Nonlinear Bayesian Filtering with **N**atural Gr**a**dient Gaussia**n** Appr**o**ximation

Wenhan Cao, Tianyi Zhang, Zeju Sun, Chang Liu, Stephen S.-T. Yau, Shengbo Eben Li

Abstract—Practical Bayes filters often assume the state distribution of each time step to be Gaussian for computational tractability, resulting in the so-called Gaussian filters. When facing nonlinear systems, Gaussian filters such as extended Kalman filter (EKF) or unscented Kalman filter (UKF) typically rely on certain linearization techniques, which can introduce large estimation errors. To address this issue, this paper reconstructs the prediction and update steps of Gaussian filtering as solutions to two distinct optimization problems, whose optimal conditions are found to have analytical forms from Stein's lemma. It is observed that the stationary point for the prediction step requires calculating the first two moments of the prior distribution, which is equivalent to that step in existing moment-matching filters. In the update step, instead of linearizing the model to approximate the stationary points, we propose an iterative approach to directly minimize the update step's objective to avoid linearization errors. For the purpose of performing the steepest descent on the Gaussian manifold, we derive its natural gradient that leverages Fisher information matrix to adjust the gradient direction, accounting for the curvature of the parameter space. Combining this update step with moment matching in the prediction step, we introduce a new iterative filter for nonlinear systems called **N**atural Gr**a**dient Gaussia**n** Appr**o**ximation filter, or NANO filter for short. We prove that NANO filter locally converges to the optimal Gaussian approximation at each time step. Furthermore, the estimation error is proven exponentially bounded for nearly linear measurement equation and low noise levels through constructing a supermartingale-like property across consecutive time steps. Real-world experiments demonstrate that, compared to popular Gaussian filters such as EKF, UKF, iterated EKF, and posterior linearization filter, NANO filter reduces the average root mean square error by approximately 45% while maintaining a comparable computational burden.

Index Terms—State estimation, Bayesian filtering, Gaussian filter, natural gradient descent

1 Introduction

State estimation of dynamical systems is a timely topic in fields such as astrophysics, robotics, power systems, manufacturing, and transportation. The most comprehensive framework for state estimation is Bayesian filtering, which targets the distribution of the current state given the available measurements to date. This state distribution, referred to as the posterior distribution, is often calculated iteratively through two steps, i.e., prediction and update. The prediction step uses the Chapman-Kolmogorov equation to forward-predict the state distribution using the transition probability, a conditional probability that describes the time evolution of the system state. Based on the prior distribution acquired after prediction, the update step applies Bayes' theorem to update the prior using the measurement probability, a conditional probability that describes the relationship between noisy measurements and the true state [1].

For linear Gaussian systems, directly applying Bayesian filtering results in the well-known Kalman filter (KF) [2], which analytically computes the Gaussian posterior through recursive updates of its mean and covariance. This analytical form relies on the closure property of Gaussian under linear transformations and the conjugate property of Gaussian under conditioning. Unfortunately, such an elegant structure of KF does not exist for nonlinear systems, as the closure and conjugate properties only hold in the linear Gaussian case.

Therefore, if the system is nonlinear or non-Gaussian, finding the exact solution of Bayesian filtering is often unattainable. In this case, designing an appropriate approximation of state distribution becomes a key step in Bayesian filtering. To this effect, the particle filter (PF) approximates the probability density function with a set of discrete particles, each representing a possible state of the system. These particles, when weighted and summed, form a discrete approximation of the continuous probability density function. Although PF can provide asymptotically optimal approximations, it requires a large number of particles, which results in significant computational demands and limits its application in practical systems [3], [4]. In contrast to approximating the posterior distribution with samples, an alternative choice is to approximate it as Gaussian distribution in each time step, which leads to the Gaussian filter family. Compared to PF, Gaussian filters offer higher computational efficiency and have become by far the most popular family to date [5].

It turns out that the design philosophy of Gaussian filters generally consists of two steps: (i) approximate the nonlin-

- *This study is supported by NSF China with 92582205 and Beijing Natural Science Foundation with L257002. It is also partially supported by Tsinghua University-Toyota Joint Research Center for AI Technology of Automated Vehicle.*
- *Wenhan Cao and Tianyi Zhang are with the School of Vehicle and Mobility, Tsinghua University, Beijing, China (E-mail: cwh19, zhangtia24@mails.tsinghua.edu.cn).*
- *Zeju Sun is with Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, China (E-mail: sunzeju@bimsa.cn).*
- *Chang Liu is with the Department of Advanced Manufacturing and Robotics, College of Engineering, Peking University, Beijing, China (E-mail: changliucoe@pku.edu.cn).*
- *Stephen S.-T. Yau is with the Department of Mathematical Sciences, Tsinghua University, Beijing, China (E-mail: yau@uic.edu).*
- *Shengbo Eben Li is with the School of Vehicle and Mobility and College of Artificial Intelligence, Tsinghua University, Beijing, China (E-mail: lishbo@tsinghua.edu.cn). All correspondence should be sent to S.E. Li.*

ear models to linear forms with additive Gaussian noise, and (ii) perform KF based on this linear Gaussian model. Within this framework, the differences between Gaussian filters are primarily attributed to different linearization techniques. The earliest technique involves directly using the Taylor series approximation to linearize the nonlinear function. Typically, the first-order expansion is employed to avoid the occurrence of tensors in high-order expansions, exemplified by the well-known extended Kalman filter (EKF) [6], [7]. Originally developed by NASA for navigation tasks, EKF linearizes the nonlinear function around the state estimate and performs KF using the resulting affine-form system equation. Building on this foundation, iterated extended Kalman filter (IEKF) [8] repeatedly performs linearization at each updated approximation of the posterior mean instead of the prior mean as in EKF. In fact, the iterative process of this method is a Gauss-Newton iteration that essentially linearizes the model around the maximum a posteriori estimate of the state [9].

A clear drawback of directly linearizing the state space model is its inability to capture the second moment, namely the covariance, after a nonlinear transformation, since Taylor-series linearization only utilizes the mean. To address this issue, a fundamentally different approach is to directly match the transformed mean and covariance, known as moment-matching. Essentially, moment-matching can be viewed as statistical linear regression (SLR), a linearization technique that obtains optimal affine representation of the nonlinear system with parameters minimizing the expected regression loss [10]. This so-called SLR method requires computing several integrals over Gaussian distributions. In particular, the use of established numerical integration methods such as unscented transform, Gauss–Hermite integration, and spherical cubature integration underpin the design of unscented Kalman filter (UKF) [11], Gauss–Hermite KF [12], and cubature KF [13]. In addition, a new algorithm called posterior linearization filter (PLF) is proposed to perform SLR at the posterior rather than the prior during Bayesian updates to improve filtering accuracy [10].

This framework of first linearizing the model and then performing KF is termed *enabling approximation* [10]. Although it has been widely applied since the 1960s and may seem like a natural choice, one critical question cannot be overlooked: Is this *enabling approximation* framework sufficient to find the optimal Gaussian approximation of Bayesian filtering? Unfortunately, no existing works discuss this issue. In this paper, we argue that applying this framework in the update step may not yield an exact solution for nonlinear Gaussian filtering. This contrasts with the prediction step, where performing *enabling approximation* using moment-matching filters [11], [12], [13] is a proper choice. To this effect, we propose a new method called **N**atural gr**A**dient Gaussia**N** appr**O**ximation (NANO) filter that applies natural gradient in the update step to find the exact solution of Gaussian approximation. Specifically, the contributions of this paper are summarized as follows:

- We interpret the prediction and update steps of Bayesian filtering as solutions to two distinct optimization problems. This new perspective allows us to define optimal Gaussian approximation and identify its corresponding extremum conditions. Leveraging the Stein's lemma, we derive that the stationary point for prediction step has an explicit form, involving the calculation of the first two moments of prior distribution. This analytical form is implementable via moment-matching filters [11], [12], [13]. In contrast, the stationary point for update step is characterized by two interdependent equations, which generally has no analytical root. A special case occurs in the linear Gaussian systems, where these two equations decouple and become linear, resulting in analytical solutions for KF. For nonlinear systems, those *enabling approximation*-based Gaussian filters need to perform certain linearization technique, which inevitably introduce large estimation errors.
- To address linearization errors in the update step, we derive natural gradient iteration to minimize the optimization cost. By leveraging the Fisher information matrix that captures the curvature of the parameter space, the gradient direction is adjusted to perform the steepest descent on the Gaussian manifold. By combining this optimization procedure in the update step with moment matching in the prediction step, we develop a new iterative filter for nonlinear systems, namely NANO filter. We demonstrate that KF is equivalent to a single iteration of NANO filter for linear Gaussian systems, providing a fresh understanding of Kalman filtering.
- We prove that the NANO filter locally converges to the optimal Gaussian approximation at each time step, with accuracy up to a second-order remainder in the Taylor expansion. Additionally, the estimation error is proven exponentially bounded for nearly linear measurement equation and low noise levels through contructing a supermartingale-like property across consecutive time steps. We further show that NANO filter can naturally extend the Bayesian posterior to Gibbs posterior, enabling the use of more flexible loss functions to enhance robustness against outliers in measurement data. On its basis, three robust variants of NANO filter are introduced, each employing different robust loss function designs. Simulations and real-world experiments demonstrate that NANO filter and its robust variants significantly outperforms popular Gaussian filters, such as EKF, UKF, IEKF, and PLF for nonlinear systems.

The remainder of this paper is structured as follows: Section 2 formulates the problem, and Section 3 explores the optimal Gaussian approximation. Section 4 introduces the NANO filter algorithm, followed by theoretical analysis in Section 5. Section 6 proposes the robust variants of NANO filter. Section 7 provides a discussion on the proposed algorithm. Simulations and experiments are presented in Section 8.

Notation: All vectors are considered as column vectors. The symbol $D_{\mathrm{KL}}(p\|q) = \mathbb{E}_p\left\{\log\frac{p}{q}\right\}$ denotes the Kullback-Leibler (KL) divergence between two probability distributions p and q. Unless otherwise specified, $\|x\|$ refers to the ℓ_2-norm of the vector x, defined as $\|x\| = \sqrt{x^\top x}$. The notation $\mathcal{N}(x;\mu,\Sigma)$ represents the Gaussian probability density function for the variable x with mean μ and covariance matrix Σ. For simplicity, this may be abbreviated as $\mathcal{N}(\mu,\Sigma)$. Furthermore, $\mathbb{E}\{x\}$, $\mathbb{D}\{x\}$, and $\mathrm{Cov}(x,y)$ denote

the expectation, variance, and covariance, respectively. The notation $A \leq B$ ($A < B$) for symmetric matrices A and B indicates that $B-A$ is positive (semi-)definite. The symbol $\otimes$ denotes the Kronecker product. The notation $\mathrm{vec}(A)$ refers to the vector obtained by stacking the columns of matrix A into a single column vector. The trace of matrix A is denoted as $\mathrm{Tr}(A)$. The notation $\mathbb{I}_{n\times n}$ represents the n-dimensional identity matrix.

2 Problem Statement

Consider the following nonlinear discrete-time stochastic system:

$$\begin{aligned} x_{t+1} &= f(x_t) + \xi_t, \\ y_t &= g(x_t) + \zeta_t, \end{aligned} \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the system state, $y_t \in \mathbb{R}^m$ is the noisy measurement. The function $f : \mathbb{R}^n \to \mathbb{R}^n$ is referred to as the transition function, while $g : \mathbb{R}^n \to \mathbb{R}^m$ is called measurement function; ξ_t denotes process noise, and ζ_t represents measurement noise. Typically, the initial state x_0, the process noise $\{\xi_t\}$, and the measurement noise $\{\zeta_t\}$ are all mutually independent, with the process and measurement noises being independent and identically distributed across time. This state-space model description in (1) can be represented as a hidden Markov model:

$$\begin{aligned} x_0 &\sim p(x_0), \\ x_t &\sim p(x_t|x_{t-1}), \\ y_t &\sim p(y_t|x_t). \end{aligned} \tag{2}$$

Here, $p(x_t|x_{t-1})$ and $p(y_t|x_t)$ are the transition and output probabilities respectively while $p(x_0)$ denotes the initial state distribution. In essence, (2) and (1) are different representations of the same system model. For example, consider the transition model $x_t = Ax_{t-1} + \xi_{t-1}$, where $\xi_{t-1} \sim \mathcal{N}(\xi_{t-1}; 0, Q)$, with Q denoting the covariance matrix of the process noise. This can be equivalently described by the transition probability $p(x_t|x_{t-1}) = \mathcal{N}(x_t; Ax_{t-1}, Q)$.

The objective of state estimation is to recover the system state x_t from noisy measurements y_t. Typically, finding the optimal estimate involves two key steps: calculating the posterior distribution $p(x_t|y_{1:t})$, and determining the optimal estimate $\hat{x}_{t|t}$ using the posterior distribution. A principled framework for posterior distribution calculation is Bayesian filtering, which computes $p(x_t|y_{1:t})$ recursively through two steps:

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\,\mathrm{d}x_{t-1}, \tag{3a}$$

$$p(x_t|y_{1:t}) = \frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{\int p(y_t|x_t)p(x_t|y_{1:t-1})\,\mathrm{d}x_t}. \tag{3b}$$

Here, (3a) is called prediction step while (3b) is called update step. The former, known as the Chapman-Kolmogorov equation, utilizes the transition probability $p(x_t|x_{t-1})$ to predict the prior distribution $p(x_t|y_{1:t-1})$. Based on this prior, the update step leverages the Bayes formula to calculate the posterior distribution $p(x_t|y_{1:t})$, where the output probability serves as likelihood function. After calculating $p(x_t|y_{1:t})$, common estimation criteria like the minimum mean square error or maximum a posteriori estimation can be used to determine the optimal estimate.

As discussed in Section 1, neither (3a) nor (3b) can be calculated analytically when the transition or output probabilities are non-linear or non-Gaussian, requiring the approximation of state distributions in practice. Given the high computational burden of discrete approximations like PF, Gaussian approximations are widely adopted in industrial applications. More specifically, in Gaussian filters, both the prior and posterior distributions are approximated by Gaussian distributions:

$$\begin{aligned} p(x_t|y_{1:t-1}) &\approx \mathcal{N}(x_t; \hat{x}_{t|t-1}, P_{t|t-1}), \\ p(x_t|y_{1:t}) &\approx \mathcal{N}(x_t; \hat{x}_{t|t}, P_{t|t}), \end{aligned} \tag{4}$$

which is usually achieved through *enabling approximation* framework. In particular, this framework first approximates the transition or output probabilities with linear Gaussian models and then performs the well-known KF. Due to the closure property of Gaussians under linear transformations and the conjugate property of Gaussians under conditioning, the resulting prior and posterior distributions are naturally approximated as Gaussian. The linearization techniques in existing Gaussian filters can be categorized into two types: Taylor series expansion and stochastic linear regression. As summarized in Table 1, the former uses the Jacobian matrix $g'(\bar{x})$ to provide local affine approximations of system models at point $\bar{x}$, while the latter minimizes the expectation of the square of the residual $y - Ax - b$ to find optimal linear parameters.

While these techniques offer practical methods for Gaussian approximation, it still remains unclear whether this *enabling approximation* framework truly achieve an optimal Gaussian approximation for Bayesian filtering. Unfortunately, there is even no clear way to formally judge whether a Gaussian approximation is optimal. This paper aims to establish a framework for defining and identifying the optimal Gaussian approximation for Bayesian filtering. Specifically, we seek to address the following three questions:

Q1: What defines the optimal Gaussian approximation in Bayesian filtering, and what conditions must it meet?

Q2: How can we design a filter that effectively achieves this optimal approximation?

Q3: What theoretical guarantees can we provide for the algorithm's convergence and the boundedness of the estimation error?

3 Optimal Gaussian Approximation for Bayesian Filtering

In this section, we will address **Q1** by exploring the necessary conditions for determining the optimal Gaussian approximation for Bayesian filtering.

3.1 Optimization Viewpoint of Bayesian Filtering

Inspired by the optimization-centric view on Bayes's rule [15], we show that the prior and posterior distributions can be interpreted as solutions of two variational problems, as shown in the subsequent proposition.

Proposition 1 (Variational Problems for Bayesian filtering)**.** *The prior distribution can be regraded as the maxmizer of an variational problem:*

$$p(x_t|y_{1:t-1}) = \arg\max_{q(x_t)} \mathbb{E}_{\substack{p(x_{t-1}|y_{1:t-1}) \\ p(x_t|x_{t-1})}} \left\{\log q(x_t)\right\}. \tag{5}$$

TABLE 1
Linearization techniques for existing Gaussian filters

Linearization Technique	Basic Principle: $\mathcal{N}(y; g(x), \Sigma) \approx \mathcal{N}(y; Ax + b, \Lambda)$	Representative Algorithms
—rule Taylor Series Expansion	$g(x) = g(\bar{x}) + g'(\bar{x})(x - \bar{x})$ $\Downarrow$ $A = g'(\bar{x}), b = g(\bar{x}) - g'(\bar{x})\bar{x}, \Lambda = \Sigma$	EKF [6], IEKF [8]
—rule Statistical Linear Regression	$\arg\min_{A,b} \mathbb{E}_x \left\{(y - Ax - b)^\top (y - Ax - b)\right\}$ $\Downarrow$ $A = \mathbf{Cov}(x, y)^\top \mathbb{D}(x), b = \mathbb{E}\{y\} - A\mathbb{E}\{x\}, \Lambda = \mathbb{D}\{y\} - A\mathbb{D}\{x\} A^\top$	UKF [11], Gauss–Hermite KF [12], cubature KF [13], and PLF [14]

Similarly, the posterior distribution can be regarded as the minimizer of a functional:

$$\begin{aligned} p(x_t|y_{1:t}) = &\arg\min_{q(x_t)} \left\{\mathbb{E}_{q(x_t)} \left\{-\log p(y_t|x_t)\right\}\right. \\ &\left. + D_{\mathrm{KL}} \left(q(x_t)||p(x_t|y_{1:t-1})\right) \right\}. \end{aligned} \tag{6}$$

Note that in both (5) *and* (6), $q : \mathbb{R}^n \to \mathbb{R}$ *represents the candidate density function. Besides, we use the notation* $\mathbb{E}_{\substack{p(x)\\p(y)}} \{f(x, y)\}$ *to represent the expectation of* $f(x, y)$ *with respect to both distributions* $p(x)$ *and* $p(y)$*, i.e.,*

$$\mathbb{E}_{\substack{p(x)\\p(y)}} \{f(x, y)\} \triangleq \mathbb{E}_{p(x)} \mathbb{E}_{p(y)} \{f(x, y)\}.$$

The proof of this proposition can be found in Appendix A in the supplementary material. This proposition shows that the prior distribution $p(x_t|y_{1:t-1})$ can be viewed as the solution to a variational problem that maximizes the expected logarithm of a candidate density $q(x_t)$ over the joint distribution of the previous state and the transition probability. This reflects the idea that the prior is derived by considering all possible transitions from the previous state, and selecting the distribution that maximizes the expected log-density under these transitions. Similarly, the posterior distribution $p(x_t|y_{1:t})$ is the solution to a variational problem that minimizes a cost combining the expected negative log-likelihood of the measurement model, $-\log p(y_t|x_t)$, with the KL divergence between the candidate density and the prior distribution $D_{\mathrm{KL}}(q(x_t)||p(x_t|y_{1:t-1}))$. This captures the Bayesian update process, where the posterior distribution adjusts the prior distribution based on new measurements to balance prior knowledge with new information.

In both (5) and (6), there are no constraints on the candidate distribution, meaning we seek an optimal candidate distribution over the entire probability space. However, such variational problems generally lack analytical solutions. A tractable approach to solve this problem is to restrict the candidate distribution to a parameterizable family of distributions, with the Gaussian distribution family being the most commonly used [16]. Therefore, by utilizing Gaussian approximations in (4), the two variational problems (5) and (6) that depict Bayesian filtering can be reduced to two optimization problems:

$$\begin{aligned} \hat{x}_{t|t-1}, P_{t|t-1} =& \arg\max_{\hat{x}_t, P_t} L(\hat{x}_t, P_t), \\ L(\hat{x}_t, P_t) =& \mathbb{E}_{\substack{\mathcal{N}(x_{t-1}; \hat{x}_{t-1|t-1}, P_{t-1|t-1})\\ p(x_t|x_{t-1})}} \left\{\log \mathcal{N}(x_t; \hat{x}_t, P_t)\right\}, \end{aligned} \tag{7a}$$

$$\begin{aligned} \hat{x}_{t|t}, P_{t|t} =& \arg\min_{\hat{x}_t, P_t} J(\hat{x}_t, P_t), \\ J(\hat{x}_t, P_t) =& D_{\mathrm{KL}} \left(\mathcal{N}(x_t; \hat{x}_t, P_t)||\mathcal{N}(x_t; \hat{x}_{t|t-1}, P_{t|t-1})\right) \\ &- \mathbb{E}_{\mathcal{N}(x_t; \hat{x}_t, P_t)} \left\{\log p(y_t|x_t)\right\}. \end{aligned} \tag{7b}$$

Here, $L(\hat{x}_t, P_t)$ is called prediction cost while $J(\hat{x}_t, P_t)$ is called update cost. Compared with (5) and (6), (7a) and (7b) transform the generally unsolvable variational problems to optimization problems for Gaussian parameters. This allows us to find the optimal Gaussian approximation for Bayesian filtering by studying the optimality conditions of (7).

3.2 Structure of Stationary Points of $L(\hat{x}_t, P_t)$

To establish the optimality conditions for problems (7a) and (7b), we examine the stationary points of $L(\hat{x}_t, P_t)$ and $J(\hat{x}_t, P_t)$. Interestingly, the extreme condition of the former reduces to a straightforward moment-matching equation, while that of the latter results in two mutually coupled implicit equations, whose roots are generally intractable. The following lemma helps elucidate the structure of $L(\hat{x}_t, P_t)$:

Lemma 1 (Stationary Points for Maximum Gaussian Likelihood)**.** *For probability density function* $p(x)$*, the stationary points of a maximum expected Gaussian likelihood problem*

$$\mu^*, \Sigma^* = \arg\max_{\mu, \Sigma} \mathbb{E}_{p(x)} \left\{\log \mathcal{N}(x; \mu, \Sigma)\right\}, \tag{8}$$

can be written as

$$\mu^* = \mathbb{E}_{p(x)} \{x\}, \tag{9a}$$

$$\begin{aligned} \Sigma^* &= \mathbb{E}_{p(x)} \left\{(x - \mu^*)(x - \mu^*)^\top\right\} \\ &= \mathbb{E}_{p(x)} \left\{xx^\top\right\} - \mu^* \mu^{*\top}. \end{aligned} \tag{9b}$$

The proof of this lemma can be found in Appendix B. This lemma states that the stationary points for maximizing the expected Gaussian likelihood are achieved when the Gaussian distribution matches the mean and variance of the given distribution. Therefore, as a special case of (8), solving (7a) requires to match the first and second order moment:

$$\begin{aligned} \hat{x}_{t|t-1} &= \mathbb{E}_{\substack{\mathcal{N}(x_{t-1}; \hat{x}_{t-1|t-1}, P_{t-1|t-1})\\ p(x_t|x_{t-1})}} \{x_t\}, \\ P_{t|t-1} &= \mathbb{E}_{\substack{\mathcal{N}(x_{t-1}; \hat{x}_{t-1|t-1}, P_{t-1|t-1})\\ p(x_t|x_{t-1})}} \left\{x_t x_t^\top\right\} - \hat{x}_{t|t-1} \hat{x}_{t|t-1}^\top. \end{aligned} \tag{10}$$

For the system obeys the transition model in (1) with ξ_t being the zero mean process noise, (10) can be further expressed as

$$\begin{aligned}\hat{x}_{t|t-1} =&\mathbb{E}_{\mathcal{N}(x_{t-1};\hat{x}_{t-1|t-1},P_{t-1|t-1})}\left\{f(x_{t-1})\right\},\\ P_{t|t-1} =&\mathbb{E}_{\mathcal{N}(x_{t-1};\hat{x}_{t-1|t-1},P_{t-1|t-1})}\left\{f(x_{t-1})f^\top(x_{t-1})\right\}\\ &+\mathbb{D}\left\{\xi_{t-1}\right\}-\hat{x}_{t|t-1}\hat{x}_{t|t-1}^\top.\end{aligned} \tag{11}$$

By leveraging variable substitution and the independence between process noise and state, we find that the optimal solution in (11) essentially requires finding the expectation of a nonlinear function with respect to a Gaussian distribution. The subsequent example explains this momoment-matching operation in terms of the canonical KF:

Example 1 (Prediction step of Kalman filter)**.** *For linear Gaussian systems $x_t = Ax_{t-1} + \xi_{t-1}$ where the process noise satisfies $\xi_{t-1} \sim \mathcal{N}(\xi_{t-1}; 0, Q)$, if the posterior distribution at time $t-1$ is given by $x_{t-1} \sim \mathcal{N}(x_{t-1}; \hat{x}_{t-1|t-1}, P_{t-1|t-1})$, the prediction step using* (11) *yields the following stationary points:*

$$\begin{aligned}\hat{x}_{t|t-1} =&\mathbb{E}_{\mathcal{N}(x_{t-1};\hat{x}_{t-1|t-1},P_{t-1|t-1})}\left\{Ax_{t-1}\right\}\\ =&A\mathbb{E}_{\mathcal{N}(x_{t-1};\hat{x}_{t-1|t-1},P_{t-1|t-1})}\left\{x_{t-1}\right\}\\ =&A\hat{x}_{t-1|t-1},\\ P_{t|t-1} =&\mathbb{E}_{\mathcal{N}(x_{t-1};\hat{x}_{t-1|t-1},P_{t-1|t-1})}\left\{Ax_{t-1}x_{t-1}^\top A^\top\right\}\\ &+\mathbb{D}\left\{\xi_{t-1}\right\}-A\hat{x}_{t-1|t-1}\hat{x}_{t-1|t-1}^\top A^\top\\ =&AP_{t-1|t-1}A^\top+Q.\end{aligned}$$

Note that the prior covariance matrix satisfies $P_{t-1|t-1} = \mathbb{E}_{\mathcal{N}(x_{t-1};\hat{x}_{t-1|t-1},P_{t-1|t-1})}\left\{x_{t-1}x_{t-1}^\top\right\} - \hat{x}_{t-1|t-1}\hat{x}_{t-1|t-1}^\top$.

In linear systems as shown in Example 1, the expectation operator in (11) is allowed to be interchanged with the affine function. However, for nonlinear functions f, computing the expectation requires numerical methods. As discussed in Section 1, moment-matching KF methods, such as UKF, GHKF, and CKF, approximate this expectation using techniques like the unscented transform, Gauss–Hermite quadrature, and spherical cubature. These methods provide optimal Gaussian approximations for the prediction step by numerically solving the expectation.

3.3 Structure of Stationary Points of $J(\hat{x}_t, P_t)$

Compared to the simple form of the stationary points of $L(\hat{x}_t, P_t)$, the stationary points of $J(\hat{x}_t, P_t)$ are relatively more complicated. For simplicity in notation, we define the negation of the log-likelihood as $\ell(x_t, y_t)$, where $\ell(x_t, y_t) = -\log p(y_t|x_t)$, and refer to $\ell(x_t, y_t)$ as the measurement-dependent loss. Using the analytical form of the KL divergence for two Gaussian distributions, the update cost can be formulated as

$$\begin{aligned}&J(\hat{x}_t, P_t)\\ =&\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t,P_t)}\left\{\ell(x_t,y_t)\right\}\\ &+D_{\mathrm{KL}}\left(\mathcal{N}(x_t;\hat{x}_t,P_t)||\mathcal{N}(x_t;\hat{x}_{t|t-1},P_{t|t-1})\right)\\ =&\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t,P_t)}\left\{\ell(x_t,y_t)\right\}\\ &+\frac{1}{2}\left(\hat{x}_{t|t-1}-\hat{x}_t\right)^\top P_{t|t-1}^{-1}\left(\hat{x}_{t|t-1}-\hat{x}_t\right)\\ &+\frac{1}{2}\mathrm{Tr}\left(P_{t|t-1}^{-1}P_t\right)-\frac{1}{2}\log\frac{|P_t|}{|P_{t|t-1}|}-\frac{1}{2}n.\end{aligned} \tag{12}$$

To find the stationary points of $J(\hat{x}_t, P_t)$, we need to calculate the partial derivatives with respect to $\hat{x}_t$ and P_t. The following lemma is helpful in simplifying these partial derivative calculations:

Lemma 2 (Gradient of expectation under Gaussian distribution)**.** *Assuming that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, we have the following results:*

(i). The gradient of $\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{f(x)\right\}$ w.r.t. the mean μ satisfies

$$\begin{aligned}\frac{\partial}{\partial\mu}\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{f(x)\right\} &= \mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{\frac{\partial}{\partial x}f(x)\right\}\\ &= \Sigma^{-1}\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{(x-\mu)f(x)\right\}.\end{aligned} \tag{13}$$

(ii). The Hessian matrix of $\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{f(x)\right\}$ w.r.t. the mean μ satisfies

$$\begin{aligned}&\frac{\partial^2}{\partial\mu^2}\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{f(x)\right\}\\ =&\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{\frac{\partial^2}{\partial x^2}f(x)\right\}\\ =&-2\Sigma^{-1}\left(\frac{\partial}{\partial\Sigma^{-1}}\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{f(x)\right\}\right)\Sigma^{-1}\\ =&\Sigma^{-1}\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{(x-\mu)(x-\mu)^\top f(x)\right\}\\ &-\Sigma^{-1}\mathbb{E}_{\mathcal{N}(x;\mu,\Sigma)}\left\{f(x)\right\}.\end{aligned} \tag{14}$$

Proof. The results presented are all related to Stein's lemma [17]. Result (i) is known as Bonnet's Theorem [18] and result (ii) is referred to as Price's Theorem [19]. Detailed discussions about these results can be found in this technical report [20]. □

Before studying the structure of the stationary points, we have the following assumption:

Assumption 1. *The measurement-dependent loss function $\ell(x_t, y_t)$ is twice differentiable with respect to the state x_t.*

Based on Assumption 1 and the result of Lemma 2, we have the partial derivative of $J(\hat{x}_t, P_t)$ with respect to $\hat{x}_t$ and P_t^{-1}:

$$\begin{aligned}&\frac{\partial J(\hat{x}_t,P_t)}{\partial\hat{x}_t}\\ =&\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t,P_t)}\left\{\frac{\partial\ell(x_t,y_t)}{\partial x_t}\right\}+P_{t|t-1}^{-1}(\hat{x}_t-\hat{x}_{t|t-1}),\end{aligned} \tag{15a}$$

$$\begin{aligned}&\frac{\partial J(\hat{x}_t,P_t)}{\partial P_t^{-1}}\\ =&-\frac{1}{2}P_t\cdot\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t,P_t)}\left\{\frac{\partial^2\ell(x_t,y_t)}{\partial x_t^2}\right\}\cdot P_t\\ &-\frac{1}{2}P_tP_{t|t-1}^{-1}P_t+\frac{1}{2}P_t.\end{aligned} \tag{15b}$$

To find extrema, we could attempt to set the first derivatives to zero

$$\frac{\partial J(\hat{x}_{t|t},P_{t|t})}{\partial\hat{x}_{t|t}}=\frac{\partial J(\hat{x}_{t|t},P_{t|t})}{\partial P_{t|t}^{-1}}=0,$$

we have

$$\hat{x}_{t|t}=\hat{x}_{t|t-1}-P_{t|t-1}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial\ell(x_t,y_t)}{\partial x_t}\right\}, \tag{16a}$$

$$P_{t|t}^{-1}=P_{t|t-1}^{-1}+\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial^2\ell(x_t,y_t)}{\partial x_t^2}\right\}. \tag{16b}$$

As shown in (16), the first-order condition is generally not possible to isolate for $\hat{x}_{t|t}$ and $P_{t|t}$. An exception for it is the well-known KF, where the expectation in (16) is a constant value, as shown in the subsequent example:

Example 2 (Update step of Kalman filter)**.** *For linear Gaussian systems with output probability $p(y_t|x_t) = \mathcal{N}(y_t; Cx_t, R)$, the optimal Gaussian approximation of the posterior mean in* (16a) *can be written as*

$$\begin{aligned}
&\hat{x}_{t|t} \\
=&\hat{x}_{t|t-1} + P_{t|t-1}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial}{\partial x_t}\log\left\{\mathcal{N}(y_t; Cx_t, R)\right\}\right\} \\
=&\hat{x}_{t|t-1} - P_{t|t-1}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial}{\partial x_t}\left\{\frac{1}{2}\|y_t - Cx_t\|^2_{R^{-1}}\right\}\right\} \\
=&\hat{x}_{t|t-1} + P_{t|t-1}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{C^\top R^{-1}(y_t - Cx_t)\right\} \\
=&\hat{x}_{t|t-1} + P_{t|t-1}C^\top R^{-1}(y_t - C\hat{x}_{t|t}).
\end{aligned} \tag{17}$$

From (17)*, we can obtain*

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(y_t - C\hat{x}_{t|t-1}), \tag{18}$$

with K_t being the gain matrix of KF defined as $K_t \triangleq P_{t|t-1}C^\top(CP_{t|t-1}C^\top + R)^{-1}$. Similarly, the optimal Gaussian approximation of the posterior covariance (16b) *can be written as*

$$\begin{aligned}
&P_{t|t} \\
=&\Big(P^{-1}_{t|t-1} - \mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\{\frac{\partial^2}{\partial x_t^2}\log\left\{\mathcal{N}(y_t; Cx_t, R)\right\}\}\Big)^{-1} \\
=&\left(P^{-1}_{t|t-1} + C^\top R^{-1}C\right)^{-1} \\
=&P_{t|t-1} - P_{t|t-1}C^\top(R + CP_{t|t-1}C^\top)^{-1}CP_{t|t-1}.
\end{aligned} \tag{19}$$

Note that (18) *and* (19) *consist of the update step of the canonical KF.*

As shown in Example 2, for the special case of a linear Gaussian system, the extreme conditions are specified by two decoupled equations whose analytical solutions correspond to the analytical form of the KF update. Specifically, the expectation on the right-hand side of (16a) depends solely on the posterior mean, while the expectation in (16b) is determined entirely by the system, meaning it is unrelated to either the mean or covariance. However, for a general nonlinear or non-Gaussian system, the expectations in (16) depend on both the posterior mean and covariance, rendering the stationary points analytically intractable.

Remark 1. *As discussed in Section 2, existing Gaussian filters rely on approximation techniques that solve* (16) *by linearizing the measurement model and then performing the KF updates as* (18) *and* (19)*. We contend that this type of approximation technique inevitably introduces linearization errors. For example, the Taylor series expansion technique unavoidably results in higher-order error terms. Moreover, these approximation techniques are applicable only for Gaussian noises [10]. Therefore, there is an urgent need to develop a new method that can directly solve the Gaussian approximation for Bayesian filtering update to avoid linearization errors.*

4 Natural Gradient Gasussian Approximation

In the previous section, we observed that existing moment-matching KF methods already provide exact numerical solutions for the optimal Gaussian approximation in the prediction step. However, the update step is not sufficiently resolved in current Gaussian filters, as these methods rely on linearization to approximate the stationary point, which introduce linearization errors. In this section, we aim to tackle **Q2**. Specifically, we will design an algorithm to solve (7b).

By examining the structure of the extreme conditons defined by (16), we find that it is challenging to directly obtain an analytical form of the stationary point because it is typically impossible to isolate the updates for the mean and covariance in (16). Therefore, a more practical approach to finding the optimal solution is to directly minimize the update cost $J(\hat{x}_t, P_t)$. To find the steepest descent in optimizing the parameters of Gaussian distributions [21], [22], we derive a natural gradient iteration for finding the optimal Gaussian approximation.

For simplicity, we stack the Gaussian parameters into a single column vector v and calculate the derivative with respect to it:

$$v = \begin{bmatrix}\hat{x}_t \\ \mathrm{vec}(P_t^{-1})\end{bmatrix},\ \frac{\partial}{\partial v}J(\hat{x}_t, P_t) = \begin{bmatrix}\frac{\partial}{\partial \hat{x}_t}J(\hat{x}_t, P_t) \\ \mathrm{vec}\left(\frac{\partial}{\partial P_t^{-\top}}J(\hat{x}_t, P_t)\right)\end{bmatrix}. \tag{20}$$

Here, we consider the inverse of the covariance matrix instead of its original form. This consideration is inspired by the structure of the information filter [23], [24], an equivalent form of the KF, where the inverse of the covariance matrix is employed instead of the covariance matrix itself. This is because the inverse can potentially simplify the mathematical expression of the update step in Bayesian filtering [23], [24]. Additionally, to easily represent the iteration, we define

$$\delta v \triangleq \begin{bmatrix}\delta\hat{x}_t \\ \mathrm{vec}(\delta P_t^{-1})\end{bmatrix} = \begin{bmatrix}\hat{x}_t^{(i+1)} - \hat{x}_t^{(i)} \\ \mathrm{vec}\left(\left(P_t^{-1}\right)^{(i+1)} - \left(P_t^{-1}\right)^{(i)}\right)\end{bmatrix},$$

where i is the iteration index. Under this notation, the natural gradient parameter update can be defined as

$$\delta v = -\left[\mathcal{F}_v^{-1}\frac{\partial}{\partial v}J(\hat{x}_t, P_t)\right]_{v=v^{(i)}}, \tag{21}$$

where $\mathcal{F}_v$ is the fisher information matrix associated with the Gaussian distribution $\mathcal{N}(x_t; \hat{x}_t, P_t)$ and $v^{(i)}$ represents the value of v in the i-th iteration. The next proposition provides the formulation of Fisher information matrix:

Proposition 2. *The inverse of the Fisher information matrix $\mathcal{F}_v^{-1}$ associated with $\mathcal{N}(x_t; \hat{x}_t, P_t)$ is*

$$\mathcal{F}_v^{-1} = \begin{bmatrix}P_t & 0 \\ 0 & 2(P_t^{-1} \otimes P_t^{-1})\end{bmatrix}, \tag{22}$$

where $\otimes$ is the kronecker product.

$$\begin{aligned}\left(P_t^{-1}\right)^{(i+1)} =& P_{t|t-1}^{-1} + \left(P_t^{-1}\right)^{(i)} \cdot \mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\left(x_t-\hat{x}_t^{(i)}\right)\left(x_t-\hat{x}_t^{(i)}\right)^\top \ell(x_t,y_t)\right\}\left(P_t^{-1}\right)^{(i)} - \left(P_t^{-1}\right)^{(i)}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\ell(x_t,y_t)\right\},\\ \hat{x}_t^{(i+1)} =& \hat{x}_t^{(i)} - P_t^{(i+1)} \cdot \left(P_t^{-1}\right)^{(i)} \cdot \mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\left(x_t-x_t^{(i)}\right)\ell(x_t,y_t)\right\} - P_t^{(i+1)}P_{t|t-1}^{-1}(\hat{x}_t^{(i)}-\hat{x}_{t|t-1}).\end{aligned} \tag{26}$$

The proof of this proposition can be found in Section 2.2 of [25]. Combining (20) and (22) with (21), we have

$$\begin{aligned}\delta\hat{x}_t =& -\left[P_t\frac{\partial}{\partial\hat{x}_t}J(\hat{x}_t,P_t)\right]_{v=v^{(i)}},\\ \operatorname{vec}\left(\delta P_t^{-1}\right) =& -2\left[\left(P_t^{-1}\otimes P_t^{-1}\right)\operatorname{vec}\left(\frac{\partial}{\partial P_t^{-1}}J(\hat{x}_t,P_t)\right)\right]_{v=v^{(i)}}\end{aligned}$$

Transforming this into matrix form, we can derive the following iterative updates:

$$\begin{aligned}\left(P_t^{-1}\right)^{(i+1)} =& \left(P_t^{-1}\right)^{(i)} \\ & - 2\left(P_t^{-1}\right)^{(i)}\frac{\partial}{\partial P_t^{-1}}J(\hat{x}_t,P_t)\Big|_{v^{(i)}}\left(P_t^{-1}\right)^{(i)},\\ \hat{x}_t^{(i+1)} =& \hat{x}_t^{(i)} - P_t^{(i+1)}\frac{\partial}{\partial\hat{x}_t}J(\hat{x}_t,P_t)\Big|_{v^{(i)}}.\end{aligned} \tag{23}$$

Combining the iterative updates (23) with the formulation of partial derivative, we have

$$\begin{aligned}\left(P_t^{-1}\right)^{(i+1)} =& P_{t|t-1}^{-1} + \mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\frac{\partial^2\ell(x_t,y_t)}{\partial x_t^2}\right\},\\ \hat{x}_t^{(i+1)} =& \hat{x}_t^{(i)} - P_t^{(i+1)}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\frac{\partial\ell(x_t,y_t)}{\partial x_t}\right\}\\ & - P_t^{(i+1)}P_{t|t-1}^{-1}(\hat{x}_t^{(i)}-\hat{x}_{t|t-1}).\end{aligned} \tag{24}$$

One practical issue when performing (24) is that the derivatives of $\ell(x_t,y_t)$ can be hard to compute. To avoid the need to compute derivatives of the measurement-dependent loss, we can once again apply Lemma 2 to acquire the derivative-free formulation. By applying (13) and (14), we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\frac{\partial\ell(x_t,y_t)}{\partial x_t}\right\}\\ =& \left(P_t^{-1}\right)^{(i)}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\left(x_t-\hat{x}_t^{(i)}\right)\ell(x_t,y_t)\right\},\\ &\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\frac{\partial^2\ell(x_t,y_t)}{\partial x_t^2}\right\}\\ =& \left(P_t^{-1}\right)^{(i)}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\{(x_t-\hat{x}_t^{(i)})(x_t-\hat{x}_t^{(i)})^\top\ell(x_t,y_t)\}\\ & - \left(P_t^{-1}\right)^{(i)}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\ell(x_t,y_t)\right\}.\end{aligned} \tag{25}$$

With the result in (25), we have the derivative-free update scheme shown in (26). This update scheme is still practically intractable for two reasons. First, the expectations generally do not have analytical forms. Second, the update scheme in (26) generally cannot guarantee that the covariance matrix will be positive definite.

To address the first issue, we could use well-established numerical integration methods, such as the unscented transform [11], Gauss–Hermite integration [26], or spherical cubature integration [13] to approximate the expectation calculations. For the second issue, one possible and efficient solution is to provide a sufficiently good initialization. For example, we could solve the maximum a posterior estimation problem:

$$\begin{aligned}\hat{x}_{t|t}^{\mathrm{MAP}} =& \arg\max_{x_t}\left\{\mathcal{N}(x_t;\hat{x}_{t|t-1},P_{t|t-1})\cdot\exp\{-\ell(x_t,y_t)\}\right\}\\ =& \arg\max_{x_t}\left\{\log\mathcal{N}(x_t;\hat{x}_{t|t-1},P_{t|t-1}) - \ell(x_t,y_t)\right\},\end{aligned} \tag{27}$$

and use Laplace's approximation [27] to construct the initial mean and covariance for the iteration in (26), as shown in the subsequent equation:

$$\begin{aligned}\hat{x}_t^{(0)} =& \hat{x}_{t|t}^{\mathrm{MAP}},\\ \left(P_t^{-1}\right)^{(0)} =& \left.\frac{\partial^2\left\{-\log\mathcal{N}(x_t;\hat{x}_{t|t-1},P_{t|t-1})+\ell(x_t,y_t)\right\}}{\partial x_t^2}\right|_{x_t=\hat{x}_{t|t}^{\mathrm{MAP}}}.\end{aligned} \tag{28}$$

This method works quite well in most scenarios. Besides this initialization trick, other methods such as Cholesky decomposition [28] or square-root parameterization [29] can also be leveraged to ensure the positive definiteness of the covariance matrix.

Another important consideration for the iterative scheme is the stopping criterion. As suggested by [14], we use the KL divergence between two consecutive Gaussian distributions, specifically $\mathcal{N}^{(i)} = \mathcal{N}(x_t;\hat{x}_{t|t}^{(i)},P_{t|t}^{(i)})$ and $\mathcal{N}^{(i+1)} = \mathcal{N}(x_t;\hat{x}_{t|t}^{(i+1)},P_{t|t}^{(i+1)})$ to determine when to stop the iteration:

$$D_{\mathrm{KL}}(\mathcal{N}^{(i)}\|\mathcal{N}^{(i+1)}) < \gamma, \tag{29}$$

where γ is a predefined threshold. This approach is more effective compared to using $D_{\mathrm{KL}}(\mathcal{N}^{(i+1)}\|\mathcal{N}^{(i)}) < \gamma$ because $\mathcal{N}^{(i+1)}$ is generally more concentrated. By using $D_{\mathrm{KL}}(\mathcal{N}^{(i)}\|\mathcal{N}^{(i+1)})$, we ensure that the criterion remains sensitive to convergence while avoiding premature termination of the algorithm.

Recall that in Section 3.2, we proved that the prediction step of the moment-matching KF algorithms essentially follows the optimal solution of Gaussian filtering. By combining this step with our natural gradient descent update step (26), we developed a new iterative filter. To emphasize that natural gradient descent is our key contribution, we call it the **N**atural gr**A**dient Gaussia**N** appr**O**ximation filter, or NANO filter for short. The pseudocode of the NANO filter is summarized in Algorithm 1. Note that all the expectation computations appearing in Algorithm 1 are suggested to use the efficient unscented transform [30].

5 Theoretical Analysis

In this section, we will answer **Q3**, i.e., we will provide the convergence and stability analysis for NANO filter.

5.1 Convergence Analysis

After deriving the update scheme of natural gradient Gaussian filtering in (24), a key question arises: does it converge

Algorithm 1 NANO Filter

Input: Stopping thresohold γ
Initialization: State estimate $\hat{x}_{0|0}$ and covariance $P_{0|0}$
for each time step t **do**
  Predict:
  Calculate predicted state mean $\hat{x}_{t|t-1}$ and covariance $P_{t|t-1}$ using (11)
  Update:
  Obtain the noisy measurement y_t
  Initialize the state estimate $\hat{x}_t^{(0)}$ and covariance $P_t^{(0)}$ using (28)
  for each iteration number i **do**
    if (29) is not satisfied **then**
      Update state estimate and covariance using (26)
    end if
  end for
  $\hat{x}_{t|t} = \hat{x}_t^{(i)}, P_{t|t} = P_t^{(i)}$
end for

to the optimal solution of (7b)? The following theorem confirms the local convergence of our update scheme:

Theorem 1. *Consider the Taylor series expansion that is second order in $\delta\hat{x}_t$ and first order in δP_t^{-1}. Under this approximation, the iterative update in* (24) *guarantees convergence, i.e.,*

$$\begin{aligned} J_t^{(i+1)} \approx\ & J_t^{(i)} + \frac{\partial J}{\partial \hat{x}_t^\top}\bigg|_{v^{(i)}} \delta\hat{x}_t + \frac{1}{2}\left(\delta\hat{x}_t\right)^\top \left(\frac{\partial^2 J}{\partial \hat{x}_t^2}\bigg|_{v^{(i)}}\right)\delta\hat{x}_t \\ & + \operatorname{Tr}\left(\frac{\partial J}{\partial P_t^{-1}}\bigg|_{v^{(i)}} \delta P_t^{-1}\right) \\ \leq\ & J_t^{(i)}, \end{aligned} \tag{30}$$

where $J_t^{(i)}$ is the update cost at the i-th iteration, defined as $J_t^{(i)} \triangleq J(\hat{x}_t^{(i)}, P_t^{(i)})$. Moreover, equality in (30) *holds if and only if $\delta\hat{x}_t = 0$ and $\delta P_t^{-1} = 0$.*

The proof of this theorem is provided in Appendix D in the supplementary material. This theorem indicates that natural gradient descent iteration in the update step of the NANO filter provides a guarantee of local convergence. This guarantee is achieved by approximating the objective function with second-order accuracy around the mean and first-order accuracy around the inverse of the covariance matrix. The key idea of proof is to show that the difference in update costs between consecutive iterations can be expressed as a semi-negative definite quadratic form. Technically, achieving this semi-negative definiteness relies on the use of the Fisher information matrix, which corrects the gradient direction to provide the steepest descent on the Gaussian manifold. This adjustment makes the gradient "natural" because it aligns with the geometry of the Riemannian space of Gaussian parameters.

For linear Gaussian systems, the NANO filter achieves the optimal solution of (7b) in a single iteration, as stated in the following corollary:

Corollary 1. *For linear Gaussian systems in Example 2, the update rule given by* (24) *converges to the optimal solution of* (7b) *within one iteration. In other words, a single iteration of NANO filter is equivalent to KF.*

The proof of this corollary can be found in Appendix C in the supplementary material. Corollary 1 holds regardless of the initialization of the Gaussian parameters $\hat{x}_t^{(0)}$ and $P_t^{(0)}$, supporting the fact that the natural gradient is the steepest descent direction in the Gaussian manifold.

5.2 Stability Analysis

Next, we will analyze the stability of the proposed NANO filter. Stability is the most critical property, as it ensures that the estimation error remains bounded throughout the filtering process. Our stability analysis is conducted based on the state-space model given in (1). We focus on the case where the process and measurement noise are both zero-mean Gaussian noise, satisfying $\xi_t \sim \mathcal{N}(0, Q_t)$ and $\zeta_t \sim \mathcal{N}(0, R_t)$. In this case, the measurement-dependent loss $\ell(x_t, y_t)$ is the log-likelihood loss function that satisfies $\ell(x_t, y_t) = -\log p(y_t|x_t) = C_0 + \frac{1}{2}(y_t - g(x_t))^\top R_t^{-1}(y_t - g(x_t))$. Here, $C_0 > 0$ is a constant value irrelevant to the state and measurement. For the purpose of stability analysis, we make the following regularity assumptions on the system functions f and g.

Assumption 2. *The derivatives of the functions f and g are bounded. Specifically, there exists a constant $C > 0$ such that*

$$\left|\frac{\partial f^i}{\partial x^j}\right|, \left|\frac{\partial g^k}{\partial x^j}\right|, \left|\frac{\partial^2 g^k}{\partial x^i \partial x^j}\right| \leq C < \infty, \tag{31}$$

for all $x \in \mathbb{R}^n$, $1 \leq i, j \leq n$, $1 \leq k \leq m$. Here, x^i represents the i-th component of the state vector x; similarly, f^i and g^k denote the corresponding components of the functions f and g, respectively.

Using the standardization of the state x_t under the Gaussian distribution, we can rewrite the stationary point condition for update step (16) in the following tensor form:

$$\begin{aligned} \hat{x}_{t|t}^i =& \hat{x}_{t|t-1}^i + (P_{t|t-1})^{ij} \int (R_t^{-1})_{kl} \frac{\partial g^k}{\partial x^j}(\hat{x}_{t|t} + S_{t|t}z) \\ & \times (y_t^l - g^l(\hat{x}_{t|t} + S_{t|t}z)) \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\|z\|^2}\, \mathrm{d}z, \end{aligned} \tag{32a}$$

$$\begin{aligned} (P_{t|t}^{-1})_{ij} =& (P_{t|t-1}^{-1})_{ij} \\ & + \int (R_t^{-1})_{kl} \left[\frac{\partial g^k}{\partial x^i}(\hat{x}_{t|t} + S_{t|t}z)\frac{\partial g^l}{\partial x^j}(\hat{x}_{t|t} + S_{t|t}z)\right. \\ & \left. - \frac{\partial^2 g^k}{\partial x^i \partial x^j}(\hat{x}_{t|t} + S_{t|t}z)(y_t^l - g^l(\hat{x}_{t|t} + S_{t|t}z))\right] \\ & \times \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\|z\|^2}\, \mathrm{d}z, \end{aligned} \tag{32b}$$

where $S_{t|t}S_{t|t}^\top = P_{t|t}$ is obtained by Cholesky decomposition, and we use the Einstein summation convention [31].

Next, we will substitute the measurement model, $y_t^l = g^l(x_t) + \zeta_t^l$ into the tensor form of the stationary point condition in (32), and consider the Taylor expansion of $g(x_t)$ at $\hat{x}_{t|t}$. For the stationary point condition of $\hat{x}_{t|t}$, let us first define the auxiliary function $h_j^{kl}(x)$, for $1 \leq j \leq n$ and $1 \leq k, l \leq n$:

$$\begin{aligned} h_j^{kl}(x) =& \int \frac{\partial g^k}{\partial x^j}(\hat{x}_{t|t} + S_{t|t}z) \\ & \times (g^l(x) - g^l(\hat{x}_{t|t} + S_{t|t}z)) \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\|z\|^2}\, \mathrm{d}z, \end{aligned}$$

then (32a) can be rewritten as

$$\begin{aligned}\hat{x}^i_{t|t} =&\hat{x}^i_{t|t-1} + (P_{t|t-1})^{ij}(R_t^{-1})_{kl}h_j^{kl}(x_t)\\ &+ (P_{t|t-1})^{ij}(R_t^{-1})_{kl}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial g^k}{\partial x^j}(x_t)\right\}\zeta_t^l\\ =&\hat{x}^i_{t|t-1} + (P_{t|t-1})^{ij}(R_t^{-1})_{kl}\\ &\times\left[h_j^{kl}(\hat{x}_{t|t}) + \frac{\partial h_j^{kl}}{\partial x^q}(x_t^q - \hat{x}^q_{t|t}) + \psi_j^{kl}(x_t - \hat{x}_{t|t})\right]\\ &+ (P_{t|t-1})^{ij}(R_t^{-1})_{kl}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial g^k}{\partial x^j}(x_t)\right\}\zeta_t^l,\end{aligned} \tag{33}$$

where

$$\psi_j^{kl}(x_t - \hat{x}_{t|t}) \le \kappa_1(\|x_t - \hat{x}_{t|t}\|^2), \forall\, 0 \le \|x_t - \hat{x}_{t|t}\| < \epsilon_1,$$

for some $\kappa_1, \epsilon_1 > 0$, are high-order terms in the Taylor expansion of h_j^{kl}.

Let us denote the estimation error as $e_{t|t} = x_t - \hat{x}_{t|t}$ and $e_{t|t-1} = x_t - \hat{x}_{t|t-1}$. Regardless of the specific methods used in the prediction step, the prediction error $e_{t|t-1}$ can be expressed as

$$\begin{aligned}e^i_{t|t-1} =&f^i(x_{t-1}) + \xi_t^i - \hat{x}^i_{t|t-1}\\ =&\frac{\partial f^i}{\partial x^j}(\hat{x}_{t-1|t-1})(x^j_{t-1} - \hat{x}^j_{t-1|t-1})\\ &+ \tilde{\psi}^i(x_{t-1} - \hat{x}_{t-1|t-1}) + \xi_t^i,\end{aligned} \tag{34}$$

where

$$\begin{aligned}\tilde{\psi}^i(x_{t-1} - \hat{x}_{t-1|t-1}) \le \kappa_2(\|x_{t-1} - \hat{x}_{t-1|t-1}\|^2),\\ \forall\, 0 \le \|x_{t-1} - \hat{x}_{t-1|t-1}\| < \epsilon_2,\end{aligned}$$

for some $\kappa_2, \epsilon_2 > 0$, which are higher-order terms in the Taylor expansion of f^i. Subtracting x_t from both sides of (33) and applying (34), we obtain

$$\begin{aligned}e^i_{t|t} =&\frac{\partial f^i}{\partial x^j}(\hat{x}_{t-1|t-1})e^j_{t-1|t-1} + \tilde{\psi}^i(x_{t-1} - \hat{x}_{t-1|t-1})\\ &+ \xi_t^i - (P_{t|t-1})^{ij}(R_t^{-1})_{kl}\\ &\times\left[h_j^{kl}(\hat{x}_{t|t}) + \frac{\partial h_j^{kl}}{\partial x^q}(\hat{x}_{t|t})e^q_{t|t} + \psi_j^{kl}(x_t - \hat{x}_{t|t})\right]\\ &- (P_{t|t-1})^{ij}(R_t^{-1})_{kl}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial g^k}{\partial x^j}(x_t)\right\}\zeta_t^l.\end{aligned} \tag{35}$$

In matrix form, (35) becomes

$$e_{t|t} = F_{t-1}e_{t-1|t-1} - H_t e_{t|t} - \bar{h}_t + \Psi_t + \xi_t - G_t\zeta_t,$$

where F_t, H_t, $\bar{h}_t$ and G_t are matrix- or vector-valued functions with components given by

$$\begin{aligned}(F_{t-1})^i_j &= \frac{\partial f^i}{\partial x^j}(\hat{x}_{t-1|t-1}),\\ (H_t)^i_q &= (P_{t|t-1})^{ij}(R_t^{-1})_{kl}\frac{\partial h_j^{kl}}{\partial x^q}(\hat{x}_{t|t}),\\ \bar{h}^i_t &= (P_{t|t-1})^{ij}(R_t^{-1})_{kl}h_j^{kl}(\hat{x}_{t|t}),\\ (G_t)^i_l &= (P_{t|t-1})^{ij}(R_t^{-1})_{kl}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial g^k}{\partial x^j}(x_t)\right\},\end{aligned}$$

and

$$\Psi^i_t = \tilde{\psi}^i(e_{t-1|t-1}) - (P_{t|t-1})^{ij}(R_t^{-1})_{kl}\psi_j^{kl}(x_t - \hat{x}_{t|t})$$

are the high-order terms. Therefore, we have

$$\begin{aligned}e_{t|t} =&(I + H_t)^{-1}F_{t-1}e_{t-1|t-1} - (I + H_t)^{-1}\bar{h}_t\\ &+ (I + H_t)^{-1}\Psi_t + (I + H_t)^{-1}(\xi_t - G_t\zeta_t).\end{aligned} \tag{36}$$

Similarly, for the stationary point condition of $P_{t|t}$, let us define the auxiliary function $\tilde{h}_{ij}^{kl}(x)$, for $1 \le i, j \le n$ and $1 \le k, l \le n$:

$$\begin{aligned}\tilde{h}_{ij}^{kl}(x) = \int\Big[&\frac{\partial g^k}{\partial x^i}(\hat{x}_{t|t} + S_{t|t}z)\frac{\partial g^l}{\partial x^j}(\hat{x}_{t|t} + S_{t|t}z)\\ &- \frac{\partial^2 g^k}{\partial x^i \partial x^j}(\hat{x}_{t|t} + S_{t|t}z)(g^l(x) - g^l(\hat{x}_{t|t} + S_{t|t}z))\Big]\\ &\times\frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}\|z\|^2}\,\mathrm{d}z,\end{aligned}$$

then (32b) can be rewritten as

$$\begin{aligned}(P^{-1}_{t|t})_{ij} =&(P^{-1}_{t|t-1})_{ij} + (R_t^{-1})_{kl}\tilde{h}_{ij}^{kl}(x_t)\\ &- (R_t^{-1})_{kl}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial^2 g^k}{\partial x^i \partial x^j}(x_t)\right\}\zeta_t^l\\ =&(P^{-1}_{t|t-1})_{ij} + (R_t^{-1})_{kl}\Big[\tilde{h}_{ij}^{kl}(\hat{x}_{t|t})\\ &+ \frac{\partial \tilde{h}_{ij}^{kl}}{\partial x^q}(\hat{x}_{t|t})(x_t^q - \hat{x}^q_{t|t}) + \varphi_{ij}^{kl}(x_t - \hat{x}_{t|t})\Big]\\ &- (R_t^{-1})_{kl}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\left\{\frac{\partial^2 g^k}{\partial x^i \partial x^j}(x_t)\right\}\zeta_t^l,\end{aligned}$$

where

$$\varphi_{ij}^{kl}(x_t - \hat{x}_{t|t}) \le \kappa_3\|x_t - \hat{x}_{t|t}\|^2, \forall\, 0 \le \|x_t - \hat{x}_{t|t}\| < \epsilon_3,$$

for some $\kappa_3, \epsilon_3 > 0$, are high-order terms in the Taylor expansion of $\tilde{h}_{ij}^{kl}$.

The main idea of stability analysis is based on the application of Lyapunov functions, just as in the case of Kalman-type nonlinear filter [32], [33]. In order to construct the Lyapunov function, let us first introduce the auxiliary covariance matrix $\tilde{P}_{t|t}$ and $\tilde{P}_{t|t-1}$, which evolve according to

$$\begin{aligned}(\tilde{P}^{-1}_{t|t})_{ij} =&(\tilde{P}^{-1}_{t|t-1})_{ij} + (R_t^{-1})_{kl}\tilde{h}_{ij}^{kl}(\hat{x}_{t|t}),\\ (\tilde{P}_{t|t-1})^{ij} =&\frac{\partial f^i}{\partial x^k}(\hat{x}_{t-1|t-1})\frac{\partial f^j}{\partial x^l}(\hat{x}_{t-1|t-1})\\ &\times(\tilde{P}_{t-1|t-1})^{kl} + (Q_t)^{ij},\end{aligned}$$

and in matrix form

$$\begin{aligned}\tilde{P}^{-1}_{t|t} &= \tilde{P}^{-1}_{t|t-1} + D_t,\\ \tilde{P}_{t|t-1} &= F_{t-1}\tilde{P}_{t-1|t-1}F_{t-1}^\top + Q_t,\end{aligned} \tag{37}$$

where D_t is the matrix with each component $(D_t)_{ij} = (R_t^{-1})_{kl}\tilde{h}_{ij}^{kl}(\hat{x}_{t|t})$. Note that the evolution of $\tilde{P}_{t|t}$ and $\tilde{P}_{t|t-1}$ does not depend directly on the errors $e_{t|t}$, $e_{t|t-1}$ or the noise terms ξ_t, ζ_t. This makes the following positive definiteness and boundedness assumption largely a condition on the system itself, much like the detectable and controllable conditions for linear systems [34].

Assumption 3. *There exist constants $\underline{p}, \bar{p} > 0$, such that*

$$\begin{aligned}0 < \underline{p}I \le \tilde{P}_{t|t} \le \bar{p}I < \infty,\ \forall\, t \ge 0,\\ 0 < \underline{p}I \le \tilde{P}_{t|t-1} \le \bar{p}I < \infty,\ \forall\, t \ge 0.\end{aligned}$$

Our main stability result is stated in the following theorem. Generally speaking, this theorem proves the stability of our proposed method for those systems with almost linear measurement functions and small noise.

Theorem 2. *Under Assumption 2 and 3, the estimation error $e_{t|t}$ is exponentially bounded in the mean square for systems with almost linear measurement functions, i.e., there exist $\epsilon, \epsilon', \lambda > 0$, such that*

$$\mathbb{E}\left\|e_{t|t}\right\|^2 \leq \epsilon \|e_{0|0}\|^2 \left(\frac{1}{1+\lambda}\right)^t + \epsilon', \ \forall\, t \geq 0,$$

as long as the initial error and the strength of the system noise are small enough, that is $\|e_{0|0}\| \leq \delta$, $\mathbb{E}\left\{\xi_t \xi_t^\top\right\} \leq \delta I$ and $\mathbb{E}\left\{\zeta_t \zeta_t^\top\right\} \leq \delta I$ for some given $\delta > 0$.

The proof of this theorem can be found in Appendix E in the supplementary material.

6 Robustifying NANO Filter with Gibbs Posterior

In the update step of Bayesian filtering, as illustrated in (6), the objective is to strike a balance between the information provided by the measurement data and the prior distribution. However, obtaining reliable information from the measurement data requires a comprehensive probabilistic model of the measurement-generating process. In other words, a precise specification of the output distribution $p(y_t|x_t)$ is needed beforehand. Unfortunately, in practical scenarios, it is often challenging to specify the true model of $p(y_t|x_t)$ due to factors such as sensor malfunctions or unmodeled system dynamics. This leads to a mismatch between the true data-generating process and the assumed output probability model. Measurements influenced by such model misspecifications, which are typically referred to as outliers, require special attention to maintain the reliability of state estimates [35], [36]. One effective approach to handle this issue is to replace the likelihood function in (6) with a generalized measurement-dependent loss function, leading to the following variational problem:

$$\begin{aligned} p_{\mathrm{G}}(x_t|y_{1:t}) = \arg\min_{q(x_t)} \Big\{ & \mathbb{E}_{q(x_t)} \left\{ \ell^G(x_t, y_t) \right\} \\ & + D_{\mathrm{KL}} \left\{ q(x_t) || p(x_t|y_{1:t-1}) \right\} \Big\}. \end{aligned} \tag{38}$$

Here, $\ell^G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the generalized measurement-dependent loss function. It turns out that the solution to (38) is known as the Gibbs posterior, which has an analytical form akin to the standard Bayesian posterior:

$$p_{\mathrm{G}}(x_t|y_{1:t}) = \frac{\exp\{-\ell^G(x_t, y_t)\} p(x_t|y_{1:t-1})}{\int \exp\{-\ell^G(x_t, y_t)\} p(x_t|y_{1:t-1})\, \mathrm{d}x_t}.$$

Here, we denote the Gibbs posterior as $p_{\mathrm{G}}(x_t|y_{1:t})$ to differentiate from the standard Bayesian posterior $p(x_t|y_{1:t})$. In fact, the standard Bayesian posterior can be regarded as a special case of the Gibbs posterior by setting the loss function as the negation of the logarithm of the likelihood function, i.e., $\ell^G(x_t, y_t) = -\log p(y_t|x_t)$.

Remark 2. *A notable fact is that the analysis of the Gaussian approximation in Section 3.3 and the derivation of the NANO filter in Section 4 do not depend on the specific form of the loss function $\ell(x_t, y_t)$. Therefore, since Gibbs Bayesian filtering in* (38) *only modifies the loss function $\ell(x_t, y_t)$ in* (6) *to a more general form, $\ell^G(x_t, y_t)$, all of the previous analysis and algorithmic design can naturally extend to this case.*

Then, we provide several potential choices of the measurement-dependent loss functions that can tackle model misspecifications:

Choice 1: Composite Likelihood: One popular choice is to use the composite likelihood loss, i.e., combine multiple likelihood functions to achieve robust and adaptive performance under varying conditions. For example, for single-dimension measurement case, we can assume that the loss function is a composition of Gaussian likelihood and Laplace likelihood:

$$\begin{aligned} & \ell^{\mathrm{h}}(x_t, y_t) \\ = & \begin{cases} -\log \mathcal{N}(y_t; g(x_t), 1) & \text{if} \quad |y_t - g(x_t)| \leq \delta, \\ -\log \mathrm{Laplace}(y_t; g(x_t), \frac{1}{\delta}) & \text{otherwise.} \end{cases} \\ \overset{c}{=} & \begin{cases} \frac{1}{2}|y_t - g(x_t)|^2 & \text{if} \quad |y_t - g(x_t)| \leq \delta, \\ \delta\left(|y_t - g(x_t)| - \frac{1}{2}\delta\right) & \text{otherwise.} \end{cases} \end{aligned}$$

Here, the notation " $\overset{c}{=}$ " represents that two expressions are equivalent up to an additive constant, and $\delta > 0$ is the threshold variable. In fact, $\ell^{\mathrm{h}}(x_t, y_t)$ represents the famous Huber loss [37] used in robust regression, which is known to be less sensitive to outliers in data than the squared error loss. In practical applications, we utilize a smooth approximation of the Huber loss, known as the Pseudo-Huber loss, to improve optimization efficiency and ensure differentiability. The Pseudo-Huber loss is defined as:

$$\ell^{\mathrm{ph}}(x_t, y_t) = \delta^2 \left(\sqrt{1 + \left(y_t - g(x_t)\right)^2 / \delta^2} - 1 \right). \tag{39}$$

Choice 2: Weighted Log-Likelihood: Besides the composition of different likelihoods, one natural choice to achieve a robust loss function is to scale the log-likelihood loss function with a data-dependent weighting term:

$$\ell^{\mathrm{w}}(x_t, y_t) = -w(x_t, y_t) \cdot \log p(y_t|x_t). \tag{40}$$

Here, $w : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}_+$ is the weighting function. The philosophy behind this form of loss is that reweighting the effect of outliers in the filtering procedure can potentionally improve robustness. The choices of the weighting function can be inspired from several domains. For example, we can use the so-called inverse multi-quadratic weighting function [38]

$$w(x_t, y_t) = \left(1 + \|y_t - g(x_t)\|^2_{R_t^{-1}} / c^2\right)^{-1},$$

where $c > 0$ is a constant number. Other choices of the weighting function can also be found in Section 3.3 of [38].

Remark 3. *When $w(x_t, y_t)$ is chosen as the constant value that is smaller than 1, it can be regarded as the result of performing the so-called exponential density rescaling for convlutional Bayes filter [39], which already shows robustness for systems with outliers. In this case, $w(x_t, y_t)$ can be regarded as the Lagrange multiplier which balances the compression and recontruction in the well-known information bottleneck problem [39].*

Choice 3: Divergence-Dependent Loss Function: The final option is to utilize robust divergence to fit the data. This approach is based on the fact that minimizing the negative log-likelihood essentially amounts to minimizing the KL divergence between the true data distribution $p_{\text{true}}(y_t)$, which includes outliers, and the assumed likelihood:

$$\begin{aligned}&\arg\min_{q(x_t)} \mathbb{E}_{q(x_t)} \left\{ D_{\text{KL}} \left(p_{\text{true}}(y_t) \| p(y_t|x_t) \right) \right\} \\ =&\arg\min_{q(x_t)} \mathbb{E}_{q(x_t)} \mathbb{E}_{p_{\text{true}}(y_t)} \left\{ -\log p(y_t|x_t) \right\} \\ \approx&\arg\min_{q(x_t)} \mathbb{E}_{q(x_t)} \left\{ -\log p(y_t|x_t) \right\}.\end{aligned} \tag{41}$$

In this equation, the first equality arises because the term $p_{\text{true}}(y_t)$ is treated as a constant regarding $q(x_t)$ and thus omitted, since the objective is to find the minimizer, not the minimum value. The final approximation holds because we cannot directly access the true data distribution; instead, we use the sample y_t to approximate the expected value.

Based on this understanding, one natural idea for designing a robust loss is to replace the KL divergence with a more robust divergence, such as β-divergence or γ-divergence. For instance, replacing the KL divergence with β-divergence in (41) leads to the β loss function $\ell^{\beta}(x_t, y_t)$ [35]:

$$\ell^{\beta}(x_t, y_t) = -\frac{\beta+1}{\beta} p(y_t|x_t)^{\beta} + \int p(y|x_t)^{\beta+1} \mathrm{d}y. \tag{42}$$

Remark 4. *Formally analyzing which loss function provides superior robustness is challenging as each is grounded in different principles. A common approach is to apply Huber's robust statistics theory [40], which assesses the robustness of an estimator using its influence function. An estimator is considered robust if its influence function remains bounded as an outlier's value increases indefinitely. Previous studies have conducted case-specific analyses of robustness for Huber loss [41], weighted loss [38], and β loss [35] in linear Gaussian systems, demonstrating that adjusting the loss function can improve robustness. Since this paper focuses on applying robust loss functions to enhance the NANO filter, a detailed formal analysis is left for future work.*

7 Discussions

Natural Gradient Descent for Gaussian Approximation: The use of natural gradient descent for finding optimal Gaussian approximations is well-documented in the literature, with early works dating back to [42] with applications for Gaussian process regression. After that, related methods have been explored in various domains, including robot batch estimation [43], [44], Bayesian deep learning [45], [46], approximate inference [47], optimization [48] and more.

We utilize the natural gradient method to find the optimal Gaussian approximation for Bayesian filtering in this paper. The advantage of incorporating the natural gradient in Gaussian filtering may arise from the profound geometric properties of the statistical manifolds generated by the family of Gaussian distributions, as discussed in [49], [50]. Intuitively, the computation of the natural gradient involves second-order derivatives of the probability density functions. Thus, geometric concepts such as Riemannian metrics, curvatures, and geodesics on the statistical manifold can be engaged and reflected in the algorithm, leading to better convergence performance. Nevertheless, a precise description of the relationship between the natural gradient and Gaussian distribution remains an important research direction for further exploration.

Gradient-Based Gaussian Filters: Gradient-based Gaussian filters, which use gradient descent and its variants to solve Gaussian filtering, were initially proposed in [51], utilizing Monte Carlo techniques to approximate the exact gradient. Subsequently, cubature rules were introduced to replace the Monte Carlo method for gradient approximation, and various optimization approaches, such as the conditional gradient method, the alternating direction method of multipliers, and the natural gradient descent were employed to replace basic gradient descent for optimization [52], [53], [54], [55]. These methods have certain assumptions and limitations compared to our approach. For instance, (i) the methods in [51], [52], [53], [55] are restricted to systems with Gaussian noise; (ii) [51], [54] and [55] require performing linear approximations of the measurement model; (iii) [52] modifies the implicit hard constraints of the original problem into soft constraints, rendering it inequivalent to the original problem and (iv) [56] only finds the maximum a posterior estimate, rather than the entire optimal Gaussian distribution. Also, they both lack rigorous proofs of algorithm convergence and stability guarantees for the estimation error, which we provide in Section 5. Besides, their optimization problems are constructed based on direct minimization of forward KL divergence between the candidate distribution and the true posterior, which does not support the extension to the Gibbs posterior as proposed in our method.

Moreover, we would like to emphasize that, while these works utilize gradient-based methods for Gaussian approximation, they lack a theoretical analysis of the stationary points related to the optimal Gaussian approximation. Specifically, they neither justify the use of gradient methods in the update process nor address the Gaussian approximation in the prediction step. In contrast, our paper provides a comprehensive analysis covering both the prediction and update steps, including a detailed examination of the stationary points for the optimal Gaussian approximation.

Kalman Filtering as Natural Gradient Descent: Corollary 1 shows that the canonical KF can be interpreted as a single iteration of the natural gradient in our proposed NANO filter. This result looks similar to the findings in [57], where the equivalence between KF and online natural gradient descent is established. However, [57] focuses on parameter estimation, where the state is treated as a deterministic variable representing the parameters of the measurement model. In that context, the natural gradient is applied to maximize the likelihood with respect to the measurement model. In contrast, our result is built on a more general Bayeisan view, where the state is a random variable, and the natural gradient is used to find the optimal parameters of state distribution. Therefore, our result can be treated as a generalization of the result in [57].

8 Experiments

In this section, we validate the proposed NANO filter and its robust variants using the loss functions introduced in Section 6. We conduct simulations for both linear and

nonlinear systems, along with a real-world experiment. All the simulations and experiment are evaluated by root mean square error (RMSE) which is defined by

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T} \|x_t - \hat{x}_t\|^2}{n \cdot T}},$$

where T is the total time step for each trajectory. As it is generally impossible to compare with all existing Gaussian filters, we focus on the most popular ones as baselines for our analysis, including the KF, EKF, UKF, IEKF, and PLF. These filters are all mentioned repeatedly in the well-regarded textbook [10].

8.1 Linear System Simulation: Wiener Velocity Model

First, we perform simulations for Wiener velocity model, which is a canonical linear Gaussian model commonly employed for target tracking [10]. In this model, the state represents the position and velocity of a moving object in two dimensions. The state vector is defined as $x = [p_x\ p_y\ v_x\ v_y]^\top$, where p_x and p_y are the object's positions along the longitude and lateral directions, and v_x and v_y are the corresponding velocities. The measurements are direct, noisy observations of the position components. The state transition model with a time step $\Delta t = 0.1$ and the measurement model can be described as

$$x_{t+1} = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} x_t + \xi_t,$$
$$y_t = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} x_t + \zeta_t.$$

Here, $\xi_t \sim \mathcal{N}(\xi_t; 0, Q)$ and $\zeta_t \sim \mathcal{N}(\zeta_t; 0, R)$ are the process and measurement noises with covariance matrices given by

$$Q = \begin{bmatrix} \frac{\Delta t^3}{3} & 0 & \frac{\Delta t^2}{2} & 0 \\ 0 & \frac{\Delta t^3}{3} & 0 & \frac{\Delta t^2}{2} \\ \frac{\Delta t^2}{2} & 0 & \Delta t & 0 \\ 0 & \frac{\Delta t^2}{2} & 0 & \Delta t \end{bmatrix}, R = \mathbb{I}_{2\times 2}.$$

To validate our method, we consider two scenarios. In the first scenario, the proposed NANO filter is compared against the KF, the optimal filter for linear Gaussian systems, as well as UKF and PLF. As shown in Fig. 8 in Appendix H in the supplementary material, the NANO filter achieves the same performance as the KF family within one iteration, supporting the result in Corollary 1 that KF equals a single iteration of NANO filter. In the second scenario, we consider the case where the measurement data deviates from the measurement model due to contamination of outliers. Specifically, the measurement data has a 10% probability of being contaminated, modeled as $\zeta_t \sim 0.9 \cdot \mathcal{N}(0, R) + 0.1 \cdot \mathcal{N}(0, 1000R)$. For this case, as illustrated in Fig. 1 in Appendix H in supplementary material, we evaluate the robust variants of the NANO filter, including the Huber-NANO filter, Weight-NANO filter, and β-NANO filter, which utilize the robust loss functions ℓ^{ph} (39), ℓ^{w} (40), and ℓ^{β} (42), respectively. The box plots indicate that the robust variants of the NANO filter generally outperform the standard NANO filter using different parameters.

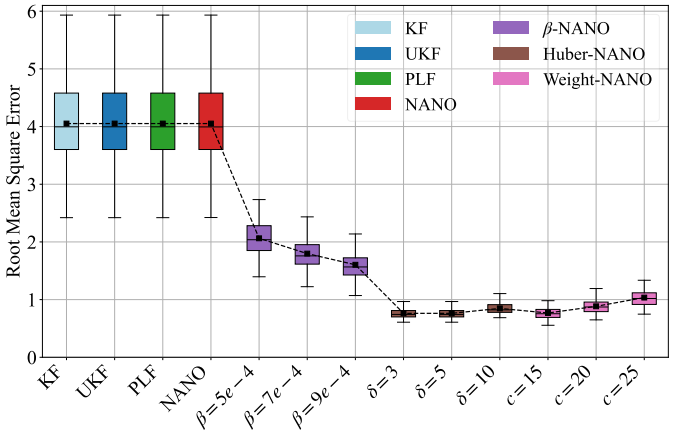

Fig. 1. Box plot of RMSE for KF, UKF, PLF, NANO filter, and its robust variants with various parameter values, for the Wiener velocity model with measurement outliers.

Notably, the Huber-NANO filter with $\delta = 3$ shows the best performance.

8.2 Nonlinear System Simulation: Lorenz System

Next, we perform simulations for a strongly nonlinear and chaotic dynamical system, namely a 6-dimensional coupled Lorenz system [58]. This system is constructed by stacking two Lorenz subsystems and introducing nearest-neighbor coupling. The full system state is defined as

$$x_t = \left(p_{x,t}^{(1)}, p_{y,t}^{(1)}, p_{z,t}^{(1)}, p_{x,t}^{(2)}, p_{y,t}^{(2)}, p_{z,t}^{(2)}\right)^\top \in \mathbb{R}^6.$$

Each subsystem evolves according to the classical Lorenz equations, while the x-component of the second subsystem receives an additional coupling term from the first:

$$\dot{p}^{(1)} = \begin{bmatrix} \sigma(p_y^{(1)} - p_x^{(1)}) \\ p_x^{(1)}(\rho - p_z^{(1)}) - p_y^{(1)} \\ p_x^{(1)} p_y^{(1)} - \beta p_z^{(1)} \end{bmatrix},$$
$$\dot{p}^{(2)} = \begin{bmatrix} \sigma(p_y^{(2)} - p_x^{(2)}) + \tilde{\sigma}(p_x^{(1)} - p_x^{(2)}) \\ p_x^{(2)}(\rho - p_z^{(2)}) - p_y^{(2)} \\ p_x^{(2)} p_y^{(2)} - \beta p_z^{(2)} \end{bmatrix}.$$

Here, $\sigma = 10$ (Prandtl number), $\rho = 28$ (Rayleigh number), and $\beta = \frac{8}{3}$ are standard parameters that yield chaotic motion in the Lorenz attractor. The additional coupling coefficient $\tilde{\sigma} = 5$ controls the strength of interaction between the two subsystems.

We discretize the dynamics using the Euler method,

$$x_{t+1} = x_t + \Delta t\, f_{\text{con}}(x_t) + \xi_t,\ \xi_t \sim \mathcal{N}(0, \mathbb{I}_{6\times 6}),$$

where f_{con} denotes the continuous-time drift function and $\Delta t = 0.01$ is the sampling period.

To further increase the nonlinear complexity of the estimation problem, we adopt a blockwise nonlinear measurement model,

$$y_t = 10 \begin{bmatrix} \frac{\sin(0.5\, p_{x,t})}{\tanh^2(p_{y,t})+2} \\ \frac{\tanh(p_{y,t})}{\exp(-0.2\, p_{z,t})+3} \\ \frac{\exp(-0.1\, p_{z,t})}{\sin^2(0.5\, p_{x,t})+5} \end{bmatrix}_{i=1,2} + \zeta_t,\ \zeta_t \sim \mathcal{N}(0, 4 \cdot \mathbb{I}_{6\times 6}).$$

Note that the measurement function contains multiple compositions of $\sin(\cdot)$, $\tanh(\cdot)$ and $\exp(\cdot)$, leading to highly nonlinear and non-monotonic observation channels that are far from the "nearly linear" assumption required in our stability analysis. This setting is therefore a stringent test for all Gaussian filters considered in this paper.

To validate the effectiveness of the NANO filter, we compare it with KF, EKF, UKF, IEKF, and PLF baselines. Besides, we also add two advanced filters, namely the iterated error state Kalman filter (IESKF) and stochastic search Kalman filter (SKF) [51]. The former is an error-state formulation of IEKF that is widely used in robotics and inertial odometry; it linearizes the error dynamics in the tangent space and iteratively refines the update [59]. The latter is a gradient-based filter that uses standard gradient descent to minimize a surrogate cost function. The original SKF algorithm relies on Monte Carlo sampling to estimate the gradient, which leads to significant computational burden and unstable gradient computation, and makes direct application impractical. To enable a fair comparison, we adopt the core gradient-based update structure of SKF while replacing the Monte Carlo gradient approximation with a more tractable deterministic alternative, namely the unscented transform.

We run 200 Monte Carlo simulations with randomized initial conditions. In each run, the same ground-truth trajectory is used for all filters, and the root mean square error (RMSE) over the full trajectory is recorded. Fig 2 summarizes the distribution of RMSE via box plots. As shown in Fig. 2, the strong nonlinearity and chaotic behavior of the coupled Lorenz system lead to poor performance for all baseline filters. Interestingly, the iterated variants such as IEKF and PLF do not improve accuracy. In fact, their repeated linearization tends to amplify local modeling errors, resulting in even larger RMSE. The IESKF also performs poorly because the Lorenz dynamics do not admit a meaningful error-state structure, making the error-state formulation ineffective in this setting. The deterministic SKF variant provides some improvement over EKF but still suffers from large estimation errors as its gradient direction does not consider the structure of Gaussian parameter space. In contrast, the proposed NANO filter achieves significantly lowest RMSE and a much tighter distribution, demonstrating its effectiveness under this challenging chaotic system.

Sensitivity Analysis: The NANO filter performs an iterative natural-gradient update in the update step, and its performance depends on both the number of inner iterations and the initialization strategy. To examine this sensitivity, we conduct an ablation study on the 6D Lorenz system by varying the number of natural-gradient iterations from 0 to 7. The iteration "0" corresponds to running the NANO filter without any natural-gradient iteration, i.e., the initialization is directly used as the final estimate in each update step. We compare two initialization methods: (i) initialization from the prior distribution and (ii) initialization from the MAP estimate (27). The average RMSE values are reported in Table 2, which shows that our proposed MAP-based initialization yields consistently lower RMSE and achieves most of its improvement with only a single natural-gradient iteration. This suggests that, especially in higher-dimensional systems where computational cost is critical, using MAP initialization together with just one natural-gradient iteration provides an excellent balance between accuracy and efficiency. In contrast, prior-based initialization requires more iterations to approach its best performance and remains less accurate overall.

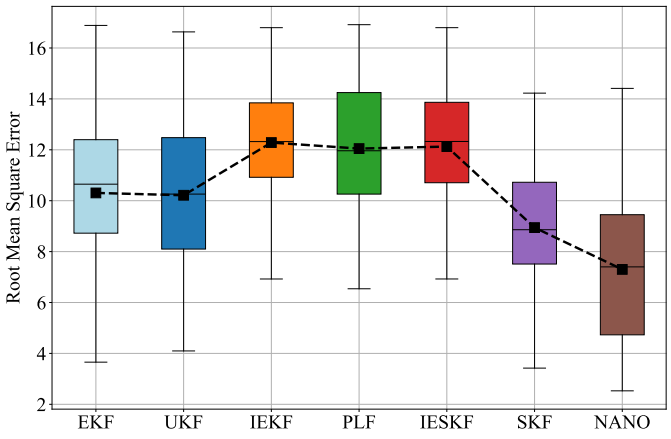

Fig. 2. Box plot of RMSE for EKF, UKF, IEKF, PLF, IESKF, SKF and NANO filter on the 6D coupled Lorenz system over 200 Monte Carlo runs. Note that the black square "■" represents the average RMSE over all the Monte Carlo experiments.

TABLE 2
Average RMSE for sensitivity analysis.

	Iterations							
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Prior	\	11.35	8.93	7.89	7.84	7.67	7.43	7.54
MAP	12.34	7.30	6.55	6.25	6.26	6.48	6.43	6.64

Scalability Analysis: To assess scalability, we further extend the coupled Lorenz system to 9D, 12D, and 15D by stacking additional subsystems with nearest-neighbor coupling and applying the same nonlinear measurement model to each block. For each dimension, 200 Monte Carlo trials are performed and the RMSE is computed for all filtering methods. The results in Figs. 9–11 (see Appendix H in the supplementary material) show that the baseline filters continue to produce relatively large estimation errors across all tested dimensions. Besides, IESKF remains a large RMSE because the Lorenz dynamics do not admit a meaningful error-state formulation. The deterministic SKF provides some improvement over EKF, but its overall accuracy remains noticeably below that of the proposed method. In contrast, the NANO filter consistently achieves the lowest RMSE and maintains one of the tightest error distributions in all cases examined. Even as the system dimension increases, NANO filter exhibits mild performance variation and retains a clear margin over all baselines, demonstrating practical scalability for higher-dimensional chaotic systems.

To further assess computational scalability, Table 3 reports the wall-clock time per natural-gradient iteration of the NANO filter. The computational cost increases moderately with the state dimension, rising from 1.65 ms at 6D to 8.05 ms at 15D. This growth is consistent with the increasing size of the Gaussian parameterization but remains well within practical limits. Notably, even at 15D the per-iteration time stays comfortably below the 10 ms integration interval used in the simulation, indicating that NANO filter

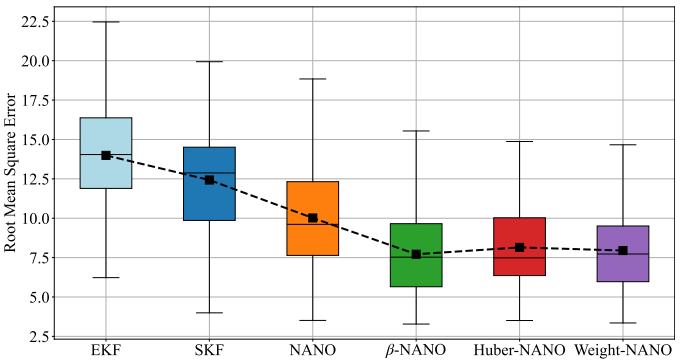

Fig. 3. Box plot of RMSE for EKF, SKF and NANO filter and its robust variants on the 6D coupled Lorenz system with contaminated measurements.

can be deployed in real-time nonlinear filtering settings for moderate-dimensional chaotic systems.

TABLE 3
Wall-clock time per iteration of the NANO filter.

Dimension	6	9	12	15
Time (ms)	1.65	3.26	5.43	8.05

Outlier Resistance: The Lorenz experiments above focus on a nominal Gaussian-noise setting in order to isolate the effects of strong nonlinearity and chaotic dynamics. Beyond this regime, the NANO filter framework naturally admits robust extensions by replacing the negative log-likelihood with general loss functions, yielding the Huber-NANO, Weight-NANO, and β-NANO variants. These methods can be interpreted as performing natural-gradient descent on a Gibbs posterior defined by a robust loss. To evaluate their robustness, we test these variants under contaminated measurements,

$$\zeta_t \sim (1-p_c)\,\mathcal{N}(0, 4\,\mathbb{I}_{6\times 6}) \;+\; p_c\,\mathcal{N}(0, 400\,\mathbb{I}_{6\times 6}), \qquad (43)$$

corresponding to a $p_c = 5\%$ probability of severe outliers with a tenfold increase in standard deviation.

Figure 3 reports the RMSE under this contaminated-measurement setting. UKF, IEKF, and IESKF diverge and are therefore omitted from the figure. Among the remaining methods, EKF, SKF, and the standard NANO filter remain numerically stable but suffer from noticeably enlarged estimation error. In contrast, the three robust NANO variants achieve substantially lower RMSE, with clear improvements over the standard NANO filter. These results demonstrate the effectiveness of incorporating robust loss functions into the NANO filter framework for mitigating the impact of heavy-tailed measurement outliers.

8.3 Real-world Experiment: Unmanned Ground Vehicle Localization

We conducted a real-world experiment to demonstrate the effectiveness of our method using an unmanned ground vehicle (UGV) equipped with a Lidar sensor for environmental perception, as shown in Fig. 4. The vehicle's state is represented by $x = \begin{bmatrix} p_x & p_y & \theta \end{bmatrix}^\top$, which includes its 2D position (p_x and p_y) and orientation angle (θ). The longitudinal velocity v and yaw rate ω serve as control inputs [60]. The Lidar has a 240-degree detection range with 0.33-degree resolution. The UGV was manually controlled throughout the experiment.

Fig. 4. The UGV and the experiment field. The three red traffic cones serve as landmarks for positioning.

The vehicle kinematics are modeled as:

$$\begin{bmatrix} p_{x,t+1} \\ p_{y,t+1} \\ \theta_{t+1} \end{bmatrix} = \begin{bmatrix} p_{x,t} \\ p_{y,t} \\ \theta_t \end{bmatrix} + \begin{bmatrix} v_t \cdot \cos\theta_t \\ v_t \cdot \sin\theta_t \\ \omega_t \end{bmatrix} \cdot \Delta t + \xi_t,$$

where $\Delta t = 0.0667$ is the sample period, v_t is the longitudinal velocity, ω_t is the yaw rate, and ξ_t is the process noise. The measurement model is given by:

$$y_t = [d_t^1\ d_t^2\ d_t^3\ \alpha_t^1\ \alpha_t^2\ \alpha_t^3]^\top + \zeta_t,$$

where d^i and α^i $(i = 1, 2, 3)$ denote the relative distance and orientation angle between the UGV and each traffic cone:

$$\begin{aligned} d^i &= \|(p_x^{\mathrm{tc},i} - p_x - l\cos\theta, p_y^{\mathrm{tc},i} - p_y - l\sin\theta)\|_2, \\ \alpha^i &= \arctan\left(\frac{p_y^{\mathrm{tc},i} - p_y - l\sin\theta}{p_x^{\mathrm{tc},i} - p_x - l\cos\theta}\right) - \theta. \end{aligned} \qquad (44)$$

Here $(p_x^{\mathrm{tc},i}, p_y^{\mathrm{tc},i})$ is the position of the i-th traffic cone, and l represents the longitudinal installation offset of the Lidar with respect to the robot center, as depicted in Fig. 5.

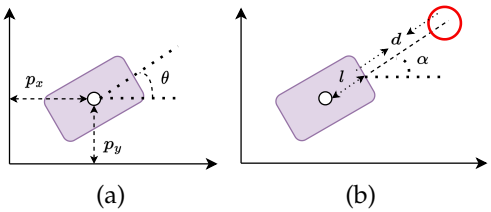

Fig. 5. (a) The diagram of the vehicle's states containing the 2D position of the vehicle p_x and p_y, and the orientation angle θ. (b) The diagram of the measurement model. Note that the red circle represents the red traffic cone.

The experiment begin by collecting raw data, including ground truth states from the high-precision motion capture system, control inputs, and Lidar point clouds. The Lidar data are processed to extract relative distances and angles to the traffic cones, resulting in a dataset comprising trajectories totaling 11 minutes. The noises are assumed to be Gaussian, whose parameters are estimated from the ground truth and measurement data using maximum likelihood estimation. The dataset is then segmented into 50 sub-trajectories, each lasting 6.67 seconds (100 time steps). Figures 6 and 7 compare the RMSE performance of the EKF,

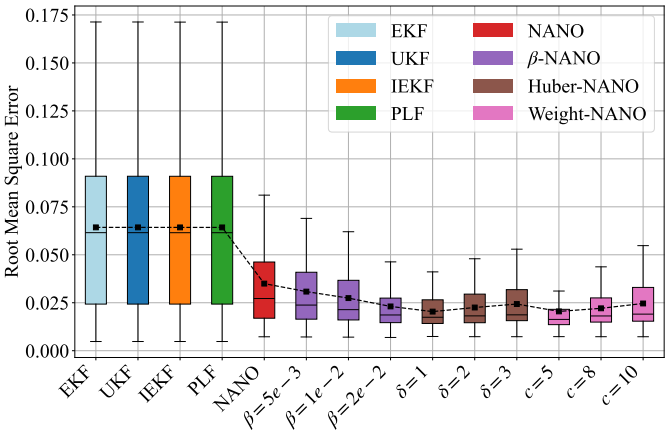

Fig. 6. Box plot of RMSE for KF, UKF, PLF, NANO and β-NANO, Huber-NANO, Weight-NANO with different values of the corresponding parameters.

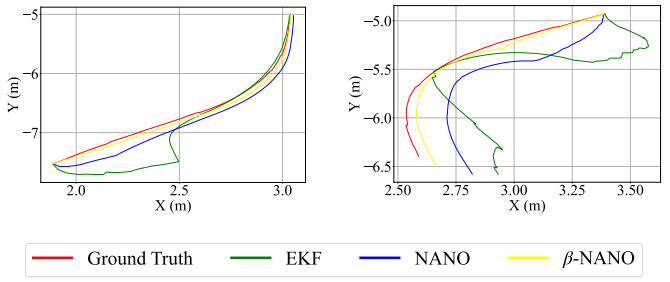

Fig. 7. Two samples of the real trajectories and their estimates.

UKF, IEKF, PLF, and NANO filters, including their robust variants. The results indicate that the NANO filter consistently outperforms traditional Gaussian filters, achieving higher accuracy. Furthermore, the robust NANO variants demonstrate improved performance, demonstrating their effectiveness in handling measurement outliers typically encountered in real-world conditions.

9 Conclusion

In this paper, we address the estimation errors commonly introduced by linearization techniques in Gaussian filters like the EKF and UKF for nonlinear systems. We reformulate the prediction and update steps of Gaussian filtering as optimization problems. While the prediction step mirrors moment-matching filters by calculating the first two moments of the prior distribution, the update step poses more challenges due to its highly nonlinear nature. To overcome these issues, we propose an iterative approach called the NANO filter, which avoids linearization by directly optimizing the update step using the natural gradient derived from the Fisher information matrix. This allows us to account for the curvature of the parameter space and ensures more accurate updates. We prove that the NANO filter converges locally to the optimal Gaussian approximation at each time step and show that it provides exponential error bounds for nearly linear measurement models with low noise. Experimental results demonstrate that the NANO filter outperforms popular filters like EKF, UKF, and others, while maintaining comparable computational efficiency.

References

[1] Z. Chen *et al.*, "Bayesian filtering: From Kalman filters to particle filters, and beyond," *Statistics*, vol. 182, no. 1, pp. 1–69, 2003.

[2] R. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[3] P. Del Moral, "Nonlinear filtering: Interacting particle resolution," *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, vol. 325, no. 6, pp. 653–658, 1997.

[4] J. S. Liu and R. Chen, "Sequential monte carlo methods for dynamic systems," *Journal of the American statistical association*, vol. 93, no. 443, pp. 1032–1044, 1998.

[5] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[6] G. L. Smith, S. F. Schmidt, and L. A. McGee, *Application of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle*. National Aeronautics and Space Administration, 1962, vol. 135.

[7] B. A. McElhoe, "An assessment of the navigation and course corrections for a manned flyby of mars or venus," *IEEE Transactions on Aerospace and Electronic Systems*, no. 4, pp. 613–623, 1966.

[8] A. Gelb *et al.*, *Applied optimal estimation*. MIT press, 1974.

[9] B. M. Bell and F. W. Cathey, "The iterated Kalman filter update as a Gauss-newton method," *IEEE Transactions on Automatic Control*, vol. 38, no. 2, pp. 294–297, 1993.

[10] S. Särkkä and L. Svensson, *Bayesian filtering and smoothing*. Cambridge university press, 2023, vol. 17.

[11] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A new approach for filtering nonlinear systems," in *Proceedings of 1995 American Control Conference-ACC'95*, vol. 3. IEEE, 1995, pp. 1628–1632.

[12] I. Arasaratnam, S. Haykin, and R. J. Elliott, "Discrete-time nonlinear filtering algorithms using Gauss–hermite quadrature," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 953–977, 2007.

[13] I. Arasaratnam and S. Haykin, "Cubature Kalman filters," *IEEE Transactions on automatic control*, vol. 54, no. 6, pp. 1254–1269, 2009.

[14] Á. F. García-Fernández, L. Svensson, M. R. Morelande, and S. Särkkä, "Posterior linearization filter: Principles and implementation using sigma points," *IEEE transactions on signal processing*, vol. 63, no. 20, pp. 5561–5573, 2015.

[15] J. Knoblauch, J. Jewson, and T. Damoulas, "Generalized variational inference: Three arguments for deriving new posteriors," *arXiv preprint arXiv:1904.02063*, 2019.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[18] G. Bonnet, "Transformations des signaux aléatoires a travers les systemes non linéaires sans mémoire," in *Annales des Télécommunications*, vol. 19. Springer, 1964, pp. 203–220.

[19] R. Price, "A useful theorem for nonlinear devices having Gaussian inputs," *IRE Transactions on Information Theory*, vol. 4, no. 2, pp. 69–72, 1958.

[20] W. Lin, M. E. Khan, and M. Schmidt, "Stein's lemma for the reparameterization trick with exponential family mixtures," *arXiv preprint arXiv:1910.13398*, 2019.

[21] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural computation*, vol. 10, no. 2, pp. 251–276, 1998.

[22] J. Martens, "New insights and perspectives on the natural gradient method," *Journal of Machine Learning Research*, vol. 21, no. 146, pp. 1–76, 2020.

[23] D. C. Fraser, "A new technique for the optimal smoothing of data," Ph.D. dissertation, Massachusetts Institute of Technology, 1967.

[24] B. D. Anderson and J. B. Moore, *Optimal filtering*. Courier Corporation, 2005.

[25] T. D. Barfoot, "Multivariate Gaussian variational inference by natural gradient descent," *arXiv preprint arXiv:2001.10025*, 2020.

[26] G. H. Golub and J. H. Welsch, "Calculation of Gauss quadrature rules," *Mathematics of computation*, vol. 23, no. 106, pp. 221–230, 1969.

[27] R. E. Kass, L. Tierney, and J. B. Kadane, "Laplace's method in Bayesian analysis," *Contemporary Mathematics*, vol. 115, pp. 89–99, 1991.

[28] H. Salimbeni, S. Eleftheriadis, and J. Hensman, "Natural gradients in practice: Non-conjugate variational inference in Gaussian process models," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 689–697.

[29] T. Glasmachers, T. Schaul, S. Yi, D. Wierstra, and J. Schmidhuber, "Exponential natural evolution strategies," in *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, 2010, pp. 393–400.

[30] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[31] A. Einstein, *The meaning of relativity*. Routledge, 2003.

[32] K. Reif, S. Gunther, E. Yaz, and R. Unbehauen, "Stochastic stability of the discrete-time extended Kalman filter," *IEEE Transactions on Automatic control*, vol. 44, no. 4, pp. 714–728, 1999.

[33] B. Xu, P. Zhang, H. Wen, and X. Wu, "Stochastic stability and performance analysis of cubature Kalman filter," *Neurocomputing*, vol. 186, pp. 218–227, 2016.

[34] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[35] W. Cao, C. Liu, Z. Lan, Y. Piao, and S. E. Li, "Generalized moving horizon estimation for nonlinear systems with robustness to measurement outliers," in *2023 American Control Conference (ACC)*. IEEE, 2023, pp. 1614–1621.

[36] A. Boustati, O. D. Akyildiz, T. Damoulas, and A. Johansen, "Generalised Bayesian filtering via sequential monte carlo," *Advances in neural information processing systems*, vol. 33, pp. 418–429, 2020.

[37] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[38] G. Duran-Martin, M. Altamirano, A. Shestopaloff, L. Sánchez-Betancourt, J. Knoblauch, M. Jones, F.-X. Briol, and K. P. Murphy, "Outlier-robust Kalman filtering through generalised Bayes," vol. 235, pp. 12 138–12 171, 21–27 Jul 2024. [Online]. Available: https://proceedings.mlr.press/v235/duran-martin24a.html

[39] W. Cao, S. Liu, C. Liu, Z. He, S. S.-T. Yau, and S. E. Li, "Convolutional Bayesian filtering," *arXiv preprint arXiv:2404.00481*, 2024.

[40] P. J. Huber, "Robust statistics," in *International encyclopedia of statistical science*. Springer, 2011, pp. 1248–1251.

[41] M. A. Gandhi and L. Mili, "Robust Kalman filter based on a generalized maximum-likelihood-type estimator," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2509–2520, 2009.

[42] M. Opper and C. Archambeau, "The variational Gaussian approximation revisited," *Neural computation*, vol. 21, no. 3, pp. 786–792, 2009.

[43] T. D. Barfoot, J. R. Forbes, and D. J. Yoon, "Exactly sparse Gaussian variational inference with application to derivative-free batch nonlinear state estimation," *The International Journal of Robotics Research*, vol. 39, no. 13, pp. 1473–1502, 2020.

[44] T. D. Barfoot, *State estimation for robotics*. Cambridge University Press, 2024.

[45] M. E. Khan and D. Nielsen, "Fast yet simple natural-gradient descent for variational inference in complex models," in *2018 International Symposium on Information Theory and Its Applications (ISITA)*. IEEE, 2018, pp. 31–35.

[46] M. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava, "Fast and scalable Bayesian deep learning by weight-perturbation in adam," in *International conference on machine learning*. PMLR, 2018, pp. 2611–2620.

[47] M. Khan and W. Lin, "Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 878–887.

[48] M. E. Khan, W. Lin, V. Tangkaratt, Z. Liu, and D. Nielsen, "The variational adaptive-newton method," in *NeurIPS Workshop on Advances in Approximate Bayesian Inference*, 2017.

[49] S.-i. Amari, *Differential-geometrical methods in statistics*. Springer Science & Business Media, 2012, vol. 28.

[50] C. Lenglet, M. Rousson, R. Deriche, and O. Faugeras, "Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor mri processing," *Journal of Mathematical Imaging and Vision*, vol. 25, pp. 423–444, 2006.

[51] S. Gultekin and J. Paisley, "Nonlinear Kalman filtering with divergence minimization," *IEEE Transactions on Signal Processing*, vol. 65, no. 23, pp. 6319–6331, 2017.

[52] S. Hu, L. Guo, and J. Zhou, "An iterative nonlinear filter based on posterior distribution approximation via penalized kullback–leibler divergence minimization," *IEEE Signal Processing Letters*, vol. 29, pp. 1137–1141, 2022.

[53] L. Guo, S. Hu, J. Zhou, and X. R. Li, "Gaussian approximation filter based on divergence minimization for nonlinear dynamic systems," in *2022 25th International Conference on Information Fusion (FUSION)*. IEEE, 2022, pp. 1–7.

[54] L. Guo, S. Hu, J. Zhou, and X. Rong Li, "Recursive nonlinear filtering via Gaussian approximation with minimized kullback–leibler divergence," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 1, pp. 965–979, 2024.

[55] H. Yumei, W. Xuezhi, P. Quan, H. Zhentao, and B. Moran, "Variational bayesian Kalman filter using natural gradient," *Chinese Journal of Aeronautics*, vol. 35, no. 5, pp. 1–10, 2022.

[56] Y. Li, Y. Cheng, X. Li, H. Wang, X. Hua, and Y. Qin, "Bayesian nonlinear filtering via information geometric optimization," *Entropy*, vol. 19, no. 12, p. 655, 2017.

[57] Y. Ollivier, "Online natural gradient as a Kalman filter," *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 2930 – 2961, 2018. [Online]. Available: https://doi.org/10.1214/18-EJS1468

[58] S. Boccaletti, J. Kurths, G. Osipov, D. Valladares, and C. Zhou, "The synchronization of chaotic systems," *Physics reports*, vol. 366, no. 1-2, pp. 1–101, 2002.

[59] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "Fast-lio2: Fast direct lidar-inertial odometry," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.

[60] J. Elfring, E. Torta, and R. van de Molengraft, "Particle filters: A hands-on tutorial," *Sensors*, vol. 21, no. 2, p. 438, 2021.

[61] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM." https://github.com/MichaelGrupp/evo, 2017.

Wenhan Cao received his B.E. degree in the School of Electrical Engineering from Beijing Jiaotong University, Beijing, China, in 2019.

He is currently a Ph.D. candidate in the School of Vehicle and Mobility, Tsinghua University, Beijing, China. His research interests include optimal filtering and reinforcement learning. He was a finalist for the Best Student Paper Award at the 2021 IFAC MECC.

Tianyi Zhang received his B.E. degree in the school of Automation Science and Electrical Engineering from Beihang University, Beijing, China, in 2024. He is currently a Ph.D. candidate in the School of Vehicle and Mobility, Tsinghua University, Beijing, China. His research interests include optimal state estimation, Bayesian inference, and reinforcement learning.

Zeju Sun (Member, IEEE) received the B.S. degree in Mathematics and Applied Mathematics, and the Ph.D. degree in Mathematics from Department of Mathematical Sciences, Tsinghua University, Beijing, China, in 2020 and 2025.

He is currently an Assistant Professor with Beijing Institute of Mathematical Sciences and Applications (BIMSA). His research interests include control theory, nonlinear filtering and artificial intelligence.

Chang Liu (Member, IEEE) received the B.S. degrees in Electronic Information Science and Technology and in Mathematics and Applied Mathematics (double degree) from the Peking University, China, in 2011, and the M.S. degrees in Mechanical Engineering and in Computer Science, and the Ph.D. degree in Mechanical Engineering from the University of California, Berkeley, USA, in 2014, 2015, and 2017, respectively. He is currently an Assistant Professor with the Department of Advanced Manufacturing and Robotics, College of Engineering, Peking University. From 2017 to 2020, he was a Postdoctoral Associate with the Cornell University, USA. He has also worked for Ford Motor Company and NVIDIA Corporation on autonomous vehicles. His research interests include robot motion planning, active sensing, and multi-robot collaboration.

Stephen S.-T. Yau (Life Fellow, IEEE) received the Ph.D. degree in mathematics from the State University of New York at Stony Brook, NY, USA in 1976.

He was a Member of the Institute of Advanced Study at Princeton from 1976-1977 and 1981-1982, and a Benjamin Pierce Assistant Professor at Harvard University during 1977-1980. After that, he joined the Department of Mathematics, Statistics and Computer Science (MSCS), University of Illinois at Chicago (UIC), and served for over 30 years. During 2005-2011, he became a joint Professor with the Department of Electrical and Computer Engineering at the MSCS, UIC. After his retirement in 2012, he joined Tsinghua University, Beijing, China, where he is a full-time professor in the Department of Mathematical Sciences. His research interests include nonlinear filtering, bioinformatics, complex algebraic geometry, CR geometry and singularity theory.

Dr. Yau is the Managing Editor and founder of the *Journal of Algebraic Geometry* since 1991, and the Editor-in-Chief and founder of *Communications in Information and Systems* from 2000 to the present. He was the General Chairman of the IEEE International Conference on Control and Information, which was held at the Chinese University of Hong Kong in 1995. He was awarded the Sloan Fellowship in 1980, the Guggenheim Fellowship in 2000, and the AMS Fellow Award in 2013. In 2005, he was entitled the UIC Distinguished Professor.

Shengbo Eben Li (Senior Member, IEEE) received his M.S. and Ph.D. degrees from Tsinghua University in 2006 and 2009. He has worked at Stanford University, University of Michigan, and UC Berkeley. He is now a professor at Tsinghua University, working on intelligent vehicles and driver assistance, embodied intelligence for robotics, deep reinforcement learning, optimal control and estimation, etc. He is the author of over 250 peer-reviewed journal/conference papers, and co-inventor of over 40 patents.

Dr. Li has received over 20 prestigious awards, including Youth Sci. & Tech Award of Ministry of Education (annually 10 receivers in China), Natural Science Award of Chinese Association of Automation (First level), National Award for Progress in Sci & Tech of China, and best (student) paper awards or finalists of IEEE ITSC, IEEE IVS, IET ITS, ICCAS, IFAC MECC, CAA CVCI, IEEE ICUS, CCCC, IEEE ITSM, L4DC, Automotive Innovation, etc. He was a member of Board Governor of IEEE ITS Society. He serves as the director of Technical Committee on AI of SAE-China, deputy director of Technical Committee on Vehicle Control and Intelligence of CAA, and the leader of AI working group in China Industry Innovation Alliance for ICVs. He also serves as Senior AE of IEEE OJ ITS, AEs of IEEE ITSM, IEEE TITS, IEEE TNNLS, IEEE TCST, and area chairs of ICLR and ICML, etc.

Supplementary Material

Appendix A
Proof of Proposition 1

Proof. By recognizing that the focus is solely on the extremizer and not on the objective value itself, it follows that for any constant $Z > 0$, the right-hand side of (5) can be rewritten as

$$\arg\min_{q(x_t)} \left\{ \mathbb{E}_{\int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\,\mathrm{d}x_{t-1}} \left\{ \log \frac{Z}{q(x_t)} \right\} \right\}. \tag{45}$$

If we choose $Z = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\,\mathrm{d}x_{t-1}$ and recall that the KL divergence reaches its minimum value uniquely when its arguments are identical, we find that the solution to (45) is indeed

$$\int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})\,\mathrm{d}x_{t-1},$$

which corresponds exactly to the prior distribution defined in (3a). Similarly, the right-hand side of (6) is equal to

$$\arg\min_{q(x_t)} \left\{ \mathbb{E}_{q(x_t)} \left\{ \log \frac{q(x_t)}{p(y_t|x_t)p(x_t|y_{1:t-1})/Z} \right\} \right\}. \tag{46}$$

If we choose $Z = \int p(y_t|x_t)p(x_t|y_{1:t-1})\,\mathrm{d}x_t$, the solution to (46) is

$$\frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{\int p(y_t|x_t)p(x_t|y_{1:t-1})\,\mathrm{d}x_t},$$

which is precisely the posterior distribution defined in (3b). □

Appendix B
Proof of Lemma 1

Proof. The log-likelihood function of a Gaussian distribution $\mathcal{N}(x;\mu,\Sigma)$ is given by

$$\log \mathcal{N}(x;\mu,\Sigma) = -\frac{1}{2}\big[\log|\Sigma| + (x-\mu)^\top\Sigma^{-1}(x-\mu) + \log(2\pi)^n\big],$$

where n is the dimensionality of x. Taking the expectation with respect to the probability density function $p(x)$, we obtain

$$\mathbb{E}_{p(x)}\{\log \mathcal{N}(x;\mu,\Sigma)\} = -\frac{1}{2}\big[\log|(2\pi)^n\Sigma| + \mathbb{E}_{p(x)}\left\{(x-\mu)^\top\Sigma^{-1}(x-\mu)\right\}\big].$$

We first compute the gradient of this expectation with respect to μ. The relevant term involving μ is the quadratic form $(x-\mu)^\top\Sigma^{-1}(x-\mu)$. Differentiating this term with respect to μ gives

$$\frac{\partial}{\partial\mu}\mathbb{E}_{p(x)}\left\{(x-\mu)^\top\Sigma^{-1}(x-\mu)\right\} = -2\Sigma^{-1}\mathbb{E}_{p(x)}\{x-\mu\}.$$

Setting this derivative equal to zero, we obtain (9a). Next, we compute the gradient with respect to Σ. The relevant terms involving Σ are the log-determinant $\log|\Sigma|$ and the quadratic form $(x-\mu)^\top\Sigma^{-1}(x-\mu)$. The gradient of the log-determinant with respect to Σ is $\frac{\partial}{\partial\Sigma}\log|\Sigma| = \Sigma^{-1}$. The gradient of the quadratic form is

$$\begin{aligned}&\frac{\partial}{\partial\Sigma}\mathbb{E}_{p(x)}\left\{(x-\mu)^\top\Sigma^{-1}(x-\mu)\right\}\\ =&-\Sigma^{-1}\mathbb{E}_{p(x)}\left\{(x-\mu)(x-\mu)^\top\right\}\Sigma^{-1}.\end{aligned}$$

Combining these results, we have

$$\begin{aligned}&\frac{\partial}{\partial\Sigma}\mathbb{E}_{p(x)}\{\log\mathcal{N}(x;\mu,\Sigma)\}\\ =&-\frac{1}{2}\left[\Sigma^{-1}-\Sigma^{-1}\mathbb{E}_{p(x)}\left\{(x-\mu)(x-\mu)^\top\right\}\Sigma^{-1}\right].\end{aligned}$$

Setting this derivative equal to zero gives

$$\Sigma^{-1} = \Sigma^{-1}\mathbb{E}_{p(x)}\left\{(x-\mu)(x-\mu)^\top\right\}\Sigma^{-1},$$

which implies (9b). □

Appendix C
Proof of Corollary 1

Proof. For linear Gaussian systems with output probability $p(y_t|x_t) = \mathcal{N}(y_t; Cx_t, R)$, the one-step iteration of the covariance matrix in (24) is given by:

$$\begin{aligned}\left(P_t^{-1}\right)^{(1)} =&P_{t|t-1}^{-1} + \mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(0)},P_t^{(0)})}\left\{\frac{\partial^2\ell(x_t,y_t)}{\partial x_t^2}\right\},\\ =&P_{t|t-1}^{-1} + C^\top R^{-1}C.\end{aligned} \tag{47}$$

Comparing (47) with KF covariance update equation (19), we observe that $P_t^{(1)} = P_{t|t}$. Then, refering to (17) and using (47), the one-step iteration of the mean vector in (24) can be expressed as:

$$\begin{aligned}\hat{x}_t^{(1)} =&\hat{x}_t^{(0)} - P_t^{(1)}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(0)},P_t^{(0)})}\left\{\frac{\partial\ell(x_t,y_t)}{\partial x_t}\right\}\\ &- P_t^{(1)}P_{t|t-1}^{-1}\left(\hat{x}_t^{(0)} - \hat{x}_{t|t-1}\right)\\ =&\hat{x}_t^{(0)} + P_{t|t}C^\top R^{-1}(y_t - C\hat{x}_t^{(0)})\\ &- P_{t|t}\left(P_{t|t}^{-1} - C^\top R^{-1}C\right)\left(\hat{x}_t^{(0)} - \hat{x}_{t|t-1}\right)\\ =&\hat{x}_{t|t-1} + P_{t|t}C^\top R^{-1}\left(y_t - C\hat{x}_{t|t-1}\right).\end{aligned} \tag{48}$$

Furthermore, noting that:

$$\begin{aligned}&P_{t|t}C^\top R^{-1}\\ =&P_{t|t-1}C^\top R^{-1}\\ &- P_{t|t-1}C^\top(R + CP_{t|t-1}C^\top)^{-1}CP_{t|t-1}C^\top R^{-1}\\ =&P_{t|t-1}C^\top\big(I - (R + CP_{t|t-1}C^\top)^{-1}(CP_{t|t-1}C^\top + R)\\ &+ (R + CP_{t|t-1}C^\top)^{-1}R\big)R^{-1}\\ =&P_{t|t-1}C^\top(R + CP_{t|t-1}C^\top)^{-1},\end{aligned}$$

we recognize that $P_{t|t}C^\top R^{-1}$ is the Kalman gain K_t. Therefore, (48) becomes:

$$\hat{x}_t^{(1)} = \hat{x}_{t|t-1} + K_t\left(y_t - C\hat{x}_{t|t-1}\right),$$

which is the standard KF mean update equation. Thus, the one-step iteration of the NANO filter in (24) yields the optimal estimate, regardless of the initialization $\hat{x}_t^{(0)}$ and $P_t^{(0)}$. □

APPENDIX D
PROOF OF THEOREM 1

Proof. This proof is inspired by Section 6.2.2 in [44]. According to (14) and (24), the second-order derivative of $J(\hat{x}_t, P_t)$ with respect to $\hat{x}_t$ satisfies

$$\begin{aligned}\left.\frac{\partial^2 J}{\partial \hat{x}_t^2}\right|_{v^{(i)}} =& \left.\frac{\partial^2}{\partial \hat{x}_t^2}\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t,P_t)}\left\{\ell(x_t,y_t)\right\}\right|_{v^{(i)}} + P_{t|t-1}^{-1}\\ =&\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_t^{(i)},P_t^{(i)})}\left\{\frac{\partial^2\ell(x_t,y_t)}{\partial x_t^2}\right\} + P_{t|t-1}^{-1}\\ =&\left(P_t^{-1}\right)^{(i+1)}.\end{aligned} \tag{49}$$

According to (23), we have

$$\begin{aligned}\left.\frac{\partial J}{\partial \hat{x}_t}\right|_{v^{(i)}} =& -\left(P_t^{-1}\right)^{(i+1)}\delta\hat{x}_t,\\ \left.\frac{\partial J}{\partial P_t^{-1}}\right|_{v^{(i)}} =& -\frac{1}{2}\left(P_t\right)^{(i)}\delta P_t^{-1}\left(P_t\right)^{(i)}.\end{aligned} \tag{50}$$

Combined with (49) and (50), the Tayor-series expansion can be expressed as

$$\begin{aligned}&J_t^{(i+1)} - J_t^{(i)}\\ \approx& -\frac{1}{2}\delta\hat{x}_t^\top\left(P_t^{-1}\right)^{(i+1)}\delta\hat{x}_t - \frac{1}{2}\mathrm{Tr}\left(P_t^{(i)}\delta P_t^{-1}P_t^{(i)}\delta P_t^{-1}\right)\\ =& -\frac{1}{2}\delta\hat{x}_t^\top\left(P_t^{-1}\right)^{(i+1)}\delta\hat{x}_t\\ &-\frac{1}{2}\mathrm{vec}(\delta P_t^{-1})^\top\left(P_t^{(i)}\otimes P_t^{(i)}\right)\mathrm{vec}(\delta P_t^{-1})\\ \leq& 0.\end{aligned}$$

□

APPENDIX E
PROOF OF THEOREM 2

Proof. Utilizing (36) and (37), $e_{t|t}^\top\tilde{P}_{t|t}^{-1}e_{t|t}$ can be computed as follows:

$$\begin{aligned}&e_{t|t}^\top\tilde{P}_{t|t}^{-1}e_{t|t}\\ =&e_{t-1|t-1}^\top F_{t-1}^\top(I+H_t)^{-1}\\ &\times\left[\left(F_{t-1}\tilde{P}_{t-1|t-1}F_{t-1}^\top+Q_t\right)^{-1}+D_t\right]\\ &\times(I+H_t)^{-1}F_{t-1}e_{t-1|t-1}\\ &-2e_{t-1|t-1}^\top F_{t-1}^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\bar{h}_t\\ &+\bar{h}_t^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\bar{h}_t\\ &+\Psi_t^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\Psi_t\\ &+\left(\xi_t-G_t\zeta_t\right)^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\left(\xi_t-G_t\zeta_t\right)\\ &+2e_{t-1|t-1}F_{t-1}^\top(I+H_t)^{-1}\\ &\times\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\left(\Psi_t+\xi_t-G_t\zeta_t\right)\\ &-2\bar{h}_t^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\left(\Psi_t+\xi_t-G_t\zeta_t\right)\\ &+2\Psi_t^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\left(\xi_t-G_t\zeta_t\right).\end{aligned} \tag{51}$$

For the first term on the right-hand side of (51), with the fact that the measurement function $h(x)$ is almost linear, there exists $\lambda_1 > 0$ such that

$$\begin{aligned}&F_{t-1}^\top(I+H_t)^{-1}\left[\left(F_{t-1}\tilde{P}_{t-1|t-1}F_{t-1}^\top+Q_t\right)^{-1}+D_t\right]\\ &\times(I+H_t)^{-1}F_{t-1}\leq\frac{1}{1+\lambda_1}\tilde{P}_{t-1|t-1}.\end{aligned} \tag{52}$$

In fact, if $g(x) = Cx$ is a linear function with some constant matrix $C\in\mathbb{R}^{m\times n}$, then

$$\begin{aligned}H_t &= P_{t|t-1}C^\top R_t^{-1}C,\\ D_t &= C^\top R_t^{-1}C,\end{aligned}$$

and $\tilde{P}_{t|t} = P_{t|t}$, $\tilde{P}_{t|t-1} = P_{t|t-1}$. In this way, we have

$$\begin{aligned}&F_{t-1}^\top(I+H_t)^{-1}\left[\left(F_{t-1}\tilde{P}_{t-1|t-1}F_{t-1}^\top+Q_t\right)^{-1}+D_t\right]\\ &\times(I+H_t)^{-1}F_{t-1}\\ =&F_{t-1}^\top(I+P_{t|t-1}C^\top R_t^{-1}C)^{-1}P_{t|t-1}^{-1}F_{t-1}\\ =&F_{t-1}^\top(I+P_{t|t-1}C^\top R_t^{-1}C)^{-1}(F_{t-1}^\top P_{t-1|t-1}F_{t-1}^\top)^{-1}\\ &\times(I+(F_{t-1}P_{t-1|t-1}F_{t-1}^\top)^{-1}Q_t)^{-1}F_{t-1}.\end{aligned}$$

By observing that

$$\begin{aligned}(I+P_{t|t-1}C^\top R_t^{-1}C)^{-1}&\leq\frac{1}{1+\lambda}I,\\ (I+(F_{t-1}P_{t-1|t-1}F_{t-1}^\top)^{-1}Q_t)^{-1}&\leq\frac{1}{1+\lambda'}I,\end{aligned}$$

with $\lambda = \lambda_{\min}(P_{t|t-1}C^\top R_t^{-1}C) > 0$ and $\lambda' = \lambda_{\min}((F_{t-1}P_{t-1|t-1}F_{t-1}^\top)^{-1}Q_t) > 0$, therefore (52) holds for some $\lambda_1 > 0$ for linear systems. Because of the continuity of (52) with respect to the measurement function g, (52) also holds for those systems with almost linear measurement functions.

Other terms on the right-hand side of (51) can also be bounded. Firstly, there exists $\delta_1 > 0$, such that

$$\bar{h}_t^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\bar{h}_t\leq\delta_1. \tag{53}$$

Because $\bar{h}_t\equiv 0$ for linear measurement functions, the constant δ_1 above can be chosen to be sufficiently small for almost linear measurement functions.

Secondly, there exist $\delta_2 > 0$, such that

$$-2e_{t-1|t-1}^\top F_{t-1}^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}\bar{h}_t\leq\delta_2\|e_{t-1|t-1}\|. \tag{54}$$

Finally, because ξ_t, ζ_t are independent Gaussian random variables, there exists $\delta_3 > 0$, such that

$$\mathbb{E}\left\{(\xi_t-G_t\zeta_t)^\top(I+H_t)^{-1}\tilde{P}_{t|t}^{-1}(I+H_t)^{-1}(\xi_t-G_t\zeta_t)\right\}\leq\delta_3. \tag{55}$$

Other terms on the right-hand side of (51) are related to the high order terms Ψ_t. Therefore, combining (52), (53) to (55), there exists $\eta > 0$, such that if $\|e_{t|t}\| < \eta$ for all $t\geq 0$, then

$$\begin{aligned}&\mathbb{E}\left\{e_{t|t}^\top\tilde{P}_{t|t}^{-1}e_{t|t}\middle|e_{t-1|t-1}\right\}\\ \leq&\frac{1}{1+\lambda_0}e_{t-1|t-1}^\top\tilde{P}_{t-1|t-1}^{-1}e_{t-1|t-1}+\delta_2\cdot\mathbb{E}\|e_{t-1|t-1}\|+\delta,\end{aligned} \tag{56}$$

for some $0 < \lambda_0 < \lambda_1$.

With the fact that $\underline{p}I \le \tilde{P}_{t|t} \le \bar{p}I$, there exist a vector $\epsilon_1 \in \mathbb{R}^n$ and a constant $\epsilon_2 > 0$, such that

$$\begin{aligned}&\mathbb{E}\left\{ (e_{t|t}-\epsilon_1)^\top \tilde{P}_{t|t}^{-1}(e_{t|t}-\epsilon_1)\Big| e_{t-1|t-1}\right\}\\ \le&\frac{1}{1+\lambda_0}(e_{t-1|t-1}-\epsilon_1)^\top \tilde{P}_{t-1|t-1}^{-1}(e_{t-1|t-1}-\epsilon_1)+\epsilon_2,\end{aligned} \tag{57}$$

as long as $\|e_{t-1|t-1}\| < \eta$.

Also, if there exists $\tilde{\eta} > 0$, such that $\tilde{\eta} < \|e_{t-1|t-1}\| < \eta$, for ϵ_1 with sufficiently small norm and sufficiently small ϵ_2, we have the following supermartingale-like property:

$$\begin{aligned}&\mathbb{E}\left\{ (e_{t|t}-\epsilon_1)^\top \tilde{P}_{t|t}^{-1}(e_{t|t}-\epsilon_1)\Big| e_{t-1|t-1}\right\}\\ &-(e_{t-1|t-1}-\epsilon_1)^\top \tilde{P}_{t-1|t-1}^{-1}(e_{t-1|t-1}-\epsilon_1)\\ \le&-\frac{\lambda_0}{2(1+\lambda_0)\bar{p}}\tilde{\eta}^2+\epsilon_2<0.\end{aligned}$$

Hence, if the initial estimation error $\|e_{0|0}\| < \eta$, then we can recursively use (57) to compute the estimation error at time t,

$$\begin{aligned}&\mathbb{E}\left\{(e_{t|t}-\epsilon_1)^\top \tilde{P}_{t|t}^{-1}(e_{t|t}-\epsilon_1)\right\}\\ \le&\epsilon_2\sum_{k=0}^{t-1}\frac{1}{(1+\lambda_0)^k}+\frac{1}{\underline{p}}\|e_{0|0}-\epsilon_1\|^2\left(\frac{1}{1+\lambda_0}\right)^t\\ \le&\frac{1+\lambda_0}{\lambda_0}\epsilon_2+\frac{1}{\underline{p}}\|e_{0|0}-\epsilon_1\|^2\left(\frac{1}{1+\lambda_0}\right)^t.\end{aligned}$$

Since

$$\frac{1}{\bar{p}}\|e_{t|t}-\epsilon_1\|^2 \le (e_{t|t}-\epsilon_1)^\top \tilde{P}_{t|t}^{-1}(e_{t|t}-\epsilon_1),$$

and thus,

$$\|e_{t|t}\| \le \sqrt{\bar{p}(|e_{t|t}|-\epsilon_1)^\top \tilde{P}_{t|t}^{-1}(|e_{t|t}|-\epsilon_1)}+|\epsilon_1|,$$

and we obtained the exponentially bounded in mean square for $e_{t|t}$, i.e., there exist $\epsilon, \epsilon' > 0$, such that

$$\mathbb{E}\|e_{t|t}\|^2 \le \epsilon\|e_{0|0}\|^2\left(\frac{1}{1+\lambda_0}\right)^t+\epsilon',\ \forall\, t\ge 0.$$

□

APPENDIX F
ROBUSTNESS ANALYSIS OF THE WEIGHTED-LOSS NANO FILTER

We provide a initial robustness analysis for the weighted-loss variant of NANO by analyzing its posterior influence function in the sense of classical robust statistics [40].

Definition 1 (Posterior influence function)**.** *Let $q_t(\cdot|y_{1:t})$ denote the (generalised) posterior at time t and let $q_t^c(\cdot|y_{1:t-1}, y_c)$ be the posterior obtained when the measurement y_t is replaced by a contaminating measurement y_c. The posterior influence function (PIF) at time t is defined by*

$$\mathrm{PIF}_t(y_c; y_{1:t}) \triangleq D_{\mathrm{KL}}\big(q_t^c(\cdot|y_{1:t-1}, y_c)\|q_t(\cdot|y_{1:t})\big).$$

We say that the filter is robust at time t if $\sup_{\|y_t-y_c\|\to\infty}\mathrm{PIF}_t(y_c; y_{1:t}) < \infty$.

For later use we record the following elementary bound as a useful lemma.

Lemma 3. *For any $0 \le \alpha < k$, there exists a constant $C_{\alpha,k} > 0$ such that*

$$\frac{s^\alpha}{(1+s)^k} \le C_{\alpha,k},\ \forall s \ge 0.$$

Proof of the lemma. For $s \in [0,1]$, we have $s^\alpha/(1+s)^k \le 1$. For $s \ge 1$, $s^\alpha/(1+s)^k \le s^{\alpha-k} \le 1$ because $\alpha - k < 0$. Thus the supremum over $[0,\infty)$ is finite. □

We now show that the weighted-loss NANO update has a bounded PIF in the linear Gaussian case.

Theorem 3. *Consider the linear Gaussian measurement model*

$$y_t = Cx_t + \zeta_t,\ \zeta_t \sim \mathcal{N}(\zeta_t; 0, R_t),$$

and the weighted log-likelihood loss

$$\begin{aligned}\ell^w(x_t, y_t) &= -w(x_t, y_t)\log\mathcal{N}\big(y_t; Cx_t, R_t\big),\\ w(x_t, y_t) &= \Big(1+\big\|y_t - Cx_t\big\|^2_{R_t^{-1}}/c^2\Big)^{-1},\end{aligned}$$

with $c > 0$ and $R_t \succ 0$. Then the NANO update obtained by minimizing update cost $J(\hat{x}_t, P_t)$ with loss ℓ^w has a bounded posterior influence function, i.e.

$$\sup_{\|y_t-y_c\|\to\infty}\mathrm{PIF}_t(y_c; y_{1:t}) < \infty.$$

Proof. We first obtain explicit expressions for the gradient and Hessian of $\ell^w(x_t, y_t)$ with respect to x_t. Define the residual and the normalized squared innovation

$$r_t(x_t) \triangleq y_t - Cx_t,\ s_t(x_t) \triangleq \frac{1}{c^2}\, r_t(x_t)^\top R_t^{-1} r_t(x_t),$$

so that the weight reads

$$w(x_t, y_t) = \big(1+s_t(x_t)\big)^{-1}.$$

Because $R_t \succ 0$, there exist positive constants $\underline{\lambda}, \overline{\lambda}$ such that

$$\underline{\lambda}\,\|r_t(x_t)\|^2 \le r_t(x_t)^\top R_t^{-1} r_t(x_t) \le \overline{\lambda}\,\|r_t(x_t)\|^2,$$

hence $s_t(x_t)$ is equivalent to $\|r_t(x_t)\|^2$ up to constants.

Write the Gaussian log-likelihood as

$$\log\mathcal{N}(y_t; Cx_t, R_t) = -\frac{1}{2}r_t(x_t)^\top R_t^{-1} r_t(x_t) - \kappa_t,$$

where $\kappa_t \triangleq \frac{1}{2}\log\big((2\pi)^m|R_t|\big)$ is constant w.r.t. x_t. Since $r_t(x_t)^\top R_t^{-1} r_t(x_t) = c^2 s_t(x_t)$, we can write

$$\ell^w(x_t, y_t) = -w(x_t, y_t)\log\mathcal{N}(y_t; Cx_t, R_t) = w(s_t)\left(\frac{c^2}{2}s_t + \kappa_t\right),$$

where $w(s) = (1+s)^{-1}$ and $s_t = s_t(x_t)$.

Introduce the shorthand

$$B_t(x_t) \triangleq C^\top R_t^{-1} r_t(x_t),\ D_t \triangleq C^\top R_t^{-1} C.$$

Then

$$\frac{\partial s_t(x_t)}{\partial x_t} = -\frac{2}{c^2}B_t(x_t),\ \frac{\partial^2 s_t(x_t)}{\partial x_t^2} = \frac{2}{c^2}D_t.$$

Viewing w as a scalar function of s,

$$w(s) = (1+s)^{-1},\ w'(s) = -(1+s)^{-2},\ w''(s) = 2(1+s)^{-3},$$

the chain rule yields

$$\nabla_{x_t} w(x_t, y_t) = w'(s_t)\,\nabla_{x_t} s_t = \frac{2}{c^2}\,(1+s_t)^{-2}B_t,$$

$$\nabla^2_{x_t} w(x_t,y_t) = w''(s_t)\,\nabla s_t\,\nabla s_t^\top + w'(s_t)\,\nabla^2 s_t$$
$$= -\frac{2}{c^2}(1+s_t)^{-2} D_t + \frac{8}{c^4}(1+s_t)^{-3} B_t B_t^\top.$$

Next define

$$\mathcal{F}(s) \triangleq \frac{c^2}{2}s + \kappa_t, \quad \text{so that} \quad \ell^w(x_t,y_t) = w(s_t)\mathcal{F}(s_t).$$

We have

$$\mathcal{F}'(s) = \frac{c^2}{2},\ \mathcal{F}''(s) = 0,$$

and therefore

$$\nabla_{x_t}\mathcal{F}(s_t) = \mathcal{F}'(s_t)\,\nabla s_t = -B_t,\ \nabla^2_{x_t}\mathcal{F}(s_t) = D_t.$$

Using the product rule,

$$\nabla_{x_t}\ell^w = \nabla w\,\mathcal{F}(s_t) + w(s_t)\,\nabla\mathcal{F}(s_t),$$

we obtain

$$\nabla_{x_t}\ell^w(x_t,y_t) = \frac{2}{c^2}(1+s_t)^{-2} B_t\,\mathcal{F}(s_t) - (1+s_t)^{-1}B_t.$$

A direct algebraic simplification shows that

$$\frac{2}{c^2}(1+s_t)^{-2}\mathcal{F}(s_t) - (1+s_t)^{-1} = (\frac{2\kappa_t}{c^2}-1)(1+s_t)^{-2},$$

and thus the gradient takes the compact form

$$\nabla_{x_t}\ell^w(x_t,y_t) = (\frac{2\kappa_t}{c^2}-1)(1+s_t(x_t))^{-2}B_t(x_t).$$

In particular, the gradient is always colinear with $B_t(x_t) = C^\top R_t^{-1}(y_t - Cx_t)$.

Similarly, the Hessian of $\ell^w = w\mathcal{F}$ is

$$\nabla^2_{x_t}\ell^w = \nabla^2 w\mathcal{F}(s_t) + \nabla w\nabla\mathcal{F}(s_t)^\top + \nabla\mathcal{F}(s_t)\nabla w^\top + w(s_t)\nabla^2\mathcal{F}(s_t).$$

Substituting the expressions above and collecting the terms in D_t and $B_tB_t^\top$, we obtain

$$\nabla^2_{x_t}\,\ell^w(x_t,y_t) = \alpha_t(s_t)\,D_t + \beta_t(s_t)\,B_tB_t^\top,$$

with scalar coefficients

$$\alpha_t(s) = -(\frac{2\kappa_t}{c^2}-1)(1+s)^{-2},\ \beta_t(s) = \frac{4}{c^2}(\frac{2\kappa_t}{c^2}-1)(1+s)^{-3}.$$

Using the spectral bounds on R_t^{-1} and C, we have

$$\|B_t(x_t)\|^2 = r_t(x_t)^\top R_t^{-1}CC^\top R_t^{-1}r_t(x_t)$$
$$\le \|CC^\top\| r_t(x_t)^\top R_t^{-1}r_t(x_t)$$
$$= \tilde{c}_1\,c^2 s_t(x_t),$$

for some constant $\tilde{c}_1 > 0$, and hence

$$\|B_t(x_t)\| \le \tilde{c}_2\sqrt{s_t(x_t)}$$

for some $\tilde{c}_2 > 0$. Moreover, $\|D_t\|$ is a finite constant independent of (x_t,y_t).

From the explicit gradient expression we obtain

$$\big\|\nabla_{x_t}\ell^w(x_t,y_t)\big\| = |\frac{2\kappa_t}{c^2}-1|\,(1+s_t)^{-2}\|B_t\|$$
$$\le |\frac{2\kappa_t}{c^2}-1|\,\tilde{c}_2\frac{\sqrt{s_t(x_t)}}{(1+s_t(x_t))^2}.$$

This is of the form $s^\alpha/(1+s)^k$ with $(\alpha,k) = (1/2,2)$, and hence uniformly bounded for $s \ge 0$ by the lemma. Therefore there exists $\mathcal{K}_1 > 0$ such that

$$\big\|\nabla_{x_t}\ell^w(x_t,y_t)\big\| \le \mathcal{K}_1,\ \forall x_t, y_t.$$

For the Hessian, using $\|D_t\| \le \tilde{c}_3$ and $\|B_tB_t^\top\| = \|B_t\|^2 \le \tilde{c}_1 c^2 s_t$, we get

$$\big\|\nabla^2_{x_t}\ell^w(x_t,y_t)\big\|$$
$$\le |\alpha_t(s_t)|\,\|D_t\| + |\beta_t(s_t)|\,\|B_tB_t^\top\|$$
$$\le \tilde{c}_3|\frac{2\kappa_t}{c^2}-1|(1+s_t)^{-2} + \tilde{c}_1c^2\frac{4}{c^2}|\frac{2\kappa_t}{c^2}-1|\frac{s_t}{(1+s_t)^3}.$$

The two terms on the right-hand side are of the form $s^\alpha/(1+s)^k$ with $(\alpha,k) = (0,2)$ and $(1,3)$, and are thus uniformly bounded over $s \ge 0$. Therefore there exists $\mathcal{K}_2 > 0$ such that

$$\big\|\nabla^2_{x_t}\ell^w(x_t,y_t)\big\| \le \mathcal{K}_2,\ \forall x_t, y_t.$$

The stationary conditions of the NANO update cost $J(\hat{x}_t, P_t)$ can be written as

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} - P_{t|t-1}\,\mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\Big\{\nabla_{x_t}\ell^w(x_t,y_t)\Big\},$$
$$P^{-1}_{t|t} = P^{-1}_{t|t-1} + \mathbb{E}_{\mathcal{N}(x_t;\hat{x}_{t|t},P_{t|t})}\Big\{\nabla^2_{x_t}\ell^w(x_t,y_t)\Big\}.$$

By the uniform bounds, we have

$$\Big\|\mathbb{E}\big\{\nabla_{x_t}\ell^w(x_t,y_t)\big\}\Big\| \le \mathcal{K}_1,\ \Big\|\mathbb{E}\big\{\nabla^2_{x_t}\ell^w(x_t,y_t)\big\}\Big\| \le \mathcal{K}_2,$$

for all y_t. Hence, for a given prior $(\hat{x}_{t|t-1}, P_{t|t-1})$, any stationary point $(\hat{x}_{t|t}, P_{t|t})$ satisfies

$$\hat{x}_{t|t} \in \hat{x}_{t|t-1} + \{-P_{t|t-1}u : \|u\| \le \mathcal{K}_1\},$$

and the eigenvalues of $P^{-1}_{t|t}$ lie in the interval obtained by shifting those of $P^{-1}_{t|t-1}$ by at most $\mathcal{K}_2$. In particular, if $P_{t|t-1}$ is positive definite with eigenvalues in $[\underline{\lambda}_P, \overline{\lambda}_P]$ independent of y_t, then all posterior covariances $P_{t|t}$ also have eigenvalues contained in a compact interval $[\lambda_{\min}, \lambda_{\max}]$ that does not depend on y_t, and the corresponding posterior means $\hat{x}_{t|t}$ remain in a bounded set.

Let $q_t(\cdot|y_{1:t}) = \mathcal{N}(\hat{x}_{t|t}, P_{t|t}), q^c_t(\cdot|y_{1:t-1}, y_c) = \mathcal{N}(\hat{x}^c_{t|t}, P^c_{t|t})$ denote the posteriors corresponding to y_t and y_c, respectively. From the discussion above, there exist constants $0 < \lambda_{\min} \le \lambda_{\max} < \infty$ and $\tilde{c} > 0$, independent of y_c, such that for all y_c,

$$\lambda_{\min} I \preceq P_{t|t}, P^c_{t|t} \preceq \lambda_{\max} I,\ \|\hat{x}_{t|t} - \hat{x}^c_{t|t}\| \le \tilde{c}.$$

The posterior influence function is the KL divergence between these Gaussians, with closed-form expression

$$\mathrm{PIF}_t(y_c; y_{1:t}) = \tfrac{1}{2}\Big(\mathrm{tr}(P^{-1}_{t|t}P^c_{t|t} - I) - \log\det(P^{-1}_{t|t}P^c_{t|t})$$
$$+ (\hat{x}_{t|t} - \hat{x}^c_{t|t})^\top P^{-1}_{t|t}(\hat{x}_{t|t} - \hat{x}^c_{t|t})\Big).$$

The spectral bounds on $P_{t|t}$ and $P^c_{t|t}$ imply that both $\mathrm{tr}(P^{-1}_{t|t}P^c_{t|t} - I)$ and $\log\det(P^{-1}_{t|t}P^c_{t|t})$ are uniformly bounded over all y_c, and the bound $\|\hat{x}_{t|t} - \hat{x}^c_{t|t}\| \le \tilde{c}$ together with $P^{-1}_{t|t} \preceq \lambda^{-1}_{\min} I$ implies that the quadratic term $(\hat{x}_{t|t} - \hat{x}^c_{t|t})^\top P^{-1}_{t|t}(\hat{x}_{t|t} - \hat{x}^c_{t|t})$ is also uniformly bounded. Therefore

$$\sup_{\|y_t - y_c\| \to \infty} \mathrm{PIF}_t(y_c; y_{1:t}) < \infty,$$

which shows that the weighted-loss NANO update has a bounded posterior influence function in the linear Gaussian case. □

This result demonstrates that, in the linear–Gaussian case, the weighted-loss NANO filter possesses a bounded PIF, offering a first theoretical justification of its robust behavior. We adopt the PIF framework—as suggested in [38]—since it characterizes the robustness of the posterior distribution rather than only the point estimate. Moreover, it is already known that the standard NANO filter (equivalent to KF in the linear–Gaussian case) has an unbounded PIF [38].

APPENDIX G
ERROR STATE FORMULATION OF NANO FILTER

Error-state formulations are widely adopted in robotic state estimation because the underlying system dynamics and measurement models often evolve on nonlinear manifolds, such as $SO(3)$ for rotations. Performing inference directly in the global state space may violate these geometric structures since the NANO filter is defined in a Euclidean vector space. In contrast, maintaining a nominal state $\bar{x}_t$ and estimating a small perturbation δx_t typically provides better numerical stability. The perturbation is defined through a retraction

$$x_t = \bar{x}_t \oplus \delta x_t,$$

where $\oplus$ denotes the manifold composition operator.

In the error-state formulation of NANO filter, the natural-gradient update is applied to the error state δx_t rather than to the global state x_t. Since the update is performed in the tangent space, all geometric constraints such as unit quaternion norms or rotation-matrix orthogonality are automatically preserved after retraction. This makes the method compatible with Lie-group-based state representations while retaining the optimization-driven nature of the natural gradient update.

To evaluate the effectiveness of this formulation in a real robotic system, we integrate the natural gradient update into the widely used **FAST-LIO2** LiDAR–inertial odometry framework [59] by replacing its IESKF update with an *Error-State NANO filter variant*. In this implementation, the natural-gradient step updates the perturbation in the tangent space, and the corrected error is subsequently retracted back to the nominal state. This modification requires no change to the rest of the FAST-LIO2 pipeline and preserves its efficient structure.

We assess performance on five sequences from the NCLT dataset and compute the absolute trajectory error (ATE) using the *evo* evaluation tool [61]. As shown in Table 4, the error-state NANO filter consistently achieves lower ATE than the original IESKF used in FAST-LIO2. These results demonstrate that the NANO filter remains effective when embedded within a full LiDAR–inertial odometry system and that it provides a viable alternative to error-state Kalman filtering.

APPENDIX H
ADDITIONAL SIMULATION RESULTS

TABLE 4
ATE comparison between IESKF (FAST-LIO2) and the Error-State NANO filter.

Dataset	IESKF (m)	NANO Filter (m)	Distance (km)
nclt_1	2.22	2.01	1.62
nclt_2	2.86	2.28	2.27
nclt_3	1.56	1.52	0.26
nclt_4	7.57	6.47	1.40
nclt_5	1.48	1.47	1.86

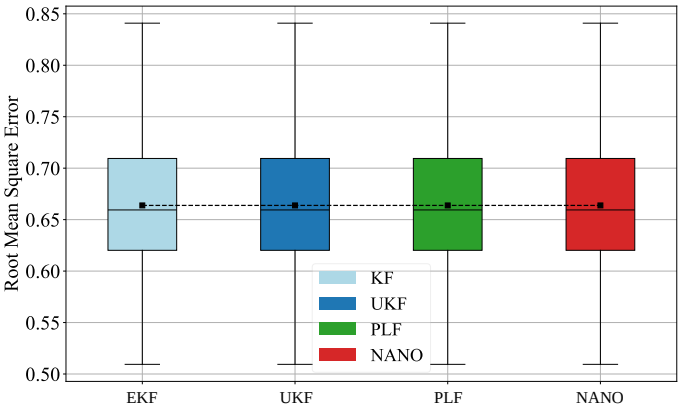

Fig. 8. Box plot of RMSE of KF, UKF, PLF and NANO filter, for the standard Wiener velocity model.

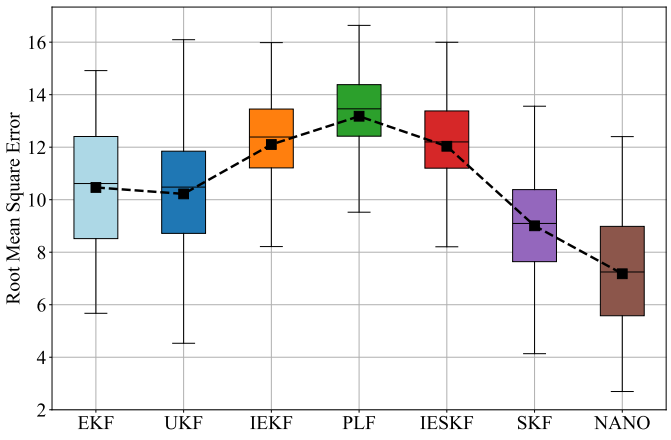

Fig. 9. Box plot of RMSE for EKF, UKF, IEKF, PLF, IESKF, SKF and NANO filter on the 9D coupled Lorenz system over 200 Monte Carlo runs.

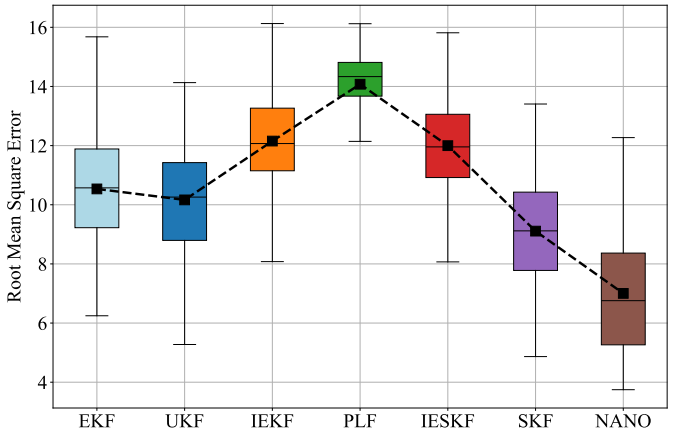

Fig. 10. Box plot of RMSE for EKF, UKF, IEKF, PLF, IESKF, SKF and NANO filter on the 12D coupled Lorenz system over 200 Monte Carlo runs.

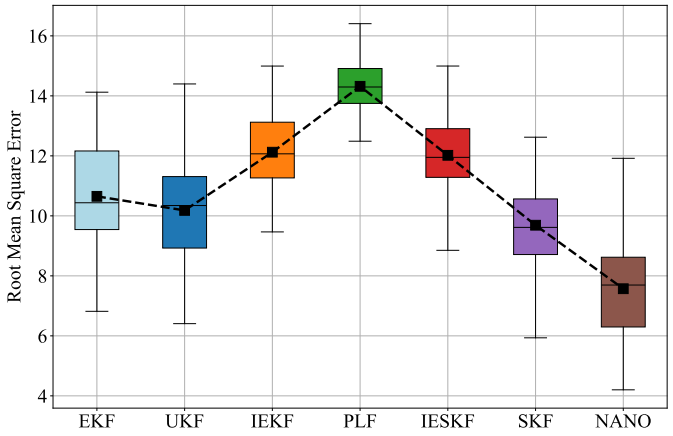

Fig. 11. Box plot of RMSE for EKF, UKF, IEKF, PLF, IESKF, SKF and NANO filter on the 15D coupled Lorenz system over 200 Monte Carlo runs.