The continuous-discrete variational Kalman filter (CD-VKF)

Marc Lambert, Silvère Bonnabel, Francis Bach

The continuous-discrete variational Kalman filter (CD-VKF)

Marc Lambert
DGA/CATOD
INRIA - ENS - PSL
marc.lambert@inria.fr

Silvère Bonnabel
ISEA
MINES ParisTech-PSL
silvere.bonnabel@mines-paristech.fr

Francis Bach
INRIA - ENS - PSL
francis.bach@inria.fr

Abstract

We consider the filtering problem of estimating the state of a continuous-time dynamical process governed by a nonlinear stochastic differential equation and observed through discrete-time measurements. As the Bayesian posterior density is difficult to compute, we use variational inference (VI) to approximate it. This is achieved by seeking the closest Gaussian density to the posterior, in the sense of the Kullback-Leibler divergence (KL). The obtained algorithm, called the continuous-discrete variational Kalman filter (CD-VKF), provides implicit formulas that solve the considered problem in closed form. Our framework avoids local linearization, and the estimation error is globally controlled at each step. We first clarify the connections between well known nonlinear Kalman filters and VI, then develop closed form approximate formulas for the CD-VKF. Our algorithm achieves state-of-the-art performances on the problem of reentry tracking of a space capsule.

1 INTRODUCTION

Continuous-discrete estimation problems naturally arise in numerous applications, such as radar tracking, guidance and navigation, see, e.g., [6]. Nonlinear dynamical models often stem from physics, which laws are continuous-time, whereas the observations come digitally in discrete time through sensors' measurements. Moreover, working with discrete-time observations allows for seamlessly accommodating timestamped observations coming from heterogeneous sensors. To account for discrepancies between the actual motion and the model, as well as sensors' errors, it is customary to assume noisy dynamics and measurements, leading to nonlinear stochastic differential equations (SDE)s. Throughout the paper, we will consider a system with a state x_t evolving continuously over time but being observed at discrete times. If the state is observed at time t, the next observation comes after a duration T and is written y_{t+T}. The state and the observation are related by a system of equations of the form:

$$dx_t = f(x_t,t)dt + L(x_t,t)d\beta \qquad \text{state-propagation} \qquad (1)$$

$$y_{t+T} = h(x_{t+T}) + \varepsilon, \qquad \text{state-observation} \qquad (2)$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ represents the dynamics (drift), $L(x_t,t)$ is the magnitude of the noise which potentially depends on the state, β is a Brownian motion such that the small increments satisfy $\delta\beta \sim \mathcal{N}(0, Q\delta t)$, $h : \mathbb{R}^d \to \mathbb{R}^p$ is the observation function, and $\varepsilon \sim \mathcal{N}(0,R)$ is the observation noise. The covariance matrix R may also depend on the state. This is the case, for example, for a radar where the range accuracy depends on the distance.

Bayesian filtering aims at estimating the distribution of the state variable x_t conditionally on past observations. When the observations are continuous the optimal filter is given by the Kushner filter [8]. In the continuous-discrete setting one separates propagation and update. Propagating the distribution of x_t from t to $t+T$ requires solving the Fokker-Planck partial differential equation. This equation may not be solved in

closed form and one needs to resort to approximations. Moreover, updating the state at time $t+T$ in the light of the observation requires conditioning upon the observation y_{t+T}. The conditional, or posterior, distribution may not be computed in a closed form and also needs to be approximated. Popular approaches consist in the well-known extended Kalman filter (EKF), the unscented Kalman filter (UKF) [10], or cubature Kalman filters (CKF) [1] that seek to sequentially approximate the maximum a posteriori (MAP) estimate of the state variable x_t and its covariance.

In the present paper, we follow a slightly different approach that seeks to approximate a distribution p by the closest Gaussian distribution q in the sense of the left Kullback-Leibler divergence [11] defined by:

$$KL(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx = H(q) - \mathbb{E}_q[\log p(x)], \tag{3}$$

where H is the negative entropy defined by $H(q) = \int q(x) \log q(x) dx$. This setting is known as variational Gaussian approximation, see [5, 19], or variational inference [9]. We will apply it for both the propagation in continuous time of the distribution of the state x_t through the Fokker-Planck equation, and the computation of the posterior distribution at state update in discrete time. Hence,

- Our first task is to approximate the SDE (1) through variational Gaussian approximation. This is a variational formulation of Gaussian assumed density approximation (see [21], Chapter 9). To our knowledge, few major advances have been done on the subject. In [4] the problem of propagating and conditioning is tackled jointly through variational optimization of a continuous-discrete process, and non-closed-form formulas are found (owing to the presence of difficult to compute continuous Lagrange multipliers). In [20], an approach based on Fokker-Planck is advocated, and then the intractable formulas are simplified through a heuristic. In this paper, we essentially reformulate the variational approach of [4], but focusing only on propagation between updates we obtain a closed-form formula.

- Our second task is to use variational inference for updates, that is, to account for observations. To this aim we use a recursive version of variational Gaussian approximation [19]. In this recursive version we seek to approximate a posterior distribution conditionally on a single observation. The Gaussian distribution minimizing the left KL divergence can be obtained in closed form and serves as a prior for the next update step in turn. This approach was proposed in earlier work [13], and is derived in a new square-root form herein.

Combining propagation and update, we obtain a novel real-time algorithm we call the continuous-discrete variational Kalman filter (CD-VKF). On the tracking problem we have considered, this filter is shown to outperform the extended Kalman filter (EKF) and to compete with the unscented Kalman filter (UKF). This filter may be less prone to divergence than other Kalman filter variants, since it optimizes a closeness to a target distribution at all times, yielding strong indications of potential stability. Beyond this contribution, we also draw connections between variational inference and various popular filters, and provide clarifications.

The paper is organized as follows: in Section 2 we formally introduce our new filter as the solution of the two mentioned variational approximation problems. In Section 3 we relate existing Kalman filters to left and right KL variational problems, and discuss how they are related to the problems we solve. In Section 4 we derive the optimal solution for the variational Gaussian propagation problem in continuous time and the variational Gaussian update in discrete time. A practical implementation is proposed in Section 5 where a new square-root form is derived. The proposed algorithm is assessed on a reentry tracking of a space capsule and compared with CD-EKF and CD-UKF.

2 Situating the problem

We now define formally the two variational problems we solve and state our main theorem.

2.1 Variational state propagation

The first problem is to propagate a Gaussian distribution throughout the nonlinear SDE (1) from time t to time $t+T$. Let's consider the Euler-Marayama discretization with time step δt of the SDE (1), yielding the transition:

$$p(x_{t+\delta t}|x_t) = \mathcal{N}\big(x_t + f(x_t)\delta t, L(t)QL(t)^T\delta t\big), \tag{4}$$

where the drift L is supposed independent of x.

Following [4], we seek to approximate (1) by a linear SDE of the form

$$dx_t = \big(A(t)x_t + b(t)\big)dt + \Lambda(t)d\beta_t. \tag{5}$$

A similar discretization yields a transition for this equation of the form

$$q(x_{t+\delta t}|x_t) = \mathcal{N}(x_t + A(t)x_t\delta t + b(t)\delta t, \Lambda(t)Q\Lambda(t)^T\delta t). \tag{6}$$

This sequence defines a recursive variational Gaussian process by letting the variational parameters $A(t)$, $b(t)$ and $\Lambda(t)$ be optimal at each time in the following sense.

Problem 1 (variational Gaussian propagation): To best approximate the nonlinear SDE (1) by an equation of the form (5), find at all time t the parameters $A(t), b(t), \Lambda(t)$ which solve

$$\min_{A(t),b(t),\Lambda(t)} KL(q(x_{t+\delta t}, x_t), p(x_{t+\delta t}, x_t)), \tag{7}$$

in the limit $\delta t \to 0$, with p corresponding to the transition (4) associated with the original nonlinear equation, and q corresponding to the transition (6) associated with its linear approximation governed by $A(t), b(t), \Lambda(t)$.

2.2 Variational state update

The second problem is to incorporate an observation at time $t+T$. Let us consider the observation model (2). Given an approximate Gaussian prior $x_{t+T|t} \sim \mathcal{N}(\mu_{t+T|t}, P_{t+T|t})$ as a result of the previous variational propagation scheme, at time $t+T$, and a single observation y_{t+T}, we seek a variational Gaussian approximation to the assumed posterior

$$p(x_{t+T}|y_{t+T}) \propto p(y_{t+T}|x_{t+T|t})q(x_{t+T|t}). \tag{8}$$

Problem 2 (variational Gaussian update): Find the Gaussian distribution conditionally on the new single observation $q(x_{t+T}|y_{t+T}) = \mathcal{N}(\mu_{t+T}, P_{t+T})$ which approximates the posterior (8):

$$\underset{\mu_{t+T}, P_{t+T}}{\arg\min} \quad KL(q(x_{t+T}|y_{t+T}), p(x_{t+T}|y_{t+T})). \tag{9}$$

This scheme has already been explored in our previous work and we called our solution the recursive variational Gaussian approximation (RVGA) [13].

2.3 Main result

As will be proved in the sequel, it turns out that formulas may be derived for both problems. This results in the following algorithm, called the continuous-discrete variational Kalman filter (CD-VKF).

Theorem 1 (CD-VKF)**.** *We consider the nonlinear setting* (1) *and* (2)*. Given an initial prior* $x_0 \sim \mathcal{N}(\mu_0, P_0)$ *at* $t = 0$*, and observations arriving at a fixed period* y_{kT}*, we can approximate the true conditional distribution process by exactly solving the approximation problems described in Problems 1 and 2 above. This is done by the real-time algorithm which we call the continuous-discrete variational Kalman filter (CD-VKF), and which is described below.*

CD-VKF Algorithm:
while $t \geq 0$ **do**

(1) Variational Gaussian propagation from t **to** $t+T$

The solution to the linear equation (5) with optimal parameters solving Problem 1 has a Gaussian distribution $q = \mathcal{N}(\mu_{t+T|t}, P_{t+T|t})$ satisfying between two successive observations

$\begin{pmatrix} \mu_{t+T|t} \\ P_{t+T|t} \end{pmatrix} = \int_t^{t+T} \begin{pmatrix} \dot{\mu}(s) \\ \dot{P}(s) \end{pmatrix} ds$ where:

$\dot{\mu}(t) = \mathbb{E}_{q(\mu,P)}[f(x,t)]$;

$\dot{P}(t) = \mathbb{E}_{q(\mu,P)}[J_f(x)]P + P\mathbb{E}_{q(\mu,P)}[J_f(x)]^T + L(t)QL(t)^T$;

where $J_f(x)$ is the (square) Jacobian matrix of the dynamics drift.

(2) Variational Gaussian update at $t+T$

The solution to the Problem 2 is $q = \mathcal{N}(\mu, P)$ st:

$v(x) = J_h(x)^T R^{-1}(y_{t+T} - h(x))$;

$\mu = \mu_{t+T|t} + P_{t+T|t}\mathbb{E}_{q(\mu,P)}[v(x)]$;

$P = P_{t+T|t} + \frac{1}{2}\mathbb{E}_{q(\mu,P)}[(x-\mu)v(x)^T]P_{t+T|t} + \frac{1}{2}P_{t+T|t}\mathbb{E}_{q(\mu,P)}[(x-\mu)v(x)^T]^T$;

where $J_h(x)$ is the (rectangular) Jacobian matrix of observation map.

$t \leftarrow t+T$;

end

The proof, along with more details and comments, will be given in Section 4. The hope that underlies our approach is that as this filter solves optimization Problems 1 and 2, the estimates are better controlled than, say, those of an EKF, so that more stability can be expected. But before getting further into CD-VKF, which will be the topic of Section 4, we would like to review and make some interesting connections between variational inference and different variants of Kalman filters.

3 A review of Kalman filters in the light of variational inference

3.1 Left vs right KL divergence

Our variational problems are stated for the left Kullback-Leibler divergence (3). In Kalman filtering the problems are generally formulated for the right one. Indeed, from definition (3) we see KL divergence is not symmetric. If q is our variational distribution the divergence (3) will be called the "left KL". In contrast, we

define the right KL as:

$$KL(p(x)||q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx = C - \mathbb{E}_p[\log q(x)], \tag{10}$$

where $C = H(p)$ is now a constant independent of our variable q. Both KLs involve expectations: in the left KL the expectation is under a Gaussian q, whereas in the right one it is under an unknown distribution we seek to approximate. Properties of integrals under Gaussian distributions make the left KL attractive and may often lead to closed form updates in particular for inference. Moreover the left KL gives a lower bound on the log-marginal likelihood and is used in variational batch optimization in which each iteration is guaranteed not to decrease the bound [9]. Finally, when considering the Gaussian approximation of a sharp distribution, the left KL is known to better catch the mode than the right one.

However, a large class of Kalman filters are based on moment matching and, as such, are implicitly defined with the right KL not the left one. To see how moment matching is related to the KL divergence, let us compute a right KL Gaussian approximation for $q(x) = \mathcal{N}(x|\mu, P)$:

$$KL(p(x)||q(x)) = C - \mathbb{E}_p[-\frac{1}{2}(x-\mu)^T P^{-1}(x-\mu)] + \frac{1}{2}\log\det P.$$

To find the optimal distribution q, we search for the critical points by cancelling the derivatives of the KL with respect to the variational parameters, here μ and P. Using the relation $\nabla_P \log\det P = P^{-1}$ and $\nabla_P x^T P^{-1} x = -P^{-1} x x^T P^{-1}$, we find:

$$\mu = \mathbb{E}_p[x] \text{ and } P = \mathbb{E}_p[(x-\mu)(x-\mu)^T], \tag{11}$$

which is exactly the moment matching between p and q. However, computing the expectations under p is not easy.

In (3), the KL divergence is defined with respect to a generic distribution $p(x)$, but the KL divergence may also be computed with respect to joint distributions $p(x,y)$ or conditional distributions $p(x|y)$. It turns out that all these variants lead to different Kalman filters. Joint and conditional divergences can be related through the following formula which will prove useful to derive our results:

$$KL(q(z,x)||p(z,x)) = KL(q(x)||p(x)) + \mathbb{E}_{q(x)}[KL(q(z|x)||p(z|x))]. \tag{12}$$

Finally, to better understand the nature of the left and right KL, we may also consider the formalism of Bayesian networks or graphical models [9]. A variational theory has been developed for graphical models [9] and is the subject of active research. The left KL divergence is associated to variational message passing algorithms [24], and the right one to the expectation-propagation algorithm [15] for factor graphs.

We now reconsider several nonlinear Kalman filters from the point of view of variational approximation, and define more precisely how our approach is related to existing methods.

3.2 Variational Gaussian propagation

If the state x follows the SDE (1), the evolution of its distribution p satisfies the Fokker-Planck equation (FPE):

$$\frac{\partial p(x,t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i} f_i(x,t) p(x,t) + \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [L(x,t) Q L(x,t)^T]_{i,j} p(x,t). \tag{13}$$

The Fokker-Planck equation (13) is a partial differential equation which may not be solved in the general case. However, since we are interested in Gaussian approximations, we need to compute only the two first

moments of the distribution. We can then derive a set of two ordinary differential equations (ODE) from the FPE using the definition of the mean and covariance matrix as an expectation under p:

$$\begin{aligned}\frac{d}{dt}\mu_t &= \mathbb{E}[f(x,t)] \\ \frac{d}{dt}P_t &= \mathbb{E}[f(x,t)(x-\mu_t)^T] + \mathbb{E}[(x-\mu_t)f(x,t)^T] + \mathbb{E}[L(x,t)QL(x,t)^T].\end{aligned} \tag{14}$$

Unfortunately, these ODEs are implicit and involve expectations under the unknown target distribution p. The CD-UKF filter [20] is based on the following heuristic: the expectations are computed under the current Gaussian q instead of the unknown distribution p (numerical computation of expectations is then obtained via quadrature rules provided in the UKF filter) [10]. Computing expectations under the Gaussian distribution q rather than p appears as a heuristic, however the resulting form resembles the one obtained by [4] using a wholly different method based on the left KL approximation of the nonlinear process. This relation suggests that there is a deeper connection between approximated Fokker-Planck equations and variational approximations, as mentioned in [20], but without further explanation. In Section 4, we will clarify this connection and lay the theoretical foundations to justify it, as this heuristic approximation done in CD-UKF corresponds exactly to the solution of the recursive variational Gaussian process as defined in (7).

3.3 Variational Gaussian update

The problem of Bayesian inference is to approximate the posterior distribution $p(x|y)$ with a Gaussian $q(x) = \mathcal{N}(x|\mu, P)$. Bayes rule gives the exact formula for the posterior distribution $p(x|y) \propto p(y|x)q_0(x)$ where $q_0(x) = \mathcal{N}(x|\mu_0, P_0)$ is the Gaussian prior on x. If the observation noise is Gaussian $\varepsilon \sim \mathcal{N}(0, R)$, the conditional distribution of observation is always Gaussian $p(y|x) \sim \mathcal{N}(h(x), R)$ but $p(y|x)p(x) = p(x,y)$ is not a Gaussian due to the potential non-linearity of h. To approximate the posterior with a Gaussian, different classes of methods may be considered:

1. Taylor-linearization of the observation

$$h(x) \approx h(\mu_0) + \nabla h_{|\mu_0} x = b + Hx. \tag{15}$$

2. Statistical linearization of the observation

$$\begin{aligned}&h(x) \approx Hx + b, \\ &\text{where } H \text{ and } b \text{ minimize the least mean square loss:} \\ &\int (h(x) - Hx - b)^T (h(x) - Hx - b) q_0(x) dx.\end{aligned} \tag{16}$$

3. Approximation of $p(x|y)$ with moment-matching

$$\begin{aligned}&\mu = \int x p(x|y) dx = \frac{1}{Z} \int x p(y|x) q_0(x) dx \\ &P = \frac{1}{Z} \int x x^T p(y|x) q_0(x) dx - \mu\mu^T \\ &\text{where } Z = \int p(y|x) q_0(x) dx.\end{aligned} \tag{17}$$

4. Approximation of $p(x,y)$ with moment-matching

$$\begin{aligned}&p(x,y) \approx \\ &\mathcal{N}\left(\begin{bmatrix}\mathbb{E}[X] \\ \mathbb{E}[Y]\end{bmatrix}, \begin{bmatrix}\mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y)\end{bmatrix}\right)\end{aligned} \tag{18}$$

where the expectations and the covariances are w.r.t. the joint distribution $(X,Y) \sim p(x,y) = p(y \mid x)q_0(x)$. Recalling $Y \sim \mathcal{N}(h(x),R)$, all the quantities in (18) are easily re-written in terms of x, $h(x)$ and $q_0(x)$ only.

The first approach is used in the celebrated extended Kalman filter (EKF), and is the most common in practice because it does not require computing any expectation. The second one is used in the quadrature Kalman filter (QKF) [3], and was associated with the Gauss-Hermite quadrature rules. We will see in Section 4.2 that this linearization technique is related to the left KL. The third approach is used in the assumed density filter [18], and was extended for batch machine learning and factor graphs through the expectation-propagation algorithms [15]. This approach is derived with the right KL (10). Finally, the last approach corresponds to filters based on the innovation process like the unscented Kalman filter (UKF) [10], or the cubature Kalman filter (CKF) [1]. They are also related to the right KL (10) but applied to the joint distribution.

When the posterior distribution $p(x|y)$ is not directly computed with moment matching (17), it can be deduced from the joint distribution $p(x,y)$ using the conditional formulas that we use in the Gaussian case:

$$\begin{aligned} \mathbb{E}[X|y] &= \mathbb{E}[X] + \mathrm{Cov}(X,Y)\mathrm{Cov}(Y,Y)^{-1}(y - \mathbb{E}[Y]) \\ \mathrm{Cov}(X|y) &= \mathrm{Cov}(X,X) - \mathrm{Cov}(X,Y)\mathrm{Cov}(Y,Y)^{-1}\mathrm{Cov}(Y,X). \end{aligned}$$

In the nonlinear case (UKF, CKF), the expectations may be approximated with quadrature rules [10] for some well-chosen sigma points x_i: $\mathbb{E}[f(x)] \approx \sum_i w_i f(x_i)$, see [14] for a survey on quadrature rules. In the linearized case (EKF or QKF) the expectations are available in closed form using the relation $y = Hx + b + \mathcal{N}(0,R)$ and the conditional formulas yield the classical Kalman update equations:

$$\begin{aligned} \mu &= \mu_0 + P_0 H^T (R + HP_0H^T)^{-1}(y - H\mu_0 - b) \\ &= \mu_0 + Kz \\ P &= P_0 - P_0 H^T (R + HP_0H^T)^{-1} HP_0 \\ &= P_0 - KHP_0, \end{aligned} \tag{19}$$

where $\mu = \mathbb{E}[x|y]$, $P = \mathrm{Cov}(x|y)$, K is the Kalman gain, and $z = y - H\mu_0 - b$ is the estimation error (the innovation). In practice, the error z is replaced by the non linearized quantity $y - h(\mu_0)$. To sum up, the moment matching (17) minimizes directly the right KL divergence for our quantity of interest $p(x|y)$ while the innovation based method (18) indirectly finds the posterior by first minimizing the right KL on the joint distribution $p(x,y)$, and then applying the Gaussian conditioning. Finally, statistical linearization (16) also first approximates the conditional distribution using a linear relation that optimizes the left KL and then using Gaussian conditioning. This is summarized in Table 3.4.

3.4 Summary

We summarize our review in Table 3.4, where M-M stands for moment matching methods, J-M-M for moment matching of the joint distribution, SLR for statistical linearization, VI for variational inference, MM-GP for moment-matching based Gaussian process and V-GP for variational (left KL) based Gaussian process. The associated equations are given in parentheses; $KL(.||p)$ stands for the left KL and $KL(p||.)$ for the right one.

4 Derivation of Theorem 1

We now prove our main theorem: first we solve the variational Gaussian propagation problem (7) in continuous time, then the variational Gaussian update problem (9) in discrete time.

Table 1: Kalman filters as variational minimizers

Method	KL	Filters	References
MM (17)	$KL(p(x\|y)\|\|.)$	ADF, E-P	[18], [15]
VI (28)	$KL(.\|\|p(x\|y))$	VGA, VB	[5], [19], [22]
		R-VGA	[13]
		CD-VKF	this paper
J-MM (18)	$KL(p(x,y)\|\|.)$	UKF, CKF	[10], [1]
SLR (16)	$KL(.\|\|p(y\|x)p(x))$	QKF	[3]
MM-GP	$KL(p(x_{t+1},x_t)\|\|.)$	ADF	[7]
V-GP (24)	$KL(.\|\|p(x_{t+1},x_t))$	SLR-GP	[23], [4]
		CD-UKF	[20]
		CD-CKF	[2]
		CD-VKF	this paper

4.1 Solution to the variational propagation problem (7)

Let us consider a small step δt such that the Euler-Marayama approximation of the SDE (1) is :

$$x_{t+1} = x_t + f(x_t)\delta t + L(x_t,t)\delta\beta_t; \quad \delta\beta_t \sim \mathcal{N}(0,\delta t Q).$$

Starting from q_t the propagated distribution is:

$$\begin{aligned} p(x_{t+1}) &= \mathbb{E}_{q_t}[p(x_{t+1}|x_t)] \\ &= \mathbb{E}_{q_t}[\mathcal{N}(x_{t+1}|x_t + f(x_t)\delta t, \delta t V(x_t))], \end{aligned}$$

where $V(x_t)$ is a compact notation for $L(x_t,t)QL(x_t,t)^T$. The variational approximation of this marginal with a Gaussian q_{t+1} for the left KL minimizes:

$$KL(q(x_{t+1})||p(x_{t+1})) = H(q_{t+1}) - \mathbb{E}_{q_{t+1}}[\log p(x_{t+1})], \tag{20}$$

where H is the negative entropy. Unfortunately, the optimal parameters are not available in closed form. To circumvent this problem, we consider the auxiliary problem which consists in approximating the process with a linear process as proposed in [3].

4.1.1 Variational linearization of the process

Our linear process for a small step δt is parametrized by $b(t)$, $A(t)$ and $\Lambda(t)$, to simplify the notation we will omit the time variable:

$$q(x_{t+1}|x_t) \sim \mathcal{N}(x_{t+1}|x_t + Ax_t\delta t + b\delta t, \delta t\Lambda).$$

The core of our approach is to consider a divergence between the following joint distributions:

$$\min_{b,A,\Lambda} KL(q(x_{t+1},x_t)||p(x_{t+1},x_t)).$$

The KL may be rewritten using the formula (12):

$$\begin{aligned} &KL(q(x_{t+1},x_t)||p(x_{t+1},x_t)) \\ &= \mathbb{E}_{q_t}[KL(q(x_{t+1}|x_t)||p(x_{t+1}|x_t))], \end{aligned}$$

where the first term in (12) vanished since the prior is supposed to follow the same Gaussian q_t. We obtain a divergence between two Gaussians which takes the form:

$$\mathbb{E}_{q_t}[\frac{1}{2}\mathrm{Tr}\Lambda V(x_t)^{-1} + \frac{1}{2}\log\det V(x_t) - \frac{1}{2}\log\det\Lambda - \frac{1}{2}d \\ + \frac{\delta t}{2}(f(x_t) - Ax_t - b)^T V(x_t)^{-1}(f(x_t) - Ax_t - b)].$$

The advantage of using the left KL clearly appears here where the mean variational parameters A and b are decoupled from the covariance variational parameter Λ. Taking the derivatives with respect to Λ, A and b, and using the relation $\nabla_\Lambda \mathrm{Tr}\Lambda V^{-1} = V^{-1}$ and $\nabla_\Lambda \log\det\Lambda = \Lambda^{-1}$, we obtain the following set of equations:

$$\Lambda = \mathbb{E}_{q_t}[(L(x_t,t)QL(x_t,t)^T)^{-1}]^{-1} \quad (21)$$
$$\mathbb{E}_{q_t}[V(x_t)^{-1}Ax_t x_t^T] = \mathbb{E}_{q_t}[V(x_t)^{-1}(f(x_t) - b)x_t^T] \quad (22)$$
$$\mathbb{E}_{q_t}[V(x_t)^{-1}Ax_t] = \mathbb{E}_{q_t}[V(x_t)^{-1}(f(x_t) - b)]. \quad (23)$$

To obtain a closed form expression, we suppose that the drift term is independent of the state x_t, i.e., $L(x_t,t) = L(t)$. The set of equations becomes:

$$\begin{aligned} \Lambda &= L(t)QL(t)^T \\ A &= \mathbb{E}_{q_t}[f(x_t)(x_t - \mu_t)^T]P^{-1} \\ b &= \mathbb{E}_{q_t}[f(x_t)] - A\mu_t. \end{aligned} \quad (24)$$

4.1.2 Variational Gaussian propagation

We now have a linear process which is given by a one-step update as:

$$x_{t+1} = x_t + Ax_t\delta t + b\delta t + \delta\beta_t; \quad \delta\beta_t \sim \mathcal{N}(0, \delta t\Lambda).$$

It is straightforward to compute its mean and covariance:

$$\begin{aligned} \mu_{t+1} &= \mathbb{E}[x_{t+1}] \\ &= \mu_t + A\mu_t\delta t + b\delta t \\ P_{t+1} &= \mathbb{E}[(x_{t+1} - \mu_{t+1})(x_{t+1} - \mu_{t+1})^T] \\ &= P_t + \delta t AP_t + \delta t P_t A^T + \delta^2 t AP_t A^T + \delta t\Lambda. \end{aligned}$$

Taking the limit $\delta t \to 0$ and replacing A, b and Λ with their expressions (24), we find a set of two coupled ODEs:

$$\begin{aligned} \dot{\mu} &= \mathbb{E}_{q(\mu,P)}[f(x,t)] \\ \dot{P} &= \mathbb{E}_{q(\mu,P)}[f(x_t)(x_t - \mu)^T] + \mathbb{E}_{q(\mu,P)}[(x_t - \mu)f(x_t)^T] + L(t)QL(t)^T. \end{aligned} \quad (25)$$

It is quite remarkable how the resulting form is close to the one derived from the Fokker-Planck equation (14), but the expectations are now on q, as proposed initially by Sarkka & al. [20] as a heuristic.

Finally, if we suppose f sufficiently smooth, we may use the properties of Gaussians $\nabla_x \mathcal{N}(x|\mu,P) = -P^{-1}(x-\mu)\mathcal{N}(x|\mu,P)$, and integration by parts to let the Jacobian matrix $\nabla_x f(x)$ appear in (25):

$$\mathbb{E}_q[f(x)(x-\mu)^T] = \mathbb{E}_q[\nabla_x f(x)]P, \quad (26)$$

which is a generalization of the Stein Lemma to multivariate functions (see [25]). Hence the form given in Theorem 1.

4.2 Solution to the variational update problem (9)

The variational Gaussian inference problem is to update a prior distribution, say, q_0, with the new observation y such that the approximated posterior satisfies:

$$(\mu, P) = \underset{q\sim\mathcal{N}(\mu,P)}{\operatorname{argmin}} \quad KL(q(x|y)||p(x|y)), \tag{27}$$

where $p(x|y) \propto p(y|x)q_0(x)$ from Bayes rule, and $p(y|x)$ corresponds to the general observation model (2) with state-dependent covariance $p(y|x) = \mathcal{N}(h(x), R(x))$.

One could follow a similar approach to the previous propagation step by minimizing $KL(q(x,y)||p(x,y))$, based on a linear relation $q(y|x) = \mathcal{N}(y|Hx+b, M)$. Similar computations then allow recovering the statistical linearization solution (16) based on least square loss minimization [3]. However, as concerns the update step, it turns out that using variational inference and Bayes rule we may tackle the initial problem directly:

$$\begin{aligned} KL(q(x|y)||p(x|y)) &= \int q(x|y) \log \frac{q(x|y)}{p(x|y)} dx \\ &= \int q(x|y) \log \frac{q(x|y)}{p(y|x)q_0(x)} dx + c \\ &= H(q) - \mathbb{E}_q[\log p(y|x)] - \mathbb{E}_q[\log q_0(x)] + c, \end{aligned} \tag{28}$$

where H denotes the negative entropy and c is a constant independent of q. This approach was developed in our previous work in the context of online machine learning [13]. Zeroing the derivatives with respect to μ and P gives the following updates (see [13], Theorem 1 for details):

$$\begin{aligned} \mu &= \mu_0 + P_0 \nabla_\mu \mathbb{E}_q[\log p(y|x)] \\ P^{-1} &= P_0^{-1} - 2\nabla_P \mathbb{E}_q[\log p(y|x)]. \end{aligned}$$

These updates can be simplified using the properties of Gaussians:

$$\begin{aligned} \nabla_\mu \mathcal{N}(x|\mu, P) &= -\nabla_x \mathcal{N}(x|\mu, P) \\ \nabla_P \mathcal{N}(x|\mu, P) &= \frac{1}{2} \nabla_x^2 \mathcal{N}(x|\mu, P). \end{aligned}$$

If we suppose p sufficiently smooth, we may use integration by parts as in Section 4.1 to obtain the derivatives of the negative likelihood. The updates are now:

$$\mu = \mu_0 + P_0 \mathbb{E}_q[\nabla_x(\log p(y|x))] \tag{29}$$

$$P^{-1} = P_0^{-1} - \mathbb{E}_q[\nabla_x^2(\log p(y|x))]. \tag{30}$$

Remark 1 (Online natural gradient)**.** *These updates can be interpreted as an averaged version of the online natural gradient which is closely related to the extended Kalman filter (EKF) as shown in [17]. In particular the update* (30) *resembles the information update in EKF but the (stochastic) Fisher information matrix* $-\nabla_x^2 \log p(y|x)$ *is averaged over x. The connexion between these updates and optimization algorithms is discussed in detail in [13] (Section 4) and in [12] (Appendix F).*

The update (30) can be reformulated in a Hessian-free version using a second integration by parts:

$$\begin{aligned} P^{-1} &= P_0^{-1} - P^{-1} \mathbb{E}_q[(x-\mu)\nabla_x \log p(y|x)^T] \\ &= P_0^{-1} (\mathbb{I} + \mathbb{E}_q[(x-\mu)\nabla_x \log p(y|x)^T])^{-1}. \end{aligned} \tag{31}$$

Inverting the above equation yields a new update in covariance:

$$\begin{aligned} P &= P_0 + \mathbb{E}_q[(x-\mu)\nabla_x \log p(y|x)^T]P_0 \\ &= P_0 + \frac{1}{2}\mathbb{E}_q[(x-\mu)\nabla_x \log p(y|x)^T]P_0 \\ &\quad + \frac{1}{2}P_0\mathbb{E}_q[(x-\mu)\nabla_x \log p(y|x)^T]^T, \end{aligned} \tag{32}$$

where the last equation comes from the fact that $\mathbb{E}_q[(x-\mu)\nabla_x \log p(y|x)^T]P_0$ is symmetric since P is symmetric. It allows us to ensure that this update always provides a symmetric matrix. We now apply this scheme to the particular case of Gaussian observations. The probability $p(y|x)$ depends on x through the observation function $h : \mathbb{R}^n \to \mathbb{R}^m$. The derivatives of $-\log p$ can be obtained using the chain rule:

$$\begin{aligned} &\nabla_x \log p(y|x) = \nabla_x \big[-\frac{1}{2}(y-h(x))^T R^{-1}(y-h(x)) \big] \\ &= J_h(x)^T R^{-1}(y-h(x)), \end{aligned}$$

where $J_h(x) \in \mathscr{M}(m,n)$ is the Jacobian matrix of h and the products are standard matrix products. Plugging the gradient in (32), we recover the update step of Theorem 1.

5 Practical implementation

We now propose an implementation of the updates described in Theorem 1.

5.1 Main difficulties

We need to tackle several difficulties before obtaining a tractable filter:

- **Solving the implicit scheme:** The theoretical scheme is implicit, since the expectation $\mathbb{E}_q$ depends on the unknown parameters μ and P. At propagation two ODEs can be vectorized to take into account the coupling between the variables, and solved by an explicit Runge-Kutta scheme to obtain the propagated moments μ_{t+T}, P_{t+T} up to the next observation date. To compute the update (29)-(30) we "open the loop" and substitute distribution q with its estimate before observation. An alternative is to use "extragradients" to approximate the implicit scheme as is done in proximal optimization and in the iterated Kalman filter (see [12] Appendix F).

- **Computing the expectations:** The expectations can be approximated with quadrature rules using sigma points as proposed in [20] and [1]. $2d+1$ sigma points are spread around the covariance as follows $x_i = \mu + c_i R e_i$, $i = 0, \dots, 2d$, where $e_i|_{i=1,\dots,n}$ are basis vectors and $e_i|_{i=n+1,\dots,2n}$ their opposite, R is the covariance square root (given by Cholesky decomposition) and c_i are spreading factors. The quadrature rules are given by $\mathbb{E}[f(x)] \approx \sum_{i=0}^{2n} w_i f(\mu + c_i R e_i)$.

- **Ensuring symmetry and positive definiteness:** The covariance matrix may no longer be symmetric or positive after propagation and update. We detail below a square root form to remedy the problem.

5.2 Proposed square-root form

Using numerical operations on the square root of the covariance matrix rather than on the full matrix helps maintaining a positive and definite covariance matrix during propagation and update.

5.2.1 Propagation

For the propagation, we use the continuous-time square-root version developed in [16] and applied in [20]. We consider a lower triangular matrix R such that $P = RR^T$. We note $\dot{P} = F(R)$ the ODE on P expressed as a function of R. For example using the sigma point approach $x_i = \mu + c_i R e_i$, we can let the variable R appear explicitly in equation (25). The ODE can be rewritten:

$$\frac{dP}{dt} = \frac{d(RR)}{dt} = \frac{dR}{dt}R^T + R\frac{dR^T}{dt} = F(R), \tag{33}$$

multiplying by R^{-1} on the left and R^{-T} on the right gives:

$$R^{-1}\frac{dR}{dt} + \frac{dR^T}{dt}R^{-T} = R^{-1}F(R)R^{-T}. \tag{34}$$

The solution is given by:

$$R^{-1}\frac{dR}{dt} = \text{Tria}(R^{-1}X(R)R^{-T}) \tag{35}$$

$$\frac{dR}{dt} = R\,\text{Tria}(R^{-1}X(R)R^{-T}), \tag{36}$$

where $\text{Tria}(A)$ gives the lower triangular matrix L corresponding to A such that $A = L + L^T$ where $L_{i,i} = \frac{1}{2}A_{i,i}$, $L_{i,j} = A_{i,j}$ if $i > j$ and $L_{i,j} = 0$ elsewhere.

5.2.2 Update

Considering the prior is available in a square-root form $P_0 = R_0 R_0^T$, we search for $P = RR^T$. The update (32) can be rewritten in a square root form as follows:

$$\begin{aligned} RR^T &= R_0R_0^T + \frac{1}{2}\mathbb{E}_q[(x-\mu)v(x)^T]R_0R_0^T \\ &+ \frac{1}{2}R_0R_0^T\mathbb{E}_q[v(x)(x-\mu)^T] \\ &\text{where } v(x) = \nabla_x \log p(y|x) = J_h(x)^T R^{-1}(y-h(x)). \end{aligned}$$

Let us "open the loop" and consider integrals under q_0 instead of q such that we can insert the sigma points $x_i = \mu_0 + c_i R_0 e_i$:

$$\begin{aligned} RR^T &= R_0R_0^T + \frac{1}{2}\sum_{i=1}^{N} w_i(c_iR_0e_i)v(\mu_0 + c_iR_0e_i)^T R_0R_0^T \\ &+ \frac{1}{2}\sum_{i=1}^{N} w_iR_0R_0^T v(\mu_0 + c_iR_0e_i)^T(c_iR_0e_i)^T. \end{aligned}$$

We can now factorize terms as follows:

$$\begin{aligned} RR^T &= R_0(\mathbb{I} + \frac{1}{2}A + \frac{1}{2}A^T)R_0^T \\ &\text{where } A = \sum_{i=1}^{N} w_ic_ie_iv(\mu + c_iR_0e_i)^T R_0. \end{aligned} \tag{37}$$

We now have a form $RR^T = R_0QR_0^T$, where Q is a symmetric matrix. We can compute the Cholesky decomposition of $Q = LL^T$ to obtain the square-root updates: $R = R_0L$.

5.3 Experimentations

We consider the reentry tracking problem described in [10]. This problem describes a space capsule which reenters the atmosphere with high velocity and drag, making the tracking challenging for a radar. The state is of dimension 5 and composed of 2D-Cartesian positions and velocities vector plus a drag coefficient a: $X = \begin{pmatrix} x(t) & y(t) & \dot{x}(t) & \dot{y}(t) & a(t) \end{pmatrix}$. a is defined such that the aerodynamic coefficient of the vehicle is $\beta = \beta_0 \exp(a)$, where β_0 is our prior value. This coefficient has an unknown dynamic and is modelled as a Brownian noise $\dot{a}(t) = w(t)$. We use the same initial conditions for the state and covariance as in [10] but to make the problem more challenging we consider sparser observations taken every $0.5s$ and higher noise: a standard deviation of $100m$ in range and $100mrad$ in angle is considered. We have made the schemes explicit without using extragradients. The ODEs are integrated with a Runge-Kutta scheme of order 4 with a step $0.25s$ (two steps between observations). All the expectations under Gaussians are computed with quadrature rules based on the UKF sigma points [10].

In Figure 1 the covariance estimated by the CD-VKF is displayed. The standard deviation is consistent with the true trajectories and the filter gives a good estimation of the drag coefficient a of the vehicle.

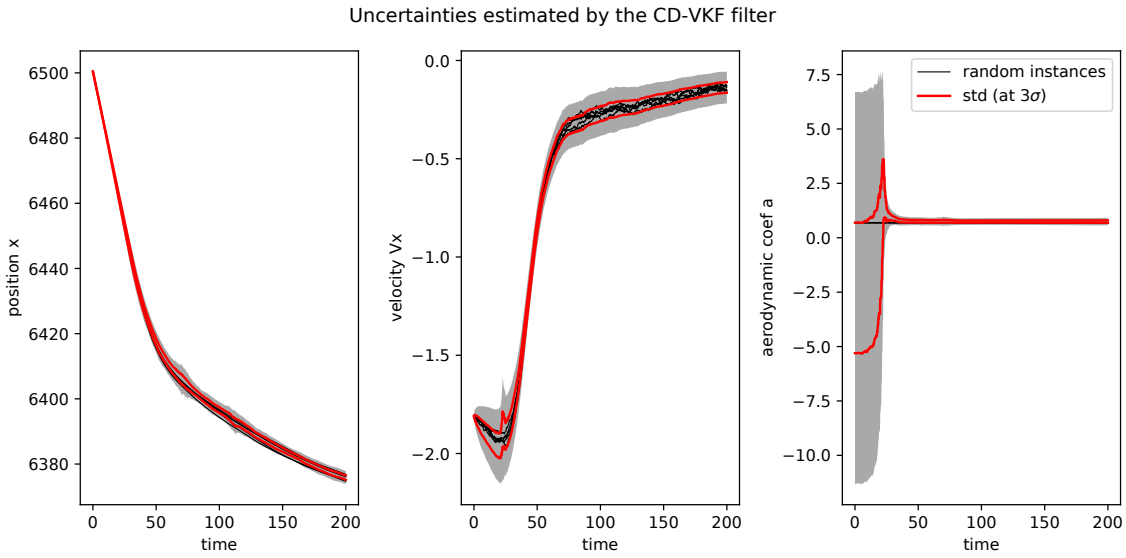

Figure 1: Uncertainties estimated by CD-VKF on the components $x(t)$, $\dot{x}(t)$ and a. The standard deviation around the mean is shown in red at 3σ and in grey at 9σ. The envelopes contain the sampled real trajectories. The drag coefficient converges to the true value at around 0.6.

In Figure 2 we show the root mean squared errors (RMSE) for the (CD-)EKF, the (CD-)UKF and the CD-VKF. The CD-VKF gives identical results as the UKF on this example, and both outperform the EKF.

The sources of the code are available on Github at the following repository:
`https://github.com/marc-h-lambert/CD-VKF`.

6 CONCLUSION

After a classification of Kalman filters as recursive minimizers for the left or right KL, we have introduced a new variational Kalman filter based on the left KL for both propagation and update. Our new algorithm is optimal for the variational optimization problems (7) and (9) and offers a new alternative to moment-matching based filters like the UKF. By optimizing a sensible criterion, it leads to a better control over the error (as compared to e.g., EKF). As a result, we anticipate great stability of this filter in practice. As a

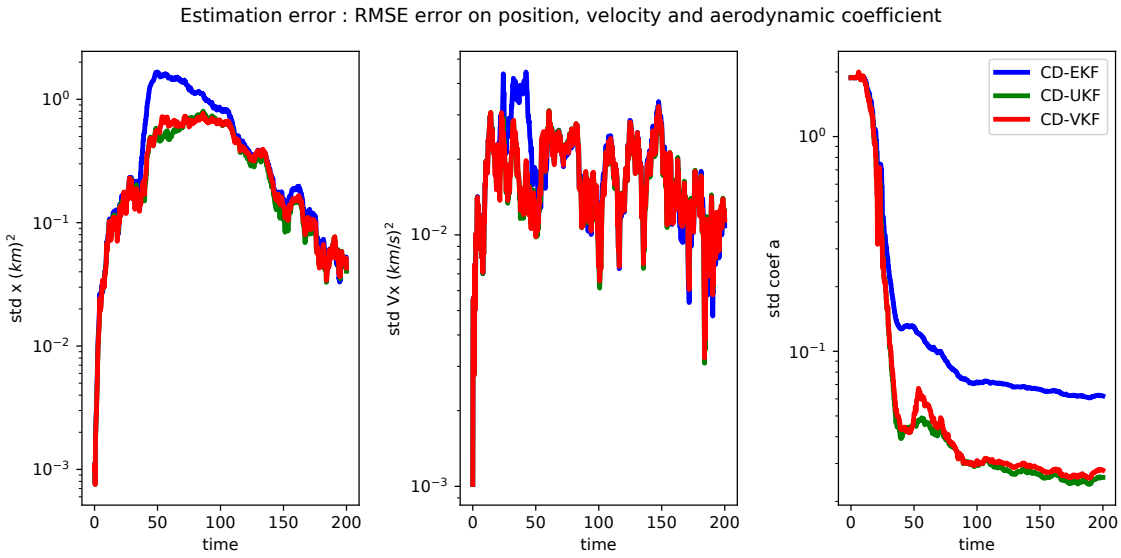

Figure 2: RMSE error achieved by the CD-EKF, CD-UKF and CD-VKF for the components $x(t)$, $\dot{x}(t)$ and a. CD-VKF and CD-UKF yield similar results and outperform the CD-EKF.

left KL minimizer, though, this new Kalman filter is implicit in essence and comes with many options for approximate implementation (implicit Runge-Kutta, Verlet method, etc.) and update (iterate filtering, mirror proximal methods, etc.). Understanding which method is best adapted to these recursive schemes is an open problem.

7 ACKNOWLEDGMENTS

This work was funded by the French Defence procurement agency (DGA) and by the French government under 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). The authors would like to thank Théophile Cantelobre for reviewing and proof reading the draft version of this manuscript.

References

[1] Haran Arasaratnam and Simon Haykin. Cubature Kalman filters. *Automatic Control, IEEE Transactions on*, pages 1254–1269, 2009.

[2] I. Arasaratnam, S. Haykin, and T. R. Hurd. Cubature Kalman filtering for continuous-discrete systems: Theory and simulations. *IEEE Transactions on Signal Processing*, 58(10):4977–4993, 2010.

[3] Ienkaran Arasaratnam, Simon Haykin, and Robert J. Elliott. Discretet-time nonlinear filtering algorithms using Gauss-Hermite quadrature. *Proceedings of the IEEE*, 95(5):953–977, 2007.

[4] Cedric Archambeau, Dan Cornford, Manfred Opper, and John. Shawe-Taylor. Gaussian process approximations of stochastic differential equations. *Journal of Machine Learning Research - Proceedings Track. 1. 1-16.*, 2007.

[5] David Barber and Christopher Bishop. Ensemble learning for multi-layer networks. In *Advances in Neural Information Processing Systems*, volume 10, 1997.

[6] Axel Barrau and Silvère Bonnabel. The invariant extended Kalman filter as a stable observer. *IEEE Transactions on Automatic Control*, 62(4):1797–1812, 2016.

[7] Damiano Brigo, Bernard Hanzon, and François Le Gland. Approximate nonlinear filtering by projection on exponential manifolds of densities. *Bernoulli*, 5(3):495–534, 1999.

[8] A. H. Jazwinski. Stochastic processes and filtering theory. *New York: Academic.*, 1970.

[9] Michael I. Jordan, Zoubin Ghahramani, and et al. An introduction to variational methods for graphical models. In *Machine Learning*, pages 183–233, 1999.

[10] S.J. Julier and J.K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

[11] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951.

[12] Marc Lambert, Silvère Bonnabel, and Francis Bach. The limited-memory recursive variational gaussian approximation (L-RVGA). *hal-03501920*, 2021.

[13] Marc Lambert, Silvère Bonnabel, and Francis Bach. The recursive variational gaussian approximation (R-VGA). *Statistics and Computing*, 32(1):10, 2022.

[14] Henrique M. T. Menegaz, João Y. Ishihara, Geovany A. Borges, and Alessandro N. Vargas. A systematization of the unscented Kalman filter theory. *IEEE Transactions on Automatic Control*, 2015.

[15] Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369, 2001.

[16] M. Morf, B. Levy, and T. Kailath. Square-root algorithms for the continuous-time linear least squares estimation problem. In *1977 IEEE Conference on Decision and Control*, pages 944–947, 1977.

[17] Yann Ollivier. Online Natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12:2930–2961, 2018.

[18] Manfred Opper. A bayesian approach to online learning. In *On-Line Learning in Neural Networks*, 2006.

[19] Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21:786–792, 2009.

[20] Simo Sarkka. On unscented Kalman filtering for state estimation of continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, 52(9):1631–1641, 2007.

[21] Simo Särkkä and Juha Sarmavuori. Gaussian filtering and smoothing for continuous-discrete dynamic systems. *Signal Processing*, 93(2):500–510, 2013.

[22] VÁclav Smidl and Anthony Quinn. Variational Bayesian filtering. *IEEE Transactions on Signal Processing*, 56(10):5020–5030, 2008.

[23] L Socha. Linearization methods for stochastic dynamic systems. *Lecture Notes in Physics*, 730, 2008.

[24] John Winn and Christopher Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 04 2005.

[25] Lin Wu, Emtiyaz Khan Mohammad, and Schmidt Mark. Stein's lemma for the reparameterization trick with exponential family mixtures. *arXiv:1910.13398*, 2019.