Variational inference via decomposable transports: algorithms for Bayesian filtering and smoothing

Alessio Spantini
MIT
Cambridge, MA 02139
spantini@mit.edu

Daniele Bigoni
MIT
Cambridge, MA 02139
dabi@mit.edu

Youssef Marzouk
MIT
Cambridge, MA 02139
ymarz@mit.edu

Abstract

We describe a variational inference method that approximates an intractable target measure as the pushforward of a tractable distribution (e.g., a Gaussian) through a transport map. We then show how such transport maps can be *decomposed*—i.e., factorized into the composition of finitely many low-dimensional maps. We use the notion of decomposable transports to derive new deterministic *online* algorithms for Bayesian filtering and smoothing in nonlinear/non-Gaussian state-space models with static parameters, and illustrate the theory on a stochastic volatility model.

1 Measure transport and variational inference

Let $\boldsymbol{Z}$ be a random variable on $\mathbb{R}^n$ endowed with an intractable continuous density π that we wish to simulate. We assume that π is available only up to a normalizing constant. For instance, π may represent the posterior density of a Bayesian inference problem, where the goal is to approximate integrals of the form $\int g(\boldsymbol{z}) \cdot \pi(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$ for some measurable $g : \mathbb{R}^n \to \mathbb{R}$. One possible approach to the problem of sampling is to seek a deterministic (transport) map $T : \mathbb{R}^n \to \mathbb{R}^n$ that couples a tractable reference random variable $\boldsymbol{X}$ of density η (e.g., a standard normal) with $\boldsymbol{Z}$ [15]. The coupling ensures that $T(\boldsymbol{X}) = \boldsymbol{Z}$ in distribution [30], or, equivalently, that T pushes forward η to π, i.e., $T_\sharp \eta = \pi$, where $T_\sharp \eta$ denotes the pushforward density of η by T. (For any invertible map T, we have $T_\sharp \eta(\boldsymbol{z}) = \eta(T^{-1}(\boldsymbol{z})) \cdot |\det \nabla T^{-1}(\boldsymbol{z})|$, where $\nabla T(\boldsymbol{z})^{-1} \in \mathbb{R}^{n \times n}$ denotes the gradient of the inverse map at $\boldsymbol{z}$.) Thus, if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ is an independent and identically distributed (iid) sample from η, then $T(\boldsymbol{X}_1), \ldots, T(\boldsymbol{X}_n)$ is an iid sample from π. In other words, T enables the generation of cheap, independent, and unweighted samples from π by pushing forward a collection of reference samples through the map. Clearly, a transport map between a tractable density η and the target π turns π into a tractable distribution and solves, at least formally, the problem of sampling.

A transport map between random variables on $\mathbb{R}^n$ exists under very weak conditions. For instance, in the example above it suffices that the law of $\boldsymbol{X}$ vanishes on subsets of (Hausdorff) dimension $n-1$ [16]. As shown in [18], the transport map can be computed via deterministic optimization by minimizing the Kullback–Leibler (KL) divergence $\mathcal{D}_{\mathrm{KL}}(\, T_\sharp \eta \,||\, \pi \,)$ over a suitable function space for the map, i.e., for $T \in \mathcal{T}$. At optimality, we have $T_\sharp \eta = \pi$. In practice, we need to represent the transport. The approach adopted in [18, 20] seeks a parametric transport map within a finite dimensional approximation space, $\mathcal{T}_h \subset \mathcal{T}$. The resulting variational problem reads as:

$$\min_{T \in \mathcal{T}_h} \mathcal{D}_{\mathrm{KL}}(\, T_\sharp \eta \,||\, \pi \,). \tag{1.1}$$

We can interpret (1.1) as seeking a density q that minimizes $\mathcal{D}_{\mathrm{KL}}(\, q \,||\, \pi \,)$ over a finite dimensional class of tractable distributions, $\mathcal{Q}$, which consists of distributions $q = T_\sharp \eta$ that can be written as the pushforward of η by a map in $\mathcal{T}_h$. Thus, (1.1) defines a particular *variational inference* method [6, 31, 2]: one that uses *measure transport* to characterize the class of approximating distributions $\mathcal{Q}$ (see [29, 22] for related methods). The richer the function space for the map, the richer $\mathcal{Q}$. In

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

particular, a general parameterization of the map can capture arbitrary probabilistic interactions [18], well beyond the usual mean-field approximation [19, 21, 27].

A key feature of this approach is that it produces a transport map T and not just an approximation, $T_\sharp\eta$, to the target density. This idea becomes very useful when T is only approximate. In this case, if the bias of approximating π with $T_\sharp\eta$ is unacceptable, one can simply evaluate (possibly up to a normalizing constant) the *pullback* density $T^\sharp\pi$, defined as $T^\sharp\pi(\boldsymbol{x}) = \pi(T(\boldsymbol{x})) \cdot |\det \nabla T(\boldsymbol{x})|$, and rewrite the integral $\int g(\boldsymbol{z}) \cdot \pi(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}$ as $\int g(T(\boldsymbol{x})) \cdot T^\sharp\pi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$. One possibility is then to use a stochastic sampling technique, like MCMC [23], to probe $T^\sharp\pi$, which, by virtue of (1.1), will be closer (in KL divergence) to the reference density η. In particular, if η is Gaussian, then we could interpret pulling back π by T as a "Gaussianization" of the target [13], which can remove the correlations that make sampling a challenging task. Thus, we can regard an approximate T as a preconditioner for existing sampling schemes [20, 17, 32].

There are infinitely many transports that push forward one density to another [30]. An important transport for our analysis is the Knothe-Rosenblatt (KR) rearrangement in $\mathbb{R}^n$ [24, 10]. For a pair of continuous densities, η and π, the KR rearrangement is the unique monotone increasing (lower) triangular transport that pushes forward η to π [3]. A lower triangular transport $T : \mathbb{R}^n \to \mathbb{R}^n$ is a multivariate map whose kth component depends only on the first k input variables, i.e.,

$$T(\boldsymbol{x}) = \begin{bmatrix} T^1(x_1) \\ T^2(x_1, x_2) \\ \vdots \\ T^n(x_1, x_2, \ldots x_n) \end{bmatrix} \qquad \forall \boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n, \tag{1.2}$$

where T^k denotes the kth output of the map. The KR rearrangement is *general*—it can couple arbitrary continuous densities—and enjoys many attractive computational features. As shown in [18, 15], it can be characterized as the unique minimizer of $\mathcal{D}_{\mathrm{KL}}(\, T_\sharp\eta \,||\, \pi\,)$ over the cone $\mathcal{T}_\triangle$ of triangular maps that are monotone increasing with respect to the lexicographic order on $\mathbb{R}^n$. In this case, (1.1) is equivalent to:

$$\begin{aligned} &\min \quad -\mathbb{E}_\eta[\log \bar{\pi}(T(\boldsymbol{X})) + \sum_k \log \partial_{x_k} T^k(\boldsymbol{X}) - \log \eta(\boldsymbol{X})] \\ &\text{s.t.} \quad T \in \mathcal{T}_{\triangle,h} \subset \mathcal{T}_\triangle;\ \dim(\mathcal{T}_{\triangle,\mathrm{h}}) < \infty \end{aligned} \tag{1.3}$$

where $\bar{\pi}$ denotes the unnormalized target density. In particular, by using monotone parameterizations for T, we can regard (1.3) as an unconstrained stochastic program [15, 9, 12, 28].

2 Decomposable transports

The key observation of this work is that a transport map is not just *any* multivariate function on $\mathbb{R}^n$. There exist transports which inherit low-dimensional parameterizations from the Markov structure [14, 11] of the underlying target density. By considering recursive graph decompositions of a (non-complete) Markov network for π, it is possible to prove the existence of transports T that factorize *exactly* as the composition of k low-dimensional maps, $T = T_1 \circ \cdots \circ T_k$, for some finite k, where each map T_j differs from the identity function only along few components and is triangular up to a permutation of the input and output space. We call such transports *decomposable*. Clearly, a decomposable transport is easier to parameterize than a regular one. Moreover, the decomposition $T = T_1 \circ \cdots \circ T_k$ suggests that the computation of T may be broken into multiple simpler steps, each associated with the computation of a low-dimensional map T_j that accounts only for *local* features of π. Instead of detailing the general theory of decomposable transports, due to the length constraints of this manuscript we will explore this theory in the context of sequential Bayesian inference for state-space models (see Section 3). Our analysis, in this setting, will suggest new and powerful variational algorithms for the Bayesian filtering and smoothing problems.

3 Online algorithms for Bayesian filtering and smoothing

We consider the problem of sequential Bayesian inference for a discrete time, continuous, nonlinear, non-Gaussian state-space model [4] in a very general formulation that includes hyperparameters (i.e., static parameters) of the model. See Figure 1 for the corresponding Markov structure, where $(\boldsymbol{Z}_k)_{k\geq 0}$ denotes the unobserved latent Markov process (each $\boldsymbol{Z}_k$ is a random variable on $\mathbb{R}^n$), $(\boldsymbol{Y}_k)_{k\geq 0}$ denotes the observed process, and where $\boldsymbol{\Theta}$ represents the hyperparameters of the model, which are treated as a random variable on $\mathbb{R}^p$. The state-space model is then fully specified in terms of the conditional densities $\pi_{\boldsymbol{Y}_k|\boldsymbol{Z}_k,\boldsymbol{\Theta}}$, $\pi_{\boldsymbol{Z}_{k+1}|\boldsymbol{Z}_k,\boldsymbol{\Theta}}$, $\pi_{\boldsymbol{Z}_0|\boldsymbol{\Theta}}$, and the marginal density $\pi_{\boldsymbol{\Theta}}$, together with the observed data $(\boldsymbol{y}_k)_{k\geq 0}$. We assume these are all given.

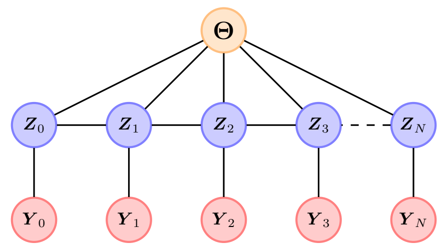

Figure 1: Markov structure of $\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:N}|\boldsymbol{Y}_{0:N}}$ for a fixed realization of the observed process.

We wish to jointly infer the hidden states and the hyperparameters of the model as observations of the process $(\boldsymbol{Y}_k)_{k\geq 0}$ become available over time. That is, the goal of inference is to characterize—*sequentially* in time and via a *recursive* algorithm—the posterior distribution,

$$\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_0,\ldots,\boldsymbol{Z}_k|\boldsymbol{Y}_0,\ldots,\boldsymbol{Y}_k} \tag{3.1}$$

for all $k \geq 0$, from which any filtering distributions $\pi_{\boldsymbol{Z}_k|\boldsymbol{Y}_{0:k}}$ or smoothing distributions $\pi_{\boldsymbol{Z}_j|\boldsymbol{Y}_{0:k}}$ with $j < k$, along with the parameter marginals $\pi_{\boldsymbol{\Theta}|\boldsymbol{Y}_{0:k}}$, are readily available [4, 26].

The key result of this section is a new *deterministic* and recursive algorithm for online inference with transport maps, which, in a *single forward pass*, computes a sequence of triangular maps of fixed dimension $(2n + p)$ that, properly composed, are capable of sampling (3.1) for all $k \geq 0$. Unlike most smoothing algorithms, the present algorithm does not resort to any *backward pass* that touches the state-space model. This is essentially the content of the forthcoming theorem (see Appendix B for a proof). In what follows, let $(\eta_{\boldsymbol{X}_k})_{k\geq 0}$ be a sequence of independent reference densities on $\mathbb{R}^n$ (e.g., standard normals) and let $\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}}$ be a reference density on $\mathbb{R}^p$. Moreover, let (η^k) and $(\widetilde{\pi}^k)$ be sequences of densities on $\mathbb{R}^{2n+p}$ defined as follows in terms of the state-space model: $\eta^k := \eta_{\boldsymbol{X}_{\boldsymbol{\Theta}},\boldsymbol{X}_k,\boldsymbol{X}_{k+1}}$ for $k \geq 0$, $\widetilde{\pi}^0 := \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_0,\boldsymbol{Z}_1|\boldsymbol{Y}_0,\boldsymbol{Y}_1}$, and $\widetilde{\pi}^k := \pi_{\boldsymbol{Z}_{k+1},\boldsymbol{Y}_{k+1}|\boldsymbol{\Theta},\boldsymbol{Z}_k}$ for $k > 0$.

Theorem 3.1 *Consider a sequence $(\mathfrak{M}_k)$ of (block) triangular maps on $\mathbb{R}^{2n+p}$ with sparsity pattern*

$$\mathfrak{M}_k(\boldsymbol{x}_\theta, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}) = \begin{bmatrix} \mathfrak{M}_k^{\boldsymbol{\Theta}}(\boldsymbol{x}_\theta) \\ \mathfrak{M}_k^0(\boldsymbol{x}_\theta, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}) \\ \mathfrak{M}_k^1(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+1}) \end{bmatrix}, \tag{3.2}$$

and defined, recursively, as follows: $\mathfrak{M}_0$ pushes forward η^0 to $\pi^0 := \widetilde{\pi}^0$; For $k \geq 1$, $\mathfrak{M}_k$ pushes forward η^k to $\pi^k(\boldsymbol{x}_\theta, \boldsymbol{x}_k, \boldsymbol{x}_{k+1}) := \eta_{\boldsymbol{X}_{\boldsymbol{\Theta}},\boldsymbol{X}_k}(\boldsymbol{x}_\theta, \boldsymbol{x}_k) \cdot \widetilde{\pi}^k(\mathfrak{T}_{k-1}^{\boldsymbol{\Theta}}(\boldsymbol{x}_\theta), \mathfrak{M}_{k-1}^1(\boldsymbol{x}_\theta, \boldsymbol{x}_k), \boldsymbol{x}_{k+1})/\mathfrak{c}_k$, where $\mathfrak{c}_k$ is a normalizing constant, and where $\mathfrak{T}_j^{\boldsymbol{\Theta}} := \mathfrak{M}_0^{\boldsymbol{\Theta}} \circ \cdots \circ \mathfrak{M}_j^{\boldsymbol{\Theta}}$ for all $j \geq 0$. Then, for all $k \geq 0$, the composition of transports $\mathfrak{T}_k := T_0 \circ \cdots \circ T_k$, where each T_j is defined (blockwise) as:

$$T_j(\boldsymbol{x}_\theta, \boldsymbol{x}_0, \ldots, \boldsymbol{x}_{k+1}) = \begin{bmatrix} \mathfrak{M}_j^{\boldsymbol{\Theta}}(\boldsymbol{x}_\theta) \\ \boldsymbol{x}_0 \\ \vdots \\ \boldsymbol{x}_{j-1} \\ \mathfrak{M}_j^0(\boldsymbol{x}_\theta, \boldsymbol{x}_j, \boldsymbol{x}_{j+1}) \\ \mathfrak{M}_j^1(\boldsymbol{x}_\theta, \boldsymbol{x}_{j+1}) \\ \boldsymbol{x}_{j+2} \\ \vdots \\ \boldsymbol{x}_{k+1} \end{bmatrix}, \tag{3.3}$$

pushes forward $\eta^{0:k} := \eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \prod_{j=0}^{k+1} \eta_{\boldsymbol{X}_j}$ to the desired target density $\pi_{\boldsymbol{\Theta},\,\boldsymbol{Z}_{0:k+1}|\boldsymbol{Y}_{0:k+1}}$.

Theorem 3.1 suggests a deterministic *online* algorithm for the joint parameter and state estimation problem: compute[1] the sequence of maps $(\mathfrak{M}_j)$, each of dimension $2n + p$; embed them into higher-dimensional identity maps to form (T_j); then evaluate $\mathfrak{T}_k := T_0 \circ \cdots \circ T_k$ to sample directly from $\pi_{\boldsymbol{\Theta}, \boldsymbol{Z}_{0:k+1} | \boldsymbol{Y}_{0:k+1}}$ and obtain information about any smoothing or filtering distribution of interest. The theorem shows that each $\pi_{\boldsymbol{\Theta}, \boldsymbol{Z}_{0:k+1} | \boldsymbol{Y}_{0:k+1}}$ can be represented via a decomposable transport $\mathfrak{T}_k$; successive transports in the sequence $(\mathfrak{T}_k)_{k \geq 0}$ are *nested* and thus ideal for online inference. The variational and online character of the proposed algorithm distinguishes it from existing state-of-the-art approaches to nonlinear and non-Gaussian smoothing and joint parameter inference [1, 7].

4 Numerical example: stochastic volatility model with hyperparameters

Following [8, 25], we model the scalar log-volatility $(\boldsymbol{Z}_t)$ of the return of a financial asset at time $t = 1, \ldots, N$ using an autoregressive process of order one, which is fully specified by $\boldsymbol{Z}_{t+1} = \mu + \phi\,(\boldsymbol{Z}_t - \mu) + \eta_t$, for all $t \geq 0$, where $\eta_t \sim \mathcal{N}(0, 1)$ is independent of $\boldsymbol{Z}_t$, $\boldsymbol{Z}_0 | \mu, \phi \sim \mathcal{N}(\mu, \frac{1}{1-\phi^2})$, and where ϕ and μ represent scalar hyperparameters of the model. In particular, $\mu \sim \mathcal{N}(0, 1)$ and $\phi = 2 \exp(\phi^\star)/(1 + \exp(\phi^\star)) - 1$ with $\phi^\star \sim \mathcal{N}(3, 1)$. We define $\boldsymbol{\Theta} := (\mu, \phi)$. The process $(\boldsymbol{Z}_t)$ and parameters $\boldsymbol{\Theta}$ are unobserved and must be estimated from an observed process $(\boldsymbol{Y}_t)$, which represents the mean return of holding the asset at time t, $\boldsymbol{Y}_t = \varepsilon_t \cdot \exp(\frac{1}{2} \boldsymbol{Z}_t)$, where ε_t is a standard normal random variable independent of $\boldsymbol{Z}_t$. As a dataset, we use the $N = 100$ daily differences of the pound/dollar exchange rate starting on 1 October 1981 [25, 5]. Our goal is to sequentially characterize $\pi_{\boldsymbol{\Theta}, \boldsymbol{Z}_{0:k} | \boldsymbol{Y}_{0:k}}$, for all $k = 0, \ldots, N$, as observations of $(\boldsymbol{Y}_t)$ become available. The Markov structure of $\pi_{\boldsymbol{\Theta}, \boldsymbol{Z}_{0:N} | \boldsymbol{Y}_{0:N}}$ matches Figure 1. We solve the problem using the algorithm introduced in Section 3: we compute a sequence, $(\mathfrak{M}_j)_{j=0}^{N-1}$, of four-dimensional transport maps ($n = 1$ and $p = 2$) according to their definition in Theorem 3.1 and using the variational form (1.3). All reference densities are standard Gaussians. Then, for any $k < N$, if we want to sample from $\pi_{\boldsymbol{\Theta}, \boldsymbol{Z}_{0:k+1} | \boldsymbol{Y}_{0:k+1}}$, we simply embed $(\mathfrak{M}_j)_{j \leq k}$ into an identity map to form the $(T_j)_{j \leq k}$ defined in (3.3), and push forward reference samples from $\eta^{0:k}$ through $\mathfrak{T}_k := T_0 \circ \cdots \circ T_k$ (see Theorem 3.1). Moreover, a simple corollary of Theorem 3.1 shows that the map $\mathfrak{T}_k^{\boldsymbol{\Theta}} = \mathfrak{M}_0^{\boldsymbol{\Theta}} \circ \cdots \circ \mathfrak{M}_k^{\boldsymbol{\Theta}}$ pushes forward $\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}}$ to the marginal $\pi_{\boldsymbol{\Theta} | \boldsymbol{Y}_{0:k+1}}$, for all $k \geq 0$, whereas the map $\mathfrak{M}_k^1$ pushes forward $\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \eta_{\boldsymbol{X}_{k+1}}$ to the filtering distribution $\pi_{\boldsymbol{Z}_{k+1} | \boldsymbol{Y}_{0:k+1}}$. ($\mathfrak{M}_k^{\boldsymbol{\Theta}}$ and $\mathfrak{M}_k^1$ were defined in (3.2).) Figure 2 illustrates the solution of the inference problem, with additional results in Appendix A.

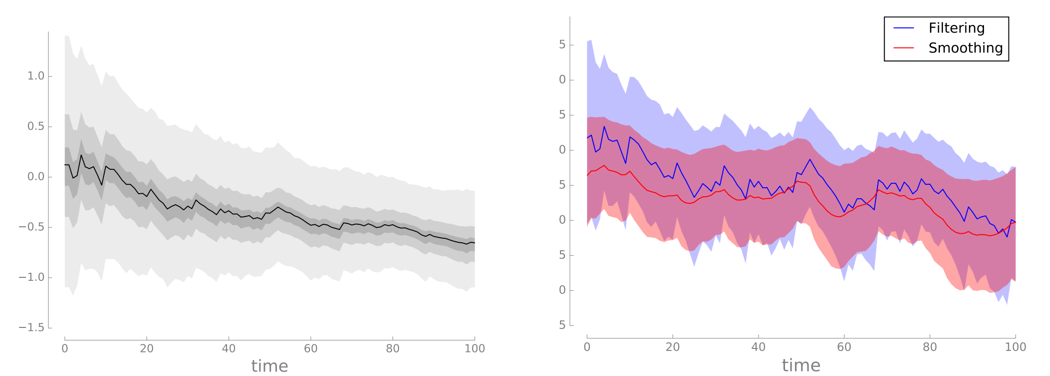

Figure 2: (*left*) At each time k, we illustrate the $\{5, 25, 40, 60, 75, 95\}$–percentiles (shaded regions) and the mean (black solid line) of the posterior distribution of the hyperparameter μ, i.e., $\pi_{\mu | \boldsymbol{Y}_{0:k}}$, for $k = 0, \ldots, N$. (*right*) Similarly, at each time k, we illustrate the mean (solid curves) and the $\{5, 95\}$–percentiles (shaded regions) of the filtering distribution $\pi_{\boldsymbol{Z}_k | \boldsymbol{Y}_{0:k}}$ (in blue) and of the marginals $\pi_{\boldsymbol{Z}_k | \boldsymbol{Y}_{0:N}}$ of the full smoothing distribution (in red), for $k = 0, \ldots, N$.

[1] Notice that $\mathfrak{M}_k$ in (3.2) is lower triangular up to a permutation of the input and output space, and thus it can be easily computed via (1.3) [15]. Its particular sparsity pattern, however, is required for the theorem to hold.

A Additional results for the stochastic volatility model of Section 4

Here we provide additional figures that illustrate the transport-based solution of the joint state–parameter inference problem described in Section 4. Captions explain each figure.

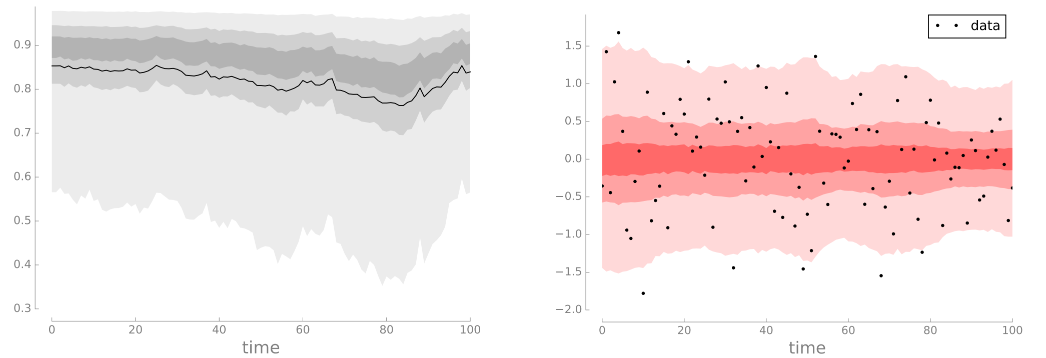

Figure 3: (*left*) Same as in Figure 2 (*left*), but for the hyperparameter ϕ. (*right*) Black dots represent the observed data $(\boldsymbol{y}_k)_{k=0}^N$. Moreover, at each time k, we illustrate the $\{5, 25, 40, 60, 75, 95\}$–percentiles (shaded regions) of the posterior predictive distribution, i.e., the distribution of $\boldsymbol{Y}_k$ conditioned on the event $\{\boldsymbol{Y}_{0:N} = \boldsymbol{y}_{0:N}\}$, for $k = 0, \ldots, N$.

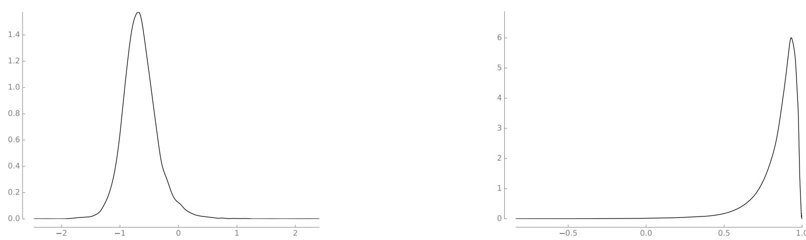

Figure 4: (*left*) Posterior marginal of μ, i.e., $\pi_{\mu|\boldsymbol{Y}_{0:N}}$. (*right*) Posterior marginal of ϕ, i.e., $\pi_{\phi|\boldsymbol{Y}_{0:N}}$.

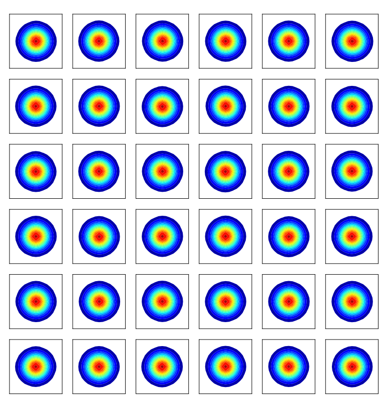

Figure 5: Randomly chosen two-dimensional conditionals of the pullback of $\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:N}|\boldsymbol{Y}_{0:N}}$ by the numerical approximation of $\mathfrak{T}_{N-1} := T_0 \circ \cdots \circ T_{N-1}$. (See the definitions of these quantities in Theorem 3.1.) Since we use a standard normal reference distribution, the numerical approximation of $\mathfrak{T}_{N-1}$ should be regarded as satisfactory if the pullback density $(\mathfrak{T}_{N-1})^\sharp \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:N}|\boldsymbol{Y}_{0:N}}$ is close to a standard normal, as it is here.

B Proofs of the main results

Proof of Theorem 3.1. Let $\mathfrak{c}_0 := \int \widetilde{\pi}^0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, and define a sequence of maps $(\widetilde{T}_k)$ as:

$$\widetilde{T}_k(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+1}) = \begin{bmatrix} \mathfrak{T}_k^{\boldsymbol{\Theta}}(\boldsymbol{x}_\theta) \\ \mathfrak{M}_k^1(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+1}) \end{bmatrix}, \qquad \widetilde{T}_k : \mathbb{R}^{n+p} \to \mathbb{R}^{n+p}, \tag{B.1}$$

for all $k \geq 0$. We first prove that $\mathfrak{c}_k < \infty$ and that $(\widetilde{T}_k)_\sharp\,(\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \eta_{\boldsymbol{X}_{k+1}}) = \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{k+1}|\boldsymbol{Y}_{0:k+1}}$, for all $k \geq 0$, using an induction argument over k. For the base case, just notice that $\mathfrak{c}_0 = 1$ since $\widetilde{\pi}^0$ is a probability density. Thus, $(\mathfrak{M}_0)_\sharp\,\eta^0 = \pi^0$ is well defined, and by (3.2) it must be that $\widetilde{T}_0$ pushes forward $\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \eta_{\boldsymbol{X}_1}$ to the marginal $\int \pi^0(\boldsymbol{x}_\theta, \boldsymbol{x}_0, \boldsymbol{x}_1)\,\mathrm{d}\boldsymbol{x}_0 = \int \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_0,\boldsymbol{Z}_1|\boldsymbol{Y}_0,\boldsymbol{Y}_1}(\boldsymbol{x}_\theta, \boldsymbol{x}_0, \boldsymbol{x}_1|\boldsymbol{y}_0, \boldsymbol{y}_1)\,\mathrm{d}\boldsymbol{x}_0 = \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_1|\boldsymbol{Y}_0,\boldsymbol{Y}_1}(\boldsymbol{x}_\theta, \boldsymbol{x}_1|\boldsymbol{y}_0, \boldsymbol{y}_1)$. (Notice that, by definition, $\mathfrak{T}_0^{\boldsymbol{\Theta}} \equiv \mathfrak{M}_0^{\boldsymbol{\Theta}}$.) Now assume that $\mathfrak{c}_k < \infty$ and that $(\widetilde{T}_k)_\sharp\,(\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \eta_{\boldsymbol{X}_{k+1}}) = \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{k+1}|\boldsymbol{Y}_{0:k+1}}$ for a fixed k. Then:

$$\begin{aligned} \mathfrak{c}_{k+1} &= \int \mathfrak{c}_{k+1} \cdot \pi^{k+1}(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+2})\,\mathrm{d}\boldsymbol{x}_\theta\,\mathrm{d}\boldsymbol{x}_{k+1}\mathrm{d}\boldsymbol{x}_{k+2} \\ &= \pi_{\boldsymbol{Y}_{k+2}|\boldsymbol{Y}_{0:k+1}}(\boldsymbol{y}_{k+2}|\boldsymbol{y}_{0:k+1}) < \infty. \end{aligned} \tag{B.2}$$

Moreover, notice that $\widetilde{T}_{k+1}$ can always be written as $\widetilde{T}_{k+1} = A_{k+1} \circ B_{k+1}$, where:

$$A_{k+1}(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+2}) = \begin{bmatrix} \mathfrak{T}_k^{\boldsymbol{\Theta}}(\boldsymbol{x}_\theta) \\ \boldsymbol{x}_{k+2} \end{bmatrix}, \qquad B_{k+1}(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+2}) = \begin{bmatrix} \mathfrak{M}_{k+1}^{\boldsymbol{\Theta}}(\boldsymbol{x}_\theta) \\ \mathfrak{M}_{k+1}^1(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+2}) \end{bmatrix}, \tag{B.3}$$

and that $(\widetilde{T}_{k+1})_\sharp\,(\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \eta_{\boldsymbol{X}_{k+2}}) = \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{k+2}|\boldsymbol{Y}_{0:k+2}}$ if and only if $(B_{k+1})_\sharp\,(\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \eta_{\boldsymbol{X}_{k+2}}) = A_{k+1}^\sharp\,\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{k+2}|\boldsymbol{Y}_{0:k+2}}$. By definition of $\mathfrak{M}_{k+1}$, B_{k+1} must push forward $\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \eta_{\boldsymbol{X}_{k+2}}$ to the marginal $\int \pi^{k+1}(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+2})\,\mathrm{d}\boldsymbol{x}_{k+1}$. A simple calculation shows that:

$$\begin{aligned} \int \pi^{k+1}(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+1}, \boldsymbol{x}_{k+2})\,\mathrm{d}\boldsymbol{x}_{k+1} &= \frac{\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{k+2}|\boldsymbol{Y}_{0:k+2}}(\boldsymbol{z}_{\boldsymbol{\Theta}}, \boldsymbol{x}_{k+2}|\boldsymbol{y}_{0:k+2})}{|\nabla(\mathfrak{T}_k^{\boldsymbol{\Theta}})^{-1}(\boldsymbol{z}_{\boldsymbol{\Theta}})|} \\ &= A_{k+1}^\sharp\,\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{k+2}|\boldsymbol{Y}_{0:k+2}}(\boldsymbol{x}_\theta, \boldsymbol{x}_{k+2}|\boldsymbol{y}_{0:k+2}), \end{aligned}$$

where $\boldsymbol{z}_{\boldsymbol{\Theta}} := \mathfrak{T}_k^{\boldsymbol{\Theta}}(\boldsymbol{x}_\theta)$, and concludes the induction argument. Since $\mathfrak{c}_k < \infty$ for all $k \geq 0$, the sequence of maps $(\mathfrak{M}_k)$ is well defined, and so is $(T_j)_{j=0}^k$. Now, we can finally prove the theorem using another induction argument over $k \geq 0$. For the base case, notice that $\mathfrak{T}_0 = T_0 = \mathfrak{M}_0$, and that, by definition, $\mathfrak{M}_0$ pushes forward η^0 to $\pi^0 = \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_0,\boldsymbol{Z}_1|\boldsymbol{Y}_0,\boldsymbol{Y}_1}$. Now assume that $\mathfrak{T}_k$ pushes forward $\eta^{0:k}$ to $\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:k+1}|\boldsymbol{Y}_{0:k+1}}$ for a fixed k, and notice that

$$\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:k+2}|\boldsymbol{Y}_{0:k+2}} = \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:k+1}|\boldsymbol{Y}_{0:k+1}} \cdot \frac{\pi_{\boldsymbol{Y}_{k+2}|\boldsymbol{Z}_{k+2},\boldsymbol{\Theta}} \cdot \pi_{\boldsymbol{Z}_{k+2}|\boldsymbol{Z}_{k+1},\boldsymbol{\Theta}}}{\pi_{\boldsymbol{Y}_{k+2}|\boldsymbol{Y}_{0:k+1}}}. \tag{B.4}$$

Thus, for a given $\mathfrak{T}_{k+1}$, it must be that

$$\begin{aligned} \mathfrak{T}_{k+1}^\sharp\,\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:k+2}|\boldsymbol{Y}_{0:k+2}} &= T_{k+1}^\sharp \left(\mathfrak{T}_k^\sharp\,\pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:k+1}|\boldsymbol{Y}_{0:k+1}} \cdot \frac{\pi^{k+1}}{\eta_{\boldsymbol{X}_{\boldsymbol{\Theta}}} \cdot \eta_{\boldsymbol{X}_{k+1}}} \right) \\ &= T_{k+1}^\sharp \left(\prod_{j=0}^{k} \eta_{\boldsymbol{X}_j} \cdot \pi^{k+1} \right) \\ &= \prod_{j=0}^{k} \eta_{\boldsymbol{X}_j} \cdot \mathfrak{M}_{k+1}^\sharp \pi^{k+1} = \eta^{0:k+1} \end{aligned} \tag{B.5}$$

Hence, $(\mathfrak{T}_{k+1})_\sharp\,\eta^{0:k+1} = \pi_{\boldsymbol{\Theta},\boldsymbol{Z}_{0:k+2}|\boldsymbol{Y}_{0:k+2}}$, concluding the induction argument and the proof of the theorem. □

Acknowledgements

This work was supported by the US Department of Energy, Office of Advanced Scientific Computing (ASCR), under grant number DE-SC0009297.

References

[1] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

[2] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: a review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

[3] G. Carlier, A. Galichon, and F. Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

[4] A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

[5] J. Durbin and S. J. Koopman. Time series analysis of non-Gaussian observations based on state-space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):3–56, 2000.

[6] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[7] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin. On particle methods for parameter estimation in state-space models. *Statistical Science*, 30(3):328–351, 2015.

[8] S. Kim, N. Shephard, and S. Chib. Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393, 1998.

[9] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[10] H. Knothe. Contributions to the theory of convex bodies. *The Michigan Mathematical Journal*, 4(1):39–52, 1957.

[11] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[12] H. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[13] V. Laparra, G. Camps-Valls, and J. Malo. Iterative gaussianization: from ICA to random rotations. *IEEE Transactions on Neural Networks*, 22(4):537–549, 2011.

[14] S. L. Lauritzen. *Graphical models*. Oxford University Press, 1996.

[15] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. An introduction to sampling via measure transport. In *Handbook of Uncertainty Quantification*. Springer, 2016. arXiv:1602.05023.

[16] R. J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–324, 1995.

[17] X. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002.

[18] T. A. Moselhy and Y. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[19] G. Parisi. *Statistical field theory*. Addison-Wesley, 1988.

[20] M. Parno and Y. Marzouk. Transport map accelerated Markov chain Monte Carlo. *arXiv preprint arXiv:1412.5492*, 2014.

[21] C. Peterson. A mean field theory learning algorithm for neural networks. *Complex systems*, 1:995–1019, 1987.

[22] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *Proceedings of The 32nd International Conference on Machine Learning*, 2015. arXiv:1505.05770.

[23] C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[24] M. Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, pages 470–472, 1952.

[25] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

[26] S. Särkkä. *Bayesian filtering and smoothing*, volume 3. Cambridge University Press, 2013.

[27] L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4(1):61–76, 1996.

[28] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

[29] D. Tran, D. Blei, and E. M. Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems*, pages 3550–3558, 2015.

[30] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[31] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[32] L. Wang and X. Meng. Warp bridge sampling: The next generation. *arXiv preprint arXiv:1609.07690*, 2016.