---

**Proceedings Paper:**
Liu, Y., Li, X., Yang, L. et al. (2 more authors) (2024) On the Gaussian filtering for nonlinear dynamic systems using variational inference. In: Proceedings of the 2024 27th International Conference on Information Fusion (FUSION). 2024 27th International Conference on Information Fusion (FUSION), 08-11 Jul 2024, Venice, Italy. Institute of Electrical and Electronics Engineers (IEEE) ISBN 9798350371420

https://doi.org/10.23919/FUSION59988.2024.10765592

---

# On the Gaussian Filtering for Nonlinear Dynamic Systems Using Variational Inference

Yi Liu[1◇], Xi Li[2◇], Le Yang[1*], Lyudmila Mihaylova[3], Ji Li[4]
1. Department of Electrical and Computer Engineering, University of Canterbury, Christchurch NZ
2. State Key Lab of Complex Electromagnetic Environment Effects on Electronics and Information Systems, National University of Defense Technology, Changsha China
3. Department of Automatic Control and System Engineering, University of Sheffield, Sheffield UK
4. Locaris Electronic Technology Co. Ltd, Zhengzhou China
◇: Equal contributors, *: corresponding author, le.yang@canterbury.ac.nz

***Abstract*—This paper introduces a new variational Gaussian filtering approach for estimating the state of a nonlinear dynamic system. We first assume that the predictive distribution of the state is Gaussian and derive an iterative method for updating the state posterior in the natural parameter space through Kullback-Leibler divergence minimization. The obtained update rule is the same as that of the conjugate-computation variational inference technique in Bayesian learning. The derivation here is simpler and more insightful. We then impose a Wishart prior on the inverse of the state prediction covariance to take into account the impact of approximating the state predictive distribution using a Gaussian density on the state posterior estimation. The prediction covariance is identified jointly with the state using variational inference and the established state posterior update rule to achieve the desired Gaussian filtering. Simulation study examines the performance of the proposed filtering framework in target tracking based on bearing and range measurements.**

## I. Introduction

Consider a dynamic system whose state evolves according to a first-order Markov process. At time step $t$, the system state $\mathbf{x}_t$ and measurement $\mathbf{z}_t$ are modeled as samples from the state transition density $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and measurement likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$, i.e.,

$$\mathbf{x}_t \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}) \text{ and } \mathbf{z}_t \sim p(\mathbf{z}_t|\mathbf{x}_t). \tag{1}$$

We are interested in inferring the state $\mathbf{x}_t$ from the measurements collected up to time step $t$, $\mathbf{z}_{1:t}$, based on the state-space model in (1). For this purpose, the filtering posterior $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ needs to be computed, which can be achieved recursively [1], [2] by first finding the predictive distribution of $\mathbf{x}_t$ using the Chapman-Kolmogorov equation

$$p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})\mathrm{d}\mathbf{x}_{t-1}, \tag{2}$$

and then performing the measurement update by evaluating

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) = \frac{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})}. \tag{3}$$

The Bayesian filtering recursion given in (2) and (3) admits a closed-form solution, the celebrated Kalman filter (KF) [3], for linear Gaussian systems. However, nonlinearity frequently arises in practice. For example, in radar tracking, the bearing and range measurements are nonlinearly related to the target position [4]. Besides, the model for the target motion may be nonlinear as well due to e.g., the turn rate being unknown [5]. The presence of nonlinearity makes solving Bayesian filtering equations analytically as in KF no longer feasible, because evaluating the integral in (2) and/or the normalization factor in (3), $p(\mathbf{z}_t|\mathbf{z}_{1:t-1}) = \int p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})\mathrm{d}\mathbf{x}_t$, is intractable.

The assumed density filtering (ADF) with Gaussian assumption, or simply Gaussian filter (GF), is popular among the methods for coping with nonlinear systems. A wide variety of GFs are available. They include the Taylor-series expansion-based extended KF (EKF) [6], [7] and posterior linearization filter [8], [9]. Through the application of different numerical integration methods [10], the unscented KF (UKF) [11], Gaussian-Hermite KF (GHKF) [12]–[14], cubature KF (CKF) [15], [16] and Gaussian-Hermite quadrature filter (GHQF) [17] have been developed. These filters approximate the state predictive distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t-1})$ in (2) and state posterior $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ in (3) using Gaussian densities $\mathcal{N}(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$ and $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$. Here, $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution in $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

This paper introduces a new Gaussian filtering algorithm that follows the existing GFs' prediction-correction framework but is different from them in two aspects. First, we derive an iterative rule for updating the *assumed* posterior $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, based on minimizing the forward Kullback-Leibler divergence (KLD) [18] between it and the 'true' posterior in (3). Noting that the Gaussian density belongs to the exponential family [19], the update equation is established in the natural parameter space using natural gradient [20]–[22]. In this way, we obtain an update rule identical to the conjugate-computation variational inference (CVI) developed in [23] for large-scale Bayesian learning. Our derivation is simpler, since it does not involve the utilization of the Bregman divergence.

Notice that in [24]–[29], a similar problem was considered. Techniques for updating the mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{\Sigma}_t$ with the gradient-based [24], [25], [27]–[29], natural gradient-like [26] and linearized alternating direction method of multipliers (ADMM) [29] algorithms were established. The development of most existing methods assumed additive Gaussian measurement noise. Moreover, to improve numerical stability, they either applied additional pre-conditioning [24]–[26], [28] or

introduced a Newton-type rule using an approximate Hessian [27]. Our method is more general and applicable to other nonlinear/non-Gaussian systems. It already includes preconditioning from the use of natural gradient, has theoretically guaranteed convergence and does not require careful selection of extra penalty parameters as in [29]. Reverse KLD minimization was adopted to optimize $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ in [24], [30]. The obtained moment matching algorithms are non-iterative but suffer from degraded performance under small measurement noise [30]. In [31], $\boldsymbol{\mu}_t$ and $\boldsymbol{\Sigma}_t$ were found by solving a constrained KLD minimization problem for approximating the optimal Bayesian smoothing.

The other difference comes from imposing a Wishart prior on the inverse of the covariance of the *assumed* state predictive distribution $\mathcal{N}(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$, which is $\boldsymbol{\Lambda}_t = \tilde{\boldsymbol{\Sigma}}_t^{-1}$. As a result, $\boldsymbol{\Lambda}_t$ needs to be estimated together with the system state $\mathbf{x}_t$. Since the joint posterior $p(\mathbf{x}_t, \boldsymbol{\Lambda}_t|\mathbf{z}_{1:t})$ is difficult to evaluate, we apply the mean-field approximation [18], and invoke variational inference and the developed state posterior update rule to realize Gaussian filtering. This modification is inspired by [32]–[34]. It adjusts in a structured way the state prediction covariance $\tilde{\boldsymbol{\Sigma}}_t$ using the measurement $\mathbf{z}_t$ to account for the impact of approximating the state prediction distribution using a single Gaussian density on the state posterior estimation. Different from [34], we do not update the mean of the state prediction $\tilde{\boldsymbol{\mu}}_t$ at the same time, which reduces the number of hyperparameters to be determined. The proposed Gaussian filtering algorithm is applied to track a maneuvering target, whose motion is modeled by a nonlinear turn model with unknown turn rate, based on bearing and range measurements. Promising results are obtained.

The rest of this paper is organized as follows. Section II derives the iterative rule for updating the assumed state posterior on the basis of forward KLD minimization and natural gradient. Section III presents the whole proposed Gaussian filtering algorithm. Simulation results are given in Section IV. Section V concludes the paper.

## II. Posterior Update via KLD Minimization

### A. Problem Formulation and Preliminaries

Suppose that at time step $t$, the state predictive distribution in (2) is Gaussian, i.e., $p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \mathcal{N}(\mathbf{x}_t, \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$. In this case, our goal is to find the Gaussian approximation $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ to the 'true' state posterior $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ in (3) that minimizes the forward KLD between them. The associated loss function is [18]

$$\begin{aligned}\mathcal{D}(q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)||p(\mathbf{x}_t|\mathbf{z}_{1:t})) &= \int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)\log\frac{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}{p(\mathbf{x}_t|\mathbf{z}_{1:t})}\mathrm{d}\mathbf{x}_t \\ &\propto \int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)\log\frac{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}\mathrm{d}\mathbf{x}_t,\end{aligned} \tag{4}$$

where we have put (3) and ignored the normalization factor $p(\mathbf{z}_t|\mathbf{z}_{1:t-1})$. The definition of $\boldsymbol{\lambda}_t$ will be given at the start of the next subsection, Section II.B. Computing directly the optimal mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{\Sigma}_t$ in closed form is difficult. We shall derive an iterative approach for updating them.

The algorithm development begins with noticing that the Gaussian density is a member of the exponential family [19]. This allows us to exploit some appealing statistical and computational properties of the exponential family. They are summarized below.

The distributions in the exponential family can be expressed in the following generic form [19], [35]

$$p(\mathbf{x}) = h(\mathbf{x})\exp\left(\boldsymbol{\eta}^T\boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\eta})\right), \tag{5}$$

where $h(\mathbf{x})$ is the base measure, $\boldsymbol{\eta}$ is the *natural parameter* and $\boldsymbol{\phi}(\mathbf{x})$ is the sufficient statistic for $\boldsymbol{\eta}$. $A(\boldsymbol{\eta})$ is the log partition function. The first-order and second-order partial derivatives of $A(\boldsymbol{\eta})$ with respect to the natural parameter $\boldsymbol{\eta}$ are equal to [35]

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = E(\boldsymbol{\phi}(\mathbf{x})) = \mathbf{m}_{\boldsymbol{\eta}}, \tag{6a}$$

$$\frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T} = \mathrm{cov}(\boldsymbol{\phi}(\mathbf{x})) = \mathrm{FIM}(\boldsymbol{\eta}), \tag{6b}$$

where $E(\boldsymbol{\phi}(\mathbf{x}))$ and $\mathrm{cov}(\boldsymbol{\phi}(\mathbf{x}))$ denote the mean and covariance of the sufficient statistic $\boldsymbol{\phi}(\mathbf{x})$. $\mathbf{m}_{\boldsymbol{\eta}}$ in (6a) is referred to as the *mean parameter* or moment parameter, while $\mathrm{FIM}(\boldsymbol{\eta})$ represents the Fisher information matrix (FIM) [36] of the natural parameter $\boldsymbol{\eta}$.

Finally, it has been shown in [19] that the Jacobian $\partial\boldsymbol{\eta}/\partial\mathbf{m}_{\boldsymbol{\eta}}$ is symmetric and equal to

$$\frac{\partial \boldsymbol{\eta}}{\partial \mathbf{m}_{\boldsymbol{\eta}}} = \mathrm{FIM}(\boldsymbol{\eta})^{-1}. \tag{7}$$

This result is expected, since combining (6a) and (6b) yields

$$\frac{\partial \mathbf{m}_{\boldsymbol{\eta}}}{\partial \boldsymbol{\eta}} = \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T} = \mathrm{FIM}(\boldsymbol{\eta}). \tag{8}$$

According to [37], [38], the multivariate Gaussian density can be written in the form (5) with base measure $h(\mathbf{x}) = 1$, natural parameter $\boldsymbol{\lambda}$, log partition function $A(\boldsymbol{\lambda}) = \frac{1}{2}\log|2\pi\boldsymbol{\Sigma}| + \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, and

$$\boldsymbol{\lambda}^T\boldsymbol{\phi}(\mathbf{x}) = \mathrm{tr}\left(\left(\boldsymbol{\lambda}^{(1)}\right)^T\boldsymbol{\phi}_1(\mathbf{x}) + \boldsymbol{\lambda}^{(2)}\boldsymbol{\phi}_2(\mathbf{x})\right). \tag{9}$$

Here, we have that for the multivariate Gaussian density, the natural parameter $\boldsymbol{\lambda}$ and sufficient statistic $\boldsymbol{\phi}(\mathbf{x})$ are given by

$$\boldsymbol{\lambda} = \begin{bmatrix}\boldsymbol{\lambda}^{(1)} \\ \mathrm{vec}(\boldsymbol{\lambda}^{(2)})\end{bmatrix}, \quad \boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix}\boldsymbol{\phi}_1(\mathbf{x}) \\ \mathrm{vec}(\boldsymbol{\phi}_2(\mathbf{x}))\end{bmatrix}, \tag{10a}$$

$$\boldsymbol{\lambda}^{(1)} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \qquad \boldsymbol{\lambda}^{(2)} = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}, \tag{10b}$$

$$\boldsymbol{\phi}_1(\mathbf{x}) = \mathbf{x}, \qquad \boldsymbol{\phi}_2(\mathbf{x}) = \mathbf{x}\mathbf{x}^T, \tag{10c}$$

where $\mathrm{vec}(\mathbf{Y})$ represents the column-vectorised version of the matrix $\mathbf{Y}$.

As a result, the corresponding mean parameter $\mathbf{m}_{\boldsymbol{\lambda}}$, from the definition of $\boldsymbol{\phi}(\mathbf{x})$ in (10a) and (10c), would be equal to

$$\mathbf{m}_{\boldsymbol{\lambda}} = E(\boldsymbol{\phi}(\mathbf{x})) = \begin{bmatrix}\mathbf{m}_{\boldsymbol{\lambda}}^{(1)} \\ \mathrm{vec}(\mathbf{m}_{\boldsymbol{\lambda}}^{(2)})\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu} \\ \mathrm{vec}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T)\end{bmatrix}, \tag{11}$$

with $\mathbf{m}_{\boldsymbol{\lambda}}^{(1)} = \boldsymbol{\mu}$ and $\mathbf{m}_{\boldsymbol{\lambda}}^{(2)} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$.

From (10b), we obtain that the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the Gaussian density $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be recovered from its natural parameter $\boldsymbol{\lambda}$ via

$$\boldsymbol{\mu} = -\frac{1}{2}\left(\boldsymbol{\lambda}^{(2)}\right)^{-1}\boldsymbol{\lambda}^{(1)}, \text{ and } \boldsymbol{\Sigma} = -\frac{1}{2}\left(\boldsymbol{\lambda}^{(2)}\right)^{-1}. \quad (12)$$

Besides, according to (11), $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be deduced from the mean paramter $\mathbf{m}_{\boldsymbol{\lambda}}$ as well, using

$$\boldsymbol{\mu} = \mathbf{m}_{\boldsymbol{\lambda}}^{(1)}, \text{ and } \boldsymbol{\Sigma} = \mathbf{m}_{\boldsymbol{\lambda}}^{(2)} - \mathbf{m}_{\boldsymbol{\lambda}}^{(1)}(\mathbf{m}_{\boldsymbol{\lambda}}^{(1)})^T. \quad (13)$$

In other words, a Gaussian density is completely specified by either the natural parameter $\boldsymbol{\lambda}$ or mean parameter $\mathbf{m}_{\boldsymbol{\lambda}}$, when it is written in the generic form of the exponential-family distributions in (5).

### *B. Natural Gradient-based Update of the Assumed Posterior*

We express the assumed state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ and state prediction distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t-1}) = \mathcal{N}(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$, which are both Gaussian as pointed out at the beginning of this section, in the form (5) with natural parameters $\boldsymbol{\lambda}_t$ and $\tilde{\boldsymbol{\lambda}}_t$. Putting these results simplifies the forward KLD in (4) into

$$\begin{aligned}
&\mathcal{D}(q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)||p(\mathbf{x}_t|\mathbf{z}_{1:t})) \\
&\propto \int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)\log\frac{h(\mathbf{x}_t)\exp(\boldsymbol{\lambda}_t^T\boldsymbol{\phi}(\mathbf{x}_t) - A(\boldsymbol{\lambda}_t))}{p(\mathbf{z}_t|\mathbf{x}_t)h(\mathbf{x}_t)\exp(\tilde{\boldsymbol{\lambda}}_t^T\boldsymbol{\phi}(\mathbf{x}_t) - A(\tilde{\boldsymbol{\lambda}}_t))}\mathrm{d}\mathbf{x}_t \\
&\propto -E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t)) - (\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t)^T\boldsymbol{\phi}(\mathbf{x}_t)] - A(\boldsymbol{\lambda}_t), \\
&= -E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))] + (\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t)^T\mathbf{m}_{\boldsymbol{\lambda}_t} - A(\boldsymbol{\lambda}_t),
\end{aligned} \quad (14)$$

where in the second line, the factor $A(\tilde{\boldsymbol{\lambda}}_t)$ is neglected, and the last equality is obtained by applying (6a). $E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\cdot]$ denotes taking expectation with respect to the Gaussian density $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$, whose mean parameter is $\mathbf{m}_{\boldsymbol{\lambda}_t}$. This subscript is intentionally introduced to improve the clarity of the remaining algorithm derivation.

We shall present a natural gradient descent-based method for iteratively updating the natural parameter $\boldsymbol{\lambda}_t$ to minimize the KLD in (14) and achieve the desired optimization of the assumed state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$. To find the natural gradient, the gradient of the KLD with respect to $\boldsymbol{\lambda}_t$ has to be computed first. It is equal to

$$\begin{aligned}
&\frac{\partial\mathcal{D}(q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)||p(\mathbf{x}_t|\mathbf{z}_{1:t}))}{\partial\boldsymbol{\lambda}_t} \\
&= -\frac{\partial E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial\boldsymbol{\lambda}_t} + \frac{\partial\mathbf{m}_{\boldsymbol{\lambda}_t}}{\partial\boldsymbol{\lambda}_t}(\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t), \\
&= -\frac{\partial E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial\boldsymbol{\lambda}_t} + \mathrm{FIM}(\boldsymbol{\lambda}_t)\cdot(\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t),
\end{aligned} \quad (15)$$

where (6a) is applied again to arrive at the first equality, and (8) has been substituted to obtain the second equality.

Multiplying both sides of (15) with the inverse of the FIM of the natural parameter $\boldsymbol{\lambda}_t$, $\mathrm{FIM}(\boldsymbol{\lambda}_t)^{-1}$, yields the desired natural gradient, which is

$$\begin{aligned}
&\mathrm{FIM}(\boldsymbol{\lambda}_t)^{-1}\cdot\frac{\partial\mathcal{D}(q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)||p(\mathbf{x}_t|\mathbf{z}_{1:t}))}{\partial\boldsymbol{\lambda}_t} \\
&= -\mathrm{FIM}(\boldsymbol{\lambda}_t)^{-1}\cdot\frac{\partial E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial\boldsymbol{\lambda}_t} + (\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t), \\
&= -\frac{\partial\boldsymbol{\lambda}_t}{\partial\mathbf{m}_{\boldsymbol{\lambda}_t}}\cdot\frac{\partial E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial\boldsymbol{\lambda}_t} + (\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t), \\
&= -\frac{\partial E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial\mathbf{m}_{\boldsymbol{\lambda}_t}} + (\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t),
\end{aligned} \quad (16)$$

where (7) has been utilized to establish the second equality. The last equality in (16) is derived by exploiting the symmetry of the partial derivative $\partial\boldsymbol{\lambda}_t/\partial\mathbf{m}_{\boldsymbol{\lambda}_t}$ and chain rule (see Theorem 1 in [39] for an alternative duality-based proof).

The natural gradient incorporates the information geometry of the assumed state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ through the FIM of the natural parameter $\boldsymbol{\lambda}_t$. Compared with the gradient-based methods, using the natural gradient can improve convergence [20], [21], [39]. With (16), the following rule for iteratively updating $\boldsymbol{\lambda}_t$ can be obtained:

$$\begin{aligned}
\boldsymbol{\lambda}_{t,k+1} &= \boldsymbol{\lambda}_{t,k} - \alpha_k\mathrm{FIM}(\boldsymbol{\lambda}_{t,k})^{-1}\frac{\partial\mathcal{D}(q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)||p(\mathbf{x}_t|\mathbf{z}_{1:t}))}{\partial\boldsymbol{\lambda}_{t,k}}, \\
&= \boldsymbol{\lambda}_{t,k} - \alpha_k(\boldsymbol{\lambda}_{t,k} - \tilde{\boldsymbol{\lambda}}_t) + \alpha_k\frac{\partial E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial\mathbf{m}_{\boldsymbol{\lambda}_{t,k}}},
\end{aligned} \quad (17)$$

where $\boldsymbol{\lambda}_{t,k}$ denotes the estimated natural paramter $\boldsymbol{\lambda}_t$ in the $k$-th iteration, $k = 1, 2, ...$, and $\alpha_k > 0$ is the step size.

To gain more insights, we subtract $\tilde{\boldsymbol{\lambda}}_t$, the natural parameter of the state predictive distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t-1})$ that is *known*, from both sides of (17) to arrive at

$$\begin{aligned}
\boldsymbol{\lambda}_{t,k+1} - \tilde{\boldsymbol{\lambda}}_t &= (1-\alpha_k)(\boldsymbol{\lambda}_{t,k} - \tilde{\boldsymbol{\lambda}}_t) \\
&\quad + \alpha_k\frac{\partial E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial\mathbf{m}_{\boldsymbol{\lambda}_{t,k}}}.
\end{aligned} \quad (18)$$

Several observations are in order.

**Remark 1**: The theoretical development in (14)-(18) is still valid as long as the assumed state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ and state predictive distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t-1})$ both belong to the exponential family and have the same functional form in (5).

**Remark 2**: The update rule (18) is exactly the CVI technique developed in [23] using the Bregman divergence and mirror descent. Its convergence is theoretically established in [40]. On the other hand, the derivation in this paper is based on the forward KLD minimization, which is equivalent to maximizing the following evidence lower bound (ELBO) [41] with respect to the assumed state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$:

$$\int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)\log\frac{p(\mathbf{z}_t|\mathbf{x}_t)\mathcal{N}(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)}{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}\mathrm{d}\mathbf{x}_t. \quad (19)$$

Note that the ELBO in (19) is just the negative of the KLD loss in (4). Our formulation simplifies the algorithm development, and more importantly, makes the obtained update equations

easy to be incorporated into other variational methods for further improving the state estimation (See Section III).

**Remark 3**: When the iteration in (18) converges, it reduces to the fixed-point condition

$$\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t = \frac{\partial E_{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial \mathbf{m}_{\boldsymbol{\lambda}_t}}. \tag{20}$$

The term on the right hand side of (20) can be considered as the natural parameter of the exponential-family approximation of the measurement likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$. This can be understood by noting that if $p(\mathbf{z}_t|\mathbf{x}_t)$ can also be written in the form (5) with a natural parameter $\hat{\boldsymbol{\lambda}}_t$, the optimal assumed state posterior would have a natural parameter $\boldsymbol{\lambda}_t = \tilde{\boldsymbol{\lambda}}_t + \hat{\boldsymbol{\lambda}}_t$.

To complete the proposed method for optimizing the assumed state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$, we give the expression for the partial derivative in (18) by exploring the Gaussianity of $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$. Applying the chain rule and (11), we can show that

$$\frac{\partial E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial \mathbf{m}^{(1)}_{\boldsymbol{\lambda}_{t,k}}} = \frac{\partial \mathcal{L}(\boldsymbol{\lambda}_{t,k})}{\partial \boldsymbol{\mu}_{t,k}} - 2\frac{\partial \mathcal{L}(\boldsymbol{\lambda}_{t,k})}{\partial \boldsymbol{\Sigma}_{t,k}}\boldsymbol{\mu}_{t,k}, \tag{21a}$$

$$\frac{\partial E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]}{\partial \mathbf{m}^{(2)}_{\boldsymbol{\lambda}_{t,k}}} = \frac{\partial \mathcal{L}(\boldsymbol{\lambda}_{t,k})}{\partial \boldsymbol{\Sigma}_{t,k}}. \tag{21b}$$

Here, $\mathcal{L}(\boldsymbol{\lambda}_{t,k}) = E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}[\log(p(\mathbf{z}_t|\mathbf{x}_t))]$ is introduced to simplify the presentation. From the Bonnet's theorem and Price's theorem [42], we have

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}_{t,k})}{\partial \boldsymbol{\mu}_{t,k}} = E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}\left[\frac{\partial \log(p(\mathbf{z}_t|\mathbf{x}_t))}{\partial \mathbf{x}_t}\right], \tag{22a}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda}_{t,k})}{\partial \boldsymbol{\Sigma}_{t,k}} = \frac{1}{2}E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}\left[\frac{\partial^2 \log(p(\mathbf{z}_t|\mathbf{x}_t))}{\partial \mathbf{x}_t \partial \mathbf{x}_t^T}\right]. \tag{22b}$$

This completes the development of the assumed state posterior update rule in the natural parameter space.

To establish the rule for directly updating the mean $\boldsymbol{\mu}_t$ and covariance $\boldsymbol{\Sigma}_t$ of the assumed state posterior, we put (21) and (22) into (17), and apply (10b) to arrive at

$$\begin{aligned}\boldsymbol{\Sigma}^{-1}_{t,k+1} = {} & (1-\alpha_k)\boldsymbol{\Sigma}^{-1}_{t,k} + \alpha_k \tilde{\boldsymbol{\Sigma}}^{-1}_t \\ & - \alpha_k E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}\left[\frac{\partial^2 \log(p(\mathbf{z}_t|\mathbf{x}_t))}{\partial \mathbf{x}_t \partial \mathbf{x}_t^T}\right],\end{aligned} \tag{23a}$$

$$\begin{aligned}\boldsymbol{\mu}_{t,k+1} = {} & \boldsymbol{\mu}_{t,k} - \alpha_k \boldsymbol{\Sigma}_{t,k+1} \cdot \tilde{\boldsymbol{\Sigma}}^{-1}_t (\boldsymbol{\mu}_{t,k} - \tilde{\boldsymbol{\mu}}_t) \\ & + \alpha_k \boldsymbol{\Sigma}_{t,k+1} \cdot E_{q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)}\left[\frac{\partial \log(p(\mathbf{z}_t|\mathbf{x}_t))}{\partial \mathbf{x}_t}\right].\end{aligned} \tag{23b}$$

Here, $\boldsymbol{\mu}_{t,k}$ and $\boldsymbol{\Sigma}_{t,k}$ are the mean and covariance of $q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)$ obtained in the $k$-th iteration. Besides, $\tilde{\boldsymbol{\mu}}_t$ and $\tilde{\boldsymbol{\Sigma}}_t$ are the mean and covariance of the state predicition distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t-1})$, and they are known.

For the linear Gaussian model where the measurement likelihood is $p(\mathbf{z}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{z}_t; \mathbf{H}\mathbf{x}_t, \mathbf{R}_t)$, where $\mathbf{H}$ is the measurement matrix and $\mathbf{R}_t$ is the noise covariance, the update rules (23) reduce to the KF. We can verify this by setting the assumed state posterior initially as the state predictive distribution, and noting that in this case, the two partial derivatives are $-\mathbf{H}^T\mathbf{R}_t^{-1}\mathbf{H}$ and $\mathbf{H}^T\mathbf{R}_t^{-1}(\mathbf{z}_t - \mathbf{H}\tilde{\boldsymbol{\mu}}_t)$.

Different from the mean update in [24], ours in (23b) already includes the updated covariance $\boldsymbol{\Sigma}_{t,k+1}$ as the preconditioning matrix. In [27], a mean update similar to (23b) is obtained through using the Newton-type method with an approximated Hessian, but the covariance update is different from (23a). Our derivation, on the contrary, is based the *exact* expression of the natural gradient of the KLD loss in (4).

When realizing the proposed assumed state posterior update equation in (18) or (23), we set the initial natural parameter $\boldsymbol{\lambda}_{t,0}$ to be that of the state predictive distribution $\tilde{\boldsymbol{\lambda}}_t$ to start the iteration. The main challenge in algorithm implementation is the evaluation of the *expected* partial derivatives in (22). We may approximate the expectations in (22) by sampling from the posterior $q_{\boldsymbol{\lambda}_{t,k}}(\mathbf{x}_t)$ and performing Monte Carlo integration. This leads to the well-known stochastic gradient descent (SGD) technique [35]. In this work, we utilize the low-discrepancy generalized Fibonacci grid from [43] to realize *deterministic* Gaussian sampling-based integration to evaluate (22). The iterative updating process converges if the KLD between the state posteriors from two successive iterations is smaller than a certain threshold $\delta = 10^{-6}$.

## III. Proposed Variational Gaussian Filter

### A. Gaussian Filtering via Variational Inference

The method in the previous section for optimizing the assumed state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ requires that the state predictive distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t-1})$ is obtained already and is Gaussian. However, this may be an approximation, as the state transition density $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ could involve nonlinearity, making the true state predictive distribution deviate from being Gaussian.

Note that the *approximated* state predictive distribution is used as the initial state posterior estimate. To account for the impact of approximating $p(\mathbf{x}_t|\mathbf{z}_{1:t-1})$ with a Gaussian density $\mathcal{N}(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$ on the state posterior estimation, we impose an 'extra' Wishart prior on the inverse of the *obtained* covariance $\tilde{\boldsymbol{\Sigma}}_t$, $\boldsymbol{\Lambda}_t = \tilde{\boldsymbol{\Sigma}}_t^{-1}$, which is [18], [44]

$$p(\boldsymbol{\Lambda}_t) = \mathcal{W}(\boldsymbol{\Lambda}_t|\mathbf{W}_t, \nu) \propto |\boldsymbol{\Lambda}_t|^{\frac{\nu - n_x - 1}{2}} \exp\Big(-\frac{1}{2}\mathrm{tr}(\mathbf{W}_t^{-1}\boldsymbol{\Lambda}_t)\Big). \tag{24}$$

The scale matrix $\mathbf{W}_t$ and degree of freedom (DoF) $\nu > n_x - 1$ are the hyperparameters. $n_x$ is the dimensionality of the state $\mathbf{x}_t$. Note that the state prediction mean $\tilde{\boldsymbol{\mu}}_t$ is left *unchanged* once it has been found.

Because the state prediction covariance $\tilde{\boldsymbol{\Sigma}}_t$ is now stochastic, it needs to be identified jointly with the system state $\mathbf{x}_t$. This allows the adjustment of $\tilde{\boldsymbol{\Sigma}}_t$ using the current measurement $\mathbf{z}_t$, which may in turn benefit the state estimation. With the prior on the inverse of $\tilde{\boldsymbol{\Sigma}}_t$ in (24), compared with the original state-space model in (1), we in fact have an augmented model whose measurement likelihood is still $p(\mathbf{z}_t|\mathbf{x}_t)$ but the predictive distribution becomes

$$p(\mathbf{x}_t, \boldsymbol{\Lambda}_t|\mathbf{z}_{1:t-1}) \approx \mathcal{N}(\mathbf{x}_t; \tilde{\boldsymbol{\mu}}_t, \boldsymbol{\Lambda}_t^{-1}) p(\boldsymbol{\Lambda}_t). \tag{25}$$

Computing the joint filtering posterior $p(\mathbf{x}_t, \mathbf{\Lambda}_t|\mathbf{z}_{1:t})$ based on the augmented state-space model mentioned above is hard. To bypass the difficulty, we apply the following mean-field approximation [18]

$$p(\mathbf{x}_t, \mathbf{\Lambda}_t|\mathbf{z}_{1:t}) \approx q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)q(\mathbf{\Lambda}_t), \tag{26}$$

and then invoke variational inference to find the state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$, which is Gaussian with natural parameter $\boldsymbol{\lambda}_t$, and the posterior of $\mathbf{\Lambda}_t$, $q(\mathbf{\Lambda}_t)$ to achieve Gaussian filtering.

Mathematically, $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ and $q(\mathbf{\Lambda}_t)$ are calculated through maximizing the ELBO [41]

$$\int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)q(\mathbf{\Lambda}_t)\log\frac{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t, \mathbf{\Lambda}_t|\mathbf{z}_{1:t-1})}{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)q(\mathbf{\Lambda}_t)}\mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{\Lambda}_t. \tag{27}$$

We solve the maximization problem of interest through applying the coordinate ascent technique, where in each iteration, we carry out the following two optimization steps in sequence.

**Step-1**: The state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ is updated in Step-1, with the posterior of $\mathbf{\Lambda}_t$, $q(\mathbf{\Lambda}_t)$, fixed. In this case, the ELBO in (27) becomes, after substituting (25) and neglecting the terms that do not depend on $\mathbf{x}_t$,

$$\int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)\left[\log\frac{p(\mathbf{z}_t|\mathbf{x}_t)}{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}+\int q(\mathbf{\Lambda}_t)\log\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t, \mathbf{\Lambda}_t^{-1})\mathrm{d}\mathbf{\Lambda}_t\right]\mathrm{d}\mathbf{x}_t. \tag{28}$$

By the definition of the multivariate Gaussian density, we have

$$\begin{aligned}\int q(\mathbf{\Lambda}_t)\log\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,&\mathbf{\Lambda}_t^{-1})\mathrm{d}\mathbf{\Lambda}_t \propto \log\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t, \mathbf{\Omega}_t^{-1}),\\ &\propto -\frac{1}{2}(\mathbf{x}_t-\tilde{\boldsymbol{\mu}}_t)^T\mathbf{\Omega}_t(\mathbf{x}_t-\tilde{\boldsymbol{\mu}}_t),\end{aligned} \tag{29}$$

where $\mathbf{\Omega}_t = E_{q(\mathbf{\Lambda}_t)}(\mathbf{\Lambda}_t)$ is the expected value of $\mathbf{\Lambda}_t$ with respect to $q(\mathbf{\Lambda}_t)$.

Putting (29) into (28) converts the ELBO into

$$\int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)\log\frac{p(\mathbf{z}_t|\mathbf{x}_t)\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t, \mathbf{\Omega}_t^{-1})}{q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)}\mathrm{d}\mathbf{x}_t, \tag{30}$$

which has the same functional form as the one in (19). This indicates that updating the state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ by maximizing (30) can be achieved through applying the iterative method in Section II with the state predictive distribution replaced with $\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t, \mathbf{\Omega}_t^{-1})$.

**Step-2**: We fix the state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ just obtained in Step-1 and then optimize the posterior of $\mathbf{\Lambda}_t$, $q(\mathbf{\Lambda}_t)$, in Step-2. Substituting (25) and ignoring the terms not related to $\mathbf{\Lambda}_t$ simplify the ELBO in (27) into

$$\int q(\mathbf{\Lambda}_t)\left[\log\frac{p(\mathbf{\Lambda}_t)}{q(\mathbf{\Lambda}_t)} + \int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)\log\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t, \mathbf{\Lambda}_t^{-1})\mathrm{d}\mathbf{x}_t\right]\mathrm{d}\mathbf{\Lambda}_t. \tag{31}$$

With slight abuse of notations, we denote $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$, since it is constrained to be Gaussian. Again, by applying the definition of the multivariate Gaussian density, we arrive at

$$\begin{aligned}\int q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)\log\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t, \mathbf{\Lambda}_t^{-1})\mathrm{d}\mathbf{x}_t \propto &\frac{1}{2}\log|\mathbf{\Lambda}_t|\\ &-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Lambda}_t\left(\mathbf{\Sigma}_t+(\boldsymbol{\mu}_t-\tilde{\boldsymbol{\mu}}_t)(\boldsymbol{\mu}_t-\tilde{\boldsymbol{\mu}}_t)^T\right)\right).\end{aligned} \tag{32}$$

Moreover, from (24), $\log p(\mathbf{\Lambda}_t)$ is equal to

$$\log p(\mathbf{\Lambda}_t) \propto \frac{1}{2}(\nu-n_x-1)\log|\mathbf{\Lambda}_t| - \frac{1}{2}\mathrm{tr}(\mathbf{W}_t^{-1}\mathbf{\Lambda}_t). \tag{33}$$

Putting (32) and (33) into (31) reveals that the ELBO attains the maximum value of 0 when the posterior of $\mathbf{\Lambda}_t$, $q(\mathbf{\Lambda}_t)$, satisfies [34]

$$\log q(\mathbf{\Lambda}_t) \propto \frac{1}{2}(\nu-n_x)\log|\mathbf{\Lambda}_t| - \frac{1}{2}\mathrm{tr}(\tilde{\mathbf{W}}_t^{-1}\mathbf{\Lambda}_t), \tag{34}$$

where $\tilde{\mathbf{W}}_t$ is defined as

$$\tilde{\mathbf{W}}_t = \left(\mathbf{W}_t^{-1}+\mathbf{\Sigma}_t+(\boldsymbol{\mu}_t-\tilde{\boldsymbol{\mu}}_t)(\boldsymbol{\mu}_t-\tilde{\boldsymbol{\mu}}_t)^T\right)^{-1}. \tag{35}$$

Comparing (34) with (33) implies that given the Gaussian state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$, the optimal estimate for the posterior of $\mathbf{\Lambda}_t$, $q(\mathbf{\Lambda}_t)$, is still a Wishart distribution but with DoF $\nu+1$ and scale matrix $\tilde{\mathbf{W}}_t$ given in (35).

With the updated posterior $q(\mathbf{\Lambda}_t)$ in (34), the proposed Gaussian filtering algorithm goes to Step-1 again to obtain a new state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$, which will then be explored in Step-2 to further refine $q(\mathbf{\Lambda}_t)$. The above process is repeated until the KLD between the state posteriors in two successive iterations is smaller than the threshold $\delta = 10^{-6}$.

### *B. Algorithm Summary*

The main computations conducted by the proposed Gaussian filter at each time instant are summarized below.

1). At the time instant $t$, the proposed Gaussian filter first computes a Gaussian approximation $\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t, \tilde{\mathbf{\Sigma}}_t)$ to the state predictive distribution in (2) using the prediction step of a conventional GF, such as those in [6]–[17].

2). Next, we fix the prediction mean $\tilde{\boldsymbol{\mu}}_t$. The hyperparameters of $p(\mathbf{\Lambda}_t)$ in (24), namely, the DoF $\nu$ and scale matrix $\mathbf{W}_t$, are chosen to be $\nu = n_x$ and $\mathbf{W}_t = \frac{1}{\nu}\tilde{\mathbf{\Sigma}}_t^{-1}$. The initial posterior of $\mathbf{\Lambda}_t$, $q(\mathbf{\Lambda}_t)$, is set to its prior $p(\mathbf{\Lambda}_t)$.

3). Step-1 described in Section III.A is executed using the initial posterior of $\mathbf{\Lambda}_t$ and the iterative method in (18) or (23) to produce an estimate of the state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$. Step-2 developed in Section III.A is then executed to obtain an updated version of $q(\mathbf{\Lambda}_t)$, which is given in (34).

4). We replace the initial posterior of $\mathbf{\Lambda}_t$ with its updated version, and run Step-1 and Step-2 again. The above process is repeated until convergence.

5). Finally, the proposed Gaussian filter propagates the state posterior $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t)$ to the next time instant while at the same time, outputting it for estimating the system state.

### *C. Selection of the Hyperparameters $\nu$ and $\mathbf{W}_t$*

We shall justify setting the hyperparameters of the Wishart prior $p(\mathbf{\Lambda}_t)$ to be $\nu = n_x$ and $\mathbf{W}_t = \frac{1}{\nu}\tilde{\mathbf{\Sigma}}_t^{-1}$. This setup is done right after the state prediction but before the measurement update begins (see Section III.B). Consistent with the measurement update of the proposed Gaussian filter (see Section III.A), we select the hyperparameters of $p(\mathbf{\Lambda}_t)$ by maximizing the ELBO in (27). At this point, the two posteriors in (27) should be $q_{\boldsymbol{\lambda}_t}(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t, \tilde{\mathbf{\Sigma}}_t)$ and $q(\mathbf{\Lambda}_t) = p(\mathbf{\Lambda}_t)$, i.e., they are equal to their corresponding prediction/prior distributions.

Putting these results and ignoring the terms independent of the hyperparameters to be determined transform (27) into

$$\begin{aligned}&\int p(\mathbf{\Lambda}_t)\int \mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\mathbf{\Sigma}}_t)\log\mathcal{N}(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\mathbf{\Lambda}_t^{-1})\mathrm{d}\mathbf{x}_t\mathrm{d}\mathbf{\Lambda}_t\\ &\propto \int p(\mathbf{\Lambda}_t)\left[\log|\mathbf{\Lambda}_t|-\mathrm{tr}(\mathbf{\Lambda}_t\tilde{\mathbf{\Sigma}}_t)\right]\mathrm{d}\mathbf{\Lambda}_t,\\ &\propto \sum_{i=1}^{n_x}\psi\left(\frac{\nu+1-i}{2}\right)+\log|\mathbf{W}_t|-\mathrm{tr}(\nu\mathbf{W}_t\tilde{\mathbf{\Sigma}}_t),\end{aligned} \tag{36}$$

where $\psi(\cdot)$ is the digamma function. The first equality in (36) comes from substituting the expression of the multivariate Gaussian density and carrying out the inner integral. Deriving the second equality used the following results of the Wishart distribution (see (B.80) and (B.81) of [18]):

$$E_{p(\mathbf{\Lambda}_t)}(\log|\mathbf{\Lambda}_t|)\propto\sum_{i=1}^{n_x}\psi\left(\frac{\nu+1-i}{2}\right)+\log|\mathbf{W}_t|, \tag{37a}$$

$$E_{p(\mathbf{\Lambda}_t)}(\mathbf{\Lambda}_t)=\nu\mathbf{W}_t. \tag{37b}$$

Taking the partial derivative of the ELBO in (36) with respect to $\mathbf{W}_t$, using $\frac{\partial\log|\mathbf{W}_t|}{\partial\mathbf{W}_t}=\mathbf{W}_t^{-1}$ and $\frac{\partial\mathrm{tr}(\nu\mathbf{W}_t\tilde{\mathbf{\Sigma}}_t)}{\partial\mathbf{W}_t}=\nu\tilde{\mathbf{\Sigma}}_t$ from [45], and setting the result to zero yield the optimal initial value for $\mathbf{W}_t$, which is exactly our setting $\mathbf{W}_t=\frac{1}{\nu}\tilde{\mathbf{\Sigma}}_t^{-1}$.

We proceed to justify choosing the DoF $\nu$ to be $\nu=n_x$. Note that according to the algorithm summary in Section III.B, Step-1 in Section III.A, when being executed for the first time, is based on the state predictive distribution with covariance $\mathbf{\Omega}_t^{-1}=(\nu\mathbf{W}_t)^{-1}=\tilde{\mathbf{\Sigma}}_t$. In the remaining iterations, however, the state prediction covariance used for updating the state posterior becomes, from (34) and (37b),

$$\mathbf{\Omega}_t^{-1}=\frac{\nu}{\nu+1}\tilde{\mathbf{\Sigma}}_t+\frac{1}{\nu+1}\left(\mathbf{\Sigma}_t+(\boldsymbol{\mu}_t-\tilde{\boldsymbol{\mu}}_t)(\boldsymbol{\mu}_t-\tilde{\boldsymbol{\mu}}_t)^T\right). \tag{38}$$

This shows how the proposed Gaussian filter combines the prior information on the state prediction covariance, which is found by approximating (2) with a Gaussian density, with that from the state posterior obtained through measurement update. We set the DoF $v$ to be the smallest possible value $n_x$ to allow $\mathbf{\Omega}_t^{-1}$ in (38) to be better adjusted by the measurement.

## IV. Simulation Results

### A. Simulation Scenario

The simulated tracking scenario is similar to the ones used in [38], [46], [47]. A stationary sensor located at the origin measures the bearing and range of the target to estimate its trajectory. The measurements are obtained with a sampling period of 3s, and are subject to independent Gaussian noise with standard deviations $\sigma_b=0.3^o$ and $\sigma_r=50$m [48].

At the beginning of the tracking process, the target is 180km away from the sensor with a true bearning of $30^o$. It moves towards southwest with velocity $[-100,-173.2]^T$m/s. The target trajectory has five segments. In the first segment (0s to 100s), the third segment (132s to 200s) and the fifth segment (232s to 300s), the target motion follows a constant velocity (CV) model with the process noise having a standard deviation of 0.01m/s$^2$. The target makes two $90^o$ turns, one in the second segment (100s to 132s) and the other in the fourth segment (200s to 232s). Both turns have an acceleration of 1.07g and are modeled by the constant turn (CT) model with the process noise having a standard deviation of 0.01m/s$^2$. The true target trajectory in a certain ensemble run is shown in Fig. 1.

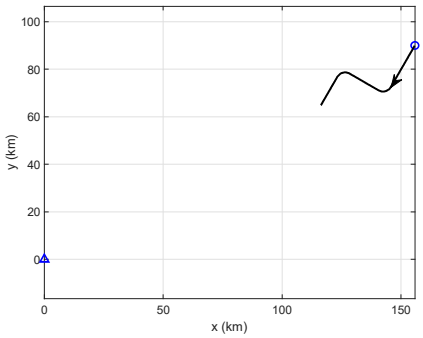


Fig. 1. True target trajectory in a certain ensemble run. The upper triangle denotes the sensor position. The circle is the initial target position. The arrow shows the target moving direction.

### B. Implementation of the Proposed Gaussian Filter

The proposed Gaussian filter assumes that the target motion is modeled using the CT model with *unknown* turn rate [5]

$$\mathbf{x}_t=\begin{bmatrix}\mathbf{F}(\omega_{t-1},T) & \mathbf{0}\\ \mathbf{0}^T & \mathrm{e}^{-\alpha T}\end{bmatrix}\mathbf{x}_{t-1}+\mathbf{G}\mathbf{w}_{t-1}, \tag{39}$$

where $\mathbf{x}_t=[x_t,y_t,v_{x,t},v_{y,t},\omega_t]^T$ is the state vector at time step $t$. $[x_t,y_t]^T$ and $[v_{x,t},v_{y,t}]^T$ are the target position and velocity at time step $t$, while $\omega_t$ is its turn rate. $\alpha=0.1$ is the inverse of the correlation time constant for the turn rate. The *nonlinear* state transition matrix $\mathbf{F}(\omega_{t-1},T)$ is defined as

$$\mathbf{F}(\omega_{t-1},T)=\begin{bmatrix}1 & 0 & \frac{\sin(\omega_{t-1}T)}{\omega_{t-1}} & -\frac{1-\cos(\omega_{t-1}T)}{\omega_{t-1}}\\ 0 & 1 & \frac{1-\cos(\omega_{t-1}T)}{\omega_{t-1}} & \frac{\sin(\omega_{t-1}T)}{\omega_{t-1}}\\ 0 & 0 & \cos(\omega_{t-1}T) & -\sin(\omega_{t-1}T)\\ 0 & 0 & \sin(\omega_{t-1}T) & \cos(\omega_{t-1}T)\end{bmatrix}. \tag{40}$$

The gain matrix $\mathbf{G}$ for the process noise $\mathbf{w}_t$ is equal to $\mathbf{G}=\mathrm{diag}([\mathbf{g}_1,\mathbf{g}_2],1)$, where $\mathbf{g}_1=[T^2/2,0,T,0]^T$, $\mathbf{g}_2=[0,T^2/2,0,T]^T$ and $\mathrm{diag}(\mathbf{A},\mathbf{B})$ denotes a block diagonal matrix with $\mathbf{A}$ and $\mathbf{B}$ being the diagonal blocks. The process noise is $\mathbf{w}_t=[w_{v_x,t},w_{v_y,t},w_\omega]^T$, whose elements correspond to the acceleration in the $x$ and $y$ directions, as well as the turn rate. $\mathbf{w}_t$ is zero-mean white Gaussian with covariance $\mathrm{diag}(0.01\mathrm{m}^2/\mathrm{s}^4,0.01\mathrm{m}^2/\mathrm{s}^4,1^o/\mathrm{s}^2)$ [4]. The sampling interval is $T=1$s, which is smaller than the measurement period of 3s to construct a smoother target trajectory. We adopt the prediction stage of GHKF [2] with 32 sigma points to obtain the (initial) state predictive distribution (see Section III.B).

For the 2D tracking problem in consideration, the logarithm of the measurement likelihood is $\log p(\mathbf{z}_t|\mathbf{x}_t)\propto-\frac{1}{2}(\mathbf{z}_t-\mathbf{h}(\mathbf{x}_t))^T\mathbf{R}^{-1}(\mathbf{z}_t-\mathbf{h}(\mathbf{x}_t))$, where $\mathbf{R}=\mathrm{diag}(\sigma_b^2,\sigma_r^2)$ is the

measurement noise covariance (see Section IV.A). $\mathbf{h}(\mathbf{x}_t)$ is the measurement function. To evaluate the expectations in (22), we apply the generalized Fibonacci grid [43] with 32 grid points to realize deterministic sampling-based integration.

### C. Results and Discussions

Figs. 2 and 3 plot the root mean square errors (RMSEs) for target position and velocity estimates generated by the proposed Gaussian filter ('Proposed') at time instants when measurements are available. The results are obtained by averaging over 2,000 ensemble runs. The estimation RMSEs of the EKF [6] ('EKF'), CKF using the 3rd-order spherical cubature rule [15] ('CKF') and GHKF with 32 sigma points for both state prediction and update [2] ('GHKF') are also included. For a fair comparison, we initialize all the filters using the first set of bearing and range measurements. Their initial velocity and turn rate estimates are set to zero. In Figs. 2 and 3, four dotted vertical lines are added to indicate the starting time instants and ending time instants of the two $90^o$ turns.

First, we observe that EKF performs the best before the first turn of the target, because its estimate of the turn rate remains closer to zero than other filters. However, the estimation accuracy of the EKF significantly degrades afterwards, showing that it has poor tracking ability. It is found (data not shown due to page limit) that this can be mitigated by increasing the variance of the process noise $\mathbf{w}_t$ in the motion model (39).

Besides, Fig. 2 shows that the proposed Gaussian filter offers evidently better target position estimation accuracy over CKF and GHKF, especially after the target makes a turn. Some improvement in the target velocity estimation performance can be seen from Fig. 3. Recall that the proposed filter uses the *same* prediction stage as the simulated GHKF to start the prediction-correction cycle when a new pair of bearing and range measurements becomes available (see Section IV.B). This indicates that the performance enhancement comes from the use of the newly developed KLD minimization-based state posterior update and state prediction covariance adjustment.

In Fig. 4, we plot the average normalized estimation error squared (NEES) over time for the four Gaussian filters considered. As the state $\mathbf{x}_t$ has five elements, the ideal NEES value is five as well [6]. It can be seen that the target turns render all the simulated filters over-confident but the proposed filter still provides the most consistent state estimation results.

In this simulation, the state posterior estimation of the proposed variational Gaussian filter (i.e., Stages 2) - 4) in Section III.B) takes on average 4.15 iterations and 0.07s to finish, which is much longer than 0.0014s of the non-iterative GHKF. Stage 3) in Section III.B finds the optimal assumed state posterior and needs on average 2.7 iterations to converge.

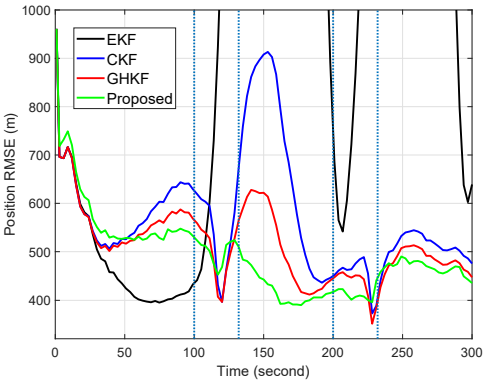


Fig. 2. Comparison of target position estimation RMSEs.

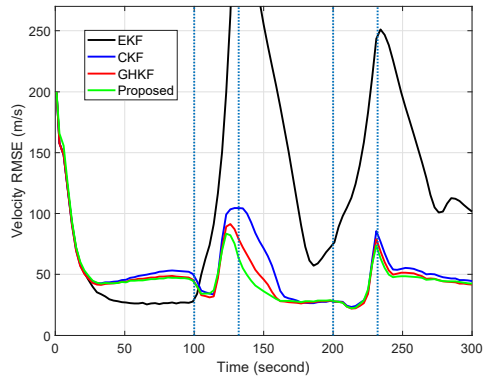


Fig. 3. Comparison of target velocity estimation RMSEs.

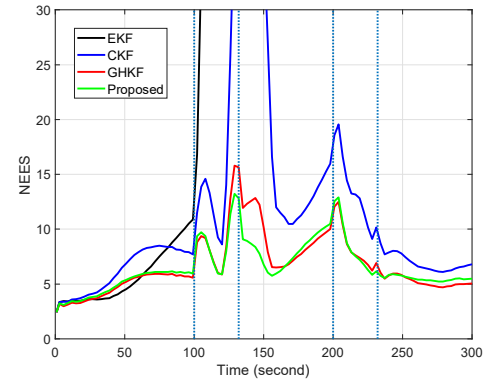


Fig. 4. Comparison of estimation NEESs.

## V. Conclusions

This paper presented a new Gaussian filtering algorithm. It continues to follow the prediction-correction framework as the conventional GFs do, and it introduces two enhancements. The first one is the KLD minimization-based state posterior update, which was derived through exploring that the Gaussian density is a member of the exponential family. The obtained update rule is identical to the CVI technique in Bayesian learning literature. But our derivation was simpler. The developed update rule was integrated into the VI method, leading to the second enhancement that adjusts the state prediction covariance using the measurements to improve performance. Implementation details of the proposed Gaussian filter were given. Simulations

demonstrated the superiority of the new algorithm over several existing GFs. In future work, we shall consider employing the $\alpha$-divergence, which includes the KLD as a special case, to derive new state posterior update rules. Another possible extension is to perform fixed-lag smoothing such that the state posteriors at the current and previous time instants may both be identified using the KLD minimization.

## REFERENCES


[1] S. Challa, M. R. Morelande, D. Mušicki, and R. J. Evans, *Fundamentals of Object Tracking*. Cambridge University Press, 2011.

[2] S. Särkkä and L. Svensson, *Bayesian Filtering and Smoothing*, 2nd ed. New York: Cambridge University Press, 2023.

[3] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Journal of Basic Engineering*, vol. 83, no. 1, pp. 95–108, Mar. 1961.

[4] Y. Bar-Shalom, P. K. Willett, and X. Tian, *Tracking and Data Fusion: A Handbook of Algorithms*. YBS Publishing, 2011.

[5] X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. Part I. Dynamic models," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1333–1364, Oct. 2003.

[6] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation: Theory, algorithms and software*. New York: Wiley, 2001.

[7] Y. Ollivier, "The extended Kalman filter is a natural gradient descent in trajectory space," *arXiv preprint*, Jan. 2019. [Online]. Available: arxiv.org/abs/1901.00696.

[8] Ángel. F. García-Fernández, L. Svensson, M. R. Morelande, and S. Särkkä, "Posterior linearization filter: Principles and implementation using sigma points," *IEEE Trans. Signal Process.*, vol. 63, pp. 5561–5573, Oct. 2015.

[9] M. Raitoharju, L. Svensson, Ángel. F. García-Fernández, and R. Piché, "Damped posterior linearization filter," *IEEE Signal Process. Lett.*, vol. 25, pp. 536–540, April 2018.

[10] Y. Wu, D. Hu, M. Wu, and X. Hu, "A numerical-integration perspective on Gaussian filters," *IEEE Trans. Signal Process.*, vol. 54, pp. 2910–2921, Aug. 2006.

[11] S. Julier, J. Uhlmann, and H. F. Durrant-Whyte, "A new method for the nonlinear transformation of means and covariances in filters and estimators," *IEEE Trans. Autom. Control*, vol. 45, pp. 477–482, Mar. 2000.

[12] K. Ito and K. Xiong, "Gaussian filters for nonlinear filtering problems," *IEEE Trans. Autom. Control*, vol. 45, pp. 910–927, May 2000.

[13] B. Jia, M. Yin, and Y. Cheng, "Sparse-grid quadrature nonlinear filtering," *Automatica*, vol. 48, pp. 327–341, Feb. 2012.

[14] H. Meng, "State estimation-Beyond Gaussian filtering," Ph.D. dissertation, University of New Orleans, 2022.

[15] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Trans. Autom. Control*, vol. 54, no. 6, pp. 1254–1269, Jun. 2009.

[16] B. Jia, M. Xin, and Y. Cheng, "High-degree cubature Kalman filter," *Automatica*, vol. 49, no. 2, pp. 510–518, Feb. 2013.

[17] I. Arasaratnam, S. Haykin, and R. J. Elliott, "Discrete-time nonlinear filtering algorithms using Gauss-Hermite quadrature," *Proc. IEEE*, vol. 95, no. 5, pp. 953–977, May. 2007.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[19] B. Efron, *Exponential Families in Theory and Practice*. Cambridge University Press, 2022.

[20] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, Feb. 1998.

[21] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, vol. 14, pp. 1303–1347, May 2013.

[22] Y. Ollivier, "Online natural gradient as a Kalman filter," *arXiv preprint*, March 2017. [Online]. Available: arxiv.org/abs/1703.00209.

[23] M. E. Khan and W. Lin, "Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models," in *Proc. Intl. Conf. Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, April 2017.

[24] S. Gultekin and J. Paisley, "Nonlinear Kalman fltering with divergence minimization," *IEEE Trans. Signal Process.*, vol. 65, pp. 6319–6331, Dec. 2017.

[25] Y. Hu, Q. Pan, Z. Guo, Z. Shi, and Z. Hu, "Iterative nonlinear Kalman filtering via variational evidence lower bound maximization," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Ottawa, Canada, Jul. 2019.

[26] Y. Hu, X. Wang, Q. Pan, Z. Hu, and B. Moran, "Variational Bayesian Kalman filter using natural gradient," *Chinese Journal of Aeronautics*, vol. 35, pp. 1–10, May 2022.

[27] E. Laz and U. Orguner, "Gaussian mixture filtering with nonlinear measurements minimizing forward Kullback-Leibler divergence," *Signal Process.*, vol. 208, July 2023.

[28] H. Lan, J. Hu, Z. Wang, and Q. Cheng, "Variational nonlinear Kalman filtering with unknown process noise covariance," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, pp. 9177–9190, Dec. 2023.

[29] L. Guo, S. Hu, J. Zhou, and X. R. Li, "Recursive nonlinear filtering via Gaussian approximation with minimized Kullback-Leibler divergence," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 60, pp. 965–979, Feb. 2024.

[30] J. E. Darling and K. J. DeMars, "Minimization of the Kullback-Leibler divergence for nonlinear estimation," *J. of Guidance, Control and Dynamics*, vol. 40, pp. 1739–1748, July 2017.

[31] J. Courts, A. G. Wills, and T. B. Schön, "Gaussian variational state estimation for nonlinear state-space models," *IEEE Trans. Signal Process.*, vol. 69, pp. 5979–5993, Oct. 2021.

[32] J. Prüber, F. Tronarp, T. Karvonen, S. Särkkä, and O. Straka, "Student-$t$ process quadrature for filtering of non-linear systems with heavy-tailed noise," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Xi'an, China, Jul. 2017.

[33] J. Prüber and O. Straka, "Gausian process quadrature moment transform," *IEEE Trans. Autom. Control*, pp. 2844–2854, Sept. 2018.

[34] H. Cui, L. S. Mihaylova, X. Wang, and S. Gao, "Uncertainty-aware variational inference for target tracking," *IEEE Trans. Aerosp. Electron. Syst.*, pp. 258–273, Feb. 2023.

[35] K. P. Murphy, *Probabilistic Machine Learning: Advanced Topics*. The MIT Press, 2023.

[36] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[37] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Prentice Hall, 1995.

[38] Y. Liu, X. Li, L. Yang, L. S. Mihaylova, and Y. Xue, "On the approximation of the quotient of two Gaussian densities for multiple-model smoothing," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Charleston, SC, USA, July 2023.

[39] M. E. Khan and D. Nielsen, "Fast yet simple natural-gradient descent for variational inference in complex models," in *Proc. Intl. Symposium on Information Theory and Its Applications (ISITA)*, Singapore, Oct. 2018.

[40] M. E. Khan, R. Babanezhad, W. Lin, M. Schmidt, and M. Sugiyama, "Faster stochastic variational inference using proximal-gradient methods with general divergence functions," in *Proc. Conf. Uncertainty in Artificial Intelligence (UAI)*, Jersey City, NJ, USA, June 2016.

[41] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variationa inference: A review for statisticians," *J. of the American Statistical Association*, vol. 112, pp. 859–877, July 2017.

[42] M. Opper and C. Archambeau, "The variational Gaussian approximation revisited," *Neural Computation*, vol. 21, pp. 786–792, March 2009.

[43] D. Frisch and U. D. Hanebeck, "The generalized Fibonacci grid as low-discrepancy point set for optimal deterministic Gaussian sampling," *J. of Advances in Information Fusion*, vol. 18, pp. 16–34, June 2023.

[44] J. W. Koch, "Bayesian approach to extended object and cluster tracking using random matrices," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 44, pp. 1042–1059, July 2008.

[45] K. B. Peterson and M. S. Peterson, *The Matrix Cookbook*, 2012. [Online]. Available: http://matrixcookbook.com

[46] X. Li, Y. Liu, L. Yang, L. S. Mihaylova, and B. Deng, "Enhanced fixed-interval smoothing for Markovian switching systems," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Sun City, South Africa, Nov. 2021.

[47] Y. Liu, X. Li, L. Yang, L. S. Mihaylova, and Y. Xue, "On the fixed-interval smoothing for jump Markov nonlinear systems," in *Proc. Intl. Conf. Information Fusion (FUSION)*, Linköping, Sweden, July 2022.

[48] M. T. Busch and S. S. Blackman, "Evaluation of IMM filtering for an air defence system application," in *Proc. SPIE Conf. Signal and Data Process. of Small Targets*, San Diego, CA, USA, Sept. 1995, pp. 435–447.