^2) \Bigg\} \\ &\leqslant P_{\theta*}\Bigg\{\int_\Theta \widetilde{q}_{\boldsymbol{\theta}}(\theta) \int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) \pi_{X^n}}{p(Y^n \mid \mu^*, X^n) \pi_{X^n}^*} dX^n d\theta - \mathbb{E}_{\theta*}\Bigg[\int_\Theta \widetilde{q}_{\boldsymbol{\theta}}(\theta) \int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) \pi_{X^n}}{p(Y^n \mid \mu^*, X^n) \pi_{X^n}^*} dX^n d\theta \Bigg] \\ &\leqslant - (D-1)(f_\lambda(n) \varepsilon_\lambda^2 + n f_\mu(n) \varepsilon_\mu^2 2) \Bigg\} \\ &\leqslant \frac{4}{(D-1)^2 (\varepsilon_\lambda^2 f_\lambda(n) + \varepsilon_\mu^2 n f_\mu(n))} \end{aligned}$$

Finally, we have

$$D\left(\widetilde{q}_{\boldsymbol{\theta}} \| p_{\theta}\right) = -[\log P_{\pi}\left[\mathcal{B}_{n}\left(\lambda^{*}, \varepsilon_{\lambda}\right)\right] + \log P_{\mu}\left[B_{n}\left(\mu^{*}, \varepsilon_{\mu}\right)\right]],$$

since for any probability measure ϕ, a measurable set A with $\phi(A) > 0$, and $\widetilde{\phi}(\cdot) = \phi(\cdot \cap A)/\phi(A)$ the restriction of ϕ to A, $D(\widetilde{\phi} \| \phi) = -\log \phi(A)$. This completes the proof sketch. A complete proof and all remaining proofs can be found in the supplemental document.

7 Discussions

In this paper, we illustrate that mean-field variational inference may lead to inconsistent parameter estimates. In particular, in the context of Bayesian state-space models, carefully designed variational families that preserve dependence among latent variables dependence are necessary to achieve parameter estimation consistency. One immediate future direction is to extend the current method and theory to time-inhomogeneous state-space models where the transition and emission matrices may evolve over time. Another interesting direction is to extend the current variational framework to other latent variable models beyond the state-space models where the underlying graphical structure of the latent variables is more complicated than a chain, for example, can be a tree or even contain loops. Last, it is also of practical importance to analyze the algorithmic properties of the aforementioned CAVI algorithms, or more generally belief propagation algorithms, by identifying a minimal set of conditions under which the algorithm converges polynomially or exponentially to the variational solution $(\widetilde{q}_{\boldsymbol{\theta},\alpha}, \widetilde{q}_{\boldsymbol{X}^n,\alpha})$.

References

Alquier, Pierre, & Ridgway, James. 2020. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, **48**(3), 1475–1497.

Barber, David, & Chiappa, Silvia. 2007. Unified Inference for Variational Bayesian Linear Gaussian State-Space Models. *Pages 81–88 of:* Schölkopf, B., Platt, J. C., & Hoffman, T. (eds), *Advances in Neural Information Processing Systems 19*. MIT Press.

Bhattacharya, Anirban, Pati, Debdeep, & Yang, Yun. 2019. Bayesian fractional posteriors. *The Annals of Statistics*, **47**(1), 39–66.

Blei, David M, Kucukelbir, Alp, & McAuliffe, Jon D. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, **112**(518), 859–877.

Bradley, Richard C. 2005. Basic properties of strong mixing conditions. A survey and some open questions. *Probability surveys*, **2**, 107–144.

De Castro, Yohann, Gassiat, Élisabeth, & Lacour, Claire. 2016. Minimax adaptive estimation of nonparametric hidden Markov models. *The Journal of Machine Learning Research*, **17**(1), 3842–3884.

Dunn, Peter K, & Smyth, Gordon K. 2005. Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing*, **15**(4), 267–280.

Ghosal, Subhashis, & Van Der Vaart, Aad. 2007. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, **35**(1), 192–223.

Ghosal, Subhashis, Ghosh, Jayanta K, & Van Der Vaart, Aad W. 2000. Convergence rates of posterior distributions. *Annals of Statistics*, 500–531.

Ghosh, Indrajit, Bhattacharya, Anirban, & Pati, Debdeep. 2020. Statistical optimality and stability of tangent transform algorithms in logit models. *arXiv preprint arXiv:2010.13039*.

Jørgensen, Bent. 1992. Exponential dispersion models and extensions: A review. *International Statistical Review/Revue Internationale de Statistique*, 5–20.

Kantas, Nicholas, Doucet, Arnaud, Singh, Sumeetpal Sindhu, & Maciejowski, Jan Marian. 2009. An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. *IFAC Proceedings Volumes*, **42**(10), 774–785.

Koller, Daphne, & Friedman, Nir. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Lehéricy, Luc. 2018. State-by-state minimax adaptive estimation for nonparametric hidden Markov models. *The Journal of Machine Learning Research*, **19**(1), 1432–1477.

Pati, Debdeep, Bhattacharya, Anirban, & Yang, Yun. 2018. On statistical optimality of variational Bayes. *Pages 1579–1588 of: International Conference on Artificial Intelligence and Statistics*. PMLR.

Pearl, Judea. 2014. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Plummer, Sean, Zhou, Shuang, Bhattacharya, Anirban, Dunson, David, & Pati, Debdeep. 2021. Statistical Guarantees for Transformation Based Models with Applications to Implicit Variational Inference. *Pages 2449–2457 of: International Conference on Artificial Intelligence and Statistics*. PMLR.

Ray, Kolyan, & Szabó, Botond. 2021. Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 1–12.

Turner, R. E., & Sahani, M. 2011. Two problems with variational expectation maximisation for time-series models. *Chap. 5, pages 109–130 of:* Barber, D., Cemgil, T., & Chiappa, S. (eds), *Bayesian Time series models*. Cambridge University Press.

Van Erven, Tim, & Harremos, Peter. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, **60**(7), 3797–3820.

Vidoni, Paolo. 1999. Exponential family state space models based on a conjugate latent process. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(1), 213–221.

Wang, Bo, & Titterington, DM. 2004. Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, **20**(3), 151–170.

Weiss, Yair. 2000. Correctness of local probability propagation in graphical models with loops. *Neural computation*, **12**(1), 1–41.

Welch, Greg, Bishop, Gary, *et al.* 1995. An introduction to the Kalman filter.

Xing, Eric P, Jordan, Michael I, & Russell, Stuart. 2012. A generalized mean field algorithm for variational inference in exponential families. *arXiv preprint arXiv:1212.2512*.

Yang, Yun, Pati, Debdeep, & Bhattacharya, Anirban. 2020. α-variational inference with statistical guarantees. *Annals of Statistics*, **48**(2), 886–905.

Yedidia, Jonathan S, Freeman, William T, Weiss, Yair, *et al.* 2000. Generalized belief propagation. *Pages 689–695 of: NIPS*, vol. 13.

Yedidia, Jonathan S, Freeman, William T, & Weiss, Yair. 2003. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, **8**, 236–239.

Zeng, Yong, & Wu, Shu. 2013. *State-space models: Applications in economics and finance*. Vol. 1. Springer.

Zhang, Fengshuo, & Gao, Chao. 2020. Convergence rates of variational posterior distributions. *The Annals of Statistics*, **48**(4), 2180–2207.

Supplementary Material: Structured Variational Inference in Bayesian State-Space Models

A Proof of Theorem 1:

Proof: Theorem 3.1 from Yang *et al.* (2020) showed that for any $\zeta \in (0,1)$, it holds with $\mathbb{P}^n_{\theta*}$ probability at least $(1-\zeta)$ that for any probability measure $q_{\boldsymbol{\theta}} \in \Gamma_{\boldsymbol{\theta}}$ with $q_{\boldsymbol{\theta}} \ll p_{\boldsymbol{\theta}}$ and any probability measure $q_{\boldsymbol{X}^n} \in \Gamma_{\boldsymbol{X}^n}$ on X^n,

$$\int D_{\alpha}^{(n)}\left(\theta, \theta^*\right) \widehat{q}_{\boldsymbol{\theta},\alpha}(\theta) d\theta \leqslant \frac{\alpha}{n(1-\alpha)} \Psi_{n,\alpha}\left(q_{\boldsymbol{\theta}}, q_{\boldsymbol{X}^n}\right) + \frac{1}{n(1-\alpha)} \log(1/\zeta),$$

where

$$\Psi_{n,\alpha}\left(q_{\boldsymbol{\theta}}, q_{\boldsymbol{X}^n}\right) := \int_{\Theta} \Big(\ell_n\left(\theta^*\right) - \widehat{\ell}_n(\theta)\Big) q_{\boldsymbol{\theta}}(d\theta) + \alpha^{-1} D\left(q_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}}\right),$$

and

$$\widehat{\ell}_n(\theta) := \int_{X^n} q_{\boldsymbol{X}^n}\left(X^n\right) \log \frac{p\left(Y^n \mid \mu, X^n\right) \pi_{X^n}}{q_{\boldsymbol{X}^n}\left(X^n\right)}.$$

In our context, we first set

$$q_{\boldsymbol{X}^n} := \widetilde{q}_{\boldsymbol{X}^n}(X^n) \propto P(Y^n \mid \mu, X^n)\, P(X^n \mid \theta^*)$$

to get the following bound

$$\int D_{\alpha}^{(n)}\left(\theta, \theta^*\right) \widehat{q}_{\theta,\alpha}(\theta) d\theta \leqslant \frac{\alpha}{n(1-\alpha)} \Bigg[-\int_{\Theta} \int_{X^n} \widetilde{q}_{\boldsymbol{X}^n}\left(X^n\right) \log \frac{P(Y^n \mid \mu, X^n) \pi_{X^n}}{P(Y^n \mid \mu^*, X^n) \pi^*_{X^n}} q_{\boldsymbol{\theta}}(d\theta) + \frac{D\left(q_{\boldsymbol{\theta}} \| p_{\theta}\right)}{\alpha} + \frac{\log(1/\zeta)}{\alpha} \Bigg].$$

The above inequality is true for any $q_{\boldsymbol{\theta}} \ll p_{\boldsymbol{\theta}}$. We specifically choose $q_{\boldsymbol{\theta}} := \widetilde{q}_{\boldsymbol{\theta}}$ as the probability density function of the probability measure $\widetilde{Q}_{\boldsymbol{\theta}}$ given by

$$\widetilde{Q}_{\boldsymbol{\theta}}(\cdot) = \frac{P_{\pi}\left[\cdot \cap \mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_{\lambda}\right)\right] \otimes P_{\mu}\left[\cdot \cap \mathcal{B}_n\left(\mu^*, \varepsilon_{\mu}\right)\right]}{P_{\pi}\left[\mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_{\lambda}\right)\right] \cdot P_{\mu}\left[\mathcal{B}_n\left(\mu^*, \varepsilon_{\mu}\right)\right]}.$$

With this choice, we proceed to further bound the right hand side of the previous display. By applying Fubini's theorem, we obtain

$$\begin{aligned} &\mathbb{E}_{\theta*}\left[\int_{\Theta} \widetilde{q}_{\boldsymbol{\theta}}(\theta) \left[\int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) p(X^n \mid \theta)}{p(Y^n \mid \mu^*, X^n) p(X^n \mid \theta^*)}\, dX^n\right] d\theta\right] \\ &= \int_{\Theta} \mathbb{E}_{\theta*}\left[\left[\int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) p(X^n \mid \theta)}{p(Y^n \mid \mu^*, X^n) p(X^n \mid \theta^*)}\, dX^n\right] \widetilde{q}_{\boldsymbol{\theta}}(\theta) d\theta \right.\end{aligned}$$

Simplifying the the inner integral over a number of steps, we obtain:

$$\mathbb{E}_{\theta*}\left[\int_{X^n} -\widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) p(X^n \mid \theta)}{p(Y^n \mid \mu^*, X^n) p(X^n \mid \theta^*)} dX^n\right] \leqslant f_{\lambda}(n) \varepsilon_{\lambda}^2 + n f_{\mu}(n) \varepsilon_{\mu}^2$$

A similar analysis leads to

$$\mathrm{Var}_{\theta*}\left[\int_{\Theta} \widetilde{q}_{\boldsymbol{\theta}}(\theta) \left[\int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n) p(X^n \mid \theta)}{p(Y^n \mid \mu^*, X^n) p(X^n \mid \theta^*)} dX^n\right] d\theta\right] \leqslant 2 f_{\lambda}(n) \varepsilon_{\lambda}^2 + 2n f_{\mu}(n) \varepsilon_{\mu}^2$$

Putting pieces together, we obtain by applying Chebyshev's inequality that

$$\begin{aligned}
&P_{\theta *}\left\{\int_{\Theta} \widetilde{q}_{\boldsymbol{\theta}}(\theta) \int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n)\pi_{X^n}}{p(Y^n \mid \mu^*, X^n)\pi^*_{X^n}} dX^n d\theta\right. \\
&\left.\leqslant - D(f_\lambda(n)\varepsilon_\lambda^2 + nf_\mu(n)\varepsilon_\mu^2)\right\} \\
&\leqslant P_{\theta *}\left\{\int_{\Theta} \widetilde{q}_{\boldsymbol{\theta}}(\theta) \int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n)\pi_{X^n}}{p(Y^n \mid \mu^*, X^n)\pi^*_{X^n}} dX^n d\theta\right. \\
&- \mathbb{E}_{\theta *}\left[\int_{\Theta} \widetilde{q}_{\boldsymbol{\theta}}(\theta) \int_{X^n} \widetilde{q}(X^n) \log \frac{p(Y^n \mid \mu, X^n)\pi_{X^n}}{p(Y^n \mid \mu^*, X^n)\pi^*_{X^n}} dX^n d\theta\right] \\
&\left.\leqslant - (D-1)(f_\lambda(n)\varepsilon_\lambda^2 + nf_\mu(n)\varepsilon_\mu^2 2)\right\} \\
&\leqslant \frac{4}{(D-1)^2(\varepsilon_\lambda^2 f_\lambda(n) + \varepsilon_\mu^2 nf_\mu(n))}
\end{aligned}$$

Finally, we have

$$D\left(\widetilde{q}_{\boldsymbol{\theta}} \| p_\theta\right) = -\left[\log P_\pi\left[\mathcal{B}_n\left(\lambda^*, \varepsilon_\lambda\right)\right] + \log P_\mu\left[B_n\left(\mu^*, \varepsilon_\mu\right)\right]\right],$$

since for any probability measure ϕ, a measurable set A with $\phi(A) > 0$, and $\widetilde{\phi}(\cdot) = \phi(\cdot \cap A)/\phi(A)$ the restriction of ϕ to A, $D(\widetilde{\phi}\|\phi) = -\log\phi(A)$. This completes the proof sketch. A complete proof and all remaining proofs can be found in the supplemental document.

B Proof of Theorem 2:

Proof: For showing the inconsistency of the $\alpha-$VI estimator under the MF regime, we can show the bayesian risk we defined at the beginning of §3 I can $r_n := \int D_\alpha^{(n)}\left(a, a^*\right) \widehat{q}_{a,\alpha}(da)$ does not go to 0 in the limit case when $a^* \in (0,1)$. For showing this, we only need to prove the poterior mean for a is not consistently converging to the true transmitting parameter a^*. Denoting the poterior mean as $\hat{a}_\alpha$. Since we set a in a normal distribution, the poterior of a is given by $\hat{q}_{a,\alpha} = \mathcal{N}(\hat{a}_\alpha, \hat{\sigma}^2_{a,\alpha})$. For simplicity, we just denote them as $(\hat{a}, \hat{\sigma}^2_a)$ as $\alpha = 1$. Under this regime, the inconsistency result $\lim\limits_{n\to\infty} |\hat{a} - a^*| > c_0$ has already been shown in Wang & Titterington (2004). For showing the inconsistency for the bayesian risk r_n in the general $\alpha-$vb case, we first need to show $\lim\limits_{n\to\infty} \hat{a} = \lim\limits_{n\to\infty} \hat{a}_\alpha$, and then we can use this result to show $\lim\limits_{n\to\infty} r_n > 0$ which is just what we want.
In order to show the inconsistency result for general α in the MF setup, we need to put the connection between the objective function in two cases. The $\alpha-$VB solver defined in our paper is

$$\left(\widehat{q}_{\boldsymbol{\theta},\alpha}, \widehat{q}_{\boldsymbol{X}^n,\alpha}\right) = \underset{\left(q_{\boldsymbol{\theta}}, q_{\boldsymbol{X}^n}\right)\in\Gamma}{\operatorname{argmin}} \ \Psi_{n,\alpha}\left(q_{\boldsymbol{\theta}}, q_{\boldsymbol{X}^n}\right),$$

which is equivalent to the formation

$$\underset{\left(q_{\boldsymbol{\theta}}, q_{\boldsymbol{X}^n}\right)\in\Gamma}{\operatorname{argmin}} \ -\int_\Theta \widehat{\ell}_n(\theta) q_{\boldsymbol{\theta}}(d\theta) + \alpha^{-1} D\left(q_{\boldsymbol{\theta}} \| p_{\boldsymbol{\theta}}\right).$$

In Wang & Titterington (2004), the objective function is just

$$-\int q(X,\theta)\log\frac{p(\theta, X, Y)}{q(X,\theta)} dXd\theta,$$

which is the case when we set $\alpha = 1$ in our configuration.
Next, we introduce the MF variational family as $q_{X_k} \sim \mathcal{N}(\mu_k, \sigma_k^2)$ and the prior for a is $p_a = \mathcal{N}(0, \sigma_A^2)$. Also, we set $X_0 \sim \mathcal{N}(0, \sigma_0^2)$. Taking derivatives over the objective function and make them all equal to 0, we can get following equations,

$$\sigma_k^2 = \frac{\sigma_0^2}{1 + \hat{a}_\alpha^2 + \hat{\sigma}^2_{a,\alpha} + b^2}, \tag{S1}$$

and

$$\sigma_n^2 = \frac{\sigma_0^2}{1+b^2}, \tag{S2}$$

and

$$\begin{pmatrix} 1+\hat{a}_\alpha^2+\hat{\sigma}_{a,\alpha}^2+b^2 & -\hat{a}_\alpha & \cdots & 0 \\ -\hat{a}_\alpha & 1+\hat{a}_\alpha^2+\hat{\sigma}_{a,\alpha}^2+b^2 & -\hat{a}_\alpha & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\hat{a}_\alpha & 1+b^2 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \vdots \\ \mu_n \end{pmatrix} = \Phi, \tag{S3}$$

where Φ is defined as $\Phi = bY^n$. Y^n is the observation vector in the scalar case. We can then optimize the objective function w.r.t. $(\hat{a}_\alpha, \hat{\sigma}_{a,\alpha}^2)$. Plugging (S1)-(S3) back, we get

$$F_{\alpha,n}(\hat{a}_\alpha, \hat{\sigma}_{a,\alpha}^2) = \alpha^{-1}\left\{\frac{1}{2}\log\hat{\sigma}_{a,\alpha}^2 - \frac{1}{2\sigma_A^2}\left\{\hat{\sigma}_{a,\alpha}^2 + \hat{a}_\alpha^2\right\}\right\} - \frac{1}{2\sigma_0^2}\left(\hat{\sigma}_{a,\alpha}^2 + \hat{a}_\alpha^2\right)\sum_{k=1}^{n-1}\left(\sigma_k^2 + \mu_k^2\right) + \frac{\hat{a}_\alpha}{\sigma_0^2}\sum_{k=1}^{n-1}\mu_k\mu_{k+1}. \tag{S4}$$

Denoting $f_{\alpha,n} := \frac{1}{n}F_{\alpha,n}$, (S4) can then be written as

$$\begin{aligned} f_{\alpha,n} &= f_n + \frac{(\alpha^{-1}-1)}{n}\left\{\frac{1}{2}\log\hat{\sigma}_{a,\alpha}^2 - \frac{1}{2\sigma_A^2}\left\{\hat{\sigma}_{a,\alpha}^2 + \hat{a}_\alpha^2\right\}\right\} \\ &= f_n + g_{\alpha,n}. \end{aligned} \tag{S5}$$

Without loss of generality, we can assume $(\hat{a}_\alpha, \hat{\sigma}_{a,\alpha}) \in D_1 := [-C_1, C_1] \times [c_1, C_1]$, where $C_1 > c_1 > 0$. Easy to verify that $g_{\alpha,n} \rightrightarrows 0$ and $g_{\alpha,n}^{'} \rightrightarrows 0$ on D_1. So, we have $\lim_{n\to\infty} f_{\alpha,n}^{'} = \lim_{n\to\infty} f_n^{'}$ on D_1. So the result $\lim_{n\to\infty}\hat{a} = \lim_{n\to\infty}\hat{a}_\alpha$ can be shown by the uniform convergence. Since $\lim_{n\to\infty}|\hat{a} - a^*| > c_0$ holds for some $c_0 > 0$. Define the region $D_2 := (\hat{a}_\alpha - c_2, \hat{a}_\alpha + c_2)$ where $c_0 > c_2 > 0$. We have $\int_{D_2}\widehat{q}_{a,\alpha}(da) > c_3$ for large n where $c_3 > 0$. So $\lim_{n\to\infty} r_n > 0$. The inconsistency then gets proved.

C Risk Bound in Scalar Case:

Let's first define two functions $S_{1(n)}$ and $S_{2(n)}$ for the convenient notation. Both will be used in later sections. They are used to quantify the KL divergence and V divergence. In this way, we can easily analyse the growth rate for these two divergence with different value taken for $\max|\lambda_{A^*}|$ as n increases. Based on this, we get the hint for how to pick the blow-up factors $f_\lambda(n)$ and $f_\mu(n)$.

$$S_{1(n)}(t) = (n + (n-1)t^2 \cdots + t^{2(n-1)}) = \frac{1-t^{2(n+1)}}{(1-t^2)^2} + \frac{n+1}{1-t^2}, \quad \text{when} \quad |t| \neq 1.$$

$$S_{1(n)}(t) = \frac{n(n+1)}{2}, \quad \text{when} \quad |t| = 1.$$

Also denote

$$S_{2(n)}(t) = \sum_{i=1}^{n}\left(\sum_{j=0}^{i-1} t^{2j}\right)^2 = \sum_{i=1}^{n}\left(\frac{1-t^{2i}}{1-t^2}\right)^2.$$

When $|t| < 1$, we have

$$S_{2(n)}(t) \leqslant \frac{1}{(1-t^2)^2}\left[\sum_{i=1}^{n}(1-t^{2i})\right] = \frac{1}{(1-t^2)^2}\left(n - \frac{t^2(1-t^{2n})}{1-t^2}\right).$$

When $|t| = 1$, we have

$$S_{2(n)}(t) = \sum_{i=1}^{n} i^2 = \frac{1}{6}n(n+1)(2n+1).$$

C.1 Proof of Corollary 1:

Proof: Direct application of Corollary 2 by setting $d_H = d_V = 1$ and $t = |a^*| \leqslant 1$.

D Proof of Lemma 1:

We first restate Lemma 1 in the main manuscript to correct a typo; we should have $f_\mu(n) \leqslant Cn^2$ instead of $f_\mu(n) \leqslant Cn$ for the $|a^*| = 1$ case. This does not change the conclusion of any Corollary.

Lemma 2. *1 (Restatement) For the scalar LGSSM, we have the following bounds on $f_\lambda(n)$ and $f_\mu(n)$ depending on the absoulte value of a:*

1. *If $|a^*| < 1$, then $f_\lambda(n) \leqslant Cn$ and $f_\mu(n) \leqslant C$;*
2. *If $|a^*| = 1$, then $f_\lambda(n) \leqslant Cn^3$ and $f_\mu(n) \leqslant Cn^2$;*
3. *If $|a^*| > 1$, then $f_\lambda(n) \geqslant e^{cn}$ and $f_\mu(n) \geqslant e^{cn}$.*

Here C, c are positive constants independent of n.

Proof: Observing the Lemma 3, Lemma 4 and (S13), we can see the growth rate for $f_\lambda(n)$ and $f_\mu(n)$ are separately

$$f_\lambda(n) = \begin{cases} Cn & |a^*| < 1 \\ Cn^3 & |a^*| = 1, \end{cases}$$

and

$$f_\mu(n) = \begin{cases} Cn & |a^*| < 1 \\ Cn^2 & |a^*| = 1. \end{cases}$$

Noticing in the scalar case with our setup, we have the equation

$$D(p(X^n \mid \theta^*) \| p(X^n \mid \theta)) = n\left(\log\frac{\sigma_0}{\sigma_0^*} + \frac{{\sigma_0^*}^2}{2\sigma_0^2} - \frac{1}{2}\right) + \frac{{\sigma_0^*}^2(a^*-a)^2}{2\sigma_0^2} S_{1(n-1)}(a^*), \tag{S6}$$

$$\mathbb{E}_{X^n|\theta*}\, D(p(Y_k \mid X_k, \theta^*) \| p(Y_k \mid X_k, \theta)) = \left(\log\frac{\sigma_0}{\sigma_0^*} + \frac{\sigma_0^{*2}}{2\sigma_0^2} - \frac{1}{2}\right) + \frac{{\sigma_0^*}^2(a^*-a)^2}{2\sigma_0^2}\left(\sum_{i=0}^{k-1} {a^*}^{2i}\right). \tag{S7}$$

Then when $|a^*| > 1$, it is necessary to pick $f_\lambda(n)$ and $f_\mu(n)$ with an exponential rate because of the last term in (S6) and (S7).

E Risk Bound in Multivariate Case:

With the formulation of a multivariate AR(1) stated in §4.2, the marginal distribution p_t for X_t is,

$$p_t := p(X_t) \sim \mathcal{N}_d(0, \Omega_t),$$

where

$$\begin{aligned} \Omega_t &= \Sigma_V + A\Omega_{t-1}A^T, \\ \Omega_1 &= \Sigma_V. \end{aligned}$$

Using the Ω_t to denote the precision matrix.

E.1 KL Divergence Bound

Given the two multivariate normal distributions,

$$\begin{aligned} p_1 &= \mathcal{N}_d(\mu_1, \Sigma_1) \\ p_2 &= \mathcal{N}_d(\mu_2, \Sigma_2), \end{aligned}$$

the KL divergence between them $D(p_1 \| p_2)$ can be written as

$$D(p_1 \| p_2) = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - d + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1)\right].$$

For our multivariate model, we can get the KL divergence bound by the following lemma.

Lemma 3. *S1(**KL Divergence Bound**) Denote* $t = \max_{1 \leqslant i \leqslant d_V} |\lambda_{A*}|$, *an upper bound for* $D(\pi^*_{X^n} \| \pi_{X^n})$ *is*

$$D(\pi^*_{X^n} \| \pi_{X^n}) \leqslant T_1 n + T_2 \|A - A^*\|_F^2 S_{1(n-1)}(t),$$

where T_1, T_2 *are defined in the proof.*

Proof: Using the chain rule for KL divergence, we have

$$D(p(X^n \mid \theta^*) \| p(X^n \mid \theta)) = D(p_1(X_1 \mid \theta^*) \| p_1(X_1 \mid \theta)) + \sum_{k=1}^{n-1} E_{p(X_k)} D(p(X_{k+1} \mid X_k, \theta^*) \| p(X_{k+1} \mid X_k, \theta)).$$

Pick p_1 and p_2 as

$$\begin{aligned} p_1 &= \mathcal{N}_d(A^* X_k, \Sigma_V^*), \\ p_2 &= \mathcal{N}_d(A X_k, \Sigma_V). \end{aligned}$$

We have

$$\begin{aligned} D(p_1 \| p_2) &= \frac{1}{2}\left[\log\frac{|\Sigma_V|}{|\Sigma_V{}^*|} - d_V + tr(\Sigma_V^* \Sigma_V^{-1}) + X_k^T (A^* - A)^T {\Sigma_V}^{-1}(A^* - A) X_k\right] \\ &= \frac{1}{2}\left[\log\frac{|\Sigma_V|}{|\Sigma_V{}^*|} - d_V + tr(\Sigma_V^* \Sigma_V^{-1}) + \mathrm{Tr}\left((A^* - A)^T {\Sigma_V^*}^{-1}(A^* - A) X_k X_k^T\right)\right]. \end{aligned}$$

Then we take the expectation,

$$\mathbb{E}_{p_k^*}(D(p_1 \| p_2)) \leqslant \frac{1}{2}\left[\left(\log \max_{1 \leqslant i \leqslant d_V} \frac{{\sigma_V}_i^2}{{\sigma^*_V}_i^2} + \max_{1 \leqslant i \leqslant d_V} \frac{{\sigma^*_V}_i^2}{{\sigma_V}_i^2} - 1\right) d_V + \mathrm{Tr}\left((A^* - A)^T {\Sigma_V}^{-1}(A^* - A)\Omega_k^*\right)\right].$$

Then we can sum it over by the chain rule,

$$\begin{aligned} D(\pi^*_{X^n} \| \pi_{X^n}) &\leqslant \frac{n d_V}{2}\left(\log \max_{1 \leqslant i \leqslant d} \frac{{\sigma_V}_i^2}{{\sigma^*_V}_i^2} + \max_{1 \leqslant i \leqslant d} \frac{{\sigma^*_V}_i^2}{{\sigma_V}_i^2} - 1\right) + \mathrm{Tr}\left[(A^* - A)^T {\Sigma_V}^{-1}(A^* - A)\left(\sum_{k=1}^{n-1} \Omega_k^*\right)\right] \\ &\leqslant I + II. \end{aligned}$$

Remind of the property that when M_1, M_2 are s.p.d. matrices, we have

$$\mathrm{Tr}(M_1 M_2) \leqslant \mathrm{Tr}(M_1)\,\mathrm{Tr}(M_2).$$

Applying this property to term II, we have

$$II \leqslant \mathrm{Tr}\left[(A^* - A)^T {\Sigma_V}^{-1}(A^* - A)\right] \mathrm{Tr}\left(\sum_{k=1}^{n-1} \Omega_k^*\right).$$

Then we can derive

$$\begin{aligned}
\mathrm{Tr}\left(\sum_{k=1}^{n-1}\Omega_k^*\right) &= \sum_{k=1}^{n-1}\mathrm{Tr}\left[(n-k)\left(A^*\right)^{k-1}\Sigma_V^*\left({A^*}^T\right)^{k-1}\right]\\
&= \sum_{k=1}^{n-1}(n-k)\,\mathrm{Tr}\left[\left({A^*}^T\right)^{k-1}\left(A^*\right)^{k-1}\Sigma_V^*\right]\\
&\leqslant \max_{1\leqslant i\leqslant d_V}\sigma_{V\,i}^{*\,2}S_{1(n-1)}(t)d_V,
\end{aligned}$$

where $t = \max_{1\leqslant i\leqslant d}|\lambda_{A*}|$. Then we have the bound

$$\begin{aligned}
D(\pi^*_{X^n}\|\pi_{X^n}) \leqslant \frac{nd_V}{2}&\left(\log\max_{1\leqslant i\leqslant d_V}\frac{\sigma_{V\,i}^{2}}{\sigma_{V\,i}^{*\,2}}+\max_{1\leqslant i\leqslant d_V}\frac{\sigma_{V\,i}^{*\,2}}{\sigma_{V\,i}^{2}}-1\right)\\
&+d_V\left(\max_{1\leqslant i\leqslant d_V}\sigma_{V\,i}^{*\,2}\right)\left(\max_{1\leqslant i\leqslant d_V}\sigma_{V\,i}^{-2}\right)\|A-A^*\|_F^2S_{1(n-1)}(t)\\
:&= T_1n+T_2\|A-A^*\|_F^2S_{1(n-1)}(t),
\end{aligned}$$

where

$$\begin{aligned}
T_1 :&= \frac{d_V}{2}\left(\log\max_{1\leqslant i\leqslant d_V}\frac{\sigma_{V\,i}^{2}}{\sigma_{V\,i}^{*\,2}}+\max_{1\leqslant i\leqslant d_V}\frac{\sigma_{V\,i}^{*\,2}}{\sigma_{V\,i}^{2}}-1\right),\\
T_2 :&= d_V\left(\max_{1\leqslant i\leqslant d_V}\sigma_{V\,i}^{*\,2}\right)\left(\max_{1\leqslant i\leqslant d_V}\sigma_{V\,i}^{-2}\right).
\end{aligned}$$

E.2 V Divergence Bound

Under the same multivariate AR(1) constructions, we can have the following lemma for bounding the V divergence between the two joint densities over the latent space.

Lemma 4. *S2(**V Divergence Bound**) Denote* $t = \max_{1\leqslant i\leqslant d_V}|\lambda_{A*}|$*, an upper bound for* $V(\pi^*_{X^n}\|\pi_{X^n})$ *is given by*

$$V(\pi^*_{X^n}\|\pi_{X^n}) \leqslant T_3n+T_4\|A-A^*\|_F^2S_{1(n-1)}(t)+T_5\|A-A^*\|_F^4S_{2(n)}(t),$$

where T_3, T_4, T_5 *are functions related to* (Σ_V, Σ_V^*) *which are defined in the proof.*

Proof: Still we can use the chain rule to get inequality

$$V(p(X^n\mid\theta^*)\|p(X^n\mid\theta)) \leqslant V(p_1(X_1\mid\theta^*)\|p_1(X_1\mid\theta))+\sum_{k=1}^{n-1}\mathbb{E}_{p(x_k)}V(p(X_{k+1}\mid X_k,\theta^*)\|p(X_{k+1}\mid X_k,\theta)).$$

We can try to get the V divergence for two single multivariate guassians. With assuming them separately as

$$\begin{aligned}
p_1 &= \mathcal{N}_d(\mu_1,\Sigma_1),\\
p_2 &= \mathcal{N}_d(\mu_2,\Sigma_2),
\end{aligned}$$

we can write the V divergence between them as

$$\begin{aligned}
V(p_1\|p_2) &= \int\frac{1}{4}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|}-(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)+(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)\right]^2p_1(X)dX\\
&= \int\frac{1}{4}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|}+(X-\mu_1)^T(\Sigma_2^{-1}-\Sigma_1^{-1})(X-\mu_1)+2(\mu_1-\mu_2)^T\Sigma_2^{-1}(X-\mu_1)+\right.\\
&\quad\left.(\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2)\right]^2p_1(X)dX\\
&\leqslant \log^2\frac{|\Sigma_2|}{|\Sigma_1|}+3\,\mathrm{Tr}\left[(\Sigma_2^{-1}\Sigma_1-I)^2\right]+4\Sigma_1\Sigma_2^{-2}(\mu_1-\mu_2)^2+(\mu_1-\mu_2)^{2^T}\Sigma_2^{-2}(\mu_1-\mu_2)^2.
\end{aligned}$$

In our model, we consider the joint distribution over the latent variables and then we can do the substitution,

$$\begin{aligned} p_1 &= \mathcal{N}_{d_V}(A^* X_k, \Sigma_V^*), \\ p_2 &= \mathcal{N}_{d_V}(A X_k, \Sigma_V). \end{aligned}$$

Then we get the following

$$\begin{aligned} V(p_1 \| p_2) \leqslant & d_V \left[d_V \log^2 \max_{1 \leqslant i \leqslant d_V} \frac{{\sigma_V}_i^2}{{\sigma_V^*}_i^2} + 3 \max_{1 \leqslant i \leqslant d_V} \left(\frac{{\sigma_V^*}_i^2}{{\sigma_V}_i^2} - 1 \right)^2 \right] + 4 \left(\max_{1 \leqslant i \leqslant d_V} \frac{{\sigma_V^*}_i^2}{{\sigma_V}_i^4} \right) \|(A - A^*) X_k\|_2^2 \\ & + \left(\max_{1 \leqslant i \leqslant d_V} {\sigma_V}_i^{-4} \right) \|(A - A^*) X_k\|_4^4 \\ = & I + II + III. \end{aligned}$$

We take the expectation with respect to p_k, and then we can get

$$\mathbb{E}_{p_k} II \leqslant 4 \left(\max_{1 \leqslant i \leqslant d} \frac{{\sigma_V^*}_i^2}{{\sigma_V}_i^4} \right) \left(\max_{1 \leqslant i \leqslant d} {\sigma_V}_i^{-4} \right) \|A - A^*\|_F^2 \operatorname{Tr}(\Omega_k^*).$$

Also we have

$$\begin{aligned} \|(A - A^*) X_k\|_4^4 := \sum_{i=1}^{d_V} \left(\sum_{j=1}^{d_V} \left(A_{ij} - A_{ij}^* \right) X_{k,j} \right)^4 & \leqslant \sum_{i=1}^{d_V} \left(\sum_{j=1}^{d_V} \left(A_{ij} - A_{ij}^* \right)^2 \right)^2 \left(\sum_{j=1}^{d_V} X_{k,j}^2 \right)^2 \\ & \leqslant d_V^2 \left(\sum_{i,j=1}^{d_V} \left(A_{ij} - A_{ij}^* \right)^4 \right) \left(\sum_{j=1}^{d_V} X_{k,j}^4 \right) \leqslant \|A - A^*\|_F^4 \left(\sum_{j=1}^{d_V} X_{k,j}^4 \right). \end{aligned} \tag{S8}$$

The above last inequality comes from the norm inequality

$$\| \cdot \|_r \leqslant d^{\frac{1}{r} - \frac{1}{p}} \| \cdot \|_p, \tag{S9}$$

where the norm $\| \cdot \|_r$ and $\| \cdot \|_p$ are all defined on vector space $\mathbb{R}^d$.
Taking expectation over (S8), we get

$$\mathbb{E}_{p_k} III \leqslant \left(\max_{1 \leqslant i \leqslant d_V} {\sigma_V}_i^{-4} \right) \|A - A^*\|_F^4 \left(\sum_{j=1}^{d_V} \mathbb{E}_{p_k} X_{k,j}^4 \right). \tag{S10}$$

Analysing the last term we have

$$\sum_{j=1}^{d_V} \mathbb{E}_{p_k} X_{k,j}^4 = 3 \sum_{j=1}^{d_V} m_{k,(j,j)}^2 \leqslant 3 d_V^{-1} \left(\sum_{j=1}^{d_V} m_{k,(j,j)} \right)^2 \tag{S11}$$

$$= 3 d_V^{-1} \operatorname{Tr} \left(\Omega_k^* \right)^2 \leqslant 3 d_V \left(\max_{1 \leqslant i \leqslant d_V} {\sigma_V^*}_i^4 \right) \left(\sum_{i=0}^{k-1} t^{2i} \right)^2. \tag{S12}$$

The (S11) and (S12) uses the trick in (S9).
Combining (S10) and (S12), we get

$$\mathbb{E}_{p_k} III \leqslant 3 d_V \left(\max_{1 \leqslant i \leqslant d_V} {\sigma_V^*}_i^4 \right) \left(\max_{1 \leqslant i \leqslant d_V} {\sigma_V}_i^{-4} \right) \|A - A^*\|_F^4 \left(\sum_{i=0}^{k-1} t^{2i} \right)^2.$$

Combining all above, we get

$$
\begin{aligned}
V(\pi^*_{X^n}\|\pi_{X^n}) \leqslant{} & d_V\left[d_V\log^2\max_{1\leqslant i\leqslant d_V}\frac{{\sigma_V}_i^2}{{\sigma^*_V}_i^2}+3\max_{1\leqslant i\leqslant d_V}\left(\frac{{\sigma^*_V}_i^2}{{\sigma_V}_i^2}-1\right)^2\right]\\
&+4d_V\left(\max_{1\leqslant i\leqslant d}\frac{{\sigma^*_V}_i^2}{{\sigma_V}_i^4}\right)\left(\max_{1\leqslant i\leqslant d}{\sigma_V}_i^{-4}\right)\left(\max_{1\leqslant i\leqslant d_V}{\sigma^*_V}_i^2\right)\|A-A^*\|_F^2S_{1(n-1)}(t)\\
&+3d_V\left(\max_{1\leqslant i\leqslant d_V}{\sigma^*_V}_i^4\right)\left(\max_{1\leqslant i\leqslant d_V}{\sigma_V}_i^{-4}\right)\|A-A^*\|_F^4S_{2(n)}(t)\\
:={} & T_3n+T_2\|A-A^*\|_F^2S_{1(n-1)}(t)+T_4\|A-A^*\|_F^5S_{2(n)}(t),
\end{aligned}
$$

where

$$
\begin{aligned}
T_3 &:= d_V\left[d_V\log^2\max_{1\leqslant i\leqslant d_V}\frac{{\sigma_V}_i^2}{{\sigma^*_V}_i^2}+3\max_{1\leqslant i\leqslant d_V}\left(\frac{{\sigma^*_V}_i^2}{{\sigma_V}_i^2}-1\right)^2\right],\\
T_4 &:= 4d_V\left(\max_{1\leqslant i\leqslant d}\frac{{\sigma^*_V}_i^2}{{\sigma_V}_i^4}\right)\left(\max_{1\leqslant i\leqslant d}{\sigma_V}_i^{-4}\right)\left(\max_{1\leqslant i\leqslant d_V}{\sigma^*_V}_i^2\right),\\
T_5 &:= 3d_V\left(\max_{1\leqslant i\leqslant d_V}{\sigma^*_V}_i^4\right)\left(\max_{1\leqslant i\leqslant d_V}{\sigma_V}_i^{-4}\right).
\end{aligned}
$$

E.3 Bound w.r.t. μ

Combining with the result from Lemma 3 and Lemma 4, if we assume that

$$
\begin{aligned}
p_1 &= \mathcal{N}_{d_H}(B^*X_k,{\Sigma_H}^*),\\
p_2 &= \mathcal{N}_{d_H}(BX_k,\Sigma_H),
\end{aligned}
$$

then we have

$$
\mathbb{E}_{p_k^*}(D(p_1\|p_2))\leqslant T_1^{'}+T_2^{'}\|B-B^*\|_F^2\left(\sum_{i=0}^{k-1}t^{2i}\right),
$$

and

$$
\mathbb{E}_{p_k^*}(V(p_1\|p_2))\leqslant T_1^{'}+T_2^{'}\|B-B^*\|_F^2\left(\sum_{i=0}^{k-1}t^{2i}\right)+T_3^{'}\|B-B^*\|_F^4\left(\sum_{i=0}^{k-1}t^{2i}\right)^2, \tag{S13}
$$

where

$$
\begin{aligned}
T_1^{'} &:= \frac{d_H}{2}\left(\log\max_{1\leqslant i\leqslant d_H}\frac{{\sigma_H}_i^2}{{\sigma^*_H}_i^2}+\max_{1\leqslant i\leqslant d_H}\frac{{\sigma^*_H}_i^2}{{\sigma_H}_i^2}-1\right),\\
T_2^{'} &:= d_V\left(\max_{1\leqslant i\leqslant d_H}{\sigma_H}_i^{-2}\right)\left(\max_{1\leqslant i\leqslant d_V}{\sigma_V}_i^2\right),\\
T_3^{'} &:= d_H\left[d_H\log^2\max_{1\leqslant i\leqslant d_H}\frac{{\sigma_H}_i^2}{{\sigma^*_H}_i^2}+3\max_{1\leqslant i\leqslant d_H}\left(\frac{{\sigma^*_H}_i^2}{{\sigma_H}_i^2}-1\right)^2\right],\\
T_4^{'} &:= 4d_V\left(\max_{1\leqslant i\leqslant d_H}\frac{{\sigma^*_H}_i^2}{{\sigma_H}_i^4}\right)\left(\max_{1\leqslant i\leqslant d_H}{\sigma_H}_i^{-4}\right)\left(\max_{1\leqslant i\leqslant d_V}{\sigma^*_V}_i^2\right),\\
T_5^{'} &:= 3d_V\left(\max_{1\leqslant i\leqslant d_H}{\sigma^*_H}_i^4\right)\left(\max_{1\leqslant i\leqslant d_V}{\sigma_V}_i^{-4}\right).
\end{aligned}
$$

Here we still use $t=\max_{1\leqslant i\leqslant d_V}|\lambda_{A^*}|$.

E.4 Proof of Corollary 2:

Proof: Without loss of generality, we can define

$$\mathcal{A}(c_0, C_0) := \left\{ c_0 \leqslant {\sigma_V}_i^2, {\sigma_H}_j^2 \leqslant C_0, \quad \forall\, i \in [d_V],\, j \in [d_H] \right\}$$

for some $C_0 > c_0 > 0$. Combining all the bounds, we pick

$$f_\lambda(n) = \begin{cases} n & t < 1 \\ n^3 & t = 1, \end{cases}$$

and

$$f_\mu(n) = \begin{cases} n & t < 1 \\ n^2 & t = 1. \end{cases}$$

For some $C_1, C_2 > 0$, we have (S14) holds on $\mathcal{A}(c_0, C_0)$ and find out $d_\lambda = d_V^2, d_\mu = d_V d_H$. Using the result from Lemma 5,we can show there exists $\beta, C, D \geqslant 0$, such that with $\mathbb{P}_{\theta*}^{(n)}$ probability at least $1 - D^{-2}(\log n)^{-\beta}$, it holds that

$$\int D_\alpha^{(n)}\left(\theta, \theta^*\right) \widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta) \leqslant CD\left(\frac{(\log n)^\beta}{n} \vee \frac{(d_V^2 \vee d_V d_H) \log n}{n} \right).$$

Then the Corollary 2 gets proved.

F Proof of Corollary 3

Proof: The assumptions given in the corollary 3 is as follow,

$$\begin{aligned} &\max \left\{ D\left(\pi_{X^n}^* \,\|\, \pi_{X^n}\right), V\left(\pi_{X^n}^* \,\|\, \pi_{X^n}\right) \right\} \leqslant C_1 f_\lambda(n) \|\lambda - \lambda^*\|^2, \\ &\max \max_{1 \leqslant k \leqslant n} \left\{ \mathbb{E}_{X^n|\theta*}\, D_k(\mu^*, \mu), \mathbb{E}_{X^n|\theta*}\, V_k(\mu^*, \mu) \right\} \leqslant C_2 f_\mu(n) \|\mu - \mu^*\|^2, \end{aligned} \tag{S14}$$

where $C_1, C_2 > 0$, and $f_\lambda(n)$ and $f_\mu(n)$ are two non-decreasing functions w.r.t. n. When these conditions are satisfied, we have the two neigbhourhoods constructed as

$$\begin{aligned} &\mathcal{B}_n\left(\pi_{X^n}^*, \varepsilon_\lambda\right) = \left\{ D\left(\pi_{X^n}^* \,\|\, \pi_{X^n}\right) \leqslant f_\lambda(n)\, \varepsilon_\lambda^2, V\left(\pi_{X^n}^* \,\|\, \pi_{X^n}\right) \leqslant f_\lambda(n)\, \varepsilon_\lambda^2 \right\}, \\ &\mathcal{B}_n\left(\mu^*, \varepsilon_\mu\right) = \left\{ \max_{1 \leqslant i \leqslant n} \mathbb{E}_{X^n|\theta*}\, D_i(\mu^*, \mu) \leqslant f_\mu(n)\, \varepsilon_\mu^2, \max_{1 \leqslant i \leqslant n} \mathbb{E}_{X^n|\theta*}\, V_i(\mu^*, \mu) \leqslant f_\mu(n)\, \varepsilon_\mu^2 \right\}, \end{aligned}$$

The constants C_1, C_2 are included into the blow-up factors f_λ, f_μ. Then we pick

$$\frac{f_\lambda(n)\, \varepsilon_{n\lambda}^2}{n} = f_\mu(n)\, \varepsilon_{n\mu}^2 = \frac{(\log n)^\beta}{n} \tag{S15}$$

for some $\beta > 0$. When (S14) are satisfied, the following inclusion relation holds constantly with any n,

$$\begin{aligned} &\left\{ \bigcap_{i=1}^{d_\lambda} \left\{ \|\lambda_i - \lambda_i^*\|^2 \leqslant \frac{\varepsilon_{n\lambda}^2}{d_\lambda} \right\} \right\} \subseteq \left\{ \|\lambda - \lambda^*\|^2 \leqslant \varepsilon_{n\lambda}^2 \right\} \subseteq \mathcal{B}_n\left(\pi_{X^n}^*, \varepsilon_{n\lambda}\right), \\ &\left\{ \bigcap_{i=1}^{d_\mu} \left\{ \|\mu_i - \mu_i^*\|^2 \leqslant \frac{\varepsilon_{n\mu}^2}{d_\mu} \right\} \right\} \subseteq \left\{ \|\mu - \mu^*\|^2 \leqslant \varepsilon_{n\mu}^2 \right\} \subseteq \mathcal{B}_n\left(\mu^*, \varepsilon_{n\mu}\right). \end{aligned}$$

Becasue of the neighborhoods construction and the condition that P_λ and P_μ are reimann integrable over the euclidean parameter space, we have the probability inequality as

$$P_\lambda\left(\mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_{n\lambda}\right)\right) \geqslant P_\lambda\left(\max_{1\leqslant i\leqslant d_\lambda} \|\lambda_i - \lambda_i^*\|^2 \leqslant \frac{\varepsilon^2_{n\lambda}}{d_\lambda}\right)^{d_\lambda} \geqslant C^{'}\left(\frac{(\log n)^\beta}{d_\lambda f_\lambda(n)}\right)^{\frac{d_\lambda}{2}}, \tag{S16}$$

$$P_\mu\left(\mathcal{B}_n\left(\mu^*, \varepsilon_\mu\right)\right) \geqslant P_\mu\left(\max_{1\leqslant i\leqslant d_\mu} \|\mu_i - \mu_i^*\|^2 \leqslant \frac{\varepsilon^2_{n\mu}}{d_\mu}\right)^{d_\mu} \geqslant C^{'}\left(\frac{(\log n)^\beta}{d_\mu f_\mu(n)}\right)^{\frac{d_\mu}{2}}, \tag{S17}$$

where $C^{'} > 0$ is a constant.
Then we analysing the upper bound in Theorem 1. It is composed as two part

$$\frac{D\alpha\varepsilon^2_{\lambda n} f_\lambda(n)}{(1-\alpha)n} + \left\{-\frac{1}{n(1-\alpha)}\log P_\lambda\left[B_n\left(\pi^*_{X^n}, \varepsilon_{\lambda n}\right)\right]\right\},$$

and

$$\frac{D\alpha\varepsilon^2_{\mu n} f_\mu(n)}{1-\alpha} + \left\{-\frac{1}{n(1-\alpha)}\log P_\mu\left[B_n\left(\mu^*, \varepsilon_{\mu n}\right)\right]\right\}.$$

Plugging all the picked $f_\lambda, f_\mu, \varepsilon^2_{n\lambda}, \varepsilon^2_{n\mu}$ and (S16), (S17) back into the Theorem 1, we can get the following bound. Since $(f_\lambda(n)\,\varepsilon^2_\lambda + f_\mu(n)\,n\varepsilon^2_\mu) = 2(\log n)^\beta$ and (S15), with appropriately picking the constant C, we have the bound with $\mathbb{P}^{(n)}_{\theta*}$ probability at least $1 - D^{-2}(\log n)^{-\beta}$ that

$$\int D^{(n)}_\alpha\left(\theta, \theta^*\right)\widehat{q}_{\boldsymbol{\theta},\alpha}(d\theta) \leqslant CD\left(\frac{(\log n)^\beta}{n} \vee \frac{d_\lambda \log f_\lambda(n)}{n} \vee \frac{d_\mu \log(n f_\mu(n))}{n}\right). \tag{S18}$$

The proof for corollary 3 is then completed.

F.1 Extension of Corollary 3:

Lemma 5. *S3 When the condition (S14) holds on a set $\mathcal{A}$ and the multivariate prior densities p_λ and p_μ are Riemann integrable, there exist $\beta, D, C > 0$ s.t. (S18) holds.*

Proof: According to the condition, we have

$$\begin{aligned}\mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_\lambda\right) \cap \mathcal{A} &\subseteq \mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_\lambda\right)\\ \mathcal{B}_n\left(\mu^*, \varepsilon_\mu\right) \cap \mathcal{A} &\subseteq \mathcal{B}_n\left(\mu^*, \varepsilon_\mu\right).\end{aligned}$$

Based on this, we similarly have

$$\begin{aligned}P_\lambda\left(\mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_{n\lambda}\right)\right) &\geqslant P_\lambda\left(\mathcal{B}_n\left(\pi^*_{X^n}, \varepsilon_{n\lambda} \cap \mathcal{A}\right)\right) P_\lambda\left(\max_{1\leqslant i\leqslant d_\lambda} \|\lambda_i - \lambda_i^*\|^2 \leqslant \frac{\varepsilon^2_{n\lambda}}{d_\lambda}\right)^{d_\lambda} \geqslant C^{'}\left(\frac{(\log n)^\beta}{d_\lambda f_\lambda(n)}\right)^{\frac{d_\lambda}{2}},\\ P_\mu\left(\mathcal{B}_n\left(\mu^*, \varepsilon_\mu\right)\right) &\geqslant P_\mu\left(\mathcal{B}_n\left(\mu^*, \varepsilon_\mu \cap \mathcal{A}\right)\right) \geqslant P_\mu\left(\max_{1\leqslant i\leqslant d_\mu} \|\mu_i - \mu_i^*\|^2 \leqslant \frac{\varepsilon^2_{n\mu}}{d_\mu}\right)^{d_\mu} \geqslant C^{'}\left(\frac{(\log n)^\beta}{d_\mu f_\mu(n)}\right)^{\frac{d_\mu}{2}},\end{aligned} \tag{S19}$$

for some $C^{'} > 0$. Then from (S19), we have (S18) hold.

G Simulation study

In this section, we conduct a small-scale simulation study to study the behavior of the Bayes risk for different values of a in a scalar LGSSM. The top panel of Figure 1 plots the Bayes risk versus the number of CAVI iterates for different values of the fractional parameter α. For each choice of α, we consider four different values of the true transmission parameter $a \in \{0.8, 0.9, 0.95, 0.97\}$. It is evident that the convergence is slower for larger a. We also provide the RMSE in the bottom panel.

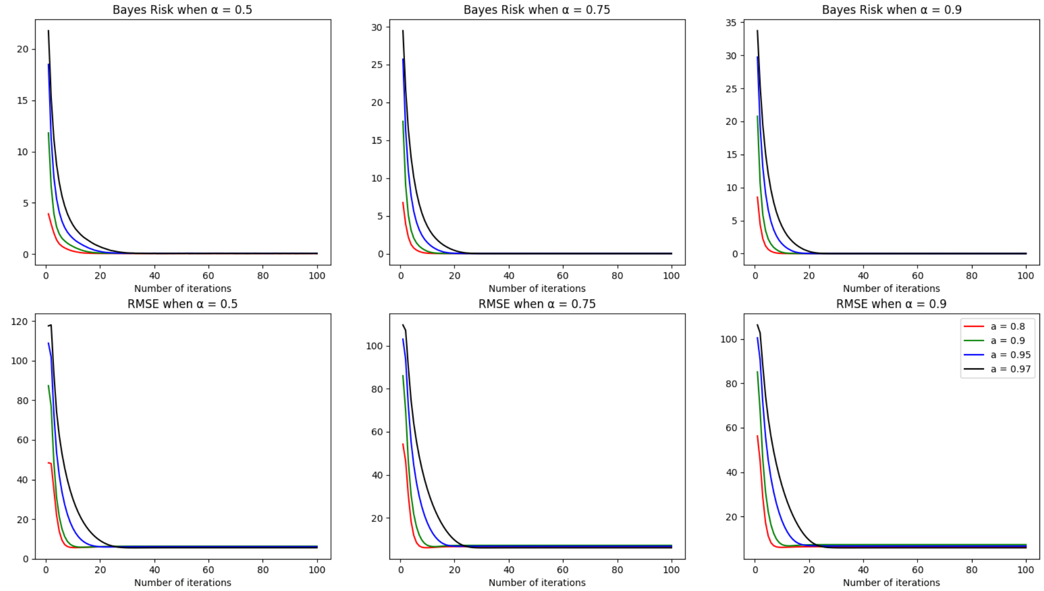

Figure 1: Plot of Bayes risk versus number of CAVI iterates.

In Figure 2, we repeat the same analysis for values of the true transmission parameter close to (and possibly exceeding) 1. The convergence slows down substantially, which is expected given the increased difficulty of the estimation problem as explained in Lemma 1.

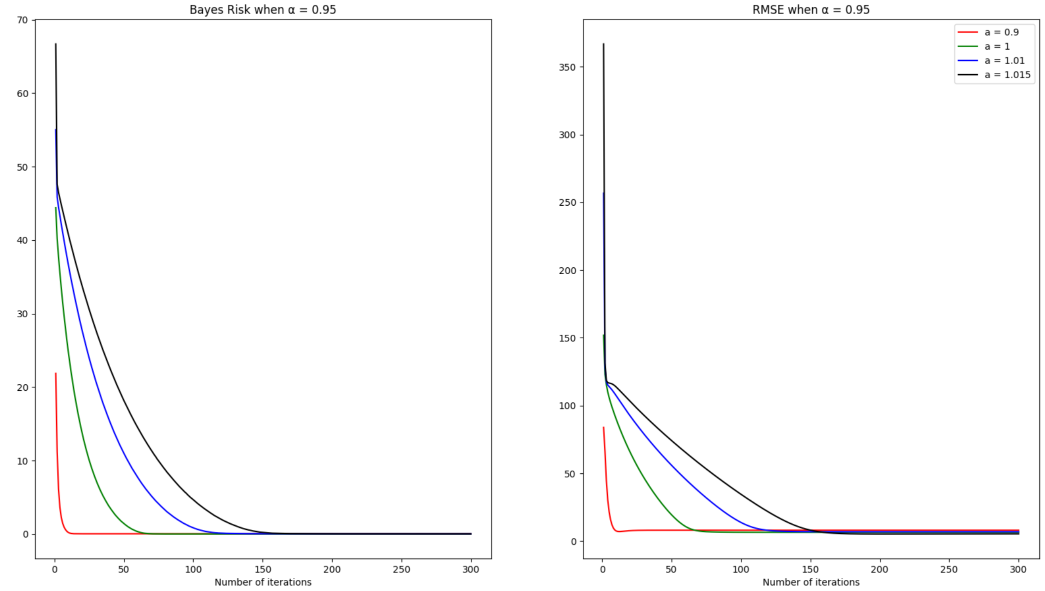

Figure 2: Plot of Bayes risk versus number of CAVI iterates when the true transmission parameter is close to 1.