Variational Gaussian approximation of the Kushner optimal filter

Marc Lambert, Silvère Bonnabel, Francis Bach

Variational Gaussian approximation of the Kushner optimal filter *

Marc Lambert
DGA/CATOD, Centre d'Analyse Technico-Opérationelle de Défense
& INRIA
marc-h.lambert@intradef.gouv.fr

Silvère Bonnabel
MINES ParisTech, PSL University, Center for robotics
silvere.bonnabel@mines-paristech.fr

Francis Bach
INRIA - Ecole Normale Supérieure - PSL Research university
francis.bach@inria.fr

Abstract

In estimation theory, the Kushner equation provides the evolution of the probability density of the state of a dynamical system given continuous-time observations. Building upon our recent work, we propose a new way to approximate the solution of the Kushner equation through tractable variational Gaussian approximations of two proximal losses associated with the propagation and Bayesian update of the probability density. The first is a proximal loss based on the Wasserstein metric and the second is a proximal loss based on the Fisher metric. The solution to this last proximal loss is given by implicit updates on the mean and covariance that we proposed earlier. These two variational updates can be fused and shown to satisfy a set of stochastic differential equations on the Gaussian's mean and covariance matrix. This Gaussian flow is consistent with the Kalman-Bucy and Riccati flows in the linear case and generalize them in the nonlinear one.

1 Introduction

We consider the general filtering problem where we aim to estimate the state x_t of a continuous-time stochastic system given noisy observations y_t. If the state follows a Langevin dynamic $f = -\nabla V$ with V a potential function and the observations occur continuously in time, the problem can be described by two stochastic differential equations (SDE) on x_t and z_t, where z_t is related to the observation by the equation $dz_t = y_t dt$:

$$dx_t = -\nabla V(x_t)dt + \sqrt{2\varepsilon}d\beta \tag{1}$$

$$dz_t = h(x_t)dt + \sqrt{R}d\eta. \tag{2}$$

β and η are independent Wiener processes and $Q = 2\varepsilon\mathbb{I}$ and R play the role of covariance matrices of the associated diffusion processes. Many dynamical systems can be rewritten in the Langevin canonical form (1), see for instance [6]. In essence (2) means "$y_t = h(x_t) + \text{noise}$", but one has to resort to (2) to avoid problems related to infinitely many observations. The optimal Bayesian filter corresponds to the conditional probability

*This document is the preprint version of the article published for the International Conference on Geometric Science of Information, page 395-404, 2023.

p_t of the state at time t given all past observations. This probability satisfies the Kushner equations which can be split into two parts:

$$dp_t = \mathscr{L}(p_t)dt + d\mathscr{H}(p_t), \tag{3}$$

where $\mathscr{L}$ is defined by the Fokker-Planck partial differential equation (PDE)

$$\mathscr{L}(p_t) = \text{div}[\nabla V p_t] + \varepsilon \Delta p_t, \tag{4}$$

whereas the second term corresponds to the Kushner stochastic PDE (SPDE):

$$d\mathscr{H}(p_t) = (h - \mathbb{E}_{p_t}[h])^T R^{-1}(dz_t - \mathbb{E}_{p_t}[h]dt)p_t,$$

where $\mathbb{E}_{p_t}[h] := \int h(x)p_t(x)dx$ and stochasticity comes from dz_t. These equations cannot be solved in the general case, and we must resort to approximation. In this paper, we consider variational Gaussian approximation, which consists in searching for the Gaussian distribution q_t closest to the optimal one p_t for a particular variational loss. Two variational losses are well suited for our problem.

Jordan-Kinderlehrer-Otto (JKO) [9] showed that the following proximal scheme:

$$\text{argmin } \mathscr{L}^{\delta t}(p) = \text{argmin } \left[KL\left(p \middle|\middle| \pi\right) + \frac{1}{2\delta t} d_w^2(p_t, p) \right], \quad \text{(JKO)} \tag{5}$$

is related to the Fokker-Planck (FP) equation associated to (1) where we denote its stationary distribution $\pi \propto \exp(-V/\varepsilon)$. Indeed, iterating this proximal algorithm yields a curve being solution to FP as $\delta t \to 0$. It is referred to as variational since it is an optimization problem over the function p, and it involves the Kullback-Leibler divergence defined by $KL(p||\pi) = \int p \log \frac{p}{\pi}$, and the Wasserstein (or optimal transport) distance $d_w^2(p_t, p)$ [2].

The variational loss associated to the Kushner PDE is the Laugesen-Mehta-Meyn-Raginsky (LMMR) proximal scheme [14] defined by:

$$\text{argmin } \mathscr{H}^{\delta t}(p) = \text{argmin } \left[\mathbb{E}_p \frac{1}{2} ||\delta z_t - h(x)\delta t||^2_{(R\delta t)^{-1}} + KL(p||p_t) \right], \text{ (LMMR)} \tag{6}$$

where $\delta z_t := z_{t+\delta t} - z_t$ comes from the Euler-Marayama discretization of the observation SDE: $\delta z_t = h(x_t)\delta t + \sqrt{R}\delta\eta$ such that $p(\delta z_t|x_t) = \mathscr{N}(h(x_t)\delta t, R\delta t)$.

For small δt those schemes generate a sequence of probability distributions that converge to the solutions of the corresponding PDE in the limit $\delta t \to 0$. We see the KLs in both schemes play a different role, though. In (5), the proximal scheme shows that the solution to the FP equation follows a gradient of the KL to the stationary distribution π. This gradient is computed with respect to the Wasserstein metric. In (6), the proximal scheme defines a gradient over the state prediction p of the expected prediction error. This gradient is computed in the sense of the metric defined by the KL around its null value, which may be related to the Fisher metric.

To approximate the solutions, we propose to constrain them to lie in the space of Gaussian distributions. That can be done by constraining in the proximal schemes the general distribution p_t to be a Gaussian distribution $q_t = \mathscr{N}(\mu, P)$. The proximal problems become finite-dimensional and boil down to minimizing $\mathscr{L}^{\delta t}$ and $\mathscr{H}^{\delta t}$ over (μ, P). The Gaussian approximation of the JKO scheme yields in the limit a set of ODEs on μ and P as shown in [13]. In this paper, we extend these results showing the Gaussian solution to the LMMR scheme corresponds to the R-VGA solution [11] which yields in the limit a set of SDEs on μ and P. Moreover, using a two-step approach, we can fuse the two Gaussian solutions to approximate the Kushner

equation (3). As shall be shown presently, we find the following SDEs for μ and P:

$$\boxed{\begin{array}{l}\textbf{The fully continuous-time variational Kalman filter}\\ d\mu_t = b_t dt + P_t dC_t\\ dP_t = A_t P_t dt + P_t A_t^T dt + \frac{1}{2} dH_t P_t + \frac{1}{2} P_t dH_t^T + 2\varepsilon \mathbb{I} dt\\ \text{where } b_t = -\mathbb{E}_{q_t}[\nabla V(x)]; \quad dC_t = \mathbb{E}_{q_t}[\nabla h(x_t)^T R^{-1}(dz_t - h(x_t)dt)]\\ A_t = -\mathbb{E}_{q_t}[\nabla^2 V(x)]; \quad dH_t = \mathbb{E}_{q_t}[(x_t - \mu_t)(dz_t - h(x_t)dt)^T R^{-1} \nabla h(x_t)].\end{array}} \tag{7}$$

The equation for P_t can be seen as a generalization of the Riccati equation in the nonlinear case. Indeed, if we replace V and h with linear functions, the ODE on P_t matches the Riccati equations and we recover the Kalman-Bucy filter, known to solve exactly the Kushner equations.

This paper is organized as follows: Section 2 is dedicated to related works on the approximation of the optimal nonlinear filter. In Section 3 we derive the variational Gaussian approximation of the LMMR scheme. In Section 4 we recall the variational Gaussian approximation of the JKO scheme proposed in our previous work. In Section 5 we combine these two results to obtain the Continuous Variational Kalman filter equations and show the equivalence with the Kalman-Bucy filter in the linear case.

2 Related works

In 1967, Kushner proposed a Gaussian assumed density filter to solve his PDE [10]. This filter is derived by keeping only the first two moments of p_t in (3) which can be computed in closed form using the Ito formula. These moments involve integrals under the unknown distribution p_t and the heuristic is to integrate them rather on the current Gaussian approximation q_t leading to a recursive scheme. A more rigorous way to do this approximation was proposed later [8, 4] with the projected filter. In this approach, the solution of the Kushner PDE is projected onto the tangent space to Gaussian distributions equipped with the Fisher information metric. This leads to ODEs that are quite different from (7). A third approach is to linearize the stochastic dynamic process to obtain a McKean Vlasov process that allows for Gaussian propagation [12, Sec 4.1.1]. The connexion between approximated SDEs and projected filters was analyzed in detail earlier in [3]. The latter approach is the one explored in the current paper, i.e., considering proximal schemes associated with the Kushner PDE where we constrain the solution to be Gaussian. It is equivalent to projecting the exact gradient flow onto the tangent space of the manifold of Gaussian distributions. This approach is preferred since it exhibits the problem's geometric structure and allows convergence guarantees to be proven. Approximation of gradient flows is an active field and several recent papers have followed this direction: the connexion between the propagation part of the Gaussian assumed density filter and the variational JKO scheme [9] was recently studied [13]; the connexion between the update part of the Gaussian assumed density filter and the variational LMMR scheme [14] was studied in [7] where a connexion with a gradient flow was first established but limited to the linear case. To the best of our knowledge, the variational approximation of the LMMR scheme in the nonlinear case has never been addressed. The various ways to obtain the ODEs (7) lead to nice connexions between geometric projection, constrained optimization, and statistical linearization. These different approaches are illustrated in Figure 1 which addresses only the approximation of dynamics (1) without measurements (i.e., propagation only) for which all methods prove equivalent.

3 Variational Gaussian approximation of the LMMR proximal

In this section, we compute the closest Gaussian solution to the LMMR problem (6). The corresponding Gaussian flow is closely related to natural gradient descent used in information geometry. This flow

Distribution p		Gaussian approx. $q(\mu, P)$
$dx_t = -\nabla V(x_t)dt + \sqrt{2\varepsilon}d\beta$ nonlinear SDE process	SDE linearization [12] $\Rightarrow$	$dx_t = A(t)(x_t - \mu_t)dt + b(t)dt + \sqrt{2\varepsilon}d\beta$ McKean–Vlasov process
$\frac{\partial p}{\partial t} = \mathrm{div}(\nabla V\, p) + \varepsilon\Delta p$ Fokker-Planck	Riemanian projection [8] $\Rightarrow$	$\dot{\mu}_t = b(t)$ $\dot{P}_t = A(t)P_t + P_tA(t)^T + 2\varepsilon\mathbb{I}$ Variational Gaussian flow
$KL(p\|\|\pi) + \frac{1}{2\delta t}d_w(p, p_t)^2$ proximal JKO	constrained optim. [13] $\Rightarrow$	$KL(q\|\|\pi) + \frac{1}{2\delta t}d_{bw}(q, q_t)^2$ proximal Bures-JKO
$\delta p = -\nabla_w KL(p\|\|\pi)$ W2 gradient flow	gradient projection [13] $\Rightarrow$	$\delta q = -\nabla_{bw} KL(q\|\|\pi)$ Bures-W2 gradient flow

where $\quad A(t) = -\mathbb{E}_{q_t}[\nabla^2 V(x)]; \qquad b(t) = -\mathbb{E}_{q_t}[\nabla V(x)]$

Figure 1: Various equivalent approaches for Gaussian approximation of the SDE (1). We denote $\pi \propto \exp(-V/\varepsilon)$ as the stationary distribution of the associated Fokker-Planck equation. The left column presents equivalent definitions of p whereas the right column corresponds to the approximated solution q in the space of Gaussian distributions. d_w denotes the Wasserstein distance whereas d_{bw} denotes the Bures-Wasserstein distance, which is its restriction to the subset of Gaussian distributions. At the last row, the tangent vector δp, respectively (resp. δq) and the gradient ∇_w (resp. ∇_{bw}) are defined with respect to the Wasserstein metric space of distribution $\left(\mathscr{P}(\mathbb{R}^d), d_w^2\right)$ (resp. the Bures-Wasserstein metric space of Gaussians $\left(\mathscr{N}(\mathbb{R}^d), d_{bw}^2\right)$). These geometries are briefly explained in Section 4.2.

approximates the Kushner optimal filter when the state is static. In the next sections, we will generalize this result to a dynamic state.

3.1 The recursive variational Gaussian approximation

The proximal LMMR problem (6) where we constrain the solution q to be a Gaussian reads (given q_t a current Gaussian distribution at time t):

$$q_{t+\delta t} = \underset{q\in\mathcal{N}(\mu,P)}{\arg\min} \quad \mathbb{E}_q \frac{1}{2}||\delta z_t - h(x)\delta t||^2_{(R\delta t)^{-1}} + KL(q||q_t) \tag{8}$$

$$= \underset{q\in\mathcal{N}(\mu,P)}{\arg\min} \quad -\int q(x)\log p(\delta z_t|x)dx + KL(q||q_t) \tag{9}$$

$$= \underset{q\in\mathcal{N}(\mu,P)}{\arg\min} \quad KL\left(q\middle|\middle|\frac{1}{Z}p(\delta z_t|x)q_t\right), \tag{10}$$

where we have introduced a normalization constant Z which does not change the problem.
Eq (10) falls into the framework of variational Gaussian approximation (R-VGA) [11]. The solution satisfies the following updates [11, Theorem 1]:

$$\mu_{t+\delta t} = \mu_t + P_t\mathbb{E}_{q_{t+\delta t}}[\nabla_x \log p(\delta z_t|x)]$$
$$P^{-1}_{t+\delta t} = P^{-1}_t - \mathbb{E}_{q_{t+\delta t}}[\nabla^2_x \log p(\delta z_t|x)],$$

where the expectations are under the Gaussian $q_{t+\delta t}\sim\mathcal{N}(\mu_{t+\delta t}, P_{t+\delta t})$ making the updates implict. In the linear case, that is, if we take $h(x) = Hx$, these updates are equivalent to the online Newton algorithm [11, Theorem 2]. Computing the Hessian $\nabla^2_x \log p$ can be avoided using integration by part:

$$P^{-1}_{t+\delta t} = P^{-1}_t - P^{-1}_{t+\delta t}\mathbb{E}_{q_{t+\delta t}}[(x-\mu_{t+\delta t})\nabla_x \log p(\delta z_t|x)^T]$$

By rearranging the terms and using that P is symmetric (see [12, Sec 4.2]) we can let appear an update on the covariance:

$$P_{t+\delta t} = P_t + \frac{1}{2}\mathbb{E}_{q_{t+\delta t}}[(x-\mu_t)\nabla_x \log p(\delta z_t|x)^T]P_t + \frac{1}{2}P_t\mathbb{E}_{q_{t+\delta t}}[\nabla_x \log p(\delta z_t|x)(x-\mu_t)^T].$$

Finally, using that $\nabla_x \log p(\delta z_t|x) = \nabla h(x)^T R^{-1}(\delta z_t - h(x)\delta t)$ we obtain:

$$\mu_{t+\delta t} = \mu_t + P_t\delta C_t, \quad P_{t+\delta t} = P_t + \frac{1}{2}\delta H_t P_t + \frac{1}{2}P_t\delta H_t^T, \tag{11}$$
where:
$$\delta C_t = \mathbb{E}_{q_{t+\delta t}}[\nabla h(x)^T R^{-1}(\delta z_t - h(x)\delta t)]$$
$$\delta H_t = \mathbb{E}_{q_{t+\delta t}}[(x-\mu_t)(\delta z_t - h(x)\delta t)^T R^{-1}\nabla h(x)].$$

Letting $\delta t\to 0$, we obtain the following SDE in the sense of Ito:

$$d\mu_t = P_t dC_t, \quad dP_t = \frac{1}{2}dH_t P_t + \frac{1}{2}P_t dH_t^T, \tag{12}$$

where it shall be noted that dH_t is non-deterministic owing to dz_t. Since the LMMR scheme has been proven to converge to the solution of the Kushner SPDE [14], this SDE describes the best Gaussian approximation of the optimal filter when the state is static.

3.2 Information geometry interpretation

We show here how the LMMR proximal scheme is related to the Fisher information geometry in the general case. Let's consider a family of densities: $S = \left\{ p(.|\theta); \theta \in \Theta; \Theta \subseteq \mathbb{R}^m \right\}$ and let

$$F(\theta) = \int \nabla_\theta \log p(x|\theta) \nabla_\theta \log p(x|\theta)^T p(x|\theta) dx,$$

be the Fisher information matrix, where θ regroups all the parameters. If we consider now the proximal LMMR on S, and if we use the second-order Taylor expansion of the KL divergence between these two distributions, we have:

$$KL(p(x|\theta)||p(x|\theta_t)) = \frac{1}{2}(\theta - \theta_t)^T F(\theta_t)(\theta - \theta_t) + o((\theta - \theta_t)^2).$$

Rather than minimizing the proximal LMMR scheme (6) in the infinite space of distributions, we now search the minimum in the finite space of parameters:

$$\theta_{t+\delta t} = \underset{\theta \in \Theta}{\arg\min} \quad \mathbb{E}_{p(x|\theta)} \left[\frac{1}{2} ||\delta z_t - h(x)\delta t||^2_{(R\delta t)^{-1}} \right] + \frac{1}{2}(\theta - \theta_t)^T F(\theta_t)(\theta - \theta_t).$$

Considering that the minimum must cancel the gradient of the above proximal loss, we obtain:

$$\begin{aligned} 0 &= \nabla_\theta \left(\mathbb{E}_{p(x|\theta)} \left[\frac{1}{2} ||\delta z_t - h(x)\delta t||^2_{(R\delta t)^{-1}} \right] \right) \Big|_{\theta_{t+\delta t}} + F(\theta_t)(\theta_{t+\delta t} - \theta_t) \\ \theta_{t+\delta t} &= \theta_t - F(\theta_t)^{-1} \nabla_\theta \left(\frac{1}{2} \mathbb{E}_{p(x|\theta)} \left[||\delta z_t - h(x)\delta t||^2_{(R\delta t)^{-1}} \right] \right) \Big|_{\theta_{t+\delta t}}, \end{aligned} \tag{13}$$

which corresponds to a gradient descent of the averaged stochastic likelihood:

$$\theta_{t+\delta t} = \theta_t - F(\theta_t)^{-1} \nabla_\theta \mathbb{E}_{p(x|\theta)}[-\log p(\delta z_t|x)] \big|_{\theta_{t+\delta t}}. \tag{14}$$

Remarkably, the optimal filer equations with a static state x are given by an implicit Bayesian variant of the natural gradient descent [1]. Indeed here x plays the role of the parameter of the likelihood distribution. The original natural gradient should be a descent with the gradient $-F(x_t)^{-1} \nabla_x \log p(\delta z_t|x) \big|_{x_t}$.

4 Variational Gaussian approximation of the JKO proximal

The canonical Langevin form (1) assumes that the drift term $f = -\nabla V$ derives from a potential V. This potential has a physical meaning in filtering (consider a gravity field for example). The evolution of the state in the filter mimics the true evolution of the physical system. It's not the case in statistical physics, where the potential is constructed such that $V = -\log \pi$ where π is the asymptotic distribution of a variable x which doesn't correspond to a physical system. We used this property in our previous work [13] and simulated a dynamic to approximate the target π with a Gaussian distribution. Here we do not want to estimate a distribution but to propagate a Gaussian through the nonlinear physical dynamic (1).

4.1 The Bures-JKO proximal

The proximal JKO problem (5) where we constrained the solution q to be a Gaussian distribution writes:

$$\min_{q \in \mathcal{N}(\mu, P)} \quad KL\left(q \middle|\middle| \pi\right) + \frac{1}{2\delta t} d_{bw}(q, q_t)^2,$$

where $d_{bw}(q, q_t)$ is the Bures distance between two Gaussians given by:

$$d_{bw}(q, q_t) = ||\mu - \mu_t||^2 + \mathscr{B}^2(P, P_t), \tag{15}$$

where $\mathscr{B}^2(P, P_t) = \mathrm{Tr}(P + P_t - 2(P^{\frac{1}{2}} P_t P^{\frac{1}{2}})^{\frac{1}{2}})$ is the squared Bures metric [5], which has a derivative available in closed form. After some computation [13, Appendix A] we can obtain implicit equations that the parameters of the optimal Gaussian solution q must satisfy:

$$\begin{aligned} \mu_{t+\delta t} &= \mu_t - \delta t . \mathbb{E}_{q_{t+\delta t}}[\nabla V(x)] \\ P_{t+\delta t} &= P_t - \delta t . \mathbb{E}_{q_{t+\delta t}}[\nabla^2 V(x)] P_t - \delta t . P_t \mathbb{E}_{q_{t+\delta t}}[\nabla^2 V(x)]^T + 2\varepsilon \delta t . \mathbb{I}, \end{aligned} \tag{16}$$

and at the limit $\delta t \to 0$, we obtain the following ODEs:

$$\begin{aligned} \dot{\mu}_t &= -\mathbb{E}_{q_t}[\nabla V(x)] := b_t \\ \dot{P}_t &= A_t P_t + P_t A_t^T + 2\varepsilon \mathbb{I} \quad \text{where } A_t := -\mathbb{E}_{q_t}[\nabla^2 V(x)]. \end{aligned} \tag{17}$$

4.2 Wasserstein geometry interpretation

The Wasserstein geometry is defined by the metric space of measure endowed with the Wasserstein distance $\left(\mathscr{P}(\mathbb{R}^d), d_w^2\right)$. The definition of a tangent vector in this space is tedious because the measure μ must satisfy the conservation of mass $\int \mu(x)dx = 1$. To handle this constraint we can use the continuity equation. This equation allows to represent any regular curves of measures with a continuous flow along a vector field $v_t \in \mathbb{L}^2$. It is closely related to the Fokker-Planck equation as we show now (see [2] for more details). The JKO proximal scheme (5) gives a sequence of distribution that satisfies at the limit the Fokker-Planck equation (4), this equation rewrites as follows:

$$\begin{aligned} \dot{p}_t &= \nabla.(\nabla V p_t) + \varepsilon \nabla.\nabla p_t = \nabla.(\nabla V p_t) + \varepsilon \nabla.(p_t \nabla \log p_t) \\ &= \nabla.(p_t(\nabla V + \varepsilon \nabla \log p_t)) = -\mathrm{div}(p_t v_t), \end{aligned} \tag{18}$$

which is a continuity equation where $v_t \in \mathbb{L}^2(\mathbb{R}^d)$ plays the role of the tangent vector δp_t along the path p_t and satisfies:

$$v_t = -\nabla V - \varepsilon \nabla \log p_t = -\nabla_w KL(p_t || \pi),$$

with $\pi \propto \exp(-V/\varepsilon)$. The last equality comes from variational calculus in the measure space: the Wasserstein gradient of a functional F is given by the Euclidian gradient of the first variation $\nabla_w F(\rho) = \nabla \delta F(\rho)$, see [2, Chapter 10].

Let's sum up what's going on: starting from a stochastic state x_t following the Langevin dynamic (1) with drift $-\nabla V$, we have rewritten the Fokker-Planck equation which describes the evolution of the density $p(x_t)$ as a continuity equation (18) where the diffusion term has disappeared. At this continuity equation correspond a deterministic ODE $\dot{x}_t = -\nabla_w KL(p_t||\pi)$. It's a nice property of the Wasserstein geometry where PDE can be described by a continuity equation that corresponds to a simple gradient flow.

Following the same track, the sequence of Gaussian distributions satisfying the ODE (17) correspond to a Wasserstein gradient flow given by the continuity equation: $\dot{q}_t = -div(q_t w_t)$, where $w_t = -\nabla_{bw} KL(q_t||\pi)$ is now a gradient with respect to the Bures-Wasserstein distance (15), see [13, Appendix B3] for the analytical expression of this gradient.

5 Variational Gaussian approximation of the Kushner optimal filter

We have tackled the two proximal problems independently but how to solve them jointly? The simplest method to do so is to alternate between propagation through dynamics (1) for a small time δt, and Bayesian update through LMMR in the light of the accumulated observations δz_t, and let $\delta t \to 0$. This is what we do presently.

5.1 The continuous variational Kalman filter

Consider one step of the Euler–Maruyama method with length δt of SDEs (1) and (2). As the Wiener processes β and η are independent, we may write:

$$p(x_t, y_{t+\delta t}, x_{t+\delta t}) = p(y_{t+\delta t}|x_{t+\delta t}, x_t)p(x_{t+\delta t}, x_t) = p(y_{t+\delta t}|x_{t+\delta t})p(x_{t+\delta t}|x_t),$$

denoting $y_{t+\delta t} = \delta z_t$. In other words, we can solve the proximal LMMR update equation (10) using as prior $q_t(x) = \mathcal{N}(\mu_{t+\delta t|t}, P_{t+\delta t|t})$, the solution of the proximal JKO. The LMMR/R-VGA discrete-time equations (11) then become:

$$\begin{aligned}
\mu_{t+\delta t} &= \mu_{t+\delta t|t} + P_{t+\delta t|t}\delta C_t \\
P_{t+\delta t} &= P_{t+\delta t|t} + \frac{1}{2}\delta H_t P_{t+\delta t|t} + \frac{1}{2}P_{t+\delta t|t}\delta H_t^T.
\end{aligned}$$

Replacing $\mu_{t+\delta t|t}$ and $P_{t+\delta t|t}$ by their expressions as the solutions to the JKO scheme (16) and putting in a residual all the terms in δt^2, we obtain:

$$\begin{aligned}
\mu_{t+\delta t} &= \mu_t + \delta t b_t + P_t \delta C_t \\
P_{t+\delta t} &= P_t + \delta t A_t P_t + \delta t P_t A_t^T + \delta t 2\varepsilon\mathbb{I} + \frac{1}{2}\delta H_t P_t + \frac{1}{2}P_t \delta H_t^T + O(\delta t^2).
\end{aligned}$$

By Ito calculus, we obtain the continuous variational Kalman updates (7).

5.2 The Kalman-Bucy filter as a particular case

Let us consider the linear case where the SDEs (1) and (2) rewrite:

$$dx_t = Fx_t dt + \sqrt{2\varepsilon}d\beta, \quad dz_t = Gx_t dt + \sqrt{R}d\eta.$$

The various expectations that appear in the proposed filter (7) apply either to quantities being independent of x_t or being linear or quadratic in x_t, yielding

$$\begin{aligned}
d\mu_t &= F\mu_t dt + P_t G^T R^{-1}(dz_t - G\mu_t dt) \\
\frac{d}{dt}P_t &= FP_t + P_t F^T - P_t G^T R^{-1} G P_t + 2\varepsilon\mathbb{I}.
\end{aligned}$$

We see we exactly recover the celebrated Kalman-Bucy filter.

Conclusion

We have approximated the Kushner optimal filter by a Gaussian filter based on variational approximations related to the JKO and LMMR proximal discrete schemes related to the Wasserstein and Fisher geometry respectively. As the dynamic and observation processes are assumed independent, we can mix the two variational solutions to form a set of SDEs on the Gaussian parameters generalizing the Riccati equations associated to the linear systems. In the linear case, the proposed filter boils down to the Kalman-Bucy optimal filter. It is still unclear, though, which global variational loss is minimized by the optimal filter.

Acknowledgements

This work was funded by the French Defence procurement agency (DGA) and by the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001(PRAIRIE 3IA Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063).

References

[1] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savare. Gradient flows in metric spaces and in the space of probability measures. *Lectures in Mathematics. ETH Zürich*, 2005.

[3] Damiano Brigo. On nonlinear SDE's whose densities evolve in a finite–dimensional family. volume 23, pages 11–19. Birkhäuser, 1997.

[4] Damiano Brigo, Bernard Hanzon, and François Gland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, 5, 1999.

[5] Donald Bures. An extension of Kakutani's theorem on infinite product measures to the tensor product of semifinite w^*-algebras. *Transactions of the American Mathematical Society*, 135:199–212, 1969.

[6] Abhishek Halder and Tryphon Georgiou. *Gradient flows in uncertainty propagation and filtering of linear Gaussian systems*. 2017.

[7] Abhishek Halder and Tryphon Georgiou. *Gradient flows in filtering and Fisher-Rao geometry*. Annual American Control Conference, 2018.

[8] Bernard Hanzon and R. Hut. New results on the projection filter. *Serie Research Memoranda 0023*, 1, 1991.

[9] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29:1–17, 1998.

[10] H. Kushner. Approximations to optimal nonlinear filters. *IEEE Transactions on Automatic Control*, 12 (5):546–556, 1967.

[11] Marc Lambert, Silvère Bonnabel, and Francis Bach. The recursive variational Gaussian approximation (R-VGA). *Statistics and Computing*, 32(1), 2022.

[12] Marc Lambert, Silvère Bonnabel, and Francis Bach. The continuous-discrete variational Kalman filter (CD-VKF). In *Conference on Decision and Control*, 2022.

[13] Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. In *Advances in Neural Information Processing Systems*, 2022.

[14] R. Laugesen, P. G. Mehta, S. P. Meyn, and M. Raginsky. Poisson's equation in nonlinear filtering. In *Conference on Decision and Control*, pages 4185–4190, 2014.