Recursive Entropic Variational Inference for Nonlinear State-Space Models

Anonymous authors
Paper under double-blind review

Abstract

We present a class of algorithms for state estimation in nonlinear, non-Gaussian state-space models. Our approach is based on a variational Lagrangian formulation that casts Bayesian inference as a sequence of entropic trust-region updates subject to dynamic consistency constraints. This framework gives rise to a family of forward-backward algorithms whose structure is determined by the chosen factorization of the variational posterior. By focusing on Gauss–Markov approximations, we derive recursive schemes with favorable computational complexity. For general nonlinear, non-Gaussian models, we close the recursions using generalized statistical linear regression and Fourier–Hermite moment matching.

1 Introduction

Accurate state estimation in partially observable dynamical phenomena is a fundamental problem across scientific and engineering disciplines, including robotics (Barfoot, 2024), economics (Jacquier et al., 2002), and biology (Murray, 2002). Bayesian inference in state-space models (SSMs) provides a principled framework, including techniques such as Kalman filtering, sequential Monte Carlo, and message passing (Särkkä & Svensson, 2023). Additionally, an important link between inference and optimization was established early in the field (Kálmán, 1960; Cox, 1964; Bell, 1994), giving rise to a suite of inference-as-optimization algorithms. More recently, this relationship has been deepened by approximate inference methods, such as variational inference (VI, Wainwright & Jordan, 2008) and expectation propagation (EP, Minka, 2001), which extend the optimization perspective beyond point estimates to structured posterior approximations.

For state-space models with conjugate prior-likelihood pairs, the corresponding Bayesian posterior distributions admit closed-form solutions via efficient recursive algorithms, such as the Rauch–Tung–Striebel (RTS) smoother for linear-Gaussian SSMs (Rauch et al., 1965). However, general nonlinear, non-Gaussian models do not usually admit closed-form recursive solutions. While sampling-based methods such as sequential Monte Carlo (Chopin & Papaspiliopoulos, 2020) and Markov chain Monte Carlo (Brooks et al., 2011) provide asymptotically exact solutions, they do so at the price of high computational complexity and limited scalability, particularly for sequences in high-dimensional spaces and long time horizons (Beskos et al., 2014). Approximate Bayesian techniques, though often suboptimal, offer a trade-off that sacrifices exactness for computational efficiency. This trade-off motivates our focus on this class of methods.

Variational inference and expectation propagation methods have been adapted for Bayesian inference in state-space models (Deisenroth & Mohamed, 2012; Chang et al., 2020; Wilkinson et al., 2020). However, despite the strong connection of both frameworks to numerical optimization (Wilkinson et al., 2023), many existing approaches extend VI and EP to state-space representations in an *ad hoc* manner. Such methods often incorporate variational principles into the structure of the Rauch–Tung–Striebel smoother template rather than deriving recursive algorithms from an explicit dynamic optimization problem. As a result, they may not fully leverage the potential benefits of a systematic approach to state estimation.

In this work, we cast approximate Bayesian inference in nonlinear, non-Gaussian state-space models as an iterative *dynamic* optimization problem over the space of candidate posteriors. Our key contributions are:

- *Dynamic optimization formulation*: We introduce a principled variational framework for Bayesian inference that formulates inference as a sequence of entropic trust-region updates in probability-density space, using a structured proximal regularizer.
- *Unified recursive algorithms*: We derive a family of efficient forward-backward recursion schemes whose structure is determined entirely by the factorization of the variational posterior — forward-, reverse-, and hybrid-Markov factorizations.
- *Adaptive step size*: We show that the step size between iterative updates corresponds to a Lagrangian multiplier whose optimal value is determined by a dual Rényi divergence objective.
- *Generalization of classical smoothers*: The proposed approach generalizes algorithms in the linear-Gaussian setting, providing a bridge between classical and modern approximate inference techniques.
- *Flexible posterior approximations*: We provide practical realizations using Gauss–Markov approximations, instantiated via generalized statistical linear regression and Fourier–Hermite moment matching, allowing for efficient inference in nonlinear and non-Gaussian models.

The paper is organized as follows. Section 2 introduces the approximate inference problem in state-space models. Section 3 provides a brief review of proximal variational optimization, which serves as the foundation for our method. In Section 4, we present our dynamic Lagrangian framework for approximate Bayesian inference. Section 5 specializes this framework to the Gaussian setting and derives practical recursive algorithms. We then connect our approach to related work in Section 6. Finally, Section 7 provides a numerical evaluation of the proposed algorithms. Proofs are deferred to the supplement.

2 Problem Statement

We consider the problem of Bayesian state estimation in nonlinear, non-Gaussian state-space models. Let $\{x_k \in \mathbb{R}^d\}_{k=0}^T$ be a latent discrete-time Markov process and $\{y_k \in \mathbb{R}^m\}_{k=1}^T$ a corresponding sequence of noisy observations. The system is governed by the state-space dynamics:

$$x_0 \sim p_0(\cdot), \qquad x_{k+1} \sim f_k(\cdot \mid x_k), \qquad y_{k+1} \sim h_{k+1}(\cdot \mid x_{k+1}), \tag{1}$$

where $p_0(x_0)$ is the initial prior probability density, $f_k(x_{k+1} \mid x_k)$ is the transition density of the latent Markov dynamics, and $h_k(y_k \mid x_k)$ is the conditional likelihood of the measurements. State estimation refers to the problems of online filtering and offline smoothing, which aim to reconstruct the posterior distributions $p(x_k \mid y_{1:k})$ and $p(x_k \mid y_{1:T})$, for all $0 \leq k \leq T$, respectively. For arbitrary transition densities $f_k(x_{k+1} \mid x_k)$ and measurement likelihoods $h_k(y_k \mid x_k)$, these posterior densities are generally intractable.

In this work, we focus on finding the best approximation $q(x_{0:T} \mid y_{1:T})$ of the smoothing posterior density within a certain variational family of joint distributions

$$q(x_{0:T} \mid y_{1:T}) \approx p(x_{0:T} \mid y_{1:T}) \propto p_0(x_0) \prod_{k=0}^{T-1} f_k(x_{k+1} \mid x_k)\, h_{k+1}(y_{k+1} \mid x_{k+1}).$$

More specifically, to leverage the temporal structure of state-space models, we consider structured Markov approximations of the smoothing posterior, leading to the following possibilities:

Assumption 1 (Forward-Markov factorization) *For state-space models of the form (1), we restrict the approximate smoothing posterior to the forward-Markov variational family*

$$\overrightarrow{q}(x_{0:T} \mid y_{1:T}) = \overrightarrow{q}_0(x_0 \mid y_{1:T}) \prod_{k=0}^{T-1} \overrightarrow{q}_k(x_{k+1} \mid x_k, y_{k+1:T}).$$

Assumption 2 (Reverse-Markov factorization) *For state-space models of the form (1), we restrict the approximate smoothing posterior to the reverse-Markov variational family*

$$\overleftarrow{q}(x_{0:T} \mid y_{1:T}) = \overleftarrow{q}_T(x_T \mid y_{1:T}) \prod_{k=1}^{T} \overleftarrow{q}_k(x_{k-1} \mid x_k, y_{1:k-1}).$$

We use $\overrightarrow{q}$ and $\overleftarrow{q}$ to distinguish forward- and reverse-Markov approximations, respectively. Although Assumption 1 and Assumption 2 explicitly state the dependence of the approximate conditionals on the measurements, we omit this dependence in subsequent sections for notational simplicity.

3 Entropic Proximal Variational Optimization

Before introducing the proposed approach to variational smoothing in state-space models, we briefly review the principles of entropic proximal variational optimization in a simpler, static setting. Consider a hidden variable $x \in \mathbb{R}^d$, observed data $y \in \mathbb{R}^m$, a likelihood $p(y \mid x)$, and a prior $p(x)$. The goal is to approximate the posterior $p(x \mid y)$ by a distribution q, belonging to a variational family of probability distributions $\mathcal{Q}$, that minimizes the Kullback–Leibler (KL) divergence:

$$q^\star(x) = \underset{q \in \mathcal{Q}}{\arg\min}\ \mathbb{D}_{\mathrm{KL}}\left[q(x) \mid\mid p(x \mid y)\right] = \underset{q \in \mathcal{Q}}{\arg\min}\ \mathbb{E}_q\left[\log \frac{q(x)}{p(x \mid y)}\right].$$

Equivalently, since the marginal likelihood $p(y)$ does not depend on $q(x)$, this problem can be written as the maximization of the evidence lower bound (ELBO) (Blei et al., 2017):

$$q^\star(x) = \underset{q \in \mathcal{Q}}{\arg\max}\ \mathcal{L}(q) \coloneqq \mathbb{E}_q\Big[\log p(x, y)\Big] - \mathbb{E}_q\Big[\log q(x)\Big] \leq \log p(y), \tag{2}$$

where $p(y)$ is the marginal likelihood or evidence. In conjugate models, the ELBO can be optimized exactly (Bishop, 2006), but for more general settings, approximate solutions are required. Black-box variational inference (Ranganath et al., 2014) provides a flexible framework for such cases. However, alternative methods such as natural and Riemannian gradient approaches (Honkela et al., 2010) and proximal variational inference (Chrétien & Hero, 2002; Khan et al., 2015; Theis & Hoffman, 2015) explicitly account for the geometry of the variational family and often lead to more stable and efficient optimization (Amari, 1998).

In this work, we focus on the entropic proximal variational optimization framework to derive recursive algorithms for structured approximate inference in nonlinear non-Gaussian state-space models. However, before directing our attention to the case of state-space models, we use the static case in (2) to illustrate the mechanics of entropic proximal variational optimization. We start by formulating the approximate inference problem within the iterative entropic proximal optimization framework proposed by Teboulle (1992); Iusem et al. (1994). Starting from the ELBO in (2), we introduce an entropic-proximal constraint on the subsequent posterior iterate in the form of a Kullback–Leibler divergence, leading to the following nonlinear program:

$$\begin{aligned} \underset{q(x)}{\text{maximize}} \quad & \mathbb{E}_q\Big[\log p(x, y)\Big] - \mathbb{E}_q\Big[\log q(x)\Big], \\ \text{subject to} \quad & \mathbb{D}_{\mathrm{KL}}\left[q(x) \mid\mid q^{[i]}(x)\right] \leq \varepsilon \quad \text{and} \quad \int q(x)\,\mathrm{d}x = 1, \end{aligned} \tag{3}$$

where $q^{[i]}(x)$ is the approximate posterior at iteration i and $\varepsilon \geq 0$ is a hyperparameter that controls the information bottleneck between the current iterate $q^{[i]}(x)$ and the solution $q^{[i+1]}(x)$. Additionally, problem (3) includes a distributional normalization constraint for $q^{[i+1]}(x)$, whereas the positivity of $q^{[i+1]}(x)$ is implied by the logarithmic function embedded within the KL constraint. The following proposition derives the solution to (3), the variational distribution iterate $q^{[i+1]}(x)$, by constructing the Lagrangian.

Proposition 1 (Damped Gibbs posterior) *The solution to the constrained nonlinear program (3), corresponding to a Bayesian model with likelihood $p(y \mid x)$ and prior $p(x)$, is given by the Gibbs posterior*

$$q^{[i+1]}(x) = \left[\mathcal{Z}^{[i+1]}(\beta)\right]^{-1} \left[p(y \mid x)\, p(x)\right]^{1-\beta} \left[q^{[i]}(x)\right]^{\beta},$$

with a damping parameter $\beta \in [0, 1)$ and a normalizing constant

$$\mathcal{Z}^{[i+1]}(\beta) = \int \left[p(y \mid x)\, p(x)\right]^{1-\beta} \left[q^{[i]}(x)\right]^{\beta} \mathrm{d}x.$$

The damping parameter β is a reparameterization of the Lagrangian multiplier $\alpha \geq 0$ associated with the Kullback–Leibler divergence constraint in (3), so that $\beta = \alpha/(1+\alpha)$. Furthermore, the optimal β is a minimizer of the dual problem

$$\underset{\beta}{\text{minimize}} \;\; \mathcal{G}(\beta) = \frac{\beta\varepsilon}{1-\beta} + \frac{1}{1-\beta}\log \mathcal{Z}^{[i+1]}(\beta), \quad \text{subject to} \;\; 0 \leq \beta < 1. \tag{4}$$

Proof. See Appendix A.

Remark 1 *The posterior iterate $q^{[i+1]}(x)$ can be interpreted as a geometric interpolation in the space of probability densities between the exact posterior $p(x \mid y) \propto p(y \mid x)\, p(x)$ and the previous iterate $q^{[i]}(x)$. As $\beta \to 0$, the update recovers the exact posterior, whereas as $\beta \to 1$, the update approaches the previous iterate $q^{[i]}(x)$ and corresponds to an infinitesimal proximal update.*

Remark 2 *The dual objective over β in (4) can be interpreted through the variational Rényi bound of Li & Turner (2016):*

$$\frac{1}{1-\beta}\log \mathcal{Z}^{[i+1]}(\beta) = \log p(y) - \mathbb{D}_\beta\left[q^{[i]}(x) \;||\; p(x \mid y)\right],$$

where $\mathbb{D}_\beta\left[\cdot \;||\; \cdot\right]$ denotes the Rényi divergence with order β. As the damping parameter β varies, the corresponding Rényi divergence traces a continuum of α-divergence geometries, each emphasizing different regions of the posterior distribution. Thus, varying β smoothly alters the global geometric structure of the optimization problem but preserves the local notion of distance defined by the Fisher information metric induced by the primal Kullback–Leibler divergence objective (Amari, 2016).

Having reviewed the principles of entropic proximal optimization, we now turn to state-space models. In the next section, we extend this framework into the dynamic setting and derive recursive algorithms for approximate Bayesian inference that exploit the structure of the assumed posterior.

4 Entropic Proximal Bayesian Smoothing

For a state-space model (1), we adapt the ELBO from (2) to the joint smoothing distribution as follows

$$\mathcal{L}(q) = \mathbb{E}_q\Big[\log p(x_{0:T}, y_{1:T})\Big] - \mathbb{E}_q\Big[\log q(x_{0:T})\Big] \leq \log p(y_{1:T}),$$

where $p(x_{0:T}, y_{1:T})$ is the joint state-measurement distribution given by

$$p(x_{0:T}, y_{1:T}) = p_0(x_0)\prod_{k=0}^{T-1} f_k(x_{k+1} \mid x_k)\, h_{k+1}(y_{k+1} \mid x_{k+1}).$$

We now formulate the entropic proximal optimization problem over the joint smoothing posterior $q(x_{0:T})$:

$$\begin{aligned} \underset{q\in\mathcal{Q}}{\text{maximize}} \quad & \mathbb{E}_q\Big[\log p(x_{0:T}, y_{1:T})\Big] - \mathbb{E}_q\Big[\log q(x_{0:T})\Big], \\ \text{subject to} \quad & \mathbb{D}_{\mathrm{KL}}\left[q(x_{0:T}) \;||\; q^{[i]}(x_{0:T})\right] \leq \varepsilon \quad \text{and} \quad \int q(x_{0:T})\,\mathrm{d}x_{0:T} = 1. \end{aligned} \tag{5}$$

This formulation does not yet impose any assumptions on the structure of the approximate posterior $q(x_{0:T})$. While (5) can be solved in a manner similar to (3), such an approach can lead to significant computational complexity due to the high dimensionality of $q(x_{0:T})$ as it extends over the state and time dimensions. To address this challenge, we introduce variations to (5) that take advantage of sparsity induced by the forward- and reverse Markov structure from Assumption 1 and Assumption 2, enabling recursive inference algorithms with linear time complexity in the horizon T.

4.1 Damped Forward-Markov Posterior

Here, we restrict the admissible posterior family in (5) to the forward-Markov decomposition stated in Assumption 1. Under this factorization, the joint smoothing posterior is represented by an initial marginal density and a sequence of forward transition conditionals. The entropic proximal update can then be decomposed into a recursive message-passing scheme, where a backward recursion first accumulates the contribution of future measurements into potential functions, and a subsequent forward recursion constructs the updated forward-Markov posterior and its marginal smoothing distributions. This leads to a generic *backward-forward* algorithm for approximate Bayesian smoothing.

Proposition 2 (Optimal forward-Markov posterior) *For state-space models of the form (1), and under Assumption 1, the approximate forward-Markov smoothing posterior that solves problem (5) within the forward-Markov variational family is characterized by the following tilted distributions:*

$$\overrightarrow{q}_0^{[i+1]}(x_0) = \left[\overrightarrow{\mathcal{Z}}_0^{[i+1]}\right]^{-1} \left[\overrightarrow{q}_0^{[i]}(x_0)\right]^{\beta} \left[\exp\left\{\overrightarrow{V}_0^{[i+1]}(x_0)\right\}\right]^{1-\beta}, \tag{6}$$

$$\overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k) = \left[\overrightarrow{\psi}_k^{[i+1]}(x_k)\right]^{-1} \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^{\beta} \times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta}, \quad 0 \le k < T, \tag{7}$$

where $\beta \in [0, 1)$ is the damping parameter associated with the Lagrangian multiplier $\alpha \ge 0$, so that $\beta = \alpha/(1+\alpha)$. The quantities $\overrightarrow{\mathcal{Z}}_0^{[i+1]}$ and $\overrightarrow{\psi}_k^{[i+1]}(x_k)$ are the corresponding normalizing factors

$$\overrightarrow{\mathcal{Z}}_0^{[i+1]} = \int \left[\overrightarrow{q}_0^{[i]}(x_0)\right]^{\beta} \left[\exp\left\{\overrightarrow{V}_0^{[i+1]}(x_0)\right\}\right]^{1-\beta} \mathrm{d}x_0, \tag{8}$$

$$\overrightarrow{\psi}_k^{[i+1]}(x_k) = \int \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^{\beta} \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \mathrm{d}x_{k+1}, \quad 0 \le k < T.$$

The potential functions $\overrightarrow{V}_k^{[i+1]}(x_k)$, for all $0 \le k \le T$, are computed recursively backwards via

$$\overrightarrow{V}_k^{[i+1]}(x_k) = \begin{cases} \log h_T(y_T \mid x_T) & \text{if } k = T, \\ \log h_k(y_k \mid x_k) + 1/(1-\beta) \log \overrightarrow{\psi}_k^{[i+1]}(x_k) & \text{if } 0 < k < T, \\ \log p_0(x_0) + 1/(1-\beta) \log \overrightarrow{\psi}_0^{[i+1]}(x_0) & \text{if } k = 0. \end{cases} \tag{9}$$

Here, each $\overrightarrow{\psi}_k^{[i+1]}(x_k)$ depends on the next potential $\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})$, so that the recursion is initialized at $k = T$ and propagated backwards to $k = 0$. Finally, the optimal damping β is the minimizer of the dual objective

$$\underset{\beta}{\text{minimize}} \;\; \overrightarrow{\mathcal{G}}(\beta) = \frac{\beta\varepsilon}{1-\beta} + \frac{1}{1-\beta} \log \overrightarrow{\mathcal{Z}}_0^{[i+1]}(\beta), \quad \text{subject to } \; 0 \le \beta < 1. \tag{10}$$

Proof. See Appendix B.

Remark 3 *The potential functions $\overrightarrow{V}_k(x_k)$ can be interpreted as damped log-space backward messages, that accumulate the contribution of the future measurements $y_{k:T}$ conditioned on x_k, while the log-normalizers $\log \overrightarrow{\psi}_k(x_k)$ and $\log \overrightarrow{\mathcal{Z}}_0$ correspond to $\log q(y_{k+1:T} \mid x_k)$ and $\log q(y_{1:T})$, respectively.*

Remark 4 *Under Assumption 1, the marginal smoothing distributions $\overrightarrow{q}_{k+1}^{[i+1]}(x_{k+1})$, for all $0 \le k < T$, are computed via forward propagation starting from $\overrightarrow{q}_0^{[i+1]}(x_0)$*

$$\overrightarrow{q}_{k+1}^{[i+1]}(x_{k+1}) = \int \overrightarrow{q}_k^{[i+1]}(x_k)\, \overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k)\, \mathrm{d}x_k,$$

where $\overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k)$ and $\overrightarrow{q}_0^{[i+1]}(x_0)$ are given by Proposition 2.

4.2 Damped Reverse-Markov Posterior

Next, we restrict the admissible posterior family in (5) to the reverse-Markov decomposition stated in Assumption 2. Under this factorization, the approximate joint smoothing posterior is parameterized by a terminal marginal density $\overleftarrow{q}_T(x_T)$ and a sequence of reverse conditionals $\overleftarrow{q}_k(x_{k-1} \mid x_k)$. The proximal update can therefore be decomposed into local tilted updates for the terminal marginal and for each reverse conditional. These local updates are coupled through potential functions that accumulate information from the initial condition and past measurements. The resulting scheme first computes these potentials by a forward recursion and then propagates the updated posterior marginals backward in time. This yields a generic *forward-backward* recursive algorithm for approximate Bayesian smoothing.

Proposition 3 (Optimal reverse-Markov posterior) *For state-space models of the form (1), and under Assumption 2, the approximate reverse-Markov smoothing posterior that solves problem (5) within the reverse-Markov variational family is characterized by the following tilted conditional and marginal distributions:*

$$\overleftarrow{q}_k^{[i+1]}(x_{k-1} \mid x_k) = \left[\overleftarrow{\psi}_k^{[i+1]}(x_k)\right]^{-1} \left[\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)\right]^{\beta} \times \left[f_{k-1}(x_k \mid x_{k-1}) \exp\left\{\overleftarrow{V}_{k-1}^{[i+1]}(x_{k-1})\right\}\right]^{1-\beta}, \quad 0 < k \leq T, \tag{11}$$

$$\overleftarrow{q}_T^{[i+1]}(x_T) = \left[\overleftarrow{\mathcal{Z}}_T^{[i+1]}\right]^{-1} \left[\overleftarrow{q}_T^{[i]}(x_T)\right]^{\beta} \left[\exp\left\{\overleftarrow{V}_T^{[i+1]}(x_T)\right\}\right]^{1-\beta}, \tag{12}$$

where $\beta \in [0, 1)$ is the damping parameter associated with the Lagrangian multiplier $\alpha \geq 0$, so that $\beta = \alpha/(1+\alpha)$. The quantities $\overleftarrow{\psi}_k^{[i+1]}(x_k)$ and $\overleftarrow{\mathcal{Z}}_T^{[i+1]}$ are the corresponding normalizing factors

$$\overleftarrow{\psi}_k^{[i+1]}(x_k) = \int \left[\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)\right]^{\beta} \left[f_{k-1}(x_k \mid x_{k-1}) \exp\left\{\overleftarrow{V}_{k-1}^{[i+1]}(x_{k-1})\right\}\right]^{1-\beta} \mathrm{d}x_{k-1}, \quad 0 < k \leq T, \tag{13}$$

$$\overleftarrow{\mathcal{Z}}_T^{[i+1]} = \int \left[\overleftarrow{q}_T^{[i]}(x_T)\right]^{\beta} \left[\exp\left\{\overleftarrow{V}_T^{[i+1]}(x_T)\right\}\right]^{1-\beta} \mathrm{d}x_T.$$

The potential functions $\overleftarrow{V}_k^{[i+1]}(x_k)$, for all $0 \leq k \leq T$, are computed recursively forward via

$$\overleftarrow{V}_k^{[i+1]}(x_k) = \begin{cases} \log p_0(x_0) & \text{if } k = 0, \\ \log h_k(y_k \mid x_k) + 1/(1-\beta) \log \overleftarrow{\psi}_k^{[i+1]}(x_k) & \text{if } 0 < k \leq T. \end{cases} \tag{14}$$

Here, each $\overleftarrow{\psi}_k^{[i+1]}(x_k)$ depends on the previous potential $\overleftarrow{V}_{k-1}^{[i+1]}(x_{k-1})$, with the recursion initialized at $k = 0$ and propagated forwards to $k = T$. Finally, the optimal damping β is the minimizer of the dual objective

$$\underset{\beta}{\text{minimize}} \quad \overleftarrow{\mathcal{G}}(\beta) = \frac{\beta\varepsilon}{1-\beta} + \frac{1}{1-\beta} \log \overleftarrow{\mathcal{Z}}_T^{[i+1]}(\beta), \quad \text{subject to } 0 \leq \beta < 1. \tag{15}$$

Proof. See Appendix C.

Remark 5 *The potential functions $\overleftarrow{V}_k(x_k)$ can be interpreted as damped log-space forward messages $\log q(x_k \mid y_{1:k})$, that accumulate the contribution of the initial condition and past measurements $y_{1:k}$ up to state x_k, while the log-normalizers $\log \overleftarrow{\psi}_k(x_k)$ and $\log \overleftarrow{\mathcal{Z}}_T$ correspond to $\log q(x_k \mid y_{1:k-1})$ and $\log q(y_{1:T})$, respectively.*

Remark 6 *Under Assumption 2, the marginal smoothing distributions $\overleftarrow{q}_{k-1}^{[i+1]}(x_{k-1})$, for all $1 \leq k \leq T$, are computed via backward propagation starting from $\overleftarrow{q}_T^{[i+1]}(x_T)$*

$$\overleftarrow{q}_{k-1}^{[i+1]}(x_{k-1}) = \int \overleftarrow{q}_k^{[i+1]}(x_k)\, \overleftarrow{q}_k^{[i+1]}(x_{k-1} \mid x_k)\, \mathrm{d}x_k,$$

where $\overleftarrow{q}_k^{[i+1]}(x_{k-1} \mid x_k)$ and $\overleftarrow{q}_T^{[i+1]}(x_T)$ are given by Proposition 3.

4.3 Hybrid Posterior Marginals

The recursions in Proposition 2 and Proposition 3 arise from two complementary factorizations of the approximate smoothing posterior. The forward-Markov factorization in Proposition 2 represents the joint trajectory through an initial marginal and forward transition conditionals, leading to a backward recursion for future-measurement potentials followed by forward marginal propagation. Conversely, the reverse-Markov factorization in Proposition 3 represents the same trajectory through a terminal marginal and reverse transition conditionals, leading to a forward recursion for past-measurement potentials followed by backward marginal propagation.

These two constructions need not be treated as mutually exclusive. When the forward- and reverse-Markov conditionals are associated with the same joint smoothing iterate $q^{[i]}(x_{0:T} \mid y_{1:T})$, they provide two alternative approximations of the same posterior. In this case, the backward potential recursion from Proposition 2 and the forward potential recursion from Proposition 3 can be combined at the marginal level. This yields a hybrid update in which each intermediate marginal receives information propagated from both temporal directions, while retaining the same entropic damping with respect to the previous marginal iterate.

Corollary 1 (Optimal hybrid marginals) *For state-space models of the form (1), suppose that the forward-Markov conditionals $\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)$ and reverse-Markov conditionals $\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)$ are the corresponding conditionals of a common joint smoothing distribution $q^{[i]}(x_{0:T} \mid y_{1:T})$, and therefore induce the same marginals $q_k^{[i]}(x_k)$. Then, the updated marginals $q_k^{[i+1]}(x_k)$ associated with problem (5) can be computed by combining the backward recursion from Proposition 2 with the forward recursion from Proposition 3*

$$q_k^{[i+1]}(x_k) \propto \begin{cases} \left[q_0^{[i]}(x_0)\right]^{\beta} \left[\exp\left\{\overrightarrow{V}_0^{[i+1]}(x_0)\right\}\right]^{1-\beta} & \text{if } k = 0, \\ \left[q_k^{[i]}(x_k)\right]^{\beta} \left[\exp\left\{\overleftarrow{V}_k^{[i+1]}(x_k)\right\} \overrightarrow{\psi}_k^{[i+1]}(x_k)\right]^{1-\beta} & \text{if } 0 < k < T, \\ \left[q_T^{[i]}(x_T)\right]^{\beta} \left[\exp\left\{\overleftarrow{V}_T^{[i+1]}(x_T)\right\}\right]^{1-\beta} & \text{if } k = T. \end{cases}$$

For $0 < k < T$, the potential $\overleftarrow{V}_k^{[i+1]}(x_k)$ carries the forward-in-time filtering contribution accumulated by the reverse-Markov recursion, while the predictive $\overrightarrow{\psi}_k^{[i+1]}(x_k)$ carries the backward-in-time likelihood contribution accumulated by the forward-Markov recursion. Therefore, the hybrid update forms a damped geometric interpolation between the previous marginal $q_k^{[i]}(x_k)$ and the product of the two directional message contributions. Equivalently, one may combine the reverse-Markov predictive $\overleftarrow{\psi}_k^{[i+1]}(x_k)$ with the forward-Markov potential $\overrightarrow{V}_k^{[i+1]}(x_k)$ to reach an equivalent construction of the marginals.

The forward–backward recursion schemes derived in Proposition 2, Proposition 3, and Corollary 1 share a general algorithmic structure that does not rely on specific assumptions about the dynamics $f_k(x_{k+1} \mid x_k)$, the measurement model $h_k(y_k \mid x_k)$, or the forms of the approximate posteriors $\overrightarrow{q}(x_{0:T})$ and $\overleftarrow{q}(x_{0:T})$. In the sections that follow, we present concrete implementations of these schemes by adopting conditionally Gaussian approximations for the forward- and reverse-Markov posterior distributions.

5 Practical Recursive Inference Algorithms

To transform the recursive schemes introduced in Section 4.1 and Section 4.2 into concrete and tractable algorithms, we now impose additional structure on the forward- and reverse-Markov posterior updates. The abstract recursions in Proposition 2 and Proposition 3 characterize the optimal tilted conditionals in (7) and (11), together with the corresponding potential functions in (9) and (14). These objects define the exact proximal update within the chosen Markov factorization, but they generally involve nonlinear integral transformations that are not analytically tractable. In this section, we obtain practical recursive algorithms by restricting the approximate joint smoothing posterior to the Gaussian family and the Markov conditionals to the Gauss–Markov family. Under this restriction, the tilted updates can be represented through finite-dimensional parameters, and the potential functions can be propagated by closed-form recursions.

Assumption 3 (Forward Gauss–Markov approximation) *Given a state-space model* (1) *and the forward-Markov factorization in Assumption 1, we restrict the approximate posterior* $\overrightarrow{q}(x_{0:T} \mid y_{1:T})$ *to the family of forward Gauss–Markov densities. Specifically, we use the parameterization*

$$\overrightarrow{q}(x_{0:T} \mid y_{1:T}) = \mathcal{N}(x_0 \mid \overrightarrow{m}_0, \overrightarrow{P}_0) \prod_{k=0}^{T-1} \mathcal{N}(x_{k+1} \mid \overrightarrow{F}_k\, x_k + \overrightarrow{d}_k, \overrightarrow{\Sigma}_k).$$

Assumption 4 (Reverse Gauss–Markov approximation) *Given a state-space model* (1) *and the reverse-Markov factorization in Assumption 2, we restrict the approximate posterior* $\overleftarrow{q}(x_{0:T} \mid y_{1:T})$ *to the family of reverse Gauss–Markov densities. Specifically, we use the parameterization*

$$\overleftarrow{q}(x_{0:T} \mid y_{1:T}) = \mathcal{N}(x_T \mid \overleftarrow{m}_T, \overleftarrow{P}_T) \prod_{k=1}^{T} \mathcal{N}(x_{k-1} \mid \overleftarrow{F}_k\, x_k + \overleftarrow{d}_k, \overleftarrow{\Sigma}_k).$$

In the following sections, we develop tractable recursive inference schemes by constructing local quadratic approximations to the potential functions and deriving update rules that preserve the Gauss–Markov structure of the approximate posterior across iteration.

5.1 Statistical Function Approximations

A key step in constructing efficient smoothing algorithms is to approximate the potential functions in (9) and (14) with tractable forms that preserve the recursive structure and admit closed-form updates. We achieve this by introducing second-order *statistical expansions* of the log-density functions associated with the latent dynamics $\log f_k(x_{k+1} \mid x_k)$ and the measurement model $\log h_k(y_k \mid x_k)$.

Given the iterative nature of the proposed optimization procedure (5), it is natural to construct these expansions locally around the current iterate $q^{[i]}(x_{0:T})$. This ensures that the approximation is tailored to the current posterior belief and captures relevant structure in the vicinity of the current iterate. Crucially, the KL-based trust-region constraint used in the entropic proximal updates controls the step size of each iteration. It prevents the distribution $q^{[i+1]}$ from deviating too far from $q^{[i]}$, thereby keeping the updates within the region where the local expansions are intended to be accurate. This interaction between local approximations and bounded updates stabilizes the optimization and maintains the fidelity of the recursive inference scheme (Teboulle, 1992; Iusem et al., 1994; Chrétien & Hero, 2002).

Definition 1 (Statistical second-order expansion) *Let* $z \sim \mathcal{N}(m, P)$ *and let* $g(z)$ *be a twice-differentiable scalar function. A second-order statistical expansion of* $g(z)$ *with respect to the random variable* $z \in \mathbb{R}^{d_z}$ *takes a quadratic form* $g(z) \approx -\frac{1}{2}\, z^\top U\, z + z^\top u + \eta$*, where* $U \in \mathbb{R}^{d_z \times d_z}$*,* $u \in \mathbb{R}^{d_z}$*, and* η *is a scalar.*

Definition 1 specifies the structure but not the computation of the expansion parameters. We return to this in Sections 5.2.1 and 5.2.2, where we describe two approximation strategies for computing (U, u, η).

Definition 2 (Quadratic expansion of log-densities) *Let* $(x_{k+1}, x_k) \sim q_k^{[i]}(x_{k+1}, x_k)$*. The statistical expansion of* $\ell_f^{[i]}(x_{k+1}, x_k) \approx \log f_k(x_{k+1} \mid x_k)$*, for all* $0 \le k < T$*, is defined as*

$$\ell_f^{[i]}(x_{k+1}, x_k) = -\frac{1}{2} \begin{bmatrix} x_{k+1}^\top & x_k^\top \end{bmatrix} \begin{bmatrix} C_{\bar{x}\bar{x},k}^{[i]} & -C_{\bar{x}x,k}^{[i]} \\ -C_{x\bar{x},k}^{[i]} & C_{xx,k}^{[i]} \end{bmatrix} \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} + \begin{bmatrix} x_{k+1}^\top & x_k^\top \end{bmatrix} \begin{bmatrix} c_{\bar{x},k}^{[i]} \\ c_{x,k}^{[i]} \end{bmatrix} + \kappa_k^{[i]}, \tag{16}$$

where $C_{\bar{x}\bar{x},k}^{[i]}, C_{xx,k}^{[i]} \in \mathbb{R}^{d \times d}$*,* $C_{\bar{x}x,k}^{[i]}, C_{x\bar{x},k}^{[i]} \in \mathbb{R}^{d \times d}$*,* $c_{\bar{x},k}^{[i]}, c_{x,k}^{[i]} \in \mathbb{R}^d$*, and* $\kappa_k^{[i]} \in \mathbb{R}$*. Further, for each* $0 < k \le T$*, the statistical expansion of* $\ell_h^{[i]}(x_k) \approx \log h_k(y_k \mid x_k)$ *is defined as*

$$\ell_h^{[i]}(x_k) = -\frac{1}{2} x_k^\top L_k^{[i]}\, x_k + x_k^\top l_k^{[i]} + \nu_k^{[i]},$$

where $L_k^{[i]} \in \mathbb{R}^{d\times d}$, $l_k^{[i]} \in \mathbb{R}^d$, *and* $\nu_k^{[i]} \in \mathbb{R}$. *Finally, for* $x_0 \sim q_0^{[i]}(x_0)$, *the statistical expansion of* $\ell_p^{[i]}(x_0) \approx \log p_0(x_0)$ *is defined as*

$$\ell_p^{[i]}(x_0) = -\frac{1}{2} x_0^\top L_0^{[i]}\, x_0 + x_0^\top l_0^{[i]} + \nu_0^{[i]},$$

where $L_0^{[i]} \in \mathbb{R}^{d\times d}$, $l_0^{[i]} \in \mathbb{R}^d$, *and* $\nu_0^{[i]} \in \mathbb{R}$.

5.2 Recursive Quadratic Potentials

In this section, we derive tractable schemes for computing the potential functions defined in (9) and (14), leveraging the statistical approximations introduced in Definition 2. We show that these approximations lead to tractable recursions over quadratic forms of $\overrightarrow{V}_k(x_k)$ and $\overleftarrow{V}_k(x_k)$ from Proposition 2 and Proposition 3. The use of quadratic potentials is consistent with the interpretations in Remark 3 and Remark 5: the forward-Markov potentials act as log-space backward messages, whereas the reverse-Markov potentials act as log-space forward messages. Under the Gaussian approximation, such messages are naturally parameterized by quadratic functions, which motivates the form adopted in our smoothing framework.

Before introducing these recursions, we first define the following parametric forms for the potential functions and the associated log-normalizing functions. Let

$$V_k^{[i+1]}(x_k) = -\frac{1}{2} x_k^\top R_k^{[i+1]}\, x_k + x_k^\top r_k^{[i+1]} + \rho_k^{[i+1]}, \tag{17}$$

where $R_k^{[i+1]} \in \mathbb{R}^{d\times d}$, $r_k^{[i+1]} \in \mathbb{R}^d$, and $\rho_k^{[i+1]} \in \mathbb{R}$. Similarly, we define

$$\log \psi_k^{[i+1]}(x_k) = -\frac{1}{2} x_k^\top S_k^{[i+1]}\, x_k + x_k^\top s_k^{[i+1]} + \xi_k^{[i+1]}, \tag{18}$$

where $S_k^{[i+1]} \in \mathbb{R}^{d\times d}$, $s_k^{[i+1]} \in \mathbb{R}^d$, and $\xi_k^{[i+1]} \in \mathbb{R}$. These quadratic parameterizations enable efficient message-passing recursions while preserving the Gaussian structure of the approximate posterior.

Proposition 4 (Recursive forward Gauss–Markov potentials) *Let* $\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)$ *be a forward Gauss–Markov conditional as defined in Assumption 3, and let* $\ell_f^{[i]}(x_{k+1}, x_k)$, $\ell_h^{[i]}(x_k)$, *and* $\ell_p^{[i]}(x_0)$ *be the second-order approximations of the log-densities from Definition 2. Then, the potentials* $\overrightarrow{V}_k^{[i+1]}(x_k)$ *in* (9) *are quadratic functions of the form* (17). *The recursion is initialized at* $k = T$ *by*

$$\overrightarrow{R}_T^{[i+1]} = L_T^{[i]}, \qquad \overrightarrow{r}_T^{[i+1]} = l_T^{[i]}.$$

For $0 \le k < T$, *the backward recursion is given by*

$$\begin{aligned}
\overrightarrow{R}_k^{[i+1]} &= L_k^{[i]} + \frac{1}{1-\beta}\, \overrightarrow{S}_k^{[i+1]}, \\
\overrightarrow{r}_k^{[i+1]} &= l_k^{[i]} + \frac{1}{1-\beta}\, \overrightarrow{s}_k^{[i+1]},
\end{aligned}$$

where the log-normalizing function $\log \overrightarrow{\psi}_k^{[i+1]}(x_k)$ *has quadratic form* (18). *Its parameters are*

$$\begin{aligned}
\overrightarrow{S}_k^{[i+1]} &= \overrightarrow{G}_{xx,k}^{[i+1]} - \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{G}_{\bar{x}x,k}^{[i+1]}, \\
\overrightarrow{s}_k^{[i+1]} &= \overrightarrow{g}_{x,k}^{[i+1]} + \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{g}_{\bar{x},k}^{[i+1]},
\end{aligned} \tag{19}$$

with intermediate quantities

$$
\begin{aligned}
\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} &\coloneqq (1-\beta)\left[C_{\bar{x}\bar{x},k}^{[i]} + \overrightarrow{R}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}, \\
\overrightarrow{G}_{xx,k}^{[i+1]} &\coloneqq (1-\beta)\, C_{xx,k}^{[i]} + \beta\left[\overrightarrow{F}_k^{[i]}\right]^{\top}\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\overrightarrow{F}_k^{[i]}, \\
\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} &\coloneqq (1-\beta)\, C_{\bar{x}x,k}^{[i]} + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\overrightarrow{F}_k^{[i]}, \\
\overrightarrow{g}_{\bar{x},k}^{[i+1]} &\coloneqq (1-\beta)\left[c_{\bar{x},k}^{[i]} + \overrightarrow{r}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\overrightarrow{d}_k^{[i]}, \\
\overrightarrow{g}_{x,k}^{[i+1]} &\coloneqq (1-\beta)\, c_{x,k}^{[i]} - \beta\left[\overrightarrow{F}_k^{[i]}\right]^{\top}\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\overrightarrow{d}_k^{[i]}.
\end{aligned}
\tag{20}
$$

Finally, given the boundary potential $\overrightarrow{V}_0^{[i+1]}(x_0)$ *and the Gaussian density* $\overrightarrow{q}_0^{[i]}(x_0) = \mathcal{N}(x_0 \mid \overrightarrow{m}_0^{[i]}, \overrightarrow{P}_0^{[i]})$*, the log-normalizing constant* $\log \overrightarrow{\mathcal{Z}}_0^{[i+1]}$ *in* (8) *is computed according to*

$$
\log \overrightarrow{\mathcal{Z}}_0^{[i+1]} = -\frac{1}{2}\left[\overrightarrow{m}_0^{[i]}\right]^{\top}\overrightarrow{U}^{[i+1]}\overrightarrow{m}_0^{[i]} + \left[\overrightarrow{m}_0^{[i]}\right]^{\top}\overrightarrow{u}^{[i+1]} + \overrightarrow{\eta}^{[i+1]},
\tag{21}
$$

where

$$
\begin{aligned}
\overrightarrow{U}^{[i+1]} &= \beta\,(1-\beta)\left[\overrightarrow{P}_0^{[i]}\right]^{-1}\left[(1-\beta)\,\overrightarrow{R}_0^{[i+1]} + \beta\left[\overrightarrow{P}_0^{[i]}\right]^{-1}\right]^{-1}\overrightarrow{R}_0^{[i+1]}, \\
\overrightarrow{u}^{[i+1]} &= -\beta\,(1-\beta)\left[\overrightarrow{P}_0^{[i]}\right]^{-1}\left[(1-\beta)\,\overrightarrow{R}_0^{[i+1]} + \beta\left[\overrightarrow{P}_0^{[i]}\right]^{-1}\right]^{-1}\overrightarrow{r}_0^{[i+1]}.
\end{aligned}
$$

Proof. See Appendix D.

Proposition 5 (Recursive reverse Gauss–Markov potentials) *Let* $\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)$ *be a reverse Gauss–Markov conditional as defined in Assumption 4 and let* $\ell_f^{[i]}(x_k, x_{k-1})$*,* $\ell_h^{[i]}(x_k)$*, and* $\ell_p^{[i]}(x_0)$ *be the second-order approximations of the log-densities in Definition 2, then the potentials* $\overleftarrow{V}_k^{[i+1]}(x_k)$ *from* (14) *are quadratic functions of the form* (17)*, computed recursively forwards starting from* $k = 0$ *with*

$$
\overleftarrow{R}_0^{[i+1]} = L_0^{[i]}, \qquad \overleftarrow{r}_0^{[i+1]} = l_0^{[i]},
$$

For $0 < k \leq T$*, the updates follow:*

$$
\begin{aligned}
\overleftarrow{R}_k^{[i+1]} &= L_k^{[i]} + \frac{1}{1-\beta}\,\overleftarrow{S}_k^{[i+1]}, \\
\overleftarrow{r}_k^{[i+1]} &= l_k^{[i]} + \frac{1}{1-\beta}\,\overleftarrow{s}_k^{[i+1]},
\end{aligned}
$$

where the log-normalizing function $\log \overleftarrow{\psi}_k^{[i+1]}(x_k)$ *has quadratic form* (18)*. Its parameters are*

$$
\begin{aligned}
\overleftarrow{S}_k^{[i+1]} &= \overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} - \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{\top}\left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}\overleftarrow{G}_{x\bar{x},k}^{[i+1]}, \\
\overleftarrow{s}_k^{[i+1]} &= \overleftarrow{g}_{\bar{x},k}^{[i+1]} + \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{\top}\left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}\overleftarrow{g}_{x,k}^{[i+1]},
\end{aligned}
\tag{22}
$$

with intermediate quantities

$$
\begin{aligned}
\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} &\coloneqq (1-\beta)\, C_{\bar{x}\bar{x},k-1}^{[i]} + \beta\left[\overleftarrow{F}_k^{[i]}\right]^{\top}\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\overleftarrow{F}_k^{[i]}, \\
\overleftarrow{G}_{xx,k}^{[i+1]} &\coloneqq (1-\beta)\left[C_{xx,k-1}^{[i]} + \overleftarrow{R}_{k-1}^{[i+1]}\right] + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}, \\
\overleftarrow{G}_{x\bar{x},k}^{[i+1]} &\coloneqq (1-\beta)\, C_{x\bar{x},k-1}^{[i]} + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\overleftarrow{F}_k^{[i]}, \\
\overleftarrow{g}_{\bar{x},k}^{[i+1]} &\coloneqq (1-\beta)\, c_{\bar{x},k-1}^{[i]} - \beta\left[\overleftarrow{F}_k^{[i]}\right]^{\top}\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\overleftarrow{d}_k^{[i]}, \\
\overleftarrow{g}_{x,k}^{[i+1]} &\coloneqq (1-\beta)\left[c_{x,k-1}^{[i]} + \overleftarrow{r}_{k-1}^{[i+1]}\right] + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\overleftarrow{d}_k^{[i]}.
\end{aligned}
\tag{23}
$$

Finally, given a quadratic potential function $\overleftarrow{V}_T^{[i+1]}(x_T)$ *of the form* (17) *and Gaussian posterior density* $\overleftarrow{q}_T^{[i]}(x_T) = \mathcal{N}(\overleftarrow{m}_T^{[i]}, \overleftarrow{P}_T^{[i]})$*, the log-normalizing constant* $\log \overleftarrow{\mathcal{Z}}_T^{[i+1]}$ *is computed according to* (13)

$$\log \overleftarrow{\mathcal{Z}}_T^{[i+1]} = -\frac{1}{2} \left[\overleftarrow{m}_T^{[i]}\right]^\top \overleftarrow{U}^{[i+1]} \overleftarrow{m}_T^{[i]} + \left[\overleftarrow{m}_T^{[i]}\right]^\top \overleftarrow{u}^{[i+1]} + \overleftarrow{\eta}^{[i+1]}, \tag{24}$$

where

$$\overleftarrow{U}^{[i+1]} = \beta\,(1-\beta) \left[\overleftarrow{P}_T^{[i]}\right]^{-1} \left[(1-\beta)\,\overleftarrow{R}_T^{[i+1]} + \beta \left[\overleftarrow{P}_T^{[i]}\right]^{-1}\right]^{-1} \overleftarrow{R}_T^{[i+1]},$$

$$\overleftarrow{u}^{[i+1]} = -\beta\,(1-\beta) \left[\overleftarrow{P}_T^{[i]}\right]^{-1} \left[(1-\beta)\,\overleftarrow{R}_T^{[i+1]} + \beta \left[\overleftarrow{P}_T^{[i]}\right]^{-1}\right]^{-1} \overleftarrow{r}_T^{[i+1]}.$$

Proof. See Appendix E.

The results in Proposition 4 and Proposition 5 rely on general second-order statistical expansions of the log-transition and log-likelihood functions. We now turn to the practical question of how these expansions can be computed, and present two distinct strategies for obtaining the required quadratic coefficients.

5.2.1 Generalized Statistical Linear Regression

Our first approach to computing the second-order statistical expansions is based on generalized statistical linear regression (GSLR, García-Fernández et al., 2015; Tronarp et al., 2018). These *enabling approximations* (Särkkä & Svensson, 2023) construct affine-Gaussian surrogate models for the transition density $f_k(x_{k+1} \mid x_k)$ and the measurement likelihood $h_k(y_k \mid x_k)$ by matching the relevant conditional moments under the current Gaussian iterate. These surrogates directly provide the coefficients required by Definition 2 for the recursive computation of the potentials and log-normalizing functions.

Definition 3 (Generalized statistical linear regression) *Let x and y be two random variables, and suppose that the first and second moments $\mathbb{E}[x]$, $\mathbb{V}[x]$, $\mathbb{E}[y]$, $\mathbb{V}[y]$, and $\mathbb{C}[y,x]$ are known. Statistical linear regression approximates the relation between x and y by $y \approx A\,x + b + \omega$, with $\omega \sim \mathcal{N}(0,\Omega)$, where*

$$A = \mathbb{C}[y,x]\,\mathbb{V}[x]^{-1}, \quad b = \mathbb{E}[y] - A\,\mathbb{E}[x], \quad \Omega = \mathbb{V}[y] - A\,\mathbb{V}[x]\,A^\top.$$

The parameters $\{A, b\}$ minimize the mean squared error objective $\mathrm{MSE}(A,b) = \mathbb{E}\big[(y - A\,x - b)^\top (y - A\,x - b)\big]$, and the covariance matrix satisfies $\Omega = \mathbb{V}[y - A\,x - b]$. When only the conditional moments $\mathbb{E}\left[y \mid x\right]$ and $\mathbb{V}\left[y \mid x\right]$ are available, generalized statistical linear regression estimates the required marginal moments via the law of total expectation and the law of total variance:

$$\mathbb{E}[y] = \mathbb{E}\Big[\mathbb{E}\left[y \mid x\right]\Big], \quad \mathbb{V}[y] = \mathbb{E}\Big[\mathbb{V}\left[y \mid x\right]\Big] + \mathbb{V}\Big[\mathbb{E}\left[y \mid x\right]\Big], \quad \mathbb{C}[y,x] = \mathbb{C}\Big[\mathbb{E}\left[y \mid x\right], x\Big],$$

where the outer expectations are evaluated with respect to a Gaussian marginal distribution $q(x)$ using numerical quadrature integration rules (Särkkä & Svensson, 2023).

Based on the recipe from Definition 3, and given the conditional moments $\mathbb{E}\left[x_{k+1} \mid x_k\right]$, $\mathbb{V}\left[x_{k+1} \mid x_k\right]$, $\mathbb{E}\left[y_k \mid x_k\right]$, and $\mathbb{V}\left[y_k \mid x_k\right]$, we can replace the conditional densities $f_k(x_{k+1} \mid x_k)$ and $h_k(y_k \mid x_k)$, at each iteration $[i+1]$, by their affine-Gaussian approximations

$$\begin{aligned} x_{k+1} &\approx A_k^{[i]}\,x_k + b_k^{[i]} + \omega_k^{[i]}, \qquad \omega_k^{[i]} \sim \mathcal{N}(0, \Omega_k^{[i]}), \qquad 0 \le k < T, \\ y_k &\approx H_k^{[i]}\,x_k + e_k^{[i]} + \delta_k^{[i]}, \qquad \delta_k^{[i]} \sim \mathcal{N}(0, \Delta_k^{[i]}), \qquad 0 < k \le T, \end{aligned} \tag{25}$$

which corresponds to generalized statistical linear regression with respect to the Gaussian marginal distributions $q_k^{[i]}(x_k)$ from iteration $[i]$

$$\begin{aligned} A_k^{[i]} &= \mathbb{C}^{[i]}[x_{k+1}, x_k]\,\mathbb{V}^{[i]}[x_k]^{-1}, \\ b_k^{[i]} &= \mathbb{E}^{[i]}[x_{k+1}] - A_k^{[i]}\,\mathbb{E}^{[i]}[x_k], \\ \Omega_k^{[i]} &= \mathbb{V}^{[i]}[x_{k+1}] - A_k^{[i]}\,\mathbb{V}^{[i]}[x_k]\left[A_k^{[i]}\right]^\top, \end{aligned} \quad \Bigg| \quad \begin{aligned} H_k^{[i]} &= \mathbb{C}^{[i]}[y_k, x_k]\,\mathbb{V}^{[i]}[x_k]^{-1}, \\ e_k^{[i]} &= \mathbb{E}^{[i]}[y_k] - H_k^{[i]}\,\mathbb{E}^{[i]}[x_k], \\ \Delta_k^{[i]} &= \mathbb{V}^{[i]}[y_k] - H_k^{[i]}\,\mathbb{V}^{[i]}[x_k]\left[H_k^{[i]}\right]^\top. \end{aligned}$$

Finally, the prior density $p_0(x_0)$ is likewise approximated by a Gaussian density by matching the first and second moments

$$p_0(x_0) \approx \mathcal{N}(\mu_0, \Lambda_0), \quad \text{with} \quad \mu_0 = \mathbb{E}\left[x_0\right], \quad \Lambda_0 = \mathbb{V}\left[x_0\right]. \tag{26}$$

Given the GSLR approximations (25) and (26), we obtain the quadratic log-density expansions required by Definition 2, enabling closed-form computation in Proposition 4 and Proposition 5. Specifically, for all $0 \leq k < T$, the resulting coefficients for $\ell_f^{[i]}(x_{k+1}, x_k)$ are

$$\begin{aligned}
C_{\bar{x}\bar{x},k}^{[i]} &= \left[\Omega_k^{[i]}\right]^{-1}, & C_{\bar{x}x,k}^{[i]} &= \left[\Omega_k^{[i]}\right]^{-1} A_k^{[i]}, & c_{\bar{x},k}^{[i]} &= \left[\Omega_k^{[i]}\right]^{-1} b_k^{[i]}, \\
C_{xx,k}^{[i]} &= \left[A_k^{[i]}\right]^\top \left[\Omega_k^{[i]}\right]^{-1} A_k^{[i]}, & C_{x\bar{x},k}^{[i]} &= \left[A_k^{[i]}\right]^\top \left[\Omega_k^{[i]}\right]^{-1}, & c_{x,k}^{[i]} &= -\left[A_k^{[i]}\right]^\top \left[\Omega_k^{[i]}\right]^{-1} b_k^{[i]}.
\end{aligned}$$

For all $0 < k \leq T$, the coefficients of $\ell_h^{[i]}(x_k)$ are

$$L_k^{[i]} = \left[H_k^{[i]}\right]^\top \left[\Delta_k^{[i]}\right]^{-1} H_k^{[i]}, \qquad l_k^{[i]} = \left[H_k^{[i]}\right]^\top \left[\Delta_k^{[i]}\right]^{-1} \left[y_k - e_k^{[i]}\right],$$

and, finally, the coefficients associated with $\ell_p^{[i]}(x_0)$ are

$$L_0^{[i]} = \Lambda_0^{-1}, \qquad l_0^{[i]} = \Lambda_0^{-1}\mu_0.$$

5.2.2 Fourier–Hermite Series Expansion

The statistical approximation provided by statistical linear regression comes with two main limitations: it imposes an explicit additive Gaussian noise model on the approximated dynamics and measurements, and neglects second-order curvature information. Alternatively, we can obtain second-order statistical approximations using Fourier–Hermite series (Sarmavuori & Särkkä, 2011; Hassan & Särkkä, 2023), which leverage Hermite polynomial bases in a Hilbert space $\mathcal{H}$ to capture higher-order effects (Janson, 1997).

Definition 4 (Fourier–Hermite expansion) *Let $g \in \mathcal{H}$ be a scalar-valued function and let $s \sim \mathcal{N}(0, I)$ be a standard Gaussian random variable. A second-order Fourier–Hermite expansion of $g(s)$ is given by*

$$g(s) \approx \mathbb{E}\left[g(s)\right] + \mathbb{E}\left[g(s)\, H_1(s)\right]^\top H_1(s) + \frac{1}{2}\operatorname{tr}\left\{\mathbb{E}\left[g(s)\, H_2(s)\right] H_2(s)\right\},$$

where $H_1(s)$ and $H_2(s)$ are the first- and second-order Hermite polynomials defined as

$$H_1(s) = s, \qquad H_2(s) = ss^\top - I.$$

This expansion generalizes to any Gaussian $\mathcal{N}(m, P)$ by letting $z = Rs + m$ and $P = RR^\top$, so that

$$\begin{aligned}
g(z) &\approx \mathbb{E}\left[g(z)\right] + \mathbb{E}\left[g(z)\, H_1(R^{-1}(z-m))\right]^\top H_1(R^{-1}(z-m)) \\
&\quad + \frac{1}{2}\operatorname{tr}\left\{\mathbb{E}\left[g(z)\, H_2(R^{-1}(z-m))\right]\left[R^{-1}(z-m)(z-m)^\top R^{-\top} - I\right]\right\} \\
&= -\frac{1}{2} z^\top U z + z^\top u + \eta.
\end{aligned}$$

The coefficients of this quadratic form are

$$\begin{aligned}
&U := -\mathbb{E}\left[\nabla_{z^2}\, g(z)\right], \quad u := \mathbb{E}\left[\nabla_z\, g(z)\right] - \mathbb{E}\left[\nabla_{z^2}\, g(z)\right] m, \\
&\eta := \mathbb{E}\left[g(z)\right] - \mathbb{E}\left[\nabla_z\, g(z)\right]^\top m + \frac{1}{2} m^\top \mathbb{E}\left[\nabla_{z^2}\, g(z)\right] m - \frac{1}{2}\operatorname{tr}\left\{R^\top \mathbb{E}\left[\nabla_{z^2}\, g(z)\right] R\right\}.
\end{aligned}$$

Here, $\nabla_z\, g(z)$ and $\nabla_{z^2}\, g(z)$ are the gradient and Hessian of $g(z)$, respectively, which, for $z \sim \mathcal{N}(m, P)$, satisfy the following identities via integration by parts (Hassan & Särkkä, 2023)

$$\mathbb{E}\left[g(z)\, H_1(R^{-1}(z-m))\right] = R^\top \mathbb{E}\left[\nabla_z\, g(z)\right], \qquad \mathbb{E}\left[g(z)\, H_2(R^{-1}(z-m))\right] = R^\top \mathbb{E}\left[\nabla_{z^2}\, g(z)\right] R.$$

The expectations involved in this approximation can be efficiently evaluated using numerical quadrature integration rules (Särkkä & Svensson, 2023).

Given a set of marginal distributions $q_k^{[i]}(x_k)$, we can apply the expansion in Definition 4 to the log-transition and log-measurement functions to obtain second-order approximations consistent with Definition 2. This approach provides an alternative to GSLR that does not impose an explicit additive Gaussian noise assumption and incorporates second-order information, offering potentially higher-fidelity approximations.

5.3 Gauss–Markov Posterior Updates

Having established the quadratic approximations that enable recursive computation of the potential functions, we now turn our attention to the computation of the tilted distributions introduced in Proposition 2 and Proposition 3. As specified in Assumption 3 and Assumption 4, our goal is to maintain the Gauss–Markov structure of the forward and reverse variational posteriors throughout the iterative optimization process. To ensure that the updated variational distributions remain within the Gauss–Markov family, we project the corresponding tilted distributions onto a marginal or conditional Gaussian form.

We begin by deriving a general moment-matching rule for tilted marginal distributions of the form in (6) and (12). These results then directly yield closed-form updates for the parameters $\overrightarrow{m}_0^{[i+1]}$, $\overrightarrow{P}_0^{[i+1]}$, $\overleftarrow{m}_T^{[i+1]}$, and $\overleftarrow{P}_T^{[i+1]}$, preserving the tractability and structure of the overall smoothing algorithms.

Lemma 1 (Tilted Gaussian moment matching) *Let $q^{[i+1]}(x)$ be a tilted distribution of the form*

$$q^{[i+1]}(x) = \left[\mathcal{Z}^{[i+1]}\right]^{-1} \left[q^{[i]}(x)\right]^{\beta} \left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta},$$

with $q^{[i]}(x) = \mathcal{N}(x \mid m^{[i]}, P^{[i]})$ and a normalizing constant

$$\mathcal{Z}^{[i+1]} = \int \left[q^{[i]}(x)\right]^{\beta} \left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta} \mathrm{d}x.$$

Then, the first and second moments of $q^{[i+1]}(x)$ are given by

$$\mathbb{E}^{[i+1]}[x] = m^{[i]} + \frac{1}{\beta} P^{[i]} \frac{\partial \log \mathcal{Z}^{[i+1]}}{\partial m^{[i]}}, \qquad \mathbb{V}^{[i+1]}[x] = \frac{1}{\beta} P^{[i]} + \frac{1}{\beta^2} P^{[i]} \frac{\partial^2 \log \mathcal{Z}^{[i+1]}}{\partial m^{[i]} \, \partial \left[m^{[i]}\right]^{\top}} P^{[i]}.$$

Proof. See Appendix F.

Corollary 2 (Forward Gauss–Markov boundary) *Let $\overrightarrow{q}_0^{[i]}(x_0)$ be a Gaussian marginal and let $\log \overrightarrow{\mathcal{Z}}_0^{[i+1]}$ be the log-normalizing constant of the tilted distribution in (21). Then, applying Lemma 1, the optimal Gaussian approximation to the tilted marginal $\overrightarrow{q}_0^{[i+1]}(x_0)$ from (6) is given by*

$$\overrightarrow{P}_0^{[i+1]} = \left[(1-\beta)\,\overrightarrow{R}_0^{[i+1]} + \beta\left[\overrightarrow{P}_0^{[i]}\right]^{-1}\right]^{-1},$$
$$\overrightarrow{m}_0^{[i+1]} = \overrightarrow{P}_0^{[i+1]}\left[(1-\beta)\,\overrightarrow{r}_0^{[i+1]} + \beta\left[\overrightarrow{P}_0^{[i]}\right]^{-1}\overrightarrow{m}_0^{[i]}\right].$$

Corollary 3 (Reverse Gauss–Markov boundary) *Let $\overleftarrow{q}_T^{[i]}(x_T)$ be a Gaussian marginal and let $\log \overleftarrow{\mathcal{Z}}_T^{[i+1]}$ be the log-normalizing constant of the tilted distribution in (24). Then, applying Lemma 1, the optimal Gaussian approximation to the tilted marginal $\overleftarrow{q}_T^{[i+1]}(x_T)$ from (12) is given by*

$$\overleftarrow{P}_T^{[i+1]} = \left[(1-\beta)\,\overleftarrow{R}_T^{[i+1]} + \beta\left[\overleftarrow{P}_T^{[i]}\right]^{-1}\right]^{-1},$$
$$\overleftarrow{m}_T^{[i+1]} = \overleftarrow{P}_T^{[i+1]}\left[(1-\beta)\,\overleftarrow{r}_T^{[i+1]} + \beta\left[\overleftarrow{P}_T^{[i]}\right]^{-1}\overleftarrow{m}_T^{[i]}\right].$$

Projecting the tilted conditionals in (7) and (11) onto the affine-Gaussian parametric family defined in Assumption 3 and Assumption 4 is more involved, as it requires deriving the conditional moments of the forward and reverse distributions $\mathbb{E}\left[x_{k+1} \mid x_k\right]$, $\mathbb{V}\left[x_{k+1} \mid x_k\right]$, $\mathbb{E}\left[x_{k-1} \mid x_k\right]$, and $\mathbb{V}\left[x_{k-1} \mid x_k\right]$. In what follows, we derive closed-form expressions for these moments and use them to update the parameters of the forward and reverse Gauss–Markov posteriors.

Proposition 6 (Forward Gauss–Markov conditionals) *Let $\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)$ be the forward Gauss–Markov conditional in Assumption 3. Additionally, suppose that $\ell_f^{[i]}(x_{k+1}, x_k)$ admits a second-order expansion as in (16), and that the log-message $\log \overrightarrow{\psi}_k^{[i+1]}(x_k)$ is quadratic as in (19). Then the tilted conditional $\overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k)$ is linear-Gaussian with conditional moments*

$$\mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}\left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} x_k + \overrightarrow{g}_{\bar{x},k}^{[i+1]}\right], \qquad \mathbb{V}^{[i+1]}\left[x_{k+1} \mid x_k\right] = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1},$$

and Gauss–Markov parameters

$$\overrightarrow{\Sigma}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}, \qquad \overrightarrow{F}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}, \qquad \overrightarrow{d}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}\overrightarrow{g}_{\bar{x},k}^{[i+1]}.$$

Substituting the parameters from (20) gives the explicit update

$$\begin{aligned}
\overrightarrow{\Sigma}_k^{[i+1]} &= \left[(1-\beta)\left[C_{\bar{x}\bar{x},k}^{[i]} + \overrightarrow{R}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\right]^{-1}, \\
\overrightarrow{F}_k^{[i+1]} &= \overrightarrow{\Sigma}_k^{[i+1]}\left[(1-\beta)\, C_{\bar{x}x,k}^{[i]} + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\overrightarrow{F}_k^{[i]}\right], \\
\overrightarrow{d}_k^{[i+1]} &= \overrightarrow{\Sigma}_k^{[i+1]}\left[(1-\beta)\left[c_{\bar{x},k}^{[i]} + \overrightarrow{r}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\overrightarrow{d}_k^{[i]}\right].
\end{aligned}$$

Proof. See Appendix G.

Proposition 7 (Reverse Gauss–Markov conditionals) *Let $\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)$ be the reverse Gauss–Markov conditional in Assumption 4. Additionally, suppose that $\ell_f^{[i]}(x_k, x_{k-1})$ admits a second-order expansion as in (16), and that the log-message $\log \overleftarrow{\psi}_k^{[i+1]}(x_k)$ is quadratic as in (22). Then the tilted conditional $\overleftarrow{q}_k^{[i+1]}(x_{k-1} \mid x_k)$ is linear-Gaussian with conditional moments*

$$\mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}\left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]} x_k + \overleftarrow{g}_{x,k}^{[i+1]}\right], \qquad \mathbb{V}^{[i+1]}\left[x_{k-1} \mid x_k\right] = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}.$$

and Gauss–Markov parameters

$$\overleftarrow{\Sigma}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}, \qquad \overleftarrow{F}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}\overleftarrow{G}_{x\bar{x},k}^{[i+1]}, \qquad \overleftarrow{d}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}\overleftarrow{g}_{x,k}^{[i+1]}.$$

Substituting the parameters from (23) gives the explicit update

$$\begin{aligned}
\overleftarrow{\Sigma}_k^{[i+1]} &= \left[(1-\beta)\left[C_{xx,k-1}^{[i]} + \overleftarrow{R}_{k-1}^{[i+1]}\right] + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\right]^{-1}, \\
\overleftarrow{F}_k^{[i+1]} &= \overleftarrow{\Sigma}_k^{[i+1]}\left[(1-\beta)\, C_{x\bar{x},k-1}^{[i]} + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\overleftarrow{F}_k^{[i]}\right], \\
\overleftarrow{d}_k^{[i+1]} &= \overleftarrow{\Sigma}_k^{[i+1]}\left[(1-\beta)\left[c_{x,k-1}^{[i]} + \overleftarrow{r}_{k-1}^{[i+1]}\right] + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\overleftarrow{d}_k^{[i]}\right].
\end{aligned}$$

Proof. See Appendix H.

5.4 Recursive Gaussian Marginals

Given the Gauss–Markov parameterization of the posterior in Assumption 3 and Assumption 4, together with the update rules introduced in Section 5.3, the smoothing posterior marginals at each iteration $[i+1]$ can be computed efficiently. Specifically, we use a forward recursion to evaluate the marginals of the forward Gauss–Markov posterior as described in Remark 4, and a backward recursion to evaluate the marginals of the reverse Gauss–Markov posterior as described in Remark 6. These marginals play a central role in the proposed iterative smoothing framework, since they are used to construct the second-order statistical expansions of the log-densities associated with the dynamics and measurement model, as discussed in Section 5.2.

Corollary 4 (Forward Gauss–Markov marginals) *Given the forward Gauss–Markov posterior from Corollary 2 and Proposition 6, the forward marginal smoothing distributions $\overrightarrow{q}_{k+1}^{[i+1]}(x_{k+1})$, for all $0 \leq k < T$, are computed in closed form via the forward recursion*

$$\begin{aligned}
\overrightarrow{m}_{k+1}^{[i+1]} &= \overrightarrow{F}_k^{[i+1]} \overrightarrow{m}_k^{[i+1]} + \overrightarrow{d}_k^{[i+1]}, \\
\overrightarrow{P}_{k+1}^{[i+1]} &= \overrightarrow{F}_k^{[i+1]} \overrightarrow{P}_k^{[i+1]} \left[\overrightarrow{F}_k^{[i+1]}\right]^\top + \overrightarrow{\Sigma}_k^{[i+1]},
\end{aligned}$$

with the initial condition $\overrightarrow{q}_0^{[i+1]}(x_0) = \mathcal{N}(x_0 \mid \overrightarrow{m}_0^{[i+1]}, \overrightarrow{P}_0^{[i+1]})$.

Corollary 5 (Reverse Gauss–Markov marginals) *Given the reverse Gauss–Markov posterior from Corollary 3 and Proposition 7, the reverse marginal smoothing distributions $\overleftarrow{q}_{k-1}^{[i+1]}(x_{k-1})$, for all $1 \leq k \leq T$, are computed in closed form via the backward recursion*

$$\begin{aligned}
\overleftarrow{m}_{k-1}^{[i+1]} &= \overleftarrow{F}_k^{[i+1]} \overleftarrow{m}_k^{[i+1]} + \overleftarrow{d}_k^{[i+1]}, \\
\overleftarrow{P}_{k-1}^{[i+1]} &= \overleftarrow{F}_k^{[i+1]} \overleftarrow{P}_k^{[i+1]} \left[\overleftarrow{F}_k^{[i+1]}\right]^\top + \overleftarrow{\Sigma}_k^{[i+1]},
\end{aligned}$$

with the terminal condition $\overleftarrow{q}_T^{[i+1]}(x_T) = \mathcal{N}(x_T \mid \overleftarrow{m}_T^{[i+1]}, \overleftarrow{P}_T^{[i+1]})$.

Corollary 6 (Hybrid Gauss–Markov marginals) *Given the recursions from Proposition 4 and Proposition 5, suppose that the current marginal is Gaussian, $q_k^{[i]}(x_k) = \mathcal{N}(x_k \mid m_k^{[i]}, P_k^{[i]})$. Then, the interior hybrid marginal smoothing distributions $q_k^{[i+1]}(x_k)$, for all $0 < k < T$, are computed according to Corollary 1 in closed form by completing the square:*

$$\begin{aligned}
P_k^{[i+1]} &= \left[(1-\beta)\left[\overleftarrow{R}_k^{[i+1]} + \overrightarrow{S}_k^{[i+1]}\right] + \beta\left[P_k^{[i]}\right]^{-1}\right]^{-1}, \\
m_k^{[i+1]} &= P_k^{[i+1]}\left[(1-\beta)\left[\overleftarrow{r}_k^{[i+1]} + \overrightarrow{s}_k^{[i+1]}\right] + \beta\left[P_k^{[i]}\right]^{-1} m_k^{[i]}\right].
\end{aligned}$$

The boundary marginals are given by the corresponding forward and reverse boundary updates, namely $q_0^{[i+1]}(x_0) = \mathcal{N}(x_0 \mid \overrightarrow{m}_0^{[i+1]}, \overrightarrow{P}_0^{[i+1]})$ and $q_T^{[i+1]}(x_T) = \mathcal{N}(x_T \mid \overleftarrow{m}_T^{[i+1]}, \overleftarrow{P}_T^{[i+1]})$.

5.5 Optimal Damping Parameter

We now address the selection of the damping parameter β, which plays a critical role in the entropic proximal update. As described in (5), β is induced by the Lagrange multiplier associated with the Kullback–Leibler trust-region constraint and may vary across iterations. Proposition 2 and Proposition 3 show that the optimal value of β at each iteration is obtained by minimizing the corresponding dual objective associated with the forward or reverse factorization.

The dual problems defined in (10) and (15) are nonlinear in β and can be approached using standard numerical optimization techniques (Nocedal & Wright, 2006). However, in practice, off-the-shelf solvers may become unstable in highly nonlinear regimes. These difficulties typically arise when the quadratic approximations used in Proposition 4 and Proposition 5 fail to preserve the positive definiteness of the

precision matrices required by the Gaussian updates. In such cases, the feasible range of β may be restricted, and numerical optimization of the dual objective requires careful handling.

To circumvent these challenges, we adopt a simple and robust alternative. We apply a bisection method to the Lagrange multiplier α, which implicitly determines the damping parameter through the relation $\beta = \alpha/(1+\alpha)$. This approach searches for a root of the derivative of the Lagrangian evaluated at the corresponding proximal update $q^{[i+1]}(x_{0:T};\alpha)$. The root condition is given by

$$\frac{\partial \mathcal{R}(\alpha^\star)}{\partial \alpha^\star} = \varepsilon - \mathbb{D}_{\mathrm{KL}}\left[q^{[i+1]}(x_{0:T};\alpha^\star) \mid\mid q^{[i]}(x_{0:T})\right] = 0, \tag{27}$$

where $\mathcal{R}(\cdot)$ denotes the Lagrangian of (5), evaluated at the updated posterior $q^{[i+1]}(x_{0:T};\alpha)$:

$$\mathcal{R}(\alpha) = \mathbb{E}_{q_\alpha^{[i+1]}}\left[\log p(x_{0:T}, y_{1:T}) - \log q^{[i+1]}(x_{0:T};\alpha)\right] + \alpha\left[\varepsilon - \mathbb{D}_{\mathrm{KL}}\left[q^{[i+1]}(x_{0:T};\alpha) \mid\mid q^{[i]}(x_{0:T})\right]\right] + \text{const.}$$

Here, $q^{[i+1]}(x_{0:T};\alpha)$ denotes the updated variational posterior that depends implicitly on β, and thus on α. Derivations of the forward and reverse Lagrangian formulations are provided in the appendix.

An outline of the bisection procedure for determining the optimal damping parameter is presented in Algorithm 3 and Algorithm 6, corresponding to the forward and reverse Gauss–Markov smoothers. When the KL constraint is active, the root condition (27) places the optimal solution $q^{[i+1]}(x_{0:T})$ on the boundary of the KL-ball centered at $q^{[i]}(x_{0:T})$ with radius ε. If the unconstrained update already lies within this ball, the optimal multiplier is $\alpha = 0$, corresponding to $\beta = 0$. This induces an automatic damping adaptation mechanism that dynamically enforces the trust-region constraint across iterations.

5.6 Proximal Variational Bayesian Smoothing Algorithms

The preceding sections provide the components needed to instantiate the general results of Proposition 2, Proposition 3, and Corollary 1 for Gaussian variational posterior approximations. In this section, we present an overview of the resulting three iterative Bayesian smoothing algorithms.

The *Forward Entropic Variational Smoother (FEVS)*, described in Algorithm 7, iteratively refines the posterior estimate through a backward-forward recursion. First, a backward pass computes the potential functions, log-normalizing terms, and updated forward conditional smoothing distributions, as detailed in Algorithm 1. This is followed by a forward pass that propagates the smoothed marginal distributions, as outlined in Algorithm 2. At each iteration, the damping parameter is adapted using Algorithm 3, which enforces the proximal trust-region constraint induced by the entropic regularization.

The *Reverse Entropic Variational Smoother (REVS)*, presented in Algorithm 8, follows the inverse recursion order. A forward pass computes the potential functions, log-normalizing terms, and updated reverse conditional smoothing distributions, as given in Algorithm 4. This is followed by a backward pass that propagates the smoothed marginal distributions, according to Algorithm 5. As in the forward smoother, the damping parameter is adaptively selected at each iteration using Algorithm 6.

The *Hybrid Entropic Variational Smoother (HEVS)*, shown in Algorithm 10, combines the two directional message recursions. In this scheme, the potential functions and log-normalizing terms are computed using both the backward pass from the forward smoother in Algorithm 1 and the forward pass from the reverse smoother in Algorithm 4. The resulting representations are then combined at the marginal level to compute the smoothed marginal distributions, as described in Corollary 1. The damping parameter can be adapted using either the forward or reverse dual objective to enforce the proximal constraint.

All proposed methods share several favorable properties. First, the recursions are formulated through log-potentials and log-normalizing functions, which improves numerical stability and mitigates underflow over long time horizons. Second, each recursion is damped through a KL trust-region constraint, which controls the deviation between successive posterior iterates and reduces sensitivity to local approximation error. Finally, for fixed state dimension and fixed cost per local approximation step, all algorithms have linear time complexity with respect to the horizon T, making them scalable for long sequences.

6 Connection to Existing Bayesian Inference Algorithms

This section provides an overview of recent research that helps situate our contribution. We highlight related research in signal processing, Markovian Gaussian processes, and approximate Bayesian inference.

6.1 Forward-Backward Smoothing Algorithms

For the forward Gauss–Markov decomposition introduced in Proposition 2 and Proposition 4, a direct conceptual connection can be made to the classical smoother proposed by Cox (1964). In his work, Cox formulates Bayesian smoothing in state-space models with additive Gaussian noise as a maximum a posteriori (MAP) optimization problem. Subject to certain non-singularity conditions, he derives a dynamic programming solution that propagates adjoint state functions backward in time, which aligns closely with the backward recursion over potential functions used in our framework. This method was later extended to nonlinear systems via iterative linearization techniques (Mortensen, 1968). Our approach generalizes Cox's formulation by moving beyond MAP estimation to target a full Gaussian approximation of the smoothing posterior. In doing so, we accommodate a broader class of nonlinear, non-Gaussian state-space models. Moreover, by incorporating entropic regularization, our method introduces a principled mechanism for trust-region control, ensuring stable and well-posed updates over the space of variational densities.

In contrast, the reverse Gauss–Markov decomposition described in Proposition 3 and Proposition 5 yields a recursion that closely parallels the structure of the Rauch–Tung–Striebel (RTS) smoother (Rauch et al., 1965). The RTS smoother operates by first performing a forward filtering pass to accumulate measurement information, followed by a backward recursion that propagates smoothed marginals through the implicit computation of reverse posterior conditionals. Our iterated reverse Gauss–Markov smoother generalizes this two-pass structure to a significantly broader class of models. It lifts the standard assumptions of linear dynamics and Gaussian noise, enabling application to nonlinear, non-Gaussian systems. Additionally, by operating directly in the log-domain of the filtered densities and incorporating entropic proximal regularization, our method enhances numerical stability and provides a principled means of controlling the update step size throughout the recursion.

Finally, the hybrid posterior decomposition introduced in Corollary 1 and Corollary 6 bears a strong resemblance to log-space two-filter smoothers (Mayne, 1966; Fraser & Potter, 1969). Originally developed for MAP-based smoothing in linear-Gaussian state-space models, these methods were later extended to nonlinear settings through Gaussian-sum filters (Kitagawa, 1987). The canonical structure involves independent forward and backward filtering recursions that are subsequently combined to form the smoothing solution. Our hybrid smoother adopts this two-filter architecture by jointly leveraging both the forward and reverse Gauss–Markov decompositions of the variational posterior. In doing so, it generalizes the classical two-filter approach to accommodate nonlinear and non-Gaussian models within a variational inference framework. To the best of our knowledge, this is the first attempt to formulate an iterated two-filter smoother that integrates proximal updates and operates in the log-domain of the densities, akin to the forward and reverse variants. A notable advantage of this hybrid scheme is the structural independence of the forward and backward recursions, which allows for parallel execution. Although this design increases the overall computational cost relative to the single-pass algorithms, it offers practical benefits in terms of speed and modularity when implemented on modern parallel hardware.

6.2 Posterior-Linearization Bayesian Smoothing

Our work is primarily inspired by posterior-linearization algorithms (García-Fernández et al., 2015; 2016; Tronarp et al., 2018), which perform approximate Gaussian Bayesian smoothing in nonlinear state-space models by alternating between posterior linearization and a Rauch-Tung-Striebel smoothing pass. These algorithms can be viewed as generalizations of the maximum-a-posteriori iterated smoother proposed by Bell (1994). While these methods offer advantages over traditional extended and unscented smoothers (Särkkä & Svensson, 2023), they have two significant limitations.

First, they impose linearity and additive noise assumptions on the dynamics and measurement likelihoods, which can be problematic in state-space models with multiplicative noise, as highlighted by Corenflos &

Abdulsamad (2023). Second, these algorithms, in their original form, fall within the class of undamped Gauss–Newton optimization methods, that rely on linear approximations of the dynamics and measurement models. Crucially, Gauss–Newton techniques require a full-rank Jacobian to ensure convergence (Nocedal & Wright, 2006), a condition that is not generally met in nonlinear settings.

To address the latter issue, a partial remedy was proposed by Raitoharju et al. (2018), who introduced an *ad hoc* damping mechanism for the mean updates in the iterated posterior-linearization filter, albeit at the cost of a computationally expensive nested optimization loop. In contrast, Lindqvist et al. (2021) proposed a posterior-linearization smoother that implements damping through Levenberg–Marquardt regularization and line-search procedures with convergence guarantees. Both approaches of Raitoharju et al. (2018) and Lindqvist et al. (2021), however, fail to exploit the information-geometric structure of the statistical manifold on which the approximate smoothing distribution is assumed to reside.

Our approach generalizes these algorithms and resolves both weaknesses. By using Fourier–Hermite expansions, we overcome the limitations of linear-Gaussian approximations, while the introduction of entropic proximal constraints provides a principled way to damp optimization over the space of densities, leading to techniques akin to trust-region approaches (Nocedal & Wright, 2006; Teboulle, 1992). This improvement, however, comes at the cost of increased algorithmic complexity.

6.3 Approximate Bayesian Inference in State-Space Models

The field of approximate Bayesian inference has seen significant progress in adapting standard techniques to exploit the temporal structure of state-space models, resulting in specialized algorithms for Bayesian smoothing. One influential approach builds on expectation propagation (EP), introduced by Minka (2001). In particular, Deisenroth & Mohamed (2012) applied EP to nonlinear state-space models with additive Gaussian noise. Their method iteratively linearizes the dynamics and measurement models around the current posterior estimate, yielding tractable forward–backward message-passing recursions. However, the construction is largely heuristic and lacks a clearly defined global objective, which complicates convergence analysis, especially in nonlinear or non-Gaussian regimes. In contrast, our method is grounded in a well-posed variational optimization problem, enabling us to draw on a body of work from convex and information-theoretic optimization (Teboulle, 1992; Chrétien & Hero, 2002), and ensuring a more principled treatment of smoothing in complex models.

A parallel line of work has explored variational inference (VI) (Wainwright & Jordan, 2008) as a foundation for approximate smoothing. For instance, Courts et al. (2021) proposed a variational Gaussian smoother that optimizes the evidence lower bound using off-the-shelf constrained nonlinear solvers. Their formulation relies on a particular parameterization of twin-marginal distributions, subject to feasibility constraints. A key limitation, however, is that the resulting inference procedure is non-recursive, and optimization is carried out directly in the parameter space of the posterior, rather than over the statistical manifold where natural gradient methods can be more efficient (Amari, 1998). Related work by Campbell et al. (2021) developed an online variational method for filtering and parameter learning in state-space models. Their approach optimizes an evidence lower bound using backward decompositions of the joint posterior approximation together with Bellman-type recursions for the objective and its gradients. This provides another connection between variational inference and recursive state estimation, although their focus is online filtering and learning rather than proximal Bayesian smoothing. In concurrent work, Tronarp (2025) introduced a recursive variational framework that derives forward–backward algorithms similar to ours, albeit without entropic regularization. Their approach invokes dynamic programming directly to formulate variational state estimation through forward and backward value-function recursions. In contrast, our dynamic programming structure emerges from an entropic dynamic optimization problem over variational posterior distributions, where the recursive equations arise as optimality conditions of a proximal variational update. This distinction leads to damped updates and an adaptive step-size mechanism through the associated dual problem.

Our framework is closely related to the structured variational inference perspective of Barfoot et al. (2020), building on the work of Opper & Archambeau (2009). Like their natural gradient-based variational Gaussian smoother, our method falls within the class of information-theoretic variational optimizers identified by Khan et al. (2015). However, it differs in two important respects. First, our method is inherently recursive. By

casting inference as a dynamic optimization problem constrained by KL-based trust regions, our method naturally exploits the temporal structure of state-space models and leads to a family of forward-backward smoothing algorithms with linear computational complexity with respect to the time horizon. Second, it is agnostic to the parameterization of the variational posterior, allowing for flexible approximation families beyond a fixed Gaussian representation.

6.4 Approximate Bayesian Inference in Temporal Gaussian Processes

There is a well-established connection between *temporal* Gaussian processes (GP) and *linear* state-space models (O'Hagan, 1978). Hartikainen & Särkkä (2010) showed that temporal GP models with Matérn kernels can be reformulated exactly as linear-Gaussian SSMs. This enables using Kalman filtering and smoothing with linear time complexity to perform inference in temporal GPs, a significant improvement over the cubic complexity in standard GP inference.

Building on this foundation, Chang et al. (2020) proposed a variational method that combines conjugate-computation variational inference (Khan & Lin, 2017) with Rauch–Tung–Striebel recursions, enabling principled inference in models with linear dynamics and nonlinear measurement models. Wilkinson et al. (2020) extended this framework by unifying variational inference and expectation propagation within Kalman smoothing. Wilkinson et al. (2023) further generalized these approaches via the Bayes–Newton framework, which interprets VI, EP, and posterior linearization as optimization algorithms akin to Newton's method. Despite these advances, a common critique of this class of algorithms is their *ad hoc* treatment of the temporal dynamics. While inference is formulated from an optimization perspective, the temporal structure is handled implicitly using linear RTS smoothing, rather than being explicitly integrated into the inference objective.

In the continuous-time setting, several variational methods have been developed. Archambeau et al. (2007) proposed one of the earliest approaches for inference in partially-observed diffusion processes with linear dynamics by approximating the posterior with a linear stochastic differential equation. This approximation can be viewed as the continuous-time limit of a Gauss–Markov factorization. Ala-Luhtala et al. (2015) extended this line of work to nonlinear dynamics through sigma-point approximations. More recently, Wildner & Koeppl (2021) and Verma et al. (2024) proposed natural-gradient-based generalizations of the method introduced by Archambeau et al. (2007). Wildner & Koeppl (2021) developed a moment-based recursive inference algorithm that reformulates the smoothing problem as an optimal control problem, thereby explicitly incorporating the dynamic structure of the setting. In contrast, Verma et al. (2024) adopts a site-based approach inspired by Minka (2001), introducing a non-recursive algorithm that leverages iterative posterior linearization to handle nonlinear dynamics. recently, Bartosh et al. (2025) proposed a continuous-time simulation-free approach based on score matching techniques that scales to high dimensional settings. Finally, Hu et al. (2026) propose a natural-gradient method for inference in stochastic differential equations with a discretized Gaussian Markovian posterior and a block-tridiagonal precision matrix. In this respect, this approach closely parallels the information-geometric sparse Gaussian variational inference perspective of Barfoot et al. (2020), albeit developed in the continuous-time setting.

7 Numerical Evaluation

In this section, we empirically evaluate the proposed Bayesian inference algorithms on three representative benchmarks. The evaluation highlights how the entropic variational smoothers behave under different posterior factorizations and approximation schemes, and compares them qualitatively with classical undamped iterated smoothers from the same algorithmic class. Additionally, we provide an open-source implementation of all algorithms in the supplementary material.

Linear-Gaussian. We begin by empirically validating that the recursions are correct. For this, we work in a controlled setting that admits exact inference. We consider the canonical linear-Gaussian setting. Concretely, the latent state $x_k \in \mathbb{R}^2$ and the observation $y_k \in \mathbb{R}^2$ evolve, for $k = 1, \ldots, T$, as

$$x_0 \sim \mathcal{N}(\mu_0, \Sigma_0), \qquad x_k = A\, x_{k-1} + \omega_k, \qquad y_k = H\, x_k + \delta_k,$$

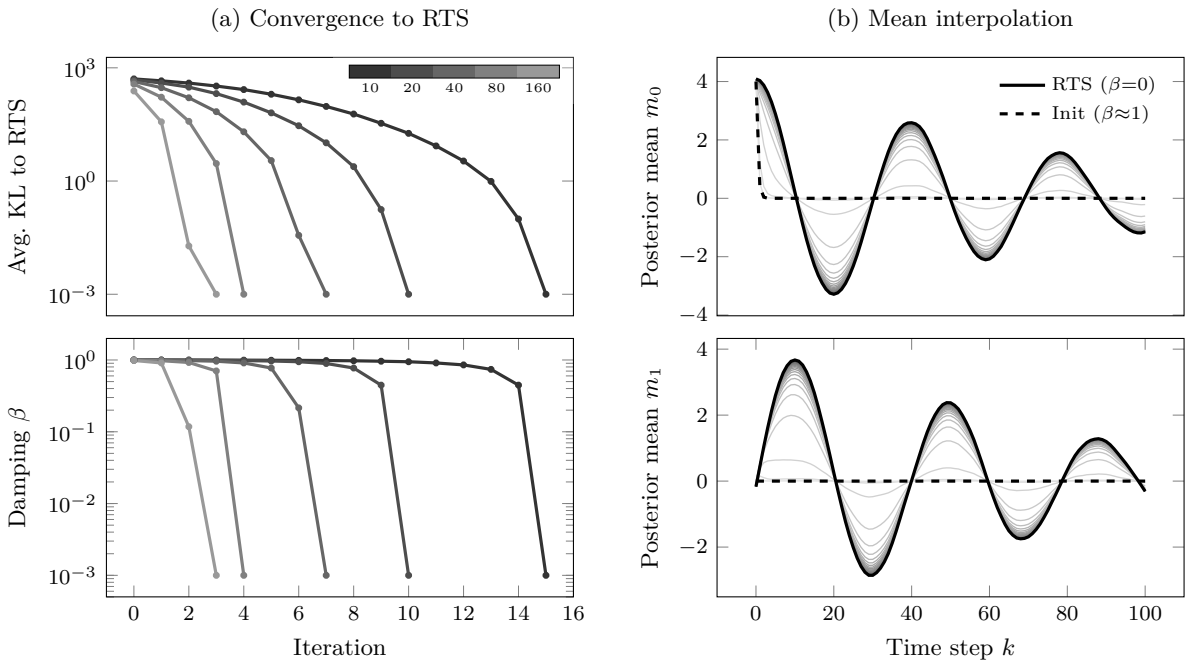

Figure 1: Linear-Gaussian problem. In this setting, the Rauch–Tung–Striebel (RTS) posterior is exact, jointly Gaussian, and lies in the variational family. **(a)** Trust-region convergence for several trust-region radii ε. For each radius, the iterated FEVS approaches the true RTS posterior as the damping β vanishes. Larger values of ε permit larger updates and reach the posterior in fewer iterations. **(b)** Geometric β-interpolation. On a two-dimensional oscillator, the damped update interpolates between the exact posterior mean ($\beta = 0$, solid) and the initialization ($\beta \approx 1$, dashed).

with $\omega_k \sim \mathcal{N}(0, \Omega)$ and $\delta_k \sim \mathcal{N}(0, \Delta)$. The transition is a contractive rotation $A = 0.985 \cdot R(0.16)$, where $R(\theta)$ is the planar rotation matrix for θ radians. The observation map is $H = I_2$, where I_2 is the identity matrix in $\mathbb{R}^{2\times 2}$. The noise covariances are $\Omega = 25 \times 10^{-4} \cdot I_2$ and $\Delta = 6.25 \times 10^{-2} \cdot I_2$, and the state is initialized at $\mu_0 = (4, 0)^\top$ with $\Sigma_0 = 1 \times 10^{-2} \cdot I_2$. We simulate sequences of length $T = 100$.

Because both the dynamics and the likelihood are linear-Gaussian, the joint distribution over $(x_{0:T}, y_{1:T})$ is Gaussian and the smoothing posterior $p(x_{0:T} \mid y_{1:T})$ is available in closed form through the Rauch–Tung–Striebel (RTS) recursion. The Gauss–Markov variational family contains this posterior, so there is no variational approximation error. We use this property to verify, in turn, that the recursions are correct and that the entropic iterations converge to the exact posterior.

For $\beta > 0$, the trust region governs how quickly the iterate approaches the true target, from an arbitrary initialization. Figure 1 illustrates this behavior for the forward-Markov variational smoother (FEVS) with a generalized statistical linear regression (GSLR) approximation. The reverse-Markov and hybrid factorizations yield indistinguishable results. Panel (a) shows that, for every trust-region radius ε, the constrained iteration reduces the average marginal KL divergence to the true RTS posterior to zero, with larger radii converging in fewer iterations. Panel (b) visualizes the geometric interpolation, showing how the posterior mean evolves from the initialization ($\beta \approx 1$) toward the exact posterior ($\beta = 0$).

Stochastic Volatility. Having established correctness when the posterior is exact, we turn to a nonlinear, non-Gaussian setting, in which the quality of approximations matters significantly. We consider a stochastic volatility model (Kim et al., 1998), in which a latent log-volatility $x_k \in \mathbb{R}$ drives the variance of a zero-mean observation $y_k \in \mathbb{R}$. For $k = 1, \ldots, T$,

$$x_0 \sim \mathcal{N}(\mu, \sigma^2/(1-\phi^2)), \qquad x_k = \mu + \phi\,(x_{k-1} - \mu) + \sigma\,\rho_k, \qquad y_k = \exp(x_k/2)\,\zeta_k,$$

with $\rho_k, \zeta_k \sim \mathcal{N}(0, 1)$. Equivalently $y_k \mid x_k \sim \mathcal{N}(0,\, e^{x_k})$. We set $\mu = -0.5$, $\phi = 0.98$, sweep over $\sigma \in [0.1, 0.3]$, and simulate sequences of length $T = 1000$.

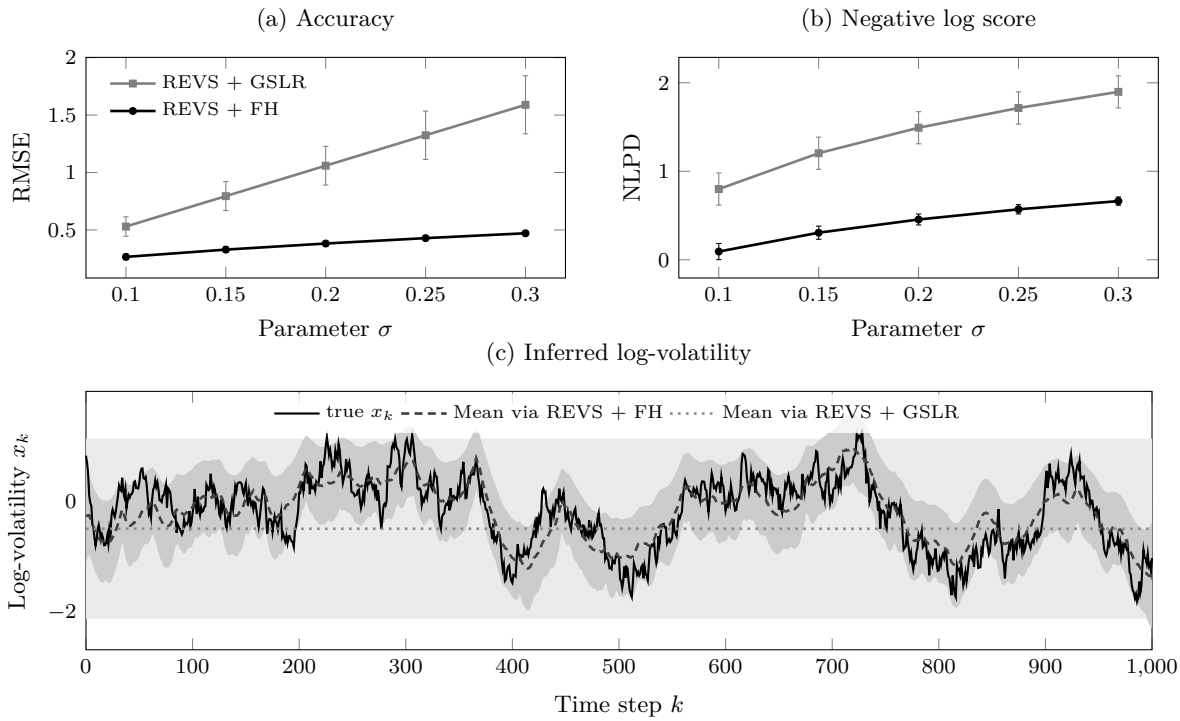

Figure 2: Stochastic volatility problem. **(a)** Root mean squared error (RMSE) and **(b)** negative log posterior density (NLPD) of the true log-volatility as a function of σ. Curves show the mean ± 1 standard deviation over 10 trials. The variant with generalized statistical linear regression (GSLR) incurs larger errors that grow rapidly with σ, whereas the Fourier–Hermite (FH) variant degrades more gradually. **(c)** An example trajectory for $T = 1000$. The GSLR variant follows the stationary prior with constant mean and variance, whereas the FH variant tracks the log-volatility while maintaining reasonable uncertainty bands.

Here, the observation is conditionally Gaussian but informative only through its variance. Its mean $\mathbb{E}[y_k \mid x_k] = 0$ is independent of the state, and all information about x_k enters through the second moment e^{x_k}. The parameter σ controls the noise scale of the latent process. As σ grows, the log-volatility covers a wider range, making the nonlinearity in the likelihood more significant and the inference problem more challenging.

This highlights a fundamental difference between the two approximation regimes. GSLR linearizes the observation by matching the conditional mean $\mathbb{E}[y_k \mid x_k] = 0$, which is constant in x_k. The resulting linear-Gaussian surrogate has observation matrix $H = 0$ and is therefore entirely uninformative, so the smoother ignores the data and reproduces the prior. The Fourier–Hermite (FH) variant, in contrast, forms a quadratic approximation to the log-likelihood and retains the information carried by the observation variance. Panel (c) in Figure 2 illustrates this issue on a representative trajectory. The GSLR variant collapses to the stationary prior dynamics, whereas the FH variant tracks the latent log-volatility with adequate uncertainty bands.

Panels (a) and (b) in Figure 2 quantify this effect. For this comparison, we use the reverse-Markov variational smoother (REVS) with a third-order Gauss–Hermite numerical integration rule, noting that the forward- and hybrid-Markov variants give nearly identical outcomes. We report the negative log posterior density (NLPD), defined as the log score of the true latent trajectory $x^\star_{0:T}$ under the approximate posterior, together with the root mean squared error (RMSE), which measures the error with respect to $x^\star_{0:T}$. Both metrics degrade more gradually for the FH variant than for the GSLR variant, highlighting the advantage of the FH approximation in capturing higher-order terms.

Cubic Sensor. We conclude with the cubic sensor problem (Katayama, 2013; Tronarp, 2025), in which a linear–Gaussian state is observed through a cubic measurement. For $k = 1, \ldots, T$,

$$x_0 \sim \mathcal{N}(\mu_0, \Sigma_0), \qquad x_k = \phi_0\, x_{k-1} + (1 - \phi_0)\, \mu_0 + \omega_k, \qquad y_k = \rho\, x_k^3 + \delta_k,$$

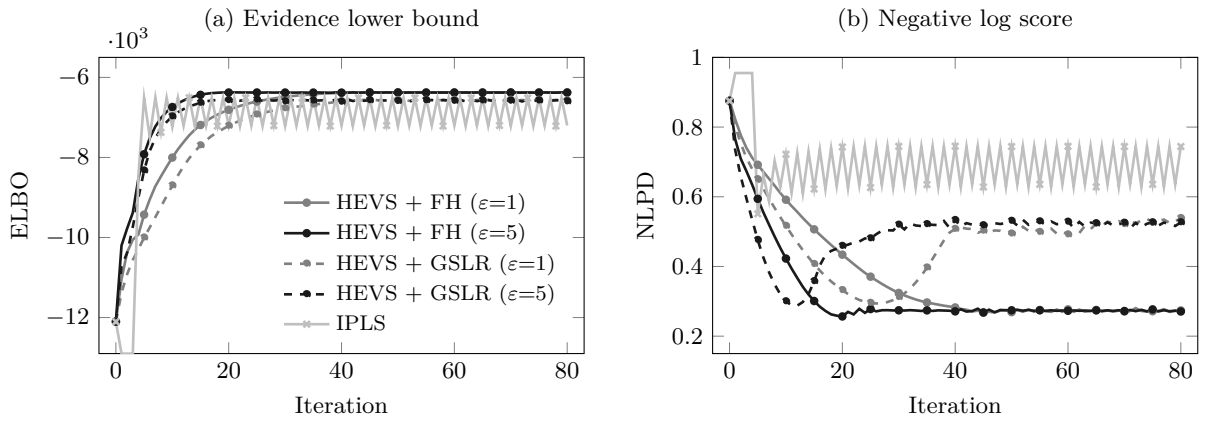

Figure 3: Cubic sensor problem. **(a)** Evidence lower bound (ELBO) and **(b)** negative log posterior density (NLPD) over iterations for the hybrid-Markov variational smoother (HEVS), using Fourier–Hermite (FH) and generalized statistical linear regression (GSLR) approximations, together with the undamped iterated posterior linearization smoother (IPLS). Both HEVS variants improve the ELBO monotonically through controlled step-size updates, while the undamped IPLS fails to converge and settles into a limit cycle.

with $\omega_k \sim \mathcal{N}(0, \Omega)$ and $\delta_k \sim \mathcal{N}(0, \Delta)$, where $\Omega = (1 - \phi_0^2)\,\Sigma_0$. We set $\mu_0 = 0.4$, $\Sigma_0 = 0.36$, $\phi_0 = 0.95$, $\rho = 1$, and $\Delta = 1$, and simulate trajectories of length $T = 4096$.

Although the dynamics are linear, the cubic sensor implies an observation log-likelihood with a degree-six polynomial in the state, $\log h(x_k) = \sum_{m=0}^{6} \ell_k^{(m)}\, x_k^m$, which presents a challenging nonlinearity.

In this evaluation, we report results from the hybrid-Markov variational smoother (HEVS) with a cubature numerical integration rule. The forward- and reverse-Markov variants produce comparable performance. Figure 3 shows the evidence lower bound (ELBO) and the negative log posterior density across iterations. Both HEVS variants, based on FH and GSLR approximations, exhibit stable ELBO convergence. However, the GSLR variant attains a worse NLPD than FH. In contrast, the undamped IPLS baseline (García-Fernández et al., 2016; Tronarp et al., 2018), which lacks an explicit mechanism to control the step size, overshoots and oscillates between two solutions. The ELBO is estimated using a second-order Fourier–Hermite expansion together with a fifth-order Gauss–Hermite quadrature rule.

8 Discussion

We have presented a variational framework for approximate Bayesian inference in general state-space models. The main idea is to formulate inference as a sequence of entropic proximal updates over posterior distributions, subject to dynamic consistency constraints. This yields a dynamic optimization problem whose optimality conditions give rise to recursive forward–backward smoothing algorithms. The resulting KL-based trust regions regularize each update relative to the previous posterior approximation, leading to damped recursions and an adaptive step size mechanism. Our framework shows how different recursive smoothers arise from different factorizations of the variational posterior. Forward, reverse, and hybrid Gauss–Markov approximations each lead to distinct algorithms with linear complexity in the time horizon. For nonlinear and non-Gaussian models, the exact local potentials are generally intractable, so we enforce the recursive structure using generalized statistical linear regression and Fourier–Hermite moment matching. Overall, this proposed framework provides a principled bridge between variational inference, proximal optimization, and recursive Bayesian inference in state-space models.

Several limitations and extensions remain. The practical algorithms rely on local quadratic approximations, and their quality depends on the accuracy of the statistical approximation procedure. In nonlinear or multi-modal problems, a Gauss–Markov approximation may be insufficient to capture the posterior complexity. Future work could extend the framework to richer variational families. Furthermore, square-root implementations of the proposed recursions may improve numerical stability, especially for long time hori-

zons (Bierman, 2006; Yaghoobi et al., 2025). Another direction is to leverage these dynamic optimization ideas within probabilistic numerics (Hennig et al., 2022), where differential equation solvers are formulated as Bayesian state estimation problems. This includes extensions to probabilistic ordinary differential equation solvers (Tronarp et al., 2019; Bosch et al., 2024) and to probabilistic solvers for time-dependent partial differential equations (Krämer et al., 2022; Iqbal et al., 2024).

References

Juha Ala-Luhtala, Simo Särkkä, and Robert Piché. Gaussian filtering and variational approximations for Bayesian smoothing in continuous-discrete stochastic dynamic systems. *Signal Processing*, 111:124–136, 2015.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Shun-ichi Amari. *Information Geometry and Its Applications*, volume 194. Springer, 2016.

Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John Shawe-Taylor. Variational inference for diffusion processes. *Advances in Neural Information Processing Systems*, 20, 2007.

Timothy D Barfoot. *State Estimation for Robotics*. Cambridge University Press, 2024.

Timothy D Barfoot, James R Forbes, and David J Yoon. Exactly sparse Gaussian variational inference with application to derivative-free batch nonlinear state estimation. *The International Journal of Robotics Research*, 39(13):1473–1502, 2020.

Grigory Bartosh, Dmitry Vetrov, and Christian A Naesseth. SDE Matching: Scalable and simulation-free training of latent stochastic differential equations. In *International Conference on Machine Learning*. PMLR, 2025.

Bradley M Bell. The iterated Kalman smoother as a Gauss–Newton method. *SIAM Journal on Optimization*, 4(3):626–636, 1994.

Alexandros Beskos, Dan Crisan, and Ajay Jasra. On the stability of sequential Monte Carlo methods in high dimensions. *The Annals of Applied Probability*, 24(4):1396 – 1445, 2014.

Gerald J Bierman. *Factorization Methods for Discrete Sequential Estimation*. Courier Corporation, 2006.

Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer, 2006.

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

Nathanael Bosch, Adrien Corenflos, Fatemeh Yaghoobi, Filip Tronarp, Philipp Hennig, and Simo Särkkä. Parallel-in-time probabilistic numerical ODE solvers. *Journal of Machine Learning Research*, 25(206): 1–27, 2024.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

Andrew Campbell, Yuyang Shi, Thomas Rainforth, and Arnaud Doucet. Online variational filtering and parameter learning. *Advances in Neural Information Processing Systems*, 34:18633–18645, 2021.

Paul E Chang, William J Wilkinson, Mohammad Emtiyaz Khan, and Arno Solin. Fast variational learning in state-space Gaussian process models. In *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6. IEEE, 2020.

Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer Cham, 2020.

Stéphane Chrétien and Alfred Hero. Kullback proximal algorithms for maximum-likelihood estimation. *IEEE Transactions on Information Theory*, 46(5):1800–1810, 2002.

Adrien Corenflos and Hany Abdulsamad. Variational Gaussian filtering via Wasserstein gradient flows. In *European Signal Processing Conference*, pp. 1838–1842. IEEE, 2023.

Jarrad Courts, Adrian G Wills, and Thomas B Schön. Gaussian variational state estimation for nonlinear state-space models. *IEEE Transactions on Signal Processing*, 69:5979–5993, 2021.

Henry Cox. On the estimation of state variables and parameters for noisy dynamic systems. *IEEE Transactions on Automatic Control*, 9(1):5–12, 1964.

Marc Deisenroth and Shakir Mohamed. Expectation propagation in Gaussian process dynamical systems. *Advances in Neural Information Processing Systems*, 25, 2012.

D Fraser and J Potter. The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control*, 14(4):387–390, 1969.

Ángel F García-Fernández, Lennart Svensson, Mark R Morelande, and Simo Särkkä. Posterior linearization filter: Principles and implementation using sigma points. *IEEE Transactions on Signal Processing*, 63 (20):5561–5573, 2015.

Ángel F García-Fernández, Lennart Svensson, and Simo Särkkä. Iterated posterior linearization smoother. *IEEE Transactions on Automatic Control*, 62(4):2056–2063, 2016.

Jouni Hartikainen and Simo Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 379–384. IEEE, 2010.

Syeda Sakira Hassan and Simo Särkkä. Fourier–Hermite dynamic programming for optimal control. *IEEE Transactions on Automatic Control*, 68(10):6377–6384, 2023.

Philipp Hennig, Michael A Osborne, and Hans P Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

Antti Honkela, Tapani Raiko, Mikael Kuusela, Matti Tornio, and Juha Karhunen. Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *The Journal of Machine Learning Research*, 11:3235–3268, 2010.

Amber Hu, Henry Smith, and Scott Linderman. SING: SDE inference via natural gradients. *Advances in Neural Information Processing Systems*, 38:153894–153947, 2026.

Sahel Iqbal, Hany Abdulsamad, Tripp Cator, Ulisses Braga-Neto, and Simo Särkkä. Parallel-in-time probabilistic solutions for time-dependent nonlinear partial differential equations. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2024.

Alfredo N Iusem, Benar Fux Svaiter, and Marc Teboulle. Entropy-like proximal methods in convex programming. *Mathematics of Operations Research*, 19(4):790–814, 1994.

Eric Jacquier, Nicholas G Polson, and Peter E Rossi. Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, 20(1):69–87, 2002.

Svante Janson. *Gaussian Hilbert Spaces*. Cambridge University Press, 1997.

Tohru Katayama. Equivalent linearization Kalman filter with application to cubic sensor problems. In *2013 European Control Conference (ECC)*, pp. 1633–1638. IEEE, 2013.

Mohammad Emtiyaz Khan and Wu Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. In *Artificial Intelligence and Statistics*, pp. 878–887. PMLR, 2017.

Mohammad Emtiyaz Khan, Pierre Baque, François Fleuret, and Pascal Fua. Kullback-Leibler proximal variational inference. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28, 2015.

Sangjoon Kim, Neil Shephard, and Siddhartha Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393, 1998.

Genshiro Kitagawa. Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1041, 1987.

Nicholas Krämer, Jonathan Schmidt, and Philipp Hennig. Probabilistic numerical method of lines for time-dependent partial differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 625–639. PMLR, 2022.

Rudolf E. Kálmán. A new approach to linear filtering and prediction problems. *Transactions of the ASME Journal of Basic Engineering*, 82:35–45, 1960.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Jakob Lindqvist, Simo Särkkä, Angel F. García-Fernández, Matti Raitoharju, and Lennart Svensson. Posterior linearisation smoothing with robust iterations. 2021.

David Q Mayne. A solution of the smoothing problem for linear dynamic systems. *Automatica*, 4(2):73–92, 1966.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Conference on Uncertainty in Artificial Intelligence*, pp. 362–369. Morgan Kaufmann Publishers Inc., 2001.

Richard E Mortensen. Maximum-likelihood recursive nonlinear filtering. *Journal of Optimization Theory and Applications*, 2(6):386–394, 1968.

James D. Murray. *Mathematical Biology: An Introduction*. Interdisciplinary Applied Mathematics. Springer, 2002.

Jorge Nocedal and Stephen J Wright. *Numerical Optimization*. Springer, 2006.

Anthony O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1):1–24, 1978.

Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

Matti Raitoharju, Lennart Svensson, Angel Froilan García-Fernández, and Robert Piché. Damped posterior linearization filter. *IEEE Signal Processing Letters*, 25(4):536–540, 2018.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black-box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822. PMLR, 2014.

Herbert E Rauch, F Tung, and Charlotte T Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

Simo Särkkä and Lennart Svensson. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2nd edition, 2023.

Juha Sarmavuori and Simo Särkkä. Fourier–Hermite Kalman filter. *IEEE Transactions on Automatic Control*, 57(6):1511–1515, 2011.

Marc Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.

Lucas Theis and Matt Hoffman. A trust-region method for stochastic variational inference with applications to streaming data. In *International Conference on Machine Learning*, pp. 2503–2511. PMLR, 2015.

Filip Tronarp. A recursive theory of variational state estimation: The dynamic programming approach, 2025. URL https://arxiv.org/abs/2511.11497.

Filip Tronarp, Angel F Garcia-Fernandez, and Simo Särkkä. Iterative filtering and smoothing in nonlinear and non-Gaussian systems using conditional moments. *IEEE Signal Processing Letters*, 25(3):408–412, 2018.

Filip Tronarp, Hans Kersting, Simo Särkkä, and Philipp Hennig. Probabilistic solutions to ordinary differential equations as nonlinear Bayesian filtering: A new perspective. *Statistics and Computing*, 29(6): 1297–1315, 2019.

Prakhar Verma, Vincent Adam, and Arno Solin. Variational Gaussian process diffusion processes. In *International Conference on Artificial Intelligence and Statistics*, pp. 1909–1917. PMLR, 2024.

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

Christian Wildner and Heinz Koeppl. Moment-based variational inference for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1918–1926. PMLR, 2021.

William Wilkinson, Paul Chang, Michael Andersen, and Arno Solin. State space expectation propagation: Efficient inference schemes for temporal Gaussian processes. In *International Conference on Machine Learning*, pp. 10270–10281. PMLR, 2020.

William Wilkinson, Simo Särkkä, and Arno Solin. Bayes–Newton methods for approximate Bayesian inference with PSD guarantees. *Journal of Machine Learning Research*, 24(83):1–50, 2023.

Fatemeh Yaghoobi, Adrien Corenflos, Sakira Hassan, and Simo Särkkä. Parallel square-root statistical linear regression for inference in nonlinear state-space models. *SIAM Journal on Scientific Computing*, 47(2): B454–B476, 2025.

A Proof of Proposition 1

To find the critical point of (3) with respect to distribution $q(x)$ given the constraints, we first construct the corresponding Lagrangian functional

$$\begin{aligned}\mathcal{R}(q,\lambda,\alpha) = \mathbb{E}_q\Big[\log p(y,x)\Big] - \mathbb{E}_q\Big[\log q(x)\Big] \\ + \lambda\left[1 - \int q(x)\,\mathrm{d}x\right] + \alpha\left[\varepsilon - \mathbb{D}_{\mathrm{KL}}\left[q(x) \mid\mid q^{[i]}(x)\right]\right],\end{aligned} \tag{28}$$

where λ and $\alpha \geq 0$ are the Lagrangian multipliers associated with the constraints. Next, we set the functional derivative of $\mathcal{R}(\cdot)$ with respect to $q(x)$ to zero

$$\frac{\partial \mathcal{R}(q,\lambda,\alpha)}{\partial q(x)} = \log p(y,x) - \log q(x) - 1 - \lambda - \alpha \log q(x) + \alpha \log q^{[i]}(x) - \alpha \coloneqq 0,$$

which in turn leads to the optimal iterate $q^{[i+1]}(x)$

$$\begin{aligned}q^{[i+1]}(x) &= \Big[\exp\big\{\lambda + (1+\alpha)\big\}\Big]^{-1/(1+\alpha)}\Big[p(y \mid x)\,p(x)\Big]^{1/(1+\alpha)}\Big[q^{[i]}(x)\Big]^{\alpha/(1+\alpha)} \\ &\propto \Big[p(y \mid x)\,p(x)\Big]^{1/(1+\alpha)}\Big[q^{[i]}(x)\Big]^{\alpha/(1+\alpha)}.\end{aligned} \tag{29}$$

Plugging the result from (29) back into the Lagrangian (28), we get the dual problem

$$\underset{\lambda,\alpha}{\text{minimize}} \quad \mathcal{G}(\lambda,\alpha) = \alpha\,\varepsilon + \lambda + (1+\alpha)\int q^{[i+1]}(x)\,\mathrm{d}x, \quad \text{subject to } \alpha \geq 0. \tag{30}$$

If we compute and set the gradient of $\mathcal{G}(\cdot)$ with respect to λ to zero, we get

$$\begin{aligned}\lambda^{[i+1]} &= (1+\alpha)\left[-1 + \log\int\Big[p(y \mid x)\,p(x)\Big]^{1/(1+\alpha)}\Big[q^{[i]}(x)\Big]^{\alpha/(1+\alpha)}\mathrm{d}x\right] \\ &= (1+\alpha)\Big[-1 + \log \mathcal{Z}^{[i+1]}(\alpha)\Big]\end{aligned} \tag{31}$$

which when plugged back into (29) leads to the normalized approximate Gibbs posterior

$$q^{[i+1]}(x) = \Big[\mathcal{Z}^{[i+1]}(\beta)\Big]^{-1}\Big[p(y \mid x)\,p(x)\Big]^{1-\beta}\Big[q^{[i]}(x)\Big]^{\beta}. \tag{32}$$

where we have defined $\beta \coloneqq \alpha/(1+\alpha)$, so that $\beta = 0$ when $\alpha = 0$ and $\beta \to 1$ as $\alpha \to \infty$. By using the results from (31) and (32), we can rewrite the dual problem (30) as follows

$$\underset{\beta}{\text{minimize}} \quad \mathcal{G}(\beta) = \frac{\beta\varepsilon}{1-\beta} + \frac{1}{1-\beta}\log \mathcal{Z}^{[i+1]}(\beta), \quad \text{subject to } 0 \leq \beta < 1.$$

B Proof of Proposition 2

Given the forward-Markov decomposition from Assumption 1, we can factorize problem (5) over time. Starting with the ELBO objective, we can write

$$\begin{aligned}\mathcal{L}(\overrightarrow{q}) &= \mathbb{E}_{\overrightarrow{q}}\Big[\log p(x_{0:T}, y_{1:T})\Big] - \mathbb{E}_{\overrightarrow{q}}\Big[\log \overrightarrow{q}(x_{0:T})\Big] \\ &= \int \overrightarrow{q_0}(x_0)\Big[\log p(x_0) - \log \overrightarrow{q_0}(x_0)\Big]\,\mathrm{d}x_0 + \sum_{k=1}^{T}\int \overrightarrow{q_k}(x_k)\log h_k(y_k \mid x_k)\,\mathrm{d}x_k \\ &\quad + \sum_{k=0}^{T-1}\int \overrightarrow{q_k}(x_k)\,\overrightarrow{q_k}(x_{k+1} \mid x_k)\Big[\log f_k(x_{k+1} \mid x_k) - \log \overrightarrow{q_k}(x_{k+1} \mid x_k)\Big]\,\mathrm{d}x_k\,\mathrm{d}x_{k+1}.\end{aligned}$$

Additionally, the Kullback–Leibler divergence factorizes forward as follows

$$
\begin{aligned}
\varepsilon &\geq \mathbb{D}_{\mathrm{KL}}\left[\overrightarrow{q}(x_{0:T}) \mid\mid \overrightarrow{q}^{[i]}(x_{0:T})\right] \\
\varepsilon &\geq \int \overrightarrow{q}_0(x_0) \prod_{k=0}^{T-1} \overrightarrow{q}_k(x_{k+1} \mid x_k) \log \frac{\overrightarrow{q}_0(x_0) \prod_{k=0}^{T-1} \overrightarrow{q}_k(x_{k+1} \mid x_k)}{\overrightarrow{q}_0^{[i]}(x_0) \prod_{k=0}^{T-1} \overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)} \,\mathrm{d}x_{0:T} \\
\varepsilon &\geq \int \overrightarrow{q}_0(x_0) \log \frac{\overrightarrow{q}_0(x_0)}{\overrightarrow{q}_0^{[i]}(x_0)} \,\mathrm{d}x_0 \\
&\quad + \sum_{k=0}^{T-1} \int \overrightarrow{q}_k(x_k) \int \overrightarrow{q}_k(x_{k+1} \mid x_k) \log \frac{\overrightarrow{q}_k(x_{k+1} \mid x_k)}{\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)} \,\mathrm{d}x_{k+1} \,\mathrm{d}x_k.
\end{aligned}
$$

Finally, the normalization constraint also factorizes to

$$
1 = \int \overrightarrow{q}_0(x_0) \,\mathrm{d}x_0, \quad 1 = \int \overrightarrow{q}_k(x_{k+1} \mid x_k) \,\mathrm{d}x_{k+1}, \quad \forall x_k,\ 0 \leq k < T.
$$

The information-theoretic proximal optimization problem (5) now has the form

$$
\begin{aligned}
\underset{\substack{\overrightarrow{q}_k(x_{k+1} \mid x_k), \\ \overrightarrow{q}_0(x_0)}}{\text{maximize}} \quad & \int \overrightarrow{q}_0(x_0) \Big[\log p(x_0) - \log \overrightarrow{q}_0(x_0)\Big] \,\mathrm{d}x_0 + \sum_{k=1}^{T} \int \overrightarrow{q}_k(x_k) \log h_k(y_k \mid x_k) \,\mathrm{d}x_k \\
& + \sum_{k=0}^{T-1} \int \overrightarrow{q}_k(x_k) \overrightarrow{q}_k(x_{k+1} \mid x_k) \Big[\log f_k(x_{k+1} \mid x_k) - \log \overrightarrow{q}_k(x_{k+1} \mid x_k)\Big] \,\mathrm{d}x_k \,\mathrm{d}x_{k+1}, \\
\text{subject to} \quad & \overrightarrow{q}_{k+1}(x_{k+1}) = \int \overrightarrow{q}_k(x_k) \overrightarrow{q}_k(x_{k+1} \mid x_k) \,\mathrm{d}x_k, \quad \forall\, x_{k+1},\ 0 \leq k < T, \\
& 1 = \int \overrightarrow{q}_0(x_0) \,\mathrm{d}x_0, \quad 1 = \int \overrightarrow{q}_k(x_{k+1} \mid x_k) \,\mathrm{d}x_{k+1}, \quad \forall x_k,\ 0 \leq k < T, \\
& \varepsilon \geq \int \overrightarrow{q}_0(x_0) \log \frac{\overrightarrow{q}_0(x_0)}{\overrightarrow{q}_0^{[i]}(x_0)} \,\mathrm{d}x_0 \\
& \quad + \sum_{k=0}^{T-1} \int \overrightarrow{q}_k(x_k) \int \overrightarrow{q}_k(x_{k+1} \mid x_k) \log \frac{\overrightarrow{q}_k(x_{k+1} \mid x_k)}{\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)} \,\mathrm{d}x_{k+1} \,\mathrm{d}x_k.
\end{aligned}
$$

Next, we construct the Lagrangian functional $\overrightarrow{\mathcal{R}}(\overrightarrow{q}_0, \overrightarrow{q}_k, \overrightarrow{\gamma}_0, \overrightarrow{\lambda}_k, \overrightarrow{V}_k, \alpha)$

$$
\begin{aligned}
\overrightarrow{\mathcal{R}}(\cdot) = & \int \overrightarrow{q}_0(x_0) \Big[\log p(x_0) - \log \overrightarrow{q}_0(x_0)\Big] \,\mathrm{d}x_0 + \sum_{k=1}^{T} \int \overrightarrow{q}_k(x_k) \log h_k(y_k \mid x_k) \,\mathrm{d}x_k \\
& + \sum_{k=0}^{T-1} \int \overrightarrow{q}_k(x_k) \overrightarrow{q}_k(x_{k+1} \mid x_k) \Big[\log f_k(x_{k+1} \mid x_k) - \log \overrightarrow{q}_k(x_{k+1} \mid x_k)\Big] \,\mathrm{d}x_k \,\mathrm{d}x_{k+1} \\
& + \sum_{k=0}^{T-1} \int \overrightarrow{V}_{k+1}(x_{k+1}) \left[\int \overrightarrow{q}_k(x_k) \overrightarrow{q}_k(x_{k+1} \mid x_k) \,\mathrm{d}x_k - \overrightarrow{q}_{k+1}(x_{k+1})\right] \mathrm{d}x_{k+1} \\
& + \sum_{k=0}^{T-1} \int \overrightarrow{\lambda}_k(x_k) \left[1 - \int \overrightarrow{q}_k(x_{k+1} \mid x_k) \,\mathrm{d}x_{k+1}\right] \mathrm{d}x_k + \overrightarrow{\gamma}_0 \left[1 - \int \overrightarrow{q}_0(x_0) \,\mathrm{d}x_0\right] \\
& + \alpha \Bigg[\varepsilon - \int \overrightarrow{q}_0(x_0) \log \frac{\overrightarrow{q}_0(x_0)}{\overrightarrow{q}_0^{[i]}(x_0)} \,\mathrm{d}x_0 \\
& \quad - \sum_{k=0}^{T-1} \int \overrightarrow{q}_k(x_k) \int \overrightarrow{q}_k(x_{k+1} \mid x_k) \log \frac{\overrightarrow{q}_k(x_{k+1} \mid x_k)}{\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)} \,\mathrm{d}x_{k+1} \,\mathrm{d}x_k\Bigg],
\end{aligned}
\tag{33}
$$

where $\overrightarrow{V}_k(x_k)$, $\overrightarrow{\lambda}_k(x_k)$, $\overrightarrow{\gamma}_0$, and $\alpha \geq 0$ are Lagrangian multipliers. To find the critical point, we set the functional derivative with respect to $\overrightarrow{q}_k(x_{k+1} \mid x_k)$ to zero

$$\overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k) = \Big[\exp\Big\{\overrightarrow{\lambda}_k(x_k)/\overrightarrow{q}_k(x_k) + (1+\alpha)\Big\}\Big]^{-1/(1+\alpha)} \times \Big[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\Big]^{\alpha/(1+\alpha)} \Big[f_k(x_{k+1} \mid x_k)\exp\Big\{\overrightarrow{V}_{k+1}(x_{k+1})\Big\}\Big]^{1/(1+\alpha)}. \tag{34}$$

Additionally, we set the functional derivative with respect to $\overrightarrow{q}_0(x_0)$ to zero and get

$$\overrightarrow{q}_0^{[i+1]}(x_0) = \Big[\exp\Big\{\overrightarrow{\gamma}_0 + (1+\alpha)\Big\}\Big]^{-1/(1+\alpha)} \Big[\overrightarrow{q}_0^{[i]}(x_0)\Big]^{\alpha/(1+\alpha)} \Big[\exp\Big\{\overrightarrow{V}_0(x_0)\Big\}\Big]^{1/(1+\alpha)}, \tag{35}$$

where, for convenience of notation, we define

$$\overrightarrow{V}_0(x_0) = \log p_0(x_0) + (1+\alpha)\log\int \Big[\overrightarrow{q}_0^{[i]}(x_1 \mid x_0)\Big]^{\alpha/(1+\alpha)} \Big[f_0(x_1 \mid x_0)\exp\Big\{\overrightarrow{V}_1(x_1)\Big\}\Big]^{1/(1+\alpha)} \mathrm{d}x_1. \tag{36}$$

Notice that it is not clear at this point that the solutions (34) and (35) are normalized densities. Establishing that requires solving for the (functional) multipliers $\overrightarrow{\lambda}_k(x_k)$ and $\overrightarrow{\gamma}_0$, which are associated with the normalization constraints. Plugging the solution (34) into the Lagrangian (33) results in the dual functional

$$\begin{aligned}\overrightarrow{\mathcal{G}}(\cdot) = \alpha\,\varepsilon &+ \int \overrightarrow{q}_0(x_0)\Big[\log p(x_0) - \log\overrightarrow{q}_0(x_0)\Big]\,\mathrm{d}x_0 + \sum_{k=1}^{T}\int \overrightarrow{q}_k(x_k)\log h_k(y_k \mid x_k)\,\mathrm{d}x_k \\ &+ \alpha\int \overrightarrow{q}_0(x_0)\Big[\log\overrightarrow{q}_0^{[i]}(x_0) - \log\overrightarrow{q}_0(x_0)\Big]\,\mathrm{d}x_0 + \overrightarrow{\gamma}_0\Big[1 - \int\overrightarrow{q}_0(x_0)\,\mathrm{d}x_0\Big] \\ &+ \sum_{k=0}^{T-1}\int\overrightarrow{\lambda}_k(x_k)\,\mathrm{d}x_k - \sum_{k=0}^{T-1}\int\overrightarrow{V}_{k+1}(x_{k+1})\overrightarrow{q}_{k+1}(x_{k+1})\,\mathrm{d}x_{k+1} \\ &+ (1+\alpha)\sum_{k=0}^{T-1}\int\overrightarrow{q}_k(x_k)\int\overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k)\,\mathrm{d}x_{k+1}\,\mathrm{d}x_k.\end{aligned}$$

Now, we can solve for the multipliers $\overrightarrow{\lambda}_k(x_k)$ by setting the associated derivatives to zero

$$\begin{aligned}\overrightarrow{\lambda}_k^{[i+1]}(x_k) &= (1+\alpha)\,\overrightarrow{q}_k(x_k)\Big[-1 + \log\int\Big[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\Big]^{\alpha/(1+\alpha)} \\ &\qquad\times\Big[f_k(x_{k+1} \mid x_k)\exp\Big\{\overrightarrow{V}_{k+1}(x_{k+1})\Big\}\Big]^{1/(1+\alpha)}\,\mathrm{d}x_{k+1}\Big] \\ &= (1+\alpha)\,\overrightarrow{q}_k(x_k)\Big[-1 + \log\overrightarrow{\psi}_k^{[i+1]}(x_k)\Big].\end{aligned} \tag{37}$$

Plugging (37) into (34) returns the normalized tilted conditional

$$\overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k) = \Big[\overrightarrow{\psi}_k^{[i+1]}(x_k)\Big]^{-1}\Big[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\Big]^{\beta} \times\Big[f_k(x_{k+1} \mid x_k)\exp\Big\{\overrightarrow{V}_{k+1}(x_{k+1})\Big\}\Big]^{1-\beta}, \tag{38}$$

where we have defined $\beta \coloneqq \alpha/(1+\alpha)$, so that $\beta = 0$ when $\alpha = 0$, and $\beta \to 1$ as $\alpha \to \infty$. Next, we use the results from (35), (37), and (38) to simplify the functional $\overrightarrow{\mathcal{G}}(\cdot)$ further

$$\begin{aligned}\overrightarrow{\mathcal{G}}(\cdot) = \alpha\,\varepsilon &+ \sum_{k=1}^{T}\int\overrightarrow{q}_k(x_k)\log h_k(y_k \mid x_k)\,\mathrm{d}x_k - \sum_{k=1}^{T}\int\overrightarrow{V}_k(x_k)\overrightarrow{q}_k(x_k)\,\mathrm{d}x_k \\ &+ (1+\alpha)\sum_{k=1}^{T-1}\int\overrightarrow{q}_k(x_k)\log\overrightarrow{\psi}_k^{[i+1]}(x_k)\,\mathrm{d}x_k + \overrightarrow{\gamma}_0 + (1+\alpha)\int\overrightarrow{q}_0^{[i+1]}(x_0)\,\mathrm{d}x_0,\end{aligned}$$

and solve for the optimal multiplier $\overrightarrow{\gamma}_0$ by zeroing the associated derivative, leading to

$$\begin{aligned}\overrightarrow{\gamma}_0^{[i+1]} &= (1+\alpha)\left[-1+\log\int\left[\overrightarrow{q}_0^{[i]}(x_0)\right]^{\alpha/(1+\alpha)}\left[\exp\left\{\overrightarrow{V}_0(x_0)\right\}\right]^{1/(1+\alpha)}\mathrm{d}x_0\right] \\ &= (1+\alpha)\left[-1+\log\overrightarrow{\mathcal{Z}}_0^{[i+1]}\right],\end{aligned} \tag{39}$$

which after plugging into (35) delivers

$$\overrightarrow{q}_0^{[i+1]}(x_0) = \left[\overrightarrow{\mathcal{Z}}_0^{[i+1]}\right]^{-1}\left[\overrightarrow{q}_0^{[i]}(x_0)\right]^{\beta}\left[\exp\left\{\overrightarrow{V}_0(x_0)\right\}\right]^{1-\beta}. \tag{40}$$

Plugging (40) and (39) back into $\overrightarrow{\mathcal{G}}(\cdot)$ leads to further simplification of the dual

$$\begin{aligned}\overrightarrow{\mathcal{G}}(\cdot) = \alpha\,\varepsilon &+ \sum_{k=1}^{T}\int\overrightarrow{q}_k(x_k)\log h_k(y_k\mid x_k)\,\mathrm{d}x_k - \sum_{k=1}^{T}\int\overrightarrow{V}_k(x_k)\overrightarrow{q}_k(x_k)\,\mathrm{d}x_k \\ &+ (1+\alpha)\sum_{k=1}^{T-1}\int\overrightarrow{q}_k(x_k)\log\overrightarrow{\psi}_k^{[i+1]}(x_k)\,\mathrm{d}x_k + (1+\alpha)\log\overrightarrow{\mathcal{Z}}_0^{[i+1]}.\end{aligned}$$

Finally, to find the optimal potentials, we zero the derivative of $\overrightarrow{\mathcal{G}}(\cdot)$ with respect to $\overrightarrow{q}_k(x_k), \forall\, 0 < k \leq T$. Combined with the results from (36) and (37), we can write

$$\overrightarrow{V}_k^{[i+1]}(x_k) = \begin{cases}\log h_T(y_T\mid x_T) & \text{if } k=T, \\ \log h_k(y_k\mid x_k) + 1/(1-\beta)\log\overrightarrow{\psi}_k^{[i+1]}(x_k) & \text{if } 0<k<T, \\ \log p_0(x_0) + 1/(1-\beta)\log\overrightarrow{\psi}_0^{[i+1]}(x_0) & \text{if } k=0,\end{cases}$$

which, when plugged back into the dual $\overrightarrow{\mathcal{G}}(\cdot)$, leads to the final simplification

$$\overrightarrow{\mathcal{G}}(\beta) = \frac{\beta\varepsilon}{1-\beta} + \frac{1}{1-\beta}\log\overrightarrow{\mathcal{Z}}_0^{[i+1]}(\beta).$$

C Proof of Proposition 3

This proof follows a similar scheme to that in Section B. However, the reverse-Markov decomposition results in novel recursions. Starting from the reverse-Markov decomposition in Assumption 2, we factorize problem (5) over time. Thus, we can write the ELBO as

$$\begin{aligned}\mathcal{L}(\overleftarrow{q}) &= \mathbb{E}_{\overleftarrow{q}}\Big[\log p(x_{0:T}, y_{1:T})\Big] - \mathbb{E}_{\overleftarrow{q}}\Big[\log\overleftarrow{q}(x_{0:T})\Big] \\ &= \int\overleftarrow{q}_0(x_0)\log p(x_0)\,\mathrm{d}x_0 - \int\overleftarrow{q}_T(x_T)\log\overleftarrow{q}_T(x_T)\,\mathrm{d}x_T + \sum_{k=1}^{T}\int\overleftarrow{q}_k(x_k)\log h_k(y_k\mid x_k)\,\mathrm{d}x_k \\ &\quad + \sum_{k=1}^{T}\int\overleftarrow{q}_k(x_k)\,\overleftarrow{q}_k(x_{k-1}\mid x_k)\Big[\log f_{k-1}(x_k\mid x_{k-1}) - \log\overleftarrow{q}_k(x_{k-1}\mid x_k)\Big]\,\mathrm{d}x_{k-1}\,\mathrm{d}x_k.\end{aligned}$$

Furthermore, the Kullback–Leibler divergence factorizes in reverse to

$$\begin{aligned}&\varepsilon \geq \mathbb{D}_{\mathrm{KL}}\left[\overleftarrow{q}(x_{0:T})\;||\;\overleftarrow{q}^{[i]}(x_{0:T})\right] \\ &\varepsilon \geq \int\overleftarrow{q}_T(x_T)\prod_{k=1}^{T}\overleftarrow{q}_k(x_{k-1}\mid x_k)\log\frac{\overleftarrow{q}_T(x_T)\prod_{k=1}^{T}\overleftarrow{q}_k(x_{k-1}\mid x_k)}{\overleftarrow{q}_T^{[i]}(x_T)\prod_{k=1}^{T}\overleftarrow{q}_k^{[i]}(x_{k-1}\mid x_k)}\,\mathrm{d}x_{0:T}\end{aligned}$$

$$\varepsilon \geq \int \overleftarrow{q}_T(x_T) \log \frac{\overleftarrow{q}_T(x_T)}{\overleftarrow{q}_T^{[i]}(x_T)} \, \mathrm{d}x_T + \sum_{k=1}^{T} \int \overleftarrow{q}_k(x_k) \int \overleftarrow{q}_k(x_{k-1} \mid x_k) \log \frac{\overleftarrow{q}_k(x_{k-1} \mid x_k)}{\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)} \, \mathrm{d}x_k \, \mathrm{d}x_{k-1}$$

Finally, the normalization constraint also factorizes to

$$1 = \int \overleftarrow{q}_T(x_T) \, \mathrm{d}x_T, \quad 1 = \int \overleftarrow{q}_k(x_{k-1} \mid x_k) \, \mathrm{d}x_{k-1}, \quad \forall x_k, \; 0 < k \leq T.$$

Now we can rewrite the optimization problem (5) as follows

$$\begin{aligned}
\underset{\substack{\overleftarrow{q}_k(x_{k-1} \mid x_k), \\ \overleftarrow{q}_T(x_T)}}{\text{maximize}} \quad & \int \overleftarrow{q}_0(x_0) \log p(x_0) \, \mathrm{d}x_0 + \sum_{k=1}^{T} \int \overleftarrow{q}_k(x_k) \log h_k(y_k \mid x_k) \, \mathrm{d}x_k \\
& + \sum_{k=1}^{T} \int \overleftarrow{q}_k(x_k) \overleftarrow{q}_k(x_{k-1} \mid x_k) \Big[\log f_{k-1}(x_k \mid x_{k-1}) - \log \overleftarrow{q}_k(x_{k-1} \mid x_k) \Big] \, \mathrm{d}x_{k-1} \, \mathrm{d}x_k \\
& - \int \overrightarrow{q}_T(x_T) \log \overleftarrow{q}_T(x_T) \, \mathrm{d}x_T \\
\text{subject to} \quad & \overleftarrow{q}_{k-1}(x_{k-1}) = \int \overleftarrow{q}_k(x_k) \overleftarrow{q}_k(x_{k-1} \mid x_k) \, \mathrm{d}x_k, \quad \forall \, x_{k-1}, \, 0 < k \leq T, \\
& 1 = \int \overleftarrow{q}_T(x_T) \, \mathrm{d}x_T, \quad 1 = \int \overleftarrow{q}_k(x_{k-1} \mid x_k) \, \mathrm{d}x_{k-1}, \quad \forall x_k, \; 0 < k \leq T, \\
& \varepsilon \geq \int \overleftarrow{q}_T(x_T) \log \frac{\overleftarrow{q}_T(x_T)}{\overleftarrow{q}_T^{[i]}(x_T)} \, \mathrm{d}x_T \\
& \quad + \sum_{k=1}^{T} \int \overleftarrow{q}_k(x_k) \int \overleftarrow{q}_k(x_{k-1} \mid x_k) \log \frac{\overleftarrow{q}_k(x_{k-1} \mid x_k)}{\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)} \, \mathrm{d}x_k \, \mathrm{d}x_{k-1},
\end{aligned}$$

and the corresponding Lagrangian functional $\overleftarrow{\mathcal{R}}(\overleftarrow{q}_T, \overleftarrow{q}_k, \overleftarrow{\gamma}_T, \overleftarrow{\lambda}_k, \overleftarrow{V}_k, \alpha)$ is

$$\begin{aligned}
\overleftarrow{\mathcal{R}}(\cdot) = & \int \overleftarrow{q}_0(x_0) \log p(x_0) \, \mathrm{d}x_0 + \sum_{k=1}^{T} \int \overleftarrow{q}_k(x_k) \log h_k(y_k \mid x_k) \, \mathrm{d}x_k - \int \overrightarrow{q}_T(x_T) \log \overleftarrow{q}_T(x_T) \, \mathrm{d}x_T \\
& + \sum_{k=1}^{T} \int \overleftarrow{q}_k(x_k) \overleftarrow{q}_k(x_{k-1} \mid x_k) \Big[\log f_{k-1}(x_k \mid x_{k-1}) - \log \overleftarrow{q}_k(x_{k-1} \mid x_k) \Big] \, \mathrm{d}x_{k-1} \, \mathrm{d}x_k \\
& + \sum_{k=1}^{T} \int \overleftarrow{V}_{k-1}(x_{k-1}) \left[\int \overleftarrow{q}_k(x_k) \overleftarrow{q}_k(x_{k-1} \mid x_k) \, \mathrm{d}x_k - \overleftarrow{q}_{k-1}(x_{k-1}) \right] \mathrm{d}x_{k-1} \\
& + \sum_{k=1}^{T} \int \overleftarrow{\lambda}_k(x_k) \left[1 - \int \overleftarrow{q}_k(x_{k-1} \mid x_k) \, \mathrm{d}x_{k-1} \right] \mathrm{d}x_k + \overleftarrow{\gamma}_T \left[1 - \int \overleftarrow{q}_T(x_T) \, \mathrm{d}x_T \right] \\
& + \alpha \Bigg[\varepsilon - \int \overleftarrow{q}_T(x_T) \log \frac{\overleftarrow{q}_T(x_T)}{\overleftarrow{q}_T^{[i]}(x_T)} \, \mathrm{d}x_T \\
& \quad - \sum_{k=1}^{T} \int \overleftarrow{q}_k(x_k) \int \overleftarrow{q}_k(x_{k-1} \mid x_k) \log \frac{\overleftarrow{q}_k(x_{k-1} \mid x_k)}{\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)} \, \mathrm{d}x_k \, \mathrm{d}x_{k-1} \Bigg],
\end{aligned} \tag{41}$$

where $\overleftarrow{V}_k(x_k)$, $\overleftarrow{\lambda}_k(x_k)$, $\overleftarrow{\gamma}_T$, and $\alpha \geq 0$ are Lagrangian multipliers. To find the solution with respect to $\overleftarrow{q}_k(x_{k-1} \mid x_k)$, we zero the associated derivatives of $\overleftarrow{\mathcal{R}}(\cdot)$

$$\begin{aligned}
\overleftarrow{q}_k^{[i+1]}(x_{k-1} \mid x_k) = & \left[\exp \left\{ \overleftarrow{\lambda}_k(x_k) / \overleftarrow{q}_k(x_k) + (1 + \alpha) \right\} \right]^{-1/(1+\alpha)} \\
& \times \left[\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k) \right]^{\alpha/(1+\alpha)} \left[f_{k-1}(x_{k+1} \mid x_k) \exp \left\{ \overleftarrow{V}_{k-1}(x_{k-1}) \right\} \right]^{1/(1+\alpha)}
\end{aligned} \tag{42}$$

Moreover, by setting the functional derivative with respect to $\overleftarrow{q}_T(x_T)$ to zero, we get

$$\overleftarrow{q}_T^{[i+1]}(x_T) = \Big[\exp\Big\{\overleftarrow{\gamma}_T + (1+\alpha)\Big\}\Big]^{-1/(1+\alpha)} \Big[\overleftarrow{q}_T^{[i]}(x_T)\Big]^{\alpha/(1+\alpha)} \Big[\exp\Big\{\overleftarrow{V}_T(x_T)\Big\}\Big]^{1/(1+\alpha)} \tag{43}$$

where, for convenience of notation, we define

$$\begin{aligned}\overleftarrow{V}_T(x_T) = \log h_T(y_T \mid x_T) + (1+\alpha)\log \int \Big[\overleftarrow{q}_{T-1}^{[i]}(x_{T-1} \mid x_T)\Big]^{\alpha/(1+\alpha)} \\ \times \Big[f_{T-1}(x_T \mid x_{T-1}) \exp\Big\{\overleftarrow{V}_{T-1}(x_1)\Big\}\Big]^{1/(1+\alpha)} \mathrm{d}x_{T-1}.\end{aligned} \tag{44}$$

Again, the solutions (42) and (43) are not *yet* normalized densities. We need to solve for the (functional) multipliers associated with the normalization constraints, $\overleftarrow{\lambda}_k(x_k)$ and $\overleftarrow{\gamma}_T$. Plugging the solution (42) into the Lagrangian (41) results in the dual functional

$$\begin{aligned}\overleftarrow{\mathcal{G}}(\cdot) = \alpha\,\varepsilon &+ \int \overleftarrow{q}_0(x_0)\log p(x_0)\,\mathrm{d}x_0 + \sum_{k=1}^{T}\int \overleftarrow{q}_k(x_k)\log h_k(y_k \mid x_k)\,\mathrm{d}x_k \\ &+ \int \overleftarrow{q}_T(x_T)\Big[\alpha\log\overleftarrow{q}_T^{[i]}(x_T) - (1+\alpha)\log\overleftarrow{q}_T(x_T)\Big]\,\mathrm{d}x_T + \overleftarrow{\gamma}_T\Big[1 - \int \overleftarrow{q}_T(x_T)\,\mathrm{d}x_T\Big] \\ &+ \sum_{k=1}^{T}\int \overleftarrow{\lambda}_k(x_k)\,\mathrm{d}x_k - \sum_{k=1}^{T}\int \overleftarrow{V}_{k-1}(x_{k-1})\overleftarrow{q}_{k-1}(x_{k-1})\,\mathrm{d}x_{k-1} \\ &+ (1+\alpha)\sum_{k=1}^{T}\int \overleftarrow{q}_k(x_k)\int \overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)\,\mathrm{d}x_k\,\mathrm{d}x_{k-1},\end{aligned}$$

which we use to solve for the multipliers $\overleftarrow{\lambda}_k(x_k)$ by zeroing the associated derivatives

$$\begin{aligned}\overleftarrow{\lambda}_k^{[i+1]}(x_k) &= (1+\alpha)\,\overleftarrow{q}_k(x_k)\Big[-1 + \log\int \Big[\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)\Big]^{\alpha/(1+\alpha)} \\ &\qquad \times \Big[f_{k-1}(x_k \mid x_{k-1})\exp\Big\{\overleftarrow{V}_{k-1}(x_{k-1})\Big\}\Big]^{1/(1+\alpha)}\,\mathrm{d}x_{k-1}\Big] \\ &= (1+\alpha)\,\overleftarrow{q}_k(x_k)\Big[-1 + \log\overleftarrow{\psi}_k^{[i+1]}(x_k)\Big].\end{aligned} \tag{45}$$

We retrieve the normalized tilted conditionals by plugging (45) into (42)

$$\begin{aligned}\overleftarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k) = \Big[\overleftarrow{\psi}_k^{[i+1]}(x_k)\Big]^{-1}\Big[\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)\Big]^{\beta} \\ \times \Big[f_{k-1}(x_k \mid x_{k-1})\exp\Big\{\overleftarrow{V}_{k-1}(x_{k-1})\Big\}\Big]^{1-\beta},\end{aligned} \tag{46}$$

where we have defined $\beta \coloneqq \alpha/(1+\alpha)$, so that $\beta = 0$ when $\alpha = 0$, and $\beta \to 1$ as $\alpha \to \infty$. Given the results (45) and (46), we simplify the functional $\overleftarrow{\mathcal{G}}(\cdot)$ further

$$\begin{aligned}\overleftarrow{\mathcal{G}}(\cdot) = \alpha\,\varepsilon &+ \int \overleftarrow{q}_0(x_0)\log p(x_0)\,\mathrm{d}x_0 + \sum_{k=1}^{T}\int \overleftarrow{q}_k(x_k)\log h_k(y_k \mid x_k)\,\mathrm{d}x_k \\ &- \sum_{k=0}^{T}\int \overleftarrow{V}_k(x_k)\overleftarrow{q}_k(x_k)\,\mathrm{d}x_k + (1+\alpha)\sum_{k=1}^{T}\int \overleftarrow{q}_k(x_k)\log\overleftarrow{\psi}_k^{[i+1]}(x_k)\,\mathrm{d}x_k \\ &+ \overleftarrow{\gamma}_T + (1+\alpha)\int \overleftarrow{q}_T^{[i+1]}(x_T)\,\mathrm{d}x_T.\end{aligned}$$

Next, we solve for the optimal multiplier $\overleftarrow{\gamma}_T$ by zeroing the associated derivative

$$\overleftarrow{\gamma}_T^{[i+1]} = (1+\alpha)\Big[-1 + \log\int \Big[\overleftarrow{q}_T^{[i]}(x_T)\Big]^{\alpha/(1+\alpha)}\Big[\exp\Big\{\overleftarrow{V}_T(x_T)\Big\}\Big]^{1/(1+\alpha)}\,\mathrm{d}x_T\Big] \tag{47}$$

$$= (1+\alpha)\left[-1+\log \overleftarrow{\mathcal{Z}}_T^{[i+1]}\right],$$

leading to the normalization of the tilted distribution (43)

$$\overleftarrow{q}_T^{[i+1]}(x_T) = \left[\overleftarrow{\mathcal{Z}}_T^{[i+1]}\right]^{-1}\left[\overleftarrow{q}_T^{[i]}(x_T)\right]^{\beta}\left[\exp\left\{\overleftarrow{V}_T(x_T)\right\}\right]^{1-\beta}. \tag{48}$$

Using (48) and (47) leads to further simplification of the dual $\overleftarrow{\mathcal{G}}(\cdot)$

$$\overleftarrow{\mathcal{G}}(\cdot) = \alpha\,\varepsilon + \int \overleftarrow{q}_0(x_0)\log p(x_0)\,\mathrm{d}x_0 + \sum_{k=1}^{T}\int \overleftarrow{q}_k(x_k)\log h_k(y_k \mid x_k)\,\mathrm{d}x_k$$
$$-\sum_{k=0}^{T}\int \overleftarrow{V}_k(x_k)\overleftarrow{q}_k(x_k)\,\mathrm{d}x_k + (1+\alpha)\sum_{k=1}^{T}\int \overleftarrow{q}_k(x_k)\log \overleftarrow{\psi}_k^{[i+1]}(x_k)\,\mathrm{d}x_k + (1+\alpha)\log \overleftarrow{\mathcal{Z}}_T^{[i+1]}.$$

Finally, by taking (44) into consideration and zeroing the derivative of $\overleftarrow{\mathcal{G}}(\cdot)$ with respect to $\overleftarrow{q}_k(x_k), \forall\, 0 \leq k < T$, we find the optimal potential functions

$$\overleftarrow{V}_k^{[i+1]}(x_k) = \begin{cases} \log p_0(x_0) & \text{if } k=0, \\ \log h_k(y_k \mid x_k) + 1/(1-\beta)\log \overleftarrow{\psi}_k^{[i+1]}(x_k) & \text{if } 0 < k \leq T. \end{cases}$$

If we plug this result back into the dual $\overleftarrow{\mathcal{G}}(\cdot)$, we retrieve the simplest form of the dual

$$\overleftarrow{\mathcal{G}}(\beta) = \frac{\beta\varepsilon}{1-\beta} + \frac{1}{1-\beta}\log \overleftarrow{\mathcal{Z}}_T^{[i+1]}(\beta).$$

D Proof of Proposition 4

At $k=T$, Definition 2 readily delivers a quadratic form for the potential function

$$\overrightarrow{V}_T^{[i+1]}(x_T) = \log h_T(y_T \mid x_T) = -\frac{1}{2}x_T^\top \overrightarrow{R}_T^{[i+1]}\,x_T + x_T^\top \overrightarrow{r}_T^{[i+1]} + \overrightarrow{\rho}_T^{[i+1]}$$
$$= -\frac{1}{2}x_T^\top L_T^{[i]}\,x_T + x_T^\top l_T^{[i]} + \nu_T^{[i]}.$$

For $\overrightarrow{V}_k^{[i+1]}(x_k)$, for all $0 \leq k < T$, we have the backward recursion

$$\overrightarrow{V}_k^{[i+1]}(x_k) = \log h_k(y_k \mid x_k) + 1/(1-\beta)\log \overrightarrow{\psi}_k^{[i+1]}(x_k)$$
$$= \log h_k(y_k \mid x_k) + 1/(1-\beta)\log \int \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^{\beta}$$
$$\times \left[f_k(x_{k+1} \mid x_k)\exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta}\mathrm{d}x_{k+1}.$$

Our aim is to show that this recursion is a tractable reverse propagation of quadratic forms. We start by examining $\overrightarrow{\psi}_k^{[i+1]}(x_k)$, the integral over x_{k+1}. Let us first drag all terms into the exponential, then we can treat the exponent as a quadratic function over x_k and x_{k+1}

$$-\frac{1}{2}\begin{bmatrix} x_{k+1}^\top & x_k^\top \end{bmatrix}\begin{bmatrix} \overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} & -\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} \\ -\overrightarrow{G}_{x\bar{x},k}^{[i+1]} & \overrightarrow{G}_{xx,k}^{[i+1]} \end{bmatrix}\begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} + \begin{bmatrix} x_{k+1}^\top & x_k^\top \end{bmatrix}\begin{bmatrix} \overrightarrow{g}_{\bar{x},k}^{[i+1]} \\ \overrightarrow{g}_{x,k}^{[i+1]} \end{bmatrix} + \overrightarrow{\theta}_k^{[i+1]}$$
$$= \beta\log \overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k) + (1-\beta)\log f_k(x_{k+1} \mid x_k) + (1-\beta)\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1}).$$

Now, given an affine-Gaussian $\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)$ (Assumption 3), a quadratic $\log f_k(x_{k+1} \mid x_k)$ (Definition 2), and a quadratic potential function $\overrightarrow{V}_{k+1}^{[i+1]}(x_k)$ of the form (17), we can match the quadratic factors between the left- and right-hand sides, leading to

$$\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} = (1-\beta)\left[C_{\bar{x}\bar{x},k}^{[i]} + \overrightarrow{R}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1},$$

$$
\begin{aligned}
\overrightarrow{G}_{xx,k}^{[i+1]} &= (1-\beta)\, C_{xx,k}^{[i]} + \beta \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{F}_k^{[i]}, \\
\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} &= (1-\beta)\, C_{\bar{x}x,k}^{[i]} + \beta \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{F}_k^{[i]}, \\
\overrightarrow{G}_{x\bar{x},k}^{[i+1]} &= (1-\beta)\, C_{x\bar{x},k}^{[i]} + \beta \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}, \\
\overrightarrow{g}_{\bar{x},k}^{[i+1]} &= (1-\beta) \left[c_{\bar{x},k}^{[i]} + \overrightarrow{r}_k^{[i+1]}\right] + \beta \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{d}_k^{[i]}, \\
\overrightarrow{g}_{x,k}^{[i+1]} &= (1-\beta)\, c_{x,k}^{[i]} - \beta \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{d}_k^{[i]}, \\
\overrightarrow{\theta}_k^{[i+1]} &= (1-\beta) \left[\kappa_k^{[i]} + \overrightarrow{\rho}_{k+1}^{[i+1]}\right] - \frac{\beta}{2} \log \left|2\pi \overrightarrow{\Sigma}_k^{[i]}\right| - \frac{\beta}{2} \left[\overrightarrow{d}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{d}_k^{[i]}.
\end{aligned}
$$

Next, we reformulate the quadratic function in the exponent explicitly as a function of x_{k+1}

$$
-\frac{1}{2} x_{k+1}^\top \overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} x_{k+1} + x_{k+1}^\top \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} x_k + \overrightarrow{g}_{\bar{x},k}^{[i+1]}\right] + \left[-\frac{1}{2} x_k^\top \overrightarrow{G}_{xx,k}^{[i+1]} x_k + x_k^\top \overrightarrow{g}_{x,k}^{[i+1]} + \overrightarrow{\theta}_k^{[i+1]}\right].
$$

Consequently, the exponential now resembles an unnormalized Gaussian distribution in *information form* with a precision matrix $\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}$. In this case, we can use the identity

$$
\log \int \exp\left\{-\frac{1}{2} x^\top U x + x^\top u + \sigma\right\} \mathrm{d}x = \frac{1}{2} \log |2\pi\, U^{-1}| + \frac{1}{2} u^\top U^{-1} u + \sigma, \tag{49}
$$

to express the log-normalizing function $\log \overrightarrow{\psi}_k^{[i+1]}(x_k)$ as a quadratic function

$$
\begin{aligned}
\log \overrightarrow{\psi}_k^{[i+1]}(x_k) &= -\frac{1}{2} x_k^\top \overrightarrow{S}_k^{[i+1]} x_k + x_k^\top \overrightarrow{s}_k^{[i+1]} + \overrightarrow{\xi}_k^{[i+1]} \\
&= \frac{1}{2} \log \left|2\pi \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}\right| + \left[-\frac{1}{2} x_k^\top \overrightarrow{G}_{xx,k}^{[i+1]} x_k + x_k^\top \overrightarrow{g}_{x,k}^{[i+1]} + \overrightarrow{\theta}_k^{[i+1]}\right] \\
&\quad + \frac{1}{2} \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} x_k + \overrightarrow{g}_{\bar{x},k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} x_k + \overrightarrow{g}_{\bar{x},k}^{[i+1]}\right]
\end{aligned}
$$

which, after matching terms, leads to

$$
\begin{aligned}
\overrightarrow{S}_k^{[i+1]} &= \overrightarrow{G}_{xx,k}^{[i+1]} - \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{G}_{\bar{x}x,k}^{[i+1]}, \\
\overrightarrow{s}_k^{[i+1]} &= \overrightarrow{g}_{x,k}^{[i+1]} + \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{g}_{\bar{x},k}^{[i+1]}, \\
\overrightarrow{\xi}_k^{[i+1]} &= \overrightarrow{\theta}_k^{[i+1]} + \frac{1}{2} \log \left|2\pi \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}\right| + \frac{1}{2} \left[\overrightarrow{g}_{\bar{x},k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{g}_{\bar{x},k}^{[i+1]}.
\end{aligned}
$$

To get the final form of the *quadratic* potential $\overrightarrow{V}_k^{[i+1]}(x_k)$, we add the contribution of the quadratic log-measurement and log-prior determined by Definition 2

$$
\overrightarrow{R}_k^{[i+1]} = L_k^{[i]} + \frac{1}{1-\beta} \overrightarrow{S}_k^{[i+1]}, \quad \overrightarrow{r}_k^{[i+1]} = l_k^{[i]} + \frac{1}{1-\beta} \overrightarrow{s}_k^{[i+1]}, \quad \overrightarrow{\rho}_k^{[i+1]} = \nu_k^{[i]} + \frac{1}{1-\beta} \overrightarrow{\xi}_k^{[i+1]}.
$$

Finally, for a Gaussian $\overrightarrow{q}_0^{[i]}(x_0)$ and a quadratic $\overrightarrow{V}_0^{[i+1]}(x_0)$, we derive a quadratic $\log \overrightarrow{\mathcal{Z}}_0^{[i+1]}$

$$
\log \overrightarrow{\mathcal{Z}}_0^{[i+1]} = \log \int \left[\overrightarrow{q}_0^{[i]}(x_0)\right]^\beta \left[\exp\left\{\overrightarrow{V}_0^{[i+1]}(x_0)\right\}\right]^{1-\beta} \mathrm{d}x_0.
$$

Again, we formulate a quadratic function over x_0 and $\overrightarrow{m}_0^{[i]}$

$$
-\frac{1}{2} \begin{bmatrix} x_0^\top & \left[\overrightarrow{m}_0^{[i]}\right]^\top \end{bmatrix} \begin{bmatrix} \overrightarrow{J}_{xx}^{[i+1]} & -\overrightarrow{J}_{xm}^{[i+1]} \\ -\overrightarrow{J}_{mx}^{[i+1]} & \overrightarrow{J}_{mm}^{[i+1]} \end{bmatrix} \begin{bmatrix} x_0 \\ \overrightarrow{m}_0^{[i]} \end{bmatrix} + \begin{bmatrix} x_0^\top & \left[\overrightarrow{m}_0^{[i]}\right]^\top \end{bmatrix} \begin{bmatrix} \overrightarrow{j}_x^{[i+1]} \\ \overrightarrow{j}_m^{[i+1]} \end{bmatrix} + \overrightarrow{\tau}^{[i+1]}
$$

$$= \beta \log \overrightarrow{q}_0^{[i]}(x_0) + (1-\beta)\, \overrightarrow{V}_0^{[i+1]}(x_0),$$

where by matching the quadratic factors, we get

$$\begin{aligned}
\overrightarrow{J}_{xx}^{[i+1]} &= (1-\beta)\, \overrightarrow{R}_0^{[i+1]} + \beta \left[\overrightarrow{P}_0^{[i]}\right]^{-1}, & \overrightarrow{J}_{xm}^{[i+1]} &= \beta \left[\overrightarrow{P}_0^{[i]}\right]^{-1}, \\
\overrightarrow{J}_{mm}^{[i+1]} &= \beta \left[\overrightarrow{P}_0^{[i]}\right]^{-1}, & \overrightarrow{J}_{mx}^{[i+1]} &= \beta \left[\overrightarrow{P}_0^{[i]}\right]^{-1}, \\
\overrightarrow{j}_x^{[i+1]} &= (1-\beta)\, \overrightarrow{r}_0^{[i+1]}, & \overrightarrow{j}_m^{[i+1]} &= 0, \\
\overrightarrow{\tau}^{[i+1]} &= (1-\beta)\, \overrightarrow{\rho}_0^{[i+1]} - \frac{\beta}{2} \log \left| 2\pi\, \overrightarrow{P}_0^{[i]} \right|.
\end{aligned}$$

Using the identity (49), we get a log-normalizer as a quadratic function over $\overrightarrow{m}_0^{[i]}$

$$\log \overrightarrow{\mathcal{Z}}_0^{[i+1]} = -\frac{1}{2} \left[\overrightarrow{m}_0^{[i]}\right]^\top \overrightarrow{U}^{[i+1]}\, \overrightarrow{m}_0^{[i]} + \left[\overrightarrow{m}_0^{[i]}\right]^\top \overrightarrow{u}^{[i+1]} + \overrightarrow{\eta}^{[i+1]},$$

where

$$\begin{aligned}
\overrightarrow{U}^{[i+1]} &= \overrightarrow{J}_{mm}^{[i+1]} - \left[\overrightarrow{J}_{xm}^{[i+1]}\right]^\top \left[\overrightarrow{J}_{xx}^{[i+1]}\right]^{-1} \overrightarrow{J}_{xm}^{[i+1]}, \\
\overrightarrow{u}^{[i+1]} &= \overrightarrow{j}_m^{[i+1]} - \left[\overrightarrow{J}_{xm}^{[i+1]}\right]^\top \left[\overrightarrow{J}_{xx}^{[i+1]}\right]^{-1} \overrightarrow{j}_x^{[i+1]}, \\
\overrightarrow{\eta}^{[i+1]} &= \overrightarrow{\tau}^{[i+1]} - \frac{1}{2} \left| 2\pi\, \overrightarrow{J}_{xx}^{[i+1]} \right]^{-1} \Big| + \frac{1}{2} \left[\overrightarrow{j}_x^{[i+1]}\right]^\top \left[\overrightarrow{J}_{xx}^{[i+1]}\right]^{-1} \left[\overrightarrow{j}_x^{[i+1]}\right].
\end{aligned}$$

Plugging in the corresponding terms leads to the final result.

E Proof of Proposition 5

For $k = 0$, the log-prior is a quadratic function per Definition 2, leading to the following parameterization of the potential

$$\begin{aligned}
\overleftarrow{V}_0^{[i+1]}(x_0) = \log p_0(x_0) &= -\frac{1}{2} x_0^\top \overleftarrow{R}_0^{[i+1]}\, x_0 + x_0^\top \overleftarrow{r}_0^{[i+1]} + \overleftarrow{\rho}_0^{[i+1]} \\
&= -\frac{1}{2} x_0^\top L_0^{[i]}\, x_0 + x_0^\top l_0^{[i]} + \nu_0^{[i]}.
\end{aligned}$$

For $\overleftarrow{V}_k^{[i+1]}(x_k)$, for all $0 < k \leq T$, we have derived a forward recursion

$$\begin{aligned}
\overleftarrow{V}_k^{[i+1]}(x_k) &= \log h_k(y_k \mid x_k) + 1/(1-\beta) \log \overleftarrow{\psi}_k^{[i+1]}(x_k) \\
&= \log h_k(y_k \mid x_k) + 1/(1-\beta) \log \int \left[\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)\right]^\beta \\
&\qquad \times \left[f_{k-1}(x_k \mid x_{k-1}) \exp\left\{ \overleftarrow{V}_{k-1}^{[i+1]}(x_{k-1}) \right\} \right]^{1-\beta} \mathrm{d}x_{k-1}.
\end{aligned}$$

We show that this recursion is a tractable forward propagation of quadratic forms. We start by moving all terms within the integral into the exponential and treat the exponent as a quadratic function over x_k and x_{k-1}

$$\begin{aligned}
&-\frac{1}{2} \begin{bmatrix} x_k^\top & x_{k-1}^\top \end{bmatrix} \begin{bmatrix} \overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} & -\overleftarrow{G}_{\bar{x}x,k}^{[i+1]} \\ -\overleftarrow{G}_{x\bar{x},k}^{[i+1]} & \overleftarrow{G}_{xx,k}^{[i+1]} \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} + \begin{bmatrix} x_k^\top & x_{k-1}^\top \end{bmatrix} \begin{bmatrix} \overleftarrow{g}_{\bar{x},k}^{[i+1]} \\ \overleftarrow{g}_{x,k}^{[i+1]} \end{bmatrix} + \overleftarrow{\theta}_k^{[i+1]} \\
&= \beta \log \overleftarrow{q}_k^{[i]}(x_k \mid x_{k-1}) + (1-\beta) \log f_{k-1}(x_k \mid x_{k-1}) + (1-\beta)\, \overleftarrow{V}_{k-1}^{[i+1]}(x_{k-1}).
\end{aligned}$$

For an affine-Gaussian $\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)$ (Assumption 4), a quadratic $\log f_{k-1}(x_k \mid x_{k-1})$ (Definition 2), and a quadratic potential function $\overleftarrow{V}_{k+1}^{[i+1]}(x_k)$ of the form (17), we get

$$\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} = (1-\beta)\, C_{\bar{x}\bar{x},k-1}^{[i]} + \beta \left[\overleftarrow{F}_k^{[i]}\right]^\top \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{F}_k^{[i]},$$

$$
\begin{aligned}
\overleftarrow{G}_{xx,k}^{[i+1]} &= (1-\beta)\left[C_{xx,k-1}^{[i]} + \overleftarrow{R}_{k-1}^{[i+1]}\right] + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}, \\
\overleftarrow{G}_{\bar{x}x,k}^{[i+1]} &= (1-\beta)\, C_{\bar{x}x,k-1}^{[i]} + \beta\left[\overleftarrow{F}_k^{[i]}\right]^\top \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}, \\
\overleftarrow{G}_{x\bar{x},k}^{[i+1]} &= (1-\beta)\, C_{x\bar{x},k-1}^{[i]} + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{F}_k^{[i]}, \\
\overleftarrow{g}_{\bar{x},k}^{[i+1]} &= (1-\beta)\, c_{\bar{x},k-1}^{[i]} - \beta\left[\overleftarrow{F}_k^{[i]}\right]^\top \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{d}_k^{[i]}, \\
\overleftarrow{g}_{x,k}^{[i+1]} &= (1-\beta)\left[c_{x,k-1}^{[i]} + \overleftarrow{r}_{k-1}^{[i+1]}\right] + \beta\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{d}_k^{[i]}, \\
\overleftarrow{\theta}_k^{[i+1]} &= (1-\beta)\left[\kappa_k^{[i]} + \overleftarrow{\rho}_{k-1}^{[i+1]}\right] - \frac{\beta}{2}\log\left|2\pi\overleftarrow{\Sigma}_k^{[i]}\right| - \frac{\beta}{2}\left[\overleftarrow{d}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{d}_k^{[i]}.
\end{aligned}
$$

By writing this quadratic function explicitly in terms of x_{k-1}

$$
-\frac{1}{2}x_{k-1}^\top \overleftarrow{G}_{xx,k}^{[i+1]}\, x_{k-1} + x_{k-1}^\top\left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\, x_k + \overleftarrow{g}_{x,k}^{[i+1]}\right] + \left[-\frac{1}{2}x_k^\top \overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\, x_k + x_k^\top \overleftarrow{g}_{\bar{x},k}^{[i+1]} + \overleftarrow{\theta}_k^{[i+1]}\right].
$$

we can make use of identity (49) to express the log-normalizing function $\log \overleftarrow{\psi}_k^{[i+1]}(x_k)$ itself as a quadratic function in x_k

$$
\begin{aligned}
\log \overleftarrow{\psi}_k^{[i+1]}(x_k) &= -\frac{1}{2}x_k^\top \overleftarrow{S}_k^{[i+1]}\, x_k + x_k^\top \overleftarrow{s}_k^{[i+1]} + \overleftarrow{\xi}_k^{[i+1]} \\
&= \frac{1}{2}\log\left|2\pi\left[\overrightarrow{G}_{xx,k}^{[i+1]}\right]^{-1}\right| + \left[-\frac{1}{2}x_k^\top \overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\, x_k + x_k^\top \overleftarrow{g}_{\bar{x},k}^{[i+1]} + \overleftarrow{\theta}_k^{[i+1]}\right] \\
&\quad + \frac{1}{2}\left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\, x_k + \overrightarrow{g}_{x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right]^{-1}\left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\, x_k + \overrightarrow{g}_{x,k}^{[i+1]}\right]
\end{aligned}
$$

which, after matching terms, leads to

$$
\begin{aligned}
\overleftarrow{S}_k^{[i+1]} &= \overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} - \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{G}_{x\bar{x},k}^{[i+1]}, \\
\overleftarrow{s}_k^{[i+1]} &= \overleftarrow{g}_{\bar{x},k}^{[i+1]} + \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{g}_{x,k}^{[i+1]}, \\
\overleftarrow{\xi}_k^{[i+1]} &= \overleftarrow{\theta}_k^{[i+1]} + \frac{1}{2}\log\left|2\pi\left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}\right| + \frac{1}{2}\left[\overleftarrow{g}_{x,k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{g}_{x,k}^{[i+1]}.
\end{aligned}
$$

Given $\log \overleftarrow{\psi}_k^{[i+1]}(x_k)$, we can now construct the *quadratic* potential function $\overleftarrow{V}_k^{[i+1]}(x_k)$ that accounts for the log-measurement contribution according to Definition 2

$$
\overleftarrow{R}_k^{[i+1]} = L_k^{[i]} + \frac{1}{1-\beta}\overleftarrow{S}_k^{[i+1]}, \quad \overleftarrow{r}_k^{[i+1]} = l_k^{[i]} + \frac{1}{1-\beta}\overleftarrow{s}_k^{[i+1]}, \quad \overleftarrow{\rho}_k^{[i+1]} = \nu_k^{[i]} + \frac{1}{1-\beta}\overleftarrow{\xi}_k^{[i+1]}.
$$

Finally, for a Gaussian $\overleftarrow{q}_T^{[i]}(x_T)$ and a quadratic $\overleftarrow{V}_T^{[i+1]}(x_T)$, we derive a quadratic $\log \overleftarrow{\mathcal{Z}}_T^{[i+1]}$

$$
\log \overleftarrow{\mathcal{Z}}_T^{[i+1]} = \log \int \left[\overleftarrow{q}_T^{[i]}(x_T)\right]^\beta \left[\exp\left\{\overleftarrow{V}_T^{[i+1]}(x_T)\right\}\right]^{1-\beta} \mathrm{d}x_T.
$$

Similar to the proof in Appendix D, we formulate a quadratic over x_T and $\overleftarrow{m}_T^{[i]}$

$$
\begin{aligned}
&-\frac{1}{2}\begin{bmatrix}\left[\overleftarrow{m}_T^{[i]}\right]^\top & x_T^\top\end{bmatrix}\begin{bmatrix}\overleftarrow{J}_{xx}^{[i+1]} & -\overleftarrow{J}_{xm}^{[i+1]} \\ -\overleftarrow{J}_{mx}^{[i+1]} & \overleftarrow{J}_{mm}^{[i+1]}\end{bmatrix}\begin{bmatrix}\overrightarrow{m}_T^{[i]} \\ x_T\end{bmatrix} + \begin{bmatrix}\left[\overleftarrow{m}_T^{[i]}\right]^\top & x_T^\top\end{bmatrix}\begin{bmatrix}\overleftarrow{j}_x^{[i+1]} \\ \overleftarrow{j}_m^{[i+1]}\end{bmatrix} + \overleftarrow{\tau}^{[i+1]} \\
&= \beta\log \overleftarrow{q}_T^{[i]}(x_T) + (1-\beta)\overleftarrow{V}_T^{[i+1]}(x_T),
\end{aligned}
$$

where by matching the quadratic factors, we get

$$
\overleftarrow{J}_{xx}^{[i+1]} = \beta\left[\overleftarrow{P}_T^{[i]}\right]^{-1}, \qquad \overleftarrow{J}_{xm}^{[i+1]} = \beta\left[\overleftarrow{P}_T^{[i]}\right]^{-1},
$$

$$\overleftarrow{J}_{mm}^{[i+1]} = (1-\beta)\,\overleftarrow{R}_T^{[i+1]} + \beta\left[\overleftarrow{P}_T^{[i]}\right]^{-1}, \qquad \overleftarrow{J}_{mx}^{[i+1]} = \beta\left[\overleftarrow{P}_T^{[i]}\right]^{-1},$$
$$\overleftarrow{j}_x^{[i+1]} = 0, \qquad \overleftarrow{j}_m^{[i+1]} = (1-\beta)\,\overleftarrow{r}_T^{[i+1]},$$
$$\overleftarrow{\tau}^{[i+1]} = (1-\beta)\,\overleftarrow{\rho}_T^{[i+1]} - \frac{\beta}{2}\log\left|2\pi\,\overleftarrow{P}_T^{[i]}\right|.$$

Using the identity (49), we get a log-normalizing constant as a quadratic function over $\overleftarrow{m}_T^{[i]}$

$$\log\overleftarrow{\mathcal{Z}}_T^{[i+1]} = -\frac{1}{2}\left[\overleftarrow{m}_T^{[i]}\right]^\top \overleftarrow{U}^{[i+1]}\overleftarrow{m}_T^{[i]} + \left[\overleftarrow{m}_T^{[i]}\right]^\top \overleftarrow{u}^{[i+1]} + \overleftarrow{\eta}^{[i+1]},$$

where

$$\overleftarrow{U}^{[i+1]} = \overleftarrow{J}_{xx}^{[i+1]} - \left[\overleftarrow{J}_{mx}^{[i+1]}\right]^\top \left[\overleftarrow{J}_{mm}^{[i+1]}\right]^{-1}\overleftarrow{J}_{mx}^{[i+1]},$$
$$\overleftarrow{u}^{[i+1]} = \overleftarrow{j}_x^{[i+1]} - \left[\overleftarrow{J}_{mx}^{[i+1]}\right]^\top \left[\overleftarrow{J}_{mm}^{[i+1]}\right]^{-1}\overleftarrow{j}_m^{[i+1]},$$
$$\overleftarrow{\eta}^{[i+1]} = \overleftarrow{\tau}^{[i+1]} - \frac{1}{2}\left|2\pi\,\overleftarrow{J}_{mm}^{[i+1]}\right]^{-1}\Big| + \frac{1}{2}\left[\overleftarrow{j}_m^{[i+1]}\right]^\top \left[\overleftarrow{J}_{mm}^{[i+1]}\right]^{-1}\left[\overleftarrow{j}_m^{[i+1]}\right].$$

Plugging in the corresponding terms leads to the final result.

F Proof of Lemma 1

For a tilted distribution of the form

$$q^{[i+1]}(x) = \left[\mathcal{Z}^{[i+1]}\right]^{-1}\left[q^{[i]}(x)\right]^{\beta}\left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta},$$

with a normalizing constant

$$\mathcal{Z}^{[i+1]} = \int\left[q^{[i]}(x)\right]^{\beta}\left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta}\mathrm{d}x,$$

where $q^{[i]}(x) = \mathcal{N}(x \mid m^{[i]}, P^{[i]})$, we compute the moments of $q^{[i+1]}(x)$ by considering the derivatives of $\mathcal{Z}^{[i+1]}$ with respect to $m^{[i]}$. Starting with the first-order derivative

$$\begin{aligned}
\frac{\partial\mathcal{Z}^{[i+1]}}{\partial m^{[i]}} &= \beta\left[P^{[i]}\right]^{-1}\int(x-m^{[i]})\left[q^{[i]}(x)\right]^{\beta}\left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta}\mathrm{d}x \\
&= \beta\left[P^{[i]}\right]^{-1}\int x\left[q^{[i]}(x)\right]^{\beta}\left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta}\mathrm{d}x \\
&\quad - \beta\left[P^{[i]}\right]^{-1} m^{[i]}\int\left[q^{[i]}(x)\right]^{\beta}\left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta}\mathrm{d}x \\
&= \beta\,\mathcal{Z}^{[i+1]}\left[P^{[i]}\right]^{-1}\left[\mathbb{E}^{[i+1]}[x] - m^{[i]}\right].
\end{aligned}$$

After rearranging the terms, we get

$$\begin{aligned}
\mathbb{E}^{[i+1]}[x] &= m^{[i]} + \frac{1}{\beta}P^{[i]}\left[\mathcal{Z}^{[i+1]}\right]^{-1}\frac{\partial\mathcal{Z}^{[i+1]}}{\partial m^{[i]}} \\
&= m^{[i]} + \frac{1}{\beta}P^{[i]}\frac{\partial\log\mathcal{Z}^{[i+1]}}{\partial m^{[i]}},
\end{aligned}$$

which is an expression for the first moment of $q^{[i+1]}(x)$.

Next, we consider the second-order derivative $\mathcal{Z}^{[i+1]}$ with respect to $m^{[i]}$

$$\begin{aligned}
\frac{\partial^2\mathcal{Z}^{[i+1]}}{\partial m^{[i]}\,\partial\left[m^{[i]}\right]^\top} &= \beta^2\left[P^{[i]}\right]^{-1}\left[\int(x-m^{[i]})(x-m^{[i]})^\top\left[q^{[i]}(x)\right]^{\beta}\left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta}\mathrm{d}x\right]\left[P^{[i]}\right]^{-1} \\
&\quad - \beta\left[P^{[i]}\right]^{-1}\int\left[q^{[i]}(x)\right]^{\beta}\left[\exp\left\{V^{[i+1]}(x)\right\}\right]^{1-\beta}\mathrm{d}x
\end{aligned}$$

$$= \beta^2\, \mathcal{Z}^{[i+1]} \left[P^{[i]}\right]^{-1} \mathbb{E}^{[i+1]} \left[x\, x^\top\right] \left[P^{[i]}\right]^{-1} + \beta^2\, \mathcal{Z}^{[i+1]} \left[P^{[i]}\right]^{-1} m^{[i]} \left[m^{[i]}\right]^\top \left[P^{[i]}\right]^{-1}$$
$$- 2\,\beta^2\, \mathcal{Z}^{[i+1]} \left[P^{[i]}\right]^{-1} \mathbb{E}^{[i+1]} [x] \left[m^{[i]}\right]^\top \left[P^{[i]}\right]^{-1} - \beta\, \mathcal{Z}^{[i+1]} \left[P^{[i]}\right]^{-1},$$

which leads to the second *raw* moment of $q^{[i+1]}(x_k)$

$$\mathbb{E}^{[i+1]} \left[x\, x^\top\right] = -\, m^{[i]} \left[m^{[i]}\right]^\top + 2\, \mathbb{E}^{[i+1]} [x] \left[m^{[i]}\right]^\top$$
$$+ \frac{1}{\beta}\, P^{[i]} + \frac{1}{\beta^2} \left[\mathcal{Z}^{[i+1]}\right]^{-1} P^{[i]}\, \frac{\partial^2 \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}\, \partial \left[m^{[i]}\right]^\top}\, P^{[i]}.$$

We can now derive the second *central* moment as follows

$$\mathbb{V}^{[i+1]} [x] = \mathbb{E}^{[i+1]} \left[x\, x^\top\right] - \mathbb{E}^{[i+1]} [x]\, \mathbb{E}^{[i+1]} [x]^\top$$
$$= m^{[i]} \left[m^{[i]}\right]^\top + 2\, \frac{1}{\beta} \left[\mathcal{Z}^{[i+1]}\right]^{-1} P^{[i]}\, \frac{\partial \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}} \left[m^{[i]}\right]^\top + \frac{1}{\beta}\, P^{[i]}$$
$$+ \frac{1}{\beta^2} \left[\mathcal{Z}^{[i+1]}\right]^{-1} P^{[i]}\, \frac{\partial^2 \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}\, \partial \left[m^{[i]}\right]^\top}\, P^{[i]} - m^{[i]} \left[m^{[i]}\right]^\top$$
$$- 2 \frac{1}{\beta} \left[\mathcal{Z}^{[i+1]}\right]^{-1} m^{[i]} \left[P^{[i]}\, \frac{\partial \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}}\right]^\top - \frac{1}{\beta^2} \left[\mathcal{Z}^{[i+1]}\right]^{-2} P^{[i]}\, \frac{\partial \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}} \left[\frac{\partial \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}}\right]^\top P^{[i]}$$
$$= \frac{1}{\beta}\, P^{[i]} + \frac{1}{\beta^2} \left[\mathcal{Z}^{[i+1]}\right]^{-1} P^{[i]}\, \frac{\partial^2 \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]} \partial \left[m^{[i]}\right]^\top}\, P^{[i]} - \frac{1}{\beta^2} \left[\mathcal{Z}^{[i+1]}\right]^{-2} P^{[i]}\, \frac{\partial \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}} \left[\frac{\partial \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}}\right]^\top P^{[i]}$$
$$= \frac{1}{\beta}\, P^{[i]} + \frac{1}{\beta^2}\, P^{[i]} \frac{\partial^2 \log \mathcal{Z}^{[i+1]}}{\partial\, m^{[i]}\, \partial \left[m^{[i]}\right]^\top}\, P^{[i]}.$$

G Proof of Proposition 6

For a tilted forward-Markov conditional distribution of the form

$$\overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k) = \left[\overrightarrow{\psi}_k^{[i+1]}(x_k)\right]^{-1} \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^{\beta} \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta},$$

with a normalizing function

$$\overrightarrow{\psi}_k^{[i+1]}(x_k) = \int \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^{\beta} \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \mathrm{d}x_{k+1},$$

where $\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k) = \mathcal{N}(x_{k+1} \mid \overrightarrow{F}_k^{[i]}\, x_k + \overrightarrow{d}_k^{[i]}, \overrightarrow{\Sigma}_k^{[i]})$ and $f_k(x_{k+1} \mid x_k) \approx \exp\left\{\ell_f^{[i]}(x_{k+1}, x_k)\right\}$

$$\ell_f^{[i]}(x_{k+1}, x_k) \approx -\frac{1}{2} \begin{bmatrix} x_{k+1}^\top & x_k^\top \end{bmatrix} \begin{bmatrix} C_{\bar{x}\bar{x},k}^{[i]} & -C_{\bar{x}x,k}^{[i]} \\ -C_{x\bar{x},k}^{[i]} & C_{xx,k}^{[i]} \end{bmatrix} \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} + \begin{bmatrix} x_{k+1}^\top & x_k^\top \end{bmatrix} \begin{bmatrix} c_{\bar{x},k}^{[i]} \\ c_{x,k}^{[i]} \end{bmatrix} + \kappa_k^{[i]}.$$

We compute the *conditional* moments of $q^{[i+1]}(y \mid x)$ from the derivatives of $\psi^{[i+1]}(x)$ with respect to x. Starting with the first-order derivative

$$\frac{\partial \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k} = \beta \int \frac{\partial \log \overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)}{\partial x_k} \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^{\beta}$$
$$\times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \mathrm{d}x_{k+1}$$
$$+ (1-\beta) \int \frac{\partial\, \log f_k(x_{k+1} \mid x_k)}{\partial x_k} \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^{\beta}$$
$$\times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \mathrm{d}x_{k+1},$$

$$
\begin{aligned}
&= \beta \int \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \left[x_{k+1} - \overrightarrow{F}_k^{[i]}\, x_k - \overrightarrow{d}_k^{[i]}\right] \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^\beta \\
&\qquad \times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \mathrm{d}x_{k+1} \\
&\quad + (1-\beta) \int \left[-C_{xx,k}^{[i]}\, x_k + C_{x\bar{x},k}^{[i]}\, x_{k+1} + c_{x,k}^{[i]}\right] \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^\beta \\
&\qquad \times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \mathrm{d}x_{k+1}, \\
&= \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[(1-\beta)\, C_{x\bar{x},k}^{[i]} + \beta \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\right] \mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] \\
&\quad - \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[(1-\beta)\, C_{xx,k}^{[i]} + \beta \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{F}_k^{[i]}\right] x_k \\
&\quad + \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[(1-\beta)\, c_{x,k}^{[i]} - \beta \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{d}_k^{[i]}\right] \\
&= \overrightarrow{\psi}_k^{[i+1]}(x_k)\, \overrightarrow{G}_{x\bar{x},k}^{[i+1]}\, \mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] - \overrightarrow{\psi}_k^{[i+1]}(x_k)\, \overrightarrow{G}_{xx,k}^{[i+1]}\, x + \overrightarrow{\psi}_k^{[i+1]}(x_k)\, \overrightarrow{g}_{x,k}^{[i+1]},
\end{aligned}
$$

where we use the definition of $\overrightarrow{G}_{x\bar{x},k}^{[i+1]}$, $\overrightarrow{G}_{xx,k}^{[i+1]}$, and $\overrightarrow{g}_{x,k}^{[i+1]}$ from Proposition 4.

Consequently, the first conditional moment of $\overrightarrow{q}_k^{[i+1]}(x_{k+1} \mid x_k)$ is

$$
\mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] = \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{G}_{xx,k}^{[i+1]}\, x_k - \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{g}_{x,k}^{[i+1]} + \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \frac{\partial \log \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k}.
$$

Now we consider the second-order derivatives of $\overrightarrow{\psi}_k^{[i+1]}(x_k)(x)$ with respect to x_k

$$
\begin{aligned}
\frac{\partial^2 \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k\, \partial x_k^\top} &= \beta^2 \int \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \left[x_{k+1} - \overrightarrow{F}_k^{[i]}\, x_k - \overrightarrow{d}_k^{[i]}\right] \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^\beta \\
&\qquad \times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \\
&\qquad \times \left[x_{k+1} - \overrightarrow{F}_k^{[i]}\, x_k - \overrightarrow{d}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{F}_k^{[i]}\, \mathrm{d}x_{k+1} \\
&\quad + \beta\,(1-\beta) \int \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \left[x_{k+1} - \overrightarrow{F}_k^{[i]}\, x_k - \overrightarrow{d}_k^{[i]}\right] \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^\beta \\
&\qquad \times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \\
&\qquad \times \left[-C_{xx,k}^{[i]}\, x + C_{x\bar{x},k}^{[i]}\, y + c_{x,k}^{[i]}\right]^\top \mathrm{d}x_{k+1} \\
&\quad + (1-\beta)^2 \int \left[-C_{xx,k}^{[i]}\, x + C_{x\bar{x},k}^{[i]}\, y + c_{x,k}^{[i]}\right] \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^\beta \\
&\qquad \times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \\
&\qquad \times \left[-C_{xx,k}^{[i]}\, x + C_{x\bar{x},k}^{[i]}\, y + c_{x,k}^{[i]}\right]^\top \mathrm{d}x_{k+1} \\
&\quad + \beta\,(1-\beta) \int \left[-C_{xx,k}^{[i]}\, x + C_{x\bar{x},k}^{[i]}\, y + c_{x,k}^{[i]}\right] \left[\overrightarrow{q}_k^{[i]}(x_{k+1} \mid x_k)\right]^\beta \\
&\qquad \times \left[f_k(x_{k+1} \mid x_k) \exp\left\{\overrightarrow{V}_{k+1}^{[i+1]}(x_{k+1})\right\}\right]^{1-\beta} \\
&\qquad \times \left[x_{k+1} - \overrightarrow{F}_k^{[i]}\, x_k - \overrightarrow{d}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{F}_k^{[i]}\, \mathrm{d}x_{k+1} \\
&\quad - \beta\, \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{F}_k^{[i]} - (1-\beta)\, \overrightarrow{\psi}_k^{[i+1]}(x_k)\, C_{xx,k}^{[i]}
\end{aligned}
$$

$$
\begin{aligned}
&= \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k+1}\, x_{k+1}^\top \mid x_k\right] \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right] \\
&\quad + \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right] \left[x_k\, x_k^\top\right] \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right]^\top \\
&\quad + \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[\overrightarrow{g}_{x,k}^{[i+1]}\right] \left[\overrightarrow{g}_{x,k}^{[i+1]}\right]^\top \\
&\quad - 2\, \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] x_k^\top \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right]^\top \\
&\quad + 2\, \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] \left[\overrightarrow{g}_{x,k}^{[i+1]}\right]^\top \\
&\quad - 2\, \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right] x_k \left[\overrightarrow{g}_{x,k}^{[i+1]}\right]^\top \\
&\quad - \overrightarrow{\psi}_k^{[i+1]}(x_k) \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right].
\end{aligned}
$$

This gives us an expression for the second *raw* conditional moment

$$
\begin{aligned}
\mathbb{E}^{[i+1]}\left[x_{k+1}\, x_{k+1}^\top \mid x_k\right] = &\left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right] \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&+ \left[\overrightarrow{\psi}_k^{[i+1]}(x_k)\right]^{-1} \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \frac{\partial^2 \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k\, \partial x_k^\top} \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&- \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right] \left[x_k\, x_k^\top\right] \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&- \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{g}_{x,k}^{[i+1]}\right] \left[\overrightarrow{g}_{x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&+ 2\, \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] x_k^\top \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&- 2\, \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] \left[\overrightarrow{g}_{x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&+ 2\, \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right] x_k \left[\overrightarrow{g}_{x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1}.
\end{aligned}
$$

Finally, we can derive the second *central* moment according to

$$
\begin{aligned}
\mathbb{V}^{[i+1]}\left[x_{k+1} \mid x_k\right] &= \mathbb{E}^{[i+1]}\left[x_{k+1}\, x_{k+1}^\top \mid x_k\right] - \mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] \mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right]^\top \\
&= \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right] \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&\quad + \left[\overrightarrow{\psi}_k^{[i+1]}(x_k)\right]^{-1} \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \frac{\partial^2 \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k\, \partial x_k^\top} \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&\quad - \left[\overrightarrow{\psi}_k^{[i+1]}(x_k)\right]^{-2} \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \frac{\partial \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k} \left[\frac{\partial \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k}\right]^\top \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \\
&= \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{xx,k}^{[i+1]}\right] \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} + \left[\overrightarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \frac{\partial^2 \log \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k\, \partial x_k^\top} \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1}.
\end{aligned}
$$

When $\log \overrightarrow{\psi}_k^{[i+1]}(x_k)$ is evaluated under the quadratic approximation from Proposition 4, its derivatives are

$$
\frac{\partial \log \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k} = -\overrightarrow{G}_{xx,k}^{[i+1]}\, x_k + \overrightarrow{g}_{x,k}^{[i+1]} + \overrightarrow{G}_{x\bar{x},k}^{[i+1]} \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\, x_k + \overrightarrow{g}_{\bar{x},k}^{[i+1]}\right],
$$

and

$$
\frac{\partial^2 \log \overrightarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k\, \partial x_k^\top} = -\overrightarrow{G}_{xx,k}^{[i+1]} + \overrightarrow{G}_{x\bar{x},k}^{[i+1]} \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{G}_{\bar{x}x,k}^{[i+1]}.
$$

Substituting these expressions into the derivative identities above gives

$$\mathbb{E}^{[i+1]}\left[x_{k+1} \mid x_k\right] = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}\left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} x_k + \overrightarrow{g}_{\bar{x},k}^{[i+1]}\right], \qquad \mathbb{V}^{[i+1]}\left[x_{k+1} \mid x_k\right] = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}.$$

Therefore, matching these moments to the affine-Gaussian conditional yields

$$\overrightarrow{\Sigma}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}, \qquad \overrightarrow{F}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}, \qquad \overrightarrow{d}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}\overrightarrow{g}_{\bar{x},k}^{[i+1]}.$$

Substituting the definitions of $\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}$, $\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}$, and $\overrightarrow{g}_{\bar{x},k}^{[i+1]}$ from Proposition 4, we obtain

$$\begin{aligned}
\overrightarrow{\Sigma}_k^{[i+1]} &= \left[(1-\beta)\left[C_{\bar{x}\bar{x},k}^{[i]} + \overrightarrow{R}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\right]^{-1}, \\
\overrightarrow{F}_k^{[i+1]} &= \overrightarrow{\Sigma}_k^{[i+1]}\left[(1-\beta)\, C_{\bar{x}x,k}^{[i]} + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\overrightarrow{F}_k^{[i]}\right], \\
\overrightarrow{d}_k^{[i+1]} &= \overrightarrow{\Sigma}_k^{[i+1]}\left[(1-\beta)\left[c_{\bar{x},k}^{[i]} + \overrightarrow{r}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}\overrightarrow{d}_k^{[i]}\right].
\end{aligned}$$

H Proof of Proposition 7

For a tilted reverse-Markov conditional distribution of the form

$$\overleftarrow{q}_k^{[i+1]}(x_{k-1} \mid x_k) = \left[\overleftarrow{\psi}_k^{[i+1]}(x_k)\right]^{-1}\left[\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)\right]^{\beta}\left[f_{k-1}(x_k \mid x_{k-1})\exp\left\{\overleftarrow{V}_{k-1}^{[i+1]}(x_{k-1})\right\}\right]^{1-\beta},$$

with a normalizing function

$$\overleftarrow{\psi}_k^{[i+1]}(x_k) = \int\left[\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k)\right]^{\beta}\left[f_{k-1}(x_k \mid x_{k-1})\exp\left\{\overleftarrow{V}_{k-1}^{[i+1]}(x_{k-1})\right\}\right]^{1-\beta}\mathrm{d}x_{k-1},$$

where $\overleftarrow{q}_k^{[i]}(x_{k-1} \mid x_k) = \mathcal{N}(x_{k-1} \mid \overleftarrow{F}_k^{[i]} x_k + \overleftarrow{d}_k^{[i]}, \overleftarrow{\Sigma}_k^{[i]})$ and $f_{k-1}(x_k \mid x_{k-1}) \approx \exp\left\{\ell_f^{[i]}(x_k, x_{k-1})\right\}$

$$\ell_f^{[i]}(x_k, x_{k-1}) \approx -\frac{1}{2}\begin{bmatrix} x_k^\top & x_{k-1}^\top \end{bmatrix}\begin{bmatrix} C_{\bar{x}\bar{x},k}^{[i]} & -C_{\bar{x}x,k}^{[i]} \\ -C_{x\bar{x},k}^{[i]} & C_{xx,k}^{[i]} \end{bmatrix}\begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} + \begin{bmatrix} x_k^\top & x_{k-1}^\top \end{bmatrix}\begin{bmatrix} c_{\bar{x},k}^{[i]} \\ c_{x,k}^{[i]} \end{bmatrix} + \kappa_k^{[i]}.$$

The following steps are analogous to those described in Appendix G, thus we will omit redundant steps to avoid repetition. Starting with the first-order derivative of $\overleftarrow{\psi}_k^{[i+1]}(x_k)$

$$\begin{aligned}
\frac{\partial\overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k} &= \overleftarrow{\psi}_k^{[i+1]}(x_k)\left[(1-\beta)\, C_{\bar{x}x,k}^{[i]} + \beta\left[\overleftarrow{F}_k^{[i]}\right]^\top\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\right]\mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] \\
&\quad - \overleftarrow{\psi}_k^{[i+1]}(x_k)\left[(1-\beta)\, C_{\bar{x}\bar{x},k}^{[i]} + \beta\left[\overleftarrow{F}_k^{[i]}\right]^\top\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\overleftarrow{F}_k^{[i]}\right] x_k \\
&\quad + \overleftarrow{\psi}_k^{[i+1]}(x_k)\left[(1-\beta)\, c_{\bar{x},k}^{[i]} - \beta\left[\overleftarrow{F}_k^{[i]}\right]^\top\left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\overleftarrow{d}_k^{[i]}\right] \\
&= \overleftarrow{\psi}_k^{[i+1]}(x_k)\,\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\,\mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] - \overleftarrow{\psi}_k^{[i+1]}(x_k)\,\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\, x + \overleftarrow{\psi}_k^{[i+1]}(x_k)\,\overleftarrow{g}_{\bar{x},k}^{[i+1]},
\end{aligned}$$

where we use the definition of $\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}$, $\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}$, and $\overleftarrow{g}_{\bar{x},k}^{[i+1]}$ from Proposition 5. Consequently, the first conditional moment of $\overleftarrow{q}_k^{[i+1]}(x_{k-1} \mid x_k)$ is

$$\mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] = \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1}\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\, x_k - \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1}\overleftarrow{g}_{\bar{x},k}^{[i+1]} + \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1}\frac{\partial\log\overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k}.$$

Now we consider the second-order derivatives of $\overleftarrow{\psi}_k^{[i+1]}(x_k)(x)$ with respect to x_k

$$\begin{aligned}
\frac{\partial^2 \overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k \, \partial x_k^\top} = & \overleftarrow{\psi}_k^{[i+1]}(x_k) \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k-1}\, x_{k-1}^\top \mid x_k\right] \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right] \\
& + \overleftarrow{\psi}_k^{[i+1]}(x_k) \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right] \left[x_k\, x_k^\top\right] \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^\top \\
& + \overleftarrow{\psi}_k^{[i+1]}(x_k) \left[\overleftarrow{g}_{\bar{x},k}^{[i+1]}\right] \left[\overleftarrow{g}_{\bar{x},k}^{[i+1]}\right]^\top \\
& - 2\, \overleftarrow{\psi}_k^{[i+1]}(x_k) \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] x_k^\top \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^\top \\
& + 2\, \overleftarrow{\psi}_k^{[i+1]}(x_k) \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] \left[\overleftarrow{g}_{\bar{x},k}^{[i+1]}\right]^\top \\
& - 2\, \overleftarrow{\psi}_k^{[i+1]}(x_k) \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right] x_k \left[\overleftarrow{g}_{\bar{x},k}^{[i+1]}\right]^\top \\
& - \overleftarrow{\psi}_k^{[i+1]}(x_k) \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right].
\end{aligned}$$

This gives us an expression for the second *raw* conditional moment

$$\begin{aligned}
\mathbb{E}^{[i+1]}\left[x_{k-1}\, x_{k-1}^\top \mid x_k\right] = & \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right] \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
& + \left[\overleftarrow{\psi}_k^{[i+1]}(x_k)\right]^{-1} \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \frac{\partial^2 \overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k \, \partial x_k^\top} \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
& - \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right] \left[x_k\, x_k^\top\right] \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
& - \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{g}_{\bar{x},k}^{[i+1]}\right] \left[\overleftarrow{g}_{\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
& + 2 \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] x_k^\top \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
& - 2 \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right] \mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] \left[\overleftarrow{g}_{\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
& + 2 \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right] x_k \left[\overleftarrow{g}_{\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1}.
\end{aligned}$$

Finally, we can derive the second *central* moment according to

$$\begin{aligned}
\mathbb{V}^{[i+1]}\left[x_{k-1} \mid x_k\right] &= \mathbb{E}^{[i+1]}\left[x_{k-1}\, x_{k-1}^\top \mid x_k\right] - \mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] \mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right]^\top \\
&= \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right] \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
&\quad + \left[\overleftarrow{\psi}_k^{[i+1]}(x_k)\right]^{-1} \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \frac{\partial^2 \overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k \, \partial x_k^\top} \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
&\quad - \left[\overleftarrow{\psi}_k^{[i+1]}(x_k)\right]^{-2} \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \frac{\partial \overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k} \left[\frac{\partial \overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k}\right]^\top \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} \\
&= \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right] \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1} + \left[\overleftarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^{-1} \frac{\partial^2 \log \overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k \, \partial x_k^\top} \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^{-1}.
\end{aligned}$$

When $\log \overleftarrow{\psi}_k^{[i+1]}(x_k)$ is evaluated under the quadratic approximation from Proposition 5, its derivatives are

$$\frac{\partial \log \overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k} = -\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\, x_k + \overleftarrow{g}_{\bar{x},k}^{[i+1]} + \overleftarrow{G}_{\bar{x}x,k}^{[i+1]} \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\, x_k + \overleftarrow{g}_{x,k}^{[i+1]}\right],$$

and

$$\frac{\partial^2 \log \overleftarrow{\psi}_k^{[i+1]}(x_k)}{\partial x_k \, \partial x_k^\top} = -\overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} + \overleftarrow{G}_{\bar{x}x,k}^{[i+1]} \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{G}_{x\bar{x},k}^{[i+1]}.$$

Substituting these expressions into the derivative identities above gives

$$\mathbb{E}^{[i+1]}\left[x_{k-1} \mid x_k\right] = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]} x_k + \overleftarrow{g}_{x,k}^{[i+1]}\right], \qquad \mathbb{V}^{[i+1]}\left[x_{k-1} \mid x_k\right] = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}.$$

Therefore, matching these moments to the affine-Gaussian conditional yields

$$\overleftarrow{\Sigma}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}, \qquad \overleftarrow{F}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{G}_{x\bar{x},k}^{[i+1]}, \qquad \overleftarrow{d}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{g}_{x,k}^{[i+1]}.$$

Substituting the definitions of $\overleftarrow{G}_{xx,k}^{[i+1]}$, $\overleftarrow{G}_{x\bar{x},k}^{[i+1]}$, and $\overleftarrow{g}_{x,k}^{[i+1]}$ from Proposition 5, we obtain

$$\begin{aligned}
\overleftarrow{\Sigma}_k^{[i+1]} &= \left[(1-\beta)\left[C_{xx,k-1}^{[i]} + \overleftarrow{R}_{k-1}^{[i+1]}\right] + \beta \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}\right]^{-1}, \\
\overleftarrow{F}_k^{[i+1]} &= \overleftarrow{\Sigma}_k^{[i+1]} \left[(1-\beta)\, C_{x\bar{x},k-1}^{[i]} + \beta \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{F}_k^{[i]}\right], \\
\overleftarrow{d}_k^{[i+1]} &= \overleftarrow{\Sigma}_k^{[i+1]} \left[(1-\beta)\left[c_{x,k-1}^{[i]} + \overleftarrow{r}_{k-1}^{[i+1]}\right] + \beta \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{d}_k^{[i]}\right].
\end{aligned}$$

I Recursive Bayesian Inference Algorithms

Algorithm 1 Backward Recursion of the Forward Entropic Variational Smoother

Require: Damping: β, first Gaussian marginal: $\overrightarrow{m}_0^{[i]}, \overrightarrow{P}_0^{[i]}$,
forward affine-Gaussian conditionals: $\overrightarrow{F}_{0:T-1}^{[i]}, \overrightarrow{d}_{0:T-1}^{[i]}, \overrightarrow{\Sigma}_{0:T-1}^{[i]}$,
log prior ℓ_p: $L_0^{[i]}, l_0^{[i]}$, log measurement ℓ_h: $L_{1:T}^{[i]}, l_{1:T}^{[i]}$,
log dynamics ℓ_f: $C_{\bar{x}\bar{x},0:T-1}^{[i]}, C_{\bar{x}x,0:T-1}^{[i]}, C_{xx,0:T-1}^{[i]}, c_{\bar{x},0:T-1}^{[i]}, c_{x,0:T-1}^{[i]}$.

$\overrightarrow{R}_T^{[i+1]} = L_T^{[i]}, \quad \overrightarrow{r}_T^{[i+1]} = l_T^{[i]}$ ▷ initialize backward recursion

for $k \leftarrow T-1, \ldots, 0$ **do**

$\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} = (1-\beta)\left[C_{\bar{x}\bar{x},k}^{[i]} + \overrightarrow{R}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1}$

$\overrightarrow{G}_{xx,k}^{[i+1]} = (1-\beta)\, C_{xx,k}^{[i]} + \beta\left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{F}_k^{[i]}$

$\overrightarrow{G}_{\bar{x}x,k}^{[i+1]} = (1-\beta)\, C_{\bar{x}x,k}^{[i]} + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{F}_k^{[i]}$

$\overrightarrow{g}_{\bar{x},k}^{[i+1]} = (1-\beta)\left[c_{\bar{x},k}^{[i]} + \overrightarrow{r}_{k+1}^{[i+1]}\right] + \beta\left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{d}_k^{[i]}$

$\overrightarrow{g}_{x,k}^{[i+1]} = (1-\beta)\, c_{x,k}^{[i]} - \beta\left[\overrightarrow{F}_k^{[i]}\right]^\top \left[\overrightarrow{\Sigma}_k^{[i]}\right]^{-1} \overrightarrow{d}_k^{[i]}$

$\overrightarrow{S}_k^{[i+1]} = \overrightarrow{G}_{xx,k}^{[i+1]} - \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{G}_{\bar{x}x,k}^{[i+1]}$ ▷ update log-normalizers

$\overrightarrow{s}_k^{[i+1]} = \overrightarrow{g}_{x,k}^{[i+1]} + \left[\overrightarrow{G}_{\bar{x}x,k}^{[i+1]}\right]^\top \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{g}_{\bar{x},k}^{[i+1]}$

$\overrightarrow{R}_k^{[i+1]} = L_k^{[i]} + 1/(1-\beta)\, \overrightarrow{S}_k^{[i+1]}$ ▷ update potential functions

$\overrightarrow{r}_k^{[i+1]} = l_k^{[i]} + 1/(1-\beta)\, \overrightarrow{s}_k^{[i+1]}$

$\overrightarrow{F}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{G}_{\bar{x}x,k}^{[i+1]}$ ▷ update conditional posteriors

$\overrightarrow{d}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1} \overrightarrow{g}_{\bar{x},k}^{[i+1]}$

$\overrightarrow{\Sigma}_k^{[i+1]} = \left[\overrightarrow{G}_{\bar{x}\bar{x},k}^{[i+1]}\right]^{-1}$

end for

$\overrightarrow{m}_0^{[i+1]} = \overrightarrow{P}_0^{[i+1]}\left[(1-\beta)\, \overrightarrow{r}_0^{[i+1]} + \beta\left[\overrightarrow{P}_0^{[i]}\right]^{-1} \overrightarrow{m}_0^{[i]}\right]$ ▷ update boundary marginal

$\overrightarrow{P}_0^{[i+1]} = \left[(1-\beta)\, \overrightarrow{R}_0^{[i+1]} + \beta\left[\overrightarrow{P}_0^{[i]}\right]^{-1}\right]^{-1}$

return $\overrightarrow{m}_0^{[i+1]}, \overrightarrow{P}_0^{[i+1]}, \overrightarrow{F}_{0:T-1}^{[i+1]}, \overrightarrow{d}_{0:T-1}^{[i+1]}, \overrightarrow{\Sigma}_{0:T-1}^{[i+1]}, \overrightarrow{R}_{0:T}^{[i+1]}, \overrightarrow{r}_{0:T}^{[i+1]}, \overrightarrow{S}_{0:T-1}^{[i+1]}, \overrightarrow{s}_{0:T-1}^{[i+1]}$.

Algorithm 2 Forward Recursion of the Forward Entropic Variational Smoother

Require: First Gaussian marginal: $\overrightarrow{m}_0^{[i+1]}, \overrightarrow{P}_0^{[i+1]}$,
forward affine-Gaussian conditionals: $\overrightarrow{F}_{0:T-1}^{[i+1]}, \overrightarrow{d}_{0:T-1}^{[i+1]}, \overrightarrow{\Sigma}_{0:T-1}^{[i+1]}$.

for $k \leftarrow 0, \ldots, T-1$ **do**
  $\overrightarrow{m}_{k+1}^{[i+1]} = \overrightarrow{F}_k^{[i+1]} \overrightarrow{m}_k^{[i+1]} + \overrightarrow{d}_k^{[i+1]}$ ▷ update marginals
  $\overrightarrow{P}_{k+1}^{[i+1]} = \overrightarrow{F}_k^{[i+1]} \overrightarrow{P}_k^{[i+1]} \left[\overrightarrow{F}_k^{[i+1]}\right]^\top + \overrightarrow{\Sigma}_k^{[i+1]}$
end for
return $\overrightarrow{m}_{0:T}^{[i+1]}, \overrightarrow{P}_{0:T}^{[i+1]}$.

Algorithm 3 Optimal Damping for the Forward Entropic Variational Smoother

Require: Maximum multiplier: $\alpha_{\max}$, minimum multiplier: $\alpha_{\min}$,
initial multiplier: α_0, first Gaussian marginal: $\overrightarrow{m}_0^{[i]}, \overrightarrow{P}_0^{[i]}$,
forward affine-Gaussian conditionals: $\overrightarrow{F}_{0:T-1}^{[i]}, \overrightarrow{d}_{0:T-1}^{[i]}, \overrightarrow{\Sigma}_{0:T-1}^{[i]}$,
log prior ℓ_p, log measurement ℓ_h, log dynamics ℓ_f.

$\alpha \leftarrow \alpha_0$
repeat
  $\beta \leftarrow \alpha/(1+\alpha)$
  $\overrightarrow{m}_0^{[i+1]}, \overrightarrow{P}_0^{[i+1]}, \overrightarrow{F}_{0:T-1}^{[i+1]}, \overrightarrow{d}_{0:T-1}^{[i+1]}, \overrightarrow{\Sigma}_{0:T-1}^{[i+1]} \leftarrow$ Forward Conditionals ▷ Algorithm 1
  $\overrightarrow{m}_{0:T}^{[i+1]}, \overrightarrow{P}_{0:T}^{[i+1]} \leftarrow$ Forward Marginals ▷ Algorithm 2
  $\Delta_{\mathrm{KL}} \leftarrow \varepsilon - \mathbb{D}_{\mathrm{KL}}\left[\overrightarrow{q}^{[i+1]} \,||\, \overrightarrow{q}^{[i]}\right]$
  if $\Delta_{\mathrm{KL}} > 0$ **then**
    $\alpha_{\max} \leftarrow \alpha$
    $\alpha \leftarrow \sqrt{\alpha \cdot \alpha_{\min}}$ ▷ reduce multiplier
  else if $\Delta_{\mathrm{KL}} < 0$ **then**
    $\alpha_{\min} \leftarrow \alpha$
    $\alpha \leftarrow \sqrt{\alpha \cdot \alpha_{\max}}$ ▷ increase multiplier
  end if
until $\Delta_{\mathrm{KL}} \approx 0$ ▷ gradient vanishes
$\beta \leftarrow \alpha/(1+\alpha)$
return β

Algorithm 4 Forward Recursion of the Reverse Entropic Variational Smoother

Require: Damping: β, last Gaussian marginal: $\overleftarrow{m}_T^{[i]}, \overleftarrow{P}_T^{[i]}$,
reverse affine-Gaussian conditionals: $\overleftarrow{F}_{1:T}^{[i]}, \overleftarrow{d}_{1:T}^{[i]}, \overleftarrow{\Sigma}_{1:T}^{[i]}$,
log prior ℓ_p: $L_0^{[i]}, l_0^{[i]}$, log measurement ℓ_h: $L_{1:T}^{[i]}, l_{1:T}^{[i]}$,
log dynamics ℓ_f: $C_{\bar{x}\bar{x},0:T-1}^{[i]}, C_{\bar{x}x,0:T-1}^{[i]}, C_{xx,0:T-1}^{[i]}, c_{\bar{x},0:T-1}^{[i]}, c_{x,0:T-1}^{[i]}$.

$\overleftarrow{R}_0^{[i+1]} = L_0^{[i]}, \quad \overleftarrow{r}_0^{[i+1]} = l_0^{[i]}$ ▷ initialize forward recursion

for $k \leftarrow 1, \dots, T$ **do**

$\quad \overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} = (1-\beta)\, C_{\bar{x}\bar{x},k-1}^{[i]} + \beta \left[\overleftarrow{F}_k^{[i]}\right]^\top \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{F}_k^{[i]}$

$\quad \overleftarrow{G}_{xx,k}^{[i+1]} = (1-\beta) \left[C_{xx,k-1}^{[i]} + \overleftarrow{R}_{k-1}^{[i+1]}\right] + \beta \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1}$

$\quad \overleftarrow{G}_{x\bar{x},k}^{[i+1]} = (1-\beta)\, C_{x\bar{x},k-1}^{[i]} + \beta \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{F}_k^{[i]}$

$\quad \overleftarrow{g}_{\bar{x},k}^{[i+1]} = (1-\beta)\, c_{\bar{x},k-1}^{[i]} - \beta \left[\overleftarrow{F}_k^{[i]}\right]^\top \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{d}_k^{[i]}$

$\quad \overleftarrow{g}_{x,k}^{[i+1]} = (1-\beta) \left[c_{x,k-1}^{[i]} + \overleftarrow{r}_{k-1}^{[i+1]}\right] + \beta \left[\overleftarrow{\Sigma}_k^{[i]}\right]^{-1} \overleftarrow{d}_k^{[i]}$

$\quad \overleftarrow{S}_k^{[i+1]} = \overleftarrow{G}_{\bar{x}\bar{x},k}^{[i+1]} - \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{G}_{x\bar{x},k}^{[i+1]}$ ▷ update log-normalizers

$\quad \overleftarrow{s}_k^{[i+1]} = \overleftarrow{g}_{\bar{x},k}^{[i+1]} + \left[\overleftarrow{G}_{x\bar{x},k}^{[i+1]}\right]^\top \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{g}_{x,k}^{[i+1]}$

$\quad \overleftarrow{R}_k^{[i+1]} = L_k^{[i]} + 1/(1-\beta)\, \overleftarrow{S}_k^{[i+1]}$ ▷ update potential functions

$\quad \overleftarrow{r}_k^{[i+1]} = l_k^{[i]} + 1/(1-\beta)\, \overleftarrow{s}_k^{[i+1]}$

$\quad \overleftarrow{F}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{G}_{x\bar{x},k}^{[i+1]}$ ▷ update conditional posteriors

$\quad \overleftarrow{d}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1} \overleftarrow{g}_{x,k}^{[i+1]}$

$\quad \overleftarrow{\Sigma}_k^{[i+1]} = \left[\overleftarrow{G}_{xx,k}^{[i+1]}\right]^{-1}$

end for

$\overleftarrow{m}_T^{[i+1]} = \overleftarrow{P}_T^{[i+1]} \left[(1-\beta)\, \overleftarrow{r}_T^{[i+1]} + \beta \left[\overleftarrow{P}_T^{[i]}\right]^{-1} \overleftarrow{m}_T^{[i]}\right]$ ▷ update boundary marginal

$\overleftarrow{P}_T^{[i+1]} = \left[(1-\beta)\, \overleftarrow{R}_T^{[i+1]} + \beta \left[\overleftarrow{P}_T^{[i]}\right]^{-1}\right]^{-1}$

return $\overleftarrow{m}_T^{[i+1]}, \overleftarrow{P}_T^{[i+1]}, \overleftarrow{F}_{1:T}^{[i+1]}, \overleftarrow{d}_{1:T}^{[i+1]}, \overleftarrow{\Sigma}_{1:T}^{[i+1]}, \overleftarrow{R}_{0:T}^{[i+1]}, \overleftarrow{r}_{0:T}^{[i+1]}, \overleftarrow{S}_{1:T}^{[i+1]}, \overleftarrow{s}_{1:T}^{[i+1]}$.

Algorithm 5 Backward Recursion of the Reverse Entropic Variational Smoother

Require: Last Gaussian marginal: $\overleftarrow{m}_T^{[i+1]}, \overleftarrow{P}_T^{[i+1]}$,
reverse affine-Gaussian conditionals: $\overleftarrow{F}_{1:T}^{[i+1]}, \overleftarrow{d}_{1:T}^{[i+1]}, \overleftarrow{\Sigma}_{1:T}^{[i+1]}$.

for $k \leftarrow T, \ldots, 1$ **do**
  $\overleftarrow{m}_{k-1}^{[i+1]} = \overleftarrow{F}_k^{[i+1]} \overleftarrow{m}_k^{[i+1]} + \overleftarrow{d}_k^{[i+1]}$ ▷ update marginals
  $\overleftarrow{P}_{k-1}^{[i+1]} = \overleftarrow{F}_k^{[i+1]} \overleftarrow{P}_k^{[i+1]} \left[\overleftarrow{F}_k^{[i+1]}\right]^\top + \overleftarrow{\Sigma}_k^{[i+1]}$
end for
return $\overleftarrow{m}_{0:T}^{[i+1]}, \overleftarrow{P}_{0:T}^{[i+1]}$.

Algorithm 6 Optimal Damping for the Reverse Entropic Variational Smoother

Require: Maximum multiplier: $\alpha_{\max}$, minimum multiplier: $\alpha_{\min}$,
initial multiplier: α_0, last Gaussian marginal: $\overleftarrow{m}_T^{[i]}, \overleftarrow{P}_T^{[i]}$,
reverse affine-Gaussian conditionals: $\overleftarrow{F}_{1:T}^{[i]}, \overleftarrow{d}_{1:T}^{[i]}, \overleftarrow{\Sigma}_{1:T}^{[i]}$,
log prior ℓ_p, log measurement ℓ_h, log dynamics ℓ_f.

$\alpha \leftarrow \alpha_0$
repeat
  $\beta \leftarrow \alpha/(1+\alpha)$
  $\overleftarrow{m}_T^{[i+1]}, \overleftarrow{P}_T^{[i+1]}, \overleftarrow{F}_{1:T}^{[i+1]}, \overleftarrow{d}_{1:T}^{[i+1]}, \overleftarrow{\Sigma}_{1:T}^{[i+1]} \leftarrow$ Reverse Conditionals ▷ Algorithm 4
  $\overleftarrow{m}_{0:T}^{[i+1]}, \overleftarrow{P}_{0:T}^{[i+1]} \leftarrow$ Reverse Marginals ▷ Algorithm 5
  $\Delta_{\mathrm{KL}} \leftarrow \varepsilon - \mathbb{D}_{\mathrm{KL}}\left[\overleftarrow{q}^{[i+1]} \,||\, \overleftarrow{q}^{[i]}\right]$
  if $\Delta_{\mathrm{KL}} > 0$ **then**
    $\alpha_{\max} \leftarrow \alpha$
    $\alpha \leftarrow \sqrt{\alpha \cdot \alpha_{\min}}$ ▷ reduce multiplier
  else if $\Delta_{\mathrm{KL}} < 0$ **then**
    $\alpha_{\min} \leftarrow \alpha$
    $\alpha \leftarrow \sqrt{\alpha \cdot \alpha_{\max}}$ ▷ increase multiplier
  end if
until $\Delta_{\mathrm{KL}} \approx 0$ ▷ gradient vanishes
$\beta \leftarrow \alpha/(1+\alpha)$
return β

Algorithm 7 Forward Entropic Variational Smoother

Require: First Gaussian marginal: $\overrightarrow{m}_0^{[0]}, \overrightarrow{P}_0^{[0]}$,
forward affine-Gaussian conditionals: $\overrightarrow{F}_{0:T-1}^{[0]}, \overrightarrow{d}_{0:T-1}^{[0]}, \overrightarrow{\Sigma}_{0:T-1}^{[0]}$.

$i \leftarrow 0$
while not converged **do**
  $\overrightarrow{m}_{0:T}^{[i]}, \overrightarrow{P}_{0:T}^{[i]} \leftarrow$ Forward Marginals ▷ Algorithm 2
  $\ell_p, \ell_h, \ell_f \leftarrow$ Approximate Model ▷ Definition 3/4
  $\beta \leftarrow$ Optimal Damping ▷ Algorithm 3
  $\overrightarrow{m}_0^{[i+1]}, \overrightarrow{P}_0^{[i+1]}, \overrightarrow{F}_{0:T-1}^{[i+1]}, \overrightarrow{d}_{0:T-1}^{[i+1]}, \overrightarrow{\Sigma}_{0:T-1}^{[i+1]} \leftarrow$ Forward Conditionals ▷ Algorithm 1
  $i \leftarrow i + 1$
end while
return $\overrightarrow{m}_{0:T}^{[\infty]}, \overrightarrow{P}_{0:T}^{[\infty]}$.

Algorithm 8 Reverse Entropic Variational Smoother

Require: Last Gaussian marginal: $\overleftarrow{m}_T^{[0]}, \overleftarrow{P}_T^{[0]}$,
reverse affine-Gaussian conditionals: $\overleftarrow{F}_{1:T}^{[0]}, \overleftarrow{d}_{1:T}^{[0]}, \overleftarrow{\Sigma}_{1:T}^{[0]}$.

$i \leftarrow 0$
while not converged **do**
  $\overleftarrow{m}_{0:T}^{[i]}, \overleftarrow{P}_{0:T}^{[i]} \leftarrow$ Reverse Marginals ▷ Algorithm 5
  $\ell_p, \ell_h, \ell_f \leftarrow$ Approximate Model ▷ Definition 3/4
  $\beta \leftarrow$ Optimal Damping ▷ Algorithm 6
  $\overleftarrow{m}_T^{[i+1]}, \overleftarrow{P}_T^{[i+1]}, \overleftarrow{F}_{1:T}^{[i+1]}, \overleftarrow{d}_{1:T}^{[i+1]}, \overleftarrow{\Sigma}_{1:T}^{[i+1]} \leftarrow$ Reverse Conditionals ▷ Algorithm 4
  $i \leftarrow i + 1$
end while
return $\overleftarrow{m}_{0:T}^{[\infty]}, \overleftarrow{P}_{0:T}^{[\infty]}$.

Algorithm 9 Marginals of the Hybrid Entropic Variational Smoother

Require: Damping: β, prior Gaussian marginal: $m_{1:T-1}^{[i]}, P_{1:T-1}^{[i]}$,
forward log-normalizing functions: $\overrightarrow{S}_{1:T-1}^{[i+1]}, \overrightarrow{s}_{1:T-1}^{[i+1]}$,
reverse potential functions: $\overleftarrow{R}_{1:T-1}^{[i+1]}, \overleftarrow{r}_{1:T-1}^{[i+1]}$.

for $k \leftarrow 1, \ldots, T-1$ **do**

$\quad P_k^{[i+1]} = \left[(1-\beta)\left[\overleftarrow{R}_k^{[i+1]} + \overrightarrow{S}_k^{[i+1]}\right] + \beta\left[P_k^{[i]}\right]^{-1}\right]^{-1}$ ▷ update marginals

$\quad m_k^{[i+1]} = P_k^{[i+1]}\left[(1-\beta)\left[\overleftarrow{r}_k^{[i+1]} + \overrightarrow{s}_k^{[i+1]}\right] + \beta\left[P_k^{[i]}\right]^{-1} m_k^{[i]}\right]$

end for

return $m_{1:T-1}^{[i+1]}, P_{1:T-1}^{[i+1]}$.

Algorithm 10 Hybrid Entropic Variational Smoother

Require: Initial Gaussian marginals: $m_{0:T}^{[0]}, P_{0:T}^{[0]}$,
forward affine-Gaussian conditionals: $\overrightarrow{F}_{0:T-1}^{[0]}, \overrightarrow{d}_{0:T-1}^{[0]}, \overrightarrow{\Sigma}_{0:T-1}^{[0]}$,
reverse affine-Gaussian conditionals: $\overleftarrow{F}_{1:T}^{[0]}, \overleftarrow{d}_{1:T}^{[0]}, \overleftarrow{\Sigma}_{1:T}^{[0]}$.

$i \leftarrow 0$

while not converged **do**

$\quad \ell_p, \ell_h, \ell_f \leftarrow$ Approximate Model ▷ Definition 3/4

$\quad \beta \leftarrow$ Optimal Damping ▷ Algorithm 3/6

$\quad m_0^{[i+1]}, P_0^{[i+1]}, \overrightarrow{F}_{0:T-1}^{[i+1]}, \overrightarrow{d}_{0:T-1}^{[i+1]}, \overrightarrow{\Sigma}_{0:T-1}^{[i+1]}, \overrightarrow{S}_{1:T-1}^{[i+1]}, \overrightarrow{s}_{1:T-1}^{[i+1]} \leftarrow$ Forward Normalizers ▷ Algorithm 1

$\quad m_T^{[i+1]}, P_T^{[i+1]}, \overleftarrow{F}_{1:T}^{[i+1]}, \overleftarrow{d}_{1:T}^{[i+1]}, \overleftarrow{\Sigma}_{1:T}^{[i+1]}, \overleftarrow{R}_{1:T-1}^{[i+1]}, \overleftarrow{r}_{1:T-1}^{[i+1]} \leftarrow$ Reverse Potentials ▷ Algorithm 4

$\quad m_{1:T-1}^{[i+1]}, P_{1:T-1}^{[i+1]} \leftarrow$ Hybrid Marginals ▷ Algorithm 9

$\quad i \leftarrow i+1$

end while

return $m_{0:T}^{[\infty]}, P_{0:T}^{[\infty]}$.