Analytical Kernel Kalman Filter with RKHS-Based Measurement Model Approximation

SANFENG HU
Yunnan University, Kunming, China

LIPING GUO, Member, IEEE
Chinese Academy of Sciences, Beijing, China

LIHUA XIE, Fellow, IEEE
Nanyang Technological University, Singapore

JI-FENG ZHANG, Fellow, IEEE
Zhongyuan University of Technology, Zhengzhou, China
Chinese Academy of Sciences, Beijing, China

***Abstract*— Kernel mean embedding has emerged as a powerful tool for nonlinear filtering by enabling linear estimation in a reproducing kernel Hilbert space (RKHS). While substantial progress has been achieved, the existing methods typically rely on empirical approximations and often involve costly preimage recovery for posterior probability density function (PDF) approximation, which may limit their accuracy and efficiency. To overcome these limitations, this paper investigates kernel mean embedding based nonlinear filtering in a model-driven setting with Gaussian kernels. An analytical kernel Kalman rule (AKKR) is proposed based on an approximated measurement function in RKHS, which avoids empirical approximations by leveraging model information to derive analytical expression of posterior kernel mean estimate. Based on the AKKR, an analytical preimage is further derived, revealing the preimage's inherent Gaussian mixture structure that enables a structure-aware regularization via convex quadratic programming for efficient PDF approximation. Based on these results, an AKKR based posterior PDF approximation algorithm and an analytical kernel Kalman filter are developed. Three representative examples demonstrate that the proposed algorithms achieve superior posterior approximation and filtering performance across various nonlinear scenarios.**

***Index Terms*— Nonlinear filtering, posterior distribution approximation, kernel mean embedding, kernel Kalman rule, preimage recovery**

This work was supported in part by CAS Special Research Assistant Project under Grant E4559305, in part by Postdoctoral Fellowship Program of CPSF under Grant GZB20230813, in part by China Postdoctoral Science Foundation under Grant 2024M763479, in part by the National Natural Science Foundation of China under Grants 62433020 and T2293770, and in part by the Fundamental Research Project of Yunnan Province under Grant 202401AU070230. *(Corresponding author: Liping Guo.)*

Sanfeng Hu is with the Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University, Kunming, Yunnan 650504, China (e-mail: sanfenghu@outlook.com). Liping Guo is with the State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lipguo@outlook.com). Lihua Xie is with School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 (e-mail: elhxie@ntu.edu.sg). Ji-Feng Zhang is with the School of Automation and Electrical Engineering, Zhongyuan University of Technology, Zhengzhou 450007, China; the State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; and the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jif@iss.ac.cn).

I. INTRODUCTION

STATE estimation serves as a cornerstone of modern engineering systems, playing a critical role in applications such as target tracking [1], [2], positioning and navigation [3], [4], and fault detection [5]. The Kalman filter and its variants have been widely employed for state estimation in linear systems due to their optimality and computational efficiency. However, real-world systems often exhibit significant nonlinearities. Although the Bayesian paradigm provides an optimal conceptual framework for nonlinear filtering, its practical applications are often limited by the intractable nature of the posterior distributions resulted from the nonlinearities of the systems [6]. Thus, various nonlinear filtering methods have been proposed to address the challenges posed by system nonlinearity over the past decades.

To achieve the simplicity and computational efficiency of a Kalman-type estimate, various linearization based approaches and linear estimation techniques have been developed. The widely adopted extended Kalman filter linearizes the system through first-order Taylor series expansion at the current estimate, and subsequently applies the Kalman filter to the linearized system [7]. The unscented Kalman filter employs the carefully selected sigma points to compute the statistical moments required by the linear estimator, thereby circumventing direct linearization of the system [8]. As an alternative to the unscented Kalman filter, the cubature Kalman filters with different algebraic degrees have been proposed based on the cubature rule [9], [10]. For purpose of improving the estimation accuracy, the third-degree moment approximation based filter has been proposed, which additionally propagates the third-degree moments based on linear estimation [11]. Besides, uncorrelated conversion based filters have been proposed based on linear estimation using the measurement augmented by its uncorrelated conversions [12]. Moreover, some iterative methods, including iterated extended Kalman filter [13] and posterior linearization filter [14], gradually approach a more accurate estimate in each time step by adjusting the linear approximation of the measurement function.

To handle increasingly complex nonlinear systems with high estimation accuracy, extensive research has focused on directly approximating the posterior probability density function (PDF) of the state. By approximating

the posterior PDF with a set of weighted particles, the well-known particle filters have been theoretically proven to be asymptotically optimal as the number of particles approaches infinity. However, particle filters often suffer from degeneracy and impoverishment. Degeneracy happens when only a few particles have significant weights, and resampling may worsen the problem by removing low-weight particles and reducing diversity. To overcome these problems, various methods have been proposed, including Markov chain Monte Carlo based refinement [15], [16], heuristic optimized particles [17], and systematic particle flow techniques [18], [19]. Beyond discrete particle based approximations, alternatively, continuous PDF based methods utilize analytic functions to approximate the posterior PDF. For example, the Gaussian PDF is usually adopted based on the Kullback-Leibler divergence minimization approach with variety of optimization strategies [20]–[22]. Furthermore, the Gaussian mixture is also a commonly used one for more higher estimation accuracy [23]. Under this case, many methods have been proposed to handle the exponentially growing size of the Gaussian components, including variational methods [24] and moment match methods [25], [26].

The recently developed kernel mean embedding based methods leverage the efficiency of linear estimation for posterior PDF approximation, making them a promising approach in terms of estimation accuracy and computational efficiency [27]. The core innovation lies in representing the posterior distribution of the state as a kernel mean within a reproducing kernel Hilbert space (RKHS), where linear estimation is performed to obtain the posterior kernel mean [28]. In [29], a nonparametric data-driven Bayesian filter was developed based on the kernel Bayes' rule. Subsequent study [30] introduced kernel Bayes' rule based filtering approaches for scenarios where the measurement model is available only through state-observation pair sample points, while the process model remains explicitly defined. This work integrates parametric Monte Carlo sampling with nonparametric learning of the measurement model [31]. Furthermore, the kernel Kalman rule was proposed using recursive least squares in RKHS to mitigate some instabilities of kernel Bayes' rule [32]. Besides these data-driven kernel methods, model-driven approaches have also been explored. Based on kernel Kalman rule, a proposal step was introduced to adjust the position of sample points, improving their alignment with the significant probability mass of the posterior PDF [31]. An alternative kernel Kalman filter was proposed by approximating the measurement function using a linear operator defined in RKHS and employing linear estimation in RKHS [22]. However, all the aforementioned methods rely on empirical approximations of the kernel mean using sample points, which may introduce errors that accumulate over time, potentially leading to performance degradation in filtering.

Furthermore, a common challenge in kernel mean embedding based filtering is the preimage recovery problem, which concerns reconstructing the posterior PDF approximation from the posterior kernel mean estimate in RKHS. This problem is inherently difficult due to the nontrivial nature of kernel mean embedding. One common approach is to approximate the posterior PDF by selecting the one within a specified PDF family whose kernel mean minimizes the distance to the posterior kernel mean estimate [27], [33]. However, this approach does not provide clear guidance on how to specify the PDF family and may often lead to non-convex optimization problems that are challenging to solve efficiently. An alternative approach, known as kernel herding [34]–[36], approximates the posterior PDF by solving a series of optimization problems to obtain a set of sample points. Nonetheless, this approach tends to incur high computational cost, which may pose challenges for real-time applications.

Motivated by the above analyses, we raise the following key questions. 1) Is it possible to obtain an analytical expression of the posterior kernel mean estimate, instead of relying on empirical approximations, in order to reduce the error accumulation caused by empirical approximations during the filtering process? 2) Is it possible to derive an analytical preimage of the posterior kernel mean estimate, thereby obtaining an approximation of the posterior PDF with significantly reduced computational cost?

This paper investigates kernel mean embedding based nonlinear filtering in a model-driven setting with Gaussian kernels. The main contributions of this paper are summarized as follows:

1) By leveraging model information and approximating the measurement function in RKHS, we derive an analytical expression of the posterior kernel mean estimate and propose an analytic kernel Kalman rule (AKKR). In contrast to the existing kernel Bayes' and kernel Kalman rules, the proposed AKKR eliminates the reliance on empirical approximations, thereby reducing empirical errors and laying a foundation for high-accuracy posterior PDF approximation and filtering.
2) Based on the proposed AKKR, we derive the analytical preimage of the posterior kernel mean estimate, which reveals the preimage's inherent Gaussian mixture structure and highlights the necessity of regularizing the posterior kernel mean estimate. Accordingly, we provide a regularization method, formulated as a convex quadratic programming problem. The regularization method fully exploits the revealed structure to efficiently obtain a Gaussian mixture approximation of the posterior PDF.
3) Based on the proposed AKKR, the derived analytical preimage, and the regularization method, we develop a posterior PDF approximation algorithm and an analytical kernel Kalman filter (AKKF). Examples demonstrate that the proposed posterior PDF approximation performs well in various nonlinear scenarios, and the proposed AKKF achieves improved estimation accuracy and reliability with acceptable

computational cost, compared with several popular filtering methods.

This paper is organized as follows. In Section II, some preliminaries about kernel mean embedding are provided. In Section III, the problem formulation and our motivations are presented. In Section IV, the AKKR is proposed. The analytical preimage, regularization and the AKKR based posterior PDF approximation algorithm are given in Section V. The AKKF is established in Section VI. Examples are shown in Section VII, followed by conclusion in Section VIII.

Notations: Scalars, vectors and matrices are denoted by lowercase letters, bold lowercase letters, and capital letters, respectively. All vectors are column vectors. Some notations are listed in Table I.

TABLE I
List of Notations

Notation	Explanation
$\mathbb{R}$	set of all real numbers
$\mathbb{Z}^+$	set of all positive integers
$\mathbb{R}^n$	set of all n-dimensional real vectors
$\mathbb{R}^{m\times n}$	set of all $m\times n$ real matrices
$\mathbf{1}$	vector with all entries one
$\|\mathbf{a}\|$	Euclidean norm of $\mathbf{a}$
$\|\mathbf{a}\|_A$	Euclidean norm of $\mathbf{a}$ weighted with $A\succ 0$, i.e., $\sqrt{\mathbf{a}^{\mathrm{T}}A\mathbf{a}}$
$\mathrm{diag}(\mathbf{a})$	diagonal matrix with elements of $\mathbf{a}$ on main diagonal
$\mathbf{a}\succeq 0$	all elements of $\mathbf{a}$ are nonnegative
$\mathbf{a}\odot\mathbf{b}$	Hadamard product of vectors $\mathbf{a}$ and $\mathbf{b}$
$A\succ 0$	A is symmetric positive definite
$A\succeq 0$	A is symmetric positive semi-definite
$\mathrm{chol}(A)$	lower triangular factor of Cholesky decomposition of $A\succ 0$
$\langle f,g\rangle_{\mathcal{H}}$	inner product of f and g in RKHS $\mathcal{H}$
$\|f\|_{\mathcal{H}}$	norm of f in RKHS $\mathcal{H}$
$\mathcal{N}(\mathbf{m},P)$	Gaussian distribution with mean $\mathbf{m}$ and covariance matrix P
$\mathcal{N}(\mathbf{x};\mathbf{m},P)$	PDF of $\mathcal{N}(\mathbf{m},P)$

II. PRELIMINARIES

This section briefly reviews the RKHS, kernel mean embedding, and covariance operator. For further details, one can refer to [27] and [28].

A. Reproducing Kernel Hilbert Space

Let $\mathcal{X}$ be an arbitrary set, and suppose that the function $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ satisfies the following conditions:

1) $k(\mathbf{x},\mathbf{y})=k(\mathbf{y},\mathbf{x}),\ \forall\mathbf{x},\mathbf{y}\in\mathcal{X}$,
2) $\sum_{i=1}^n\sum_{j=1}^n a_ia_jk(\mathbf{x}_i,\mathbf{x}_j)\geq 0,\ \forall n\in\mathbb{Z}^+,a_i\in\mathbb{R},\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}$.

Then, $k(\cdot,\cdot)$ is called a kernel on the space $\mathcal{X}$. The corresponding RKHS, denoted by $\mathcal{H}$, associated with the kernel $k(\cdot,\cdot)$ is defined as the Hilbert space obtained by completing the inner product space of functions:

$$f(\cdot)=\sum_{i=1}^n a_ik(\mathbf{x}_i,\cdot),\ n\in\mathbb{Z}^+,a_i\in\mathbb{R},\mathbf{x}_i\in\mathcal{X},$$

with the inner product satisfying

$$\langle k(\mathbf{x}_i,\cdot),k(\mathbf{x}_j,\cdot)\rangle=k(\mathbf{x}_i,\mathbf{x}_j).$$

For clarity, we occasionally employ the notation $f^{\mathrm{T}}g$ to represent the inner product $\langle f,g\rangle_{\mathcal{H}}$ for $f,g\in\mathcal{H}$ in the following context. Besides, we refer to $k(\mathbf{x},\cdot)$ as $\phi(\mathbf{x})$, known as the feature mapping, i.e., $\phi(\mathbf{x})=k(\mathbf{x},\cdot)$. It is worth noting that $\phi(\mathbf{x})$ is a function on $\mathcal{X}$ for any given $\mathbf{x}\in\mathcal{X}$. One of the most important properties of RKHS is the reproducing property:

$$f(\mathbf{x})=\langle f,\phi(\mathbf{x})\rangle_{\mathcal{H}},\ \forall f\in\mathcal{H},\mathbf{x}\in\mathcal{X},$$

which will be frequently used later.

Some commonly used kernels on space $\mathcal{X}=\mathbb{R}^n$ are listed as follows:

1) Gaussian kernel: $k(\mathbf{x},\mathbf{y})=\mathcal{N}(\mathbf{x};\mathbf{y},\Sigma)$, with $\Sigma\succ 0$;
2) Laplace kernel: $k(\mathbf{x},\mathbf{y})=\exp\{-\sigma\|\mathbf{x}-\mathbf{y}\|_1\}$, with $\sigma>0$;
3) Polynomial kernel: $k(\mathbf{x},\mathbf{y})=(\mathbf{x}^{\mathrm{T}}\mathbf{y}+c)^\alpha$, with $c\geq 0,\alpha>0$.

B. Kernel Mean Embedding and Covariance Operator

Let $\mathcal{M}(\mathcal{X})$ be the set of all probability distributions defined on $\mathcal{X}$. The kernel mean embedding, which maps a distribution $P\in\mathcal{M}(\mathcal{X})$ into the RKHS $\mathcal{H}$, is defined as

$$\iota:\mathcal{M}(\mathcal{X})\longrightarrow\mathcal{H},\quad P\longmapsto\mu_P:=\int_{\mathcal{X}}k(\mathbf{x},\cdot)\mathrm{d}P(\mathbf{x}),$$

where μ_P is referred to as the kernel mean of P. Given an independent and identically distributed sample $\{\mathbf{x}_i\}_{i=1}^n$ drawn from P, the kernel mean can be approximated empirically as follows:

$$\mu_P\approx\frac{1}{n}\sum_{i=1}^n k(\mathbf{x}_i,\cdot)=\frac{1}{n}\sum_{i=1}^n\phi(\mathbf{x}_i).\tag{1}$$

A kernel $k(\cdot,\cdot)$ is said to be characteristic if the associated mean embedding mapping ι is injective. That is, for any two probability distributions $P,Q\in\mathcal{M}(\mathcal{X})$, $P\neq Q$ implies $\mu_P\neq\mu_Q$. Intuitively, if a characteristic kernel is adopted, then the kernel mean embedding can preserve all information of a distribution. Particularly, the Gaussian kernel is one of the most commonly used characteristic kernels.

Another important concept is the covariance operator. Formally, the covariance operator $\mathcal{P}:\mathcal{H}\longrightarrow\mathcal{H}$ with respect to the probability distribution $P\in\mathcal{M}(\mathcal{X})$ is defined as

$$\mathcal{P}:=\int_{\mathcal{X}}\phi(\mathbf{x})\otimes\phi(\mathbf{x})\mathrm{d}P(\mathbf{x})-\mu_P\otimes\mu_P,$$

where "$\otimes$" denotes the tensor product operation (see Section 3.2 of [27]). The following identity is a fundamental property of the covariance operator:

$$\langle f, \mathcal{P}g\rangle_{\mathcal{H}} = \int_{\mathcal{X}} f(\mathbf{x})g(\mathbf{x})\mathrm{d}P(\mathbf{x}) - \int_{\mathcal{X}} f(\mathbf{x})\mathrm{d}P(\mathbf{x}) \cdot \int_{\mathcal{X}} g(\mathbf{x})\mathrm{d}P(\mathbf{x}),\ \forall f, g \in \mathcal{H}. \tag{2}$$

In fact, (2) implies that $\mathcal{P}$ captures the dependency structure between functions in $\mathcal{H}$.

Given an independent and identically distributed sample $\{\mathbf{x}_i\}_{i=1}^n$ drawn from the distribution P, the covariance operator can be approximated empirically as

$$\mathcal{P} \approx \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i) \otimes \phi(\mathbf{x}_i) - \left(\frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i)\right) \otimes \left(\frac{1}{n}\sum_{j=1}^{n} \phi(\mathbf{x}_j)\right). \tag{3}$$

III. PROBLEM FORMULATION AND MOTIVATIONS

A. Posterior PDF Approximation Problem

Nonlinear filtering tackles the fundamental challenge of estimating system state from noisy measurements, formulated as the posterior PDF approximation problem as follows.

Consider two random vectors: the state $\mathbf{x} \in \mathbb{R}^{n_x}$ and the measurement $\mathbf{y} \in \mathbb{R}^{n_y}$, correlated through the following nonlinear measurement model:

$$\mathbf{y} = h(\mathbf{x}) + \mathbf{v}, \tag{4}$$

where $h : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ is a known nonlinear function, $\mathbf{v} \in \mathbb{R}^{n_y}$ represents the measurement noise with zero mean and covariance matrix $R \succ 0$, statistically independent of $\mathbf{x}$. Given the prior PDF $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{m}, P)$, the posterior PDF takes the form

$$p(\mathbf{x}|\mathbf{y}) = c\exp\left\{-\frac{1}{2}\left(\|\mathbf{y} - h(\mathbf{x})\|^2_{R^{-1}} - \|\mathbf{x} - \mathbf{m}\|^2_{P^{-1}}\right)\right\}, \tag{5}$$

where c is a normalizing constant ensuring that the posterior integrates to one. However, due to the nonlinearity of $h(\cdot)$, it is generally difficult to use (5) directly for subsequent statistical inference. Typical inference tasks in filtering, such as posterior state estimation and state prediction via the Chapman-Kolmogorov equation, involve intractable integrals arising from complexity of (5). Therefore, the problem is how to approximate $p(\mathbf{x}|\mathbf{y})$ as accurately as possible under an acceptable computational cost, in a manner that facilitates subsequent statistical inference.

As is well known, the kernel mean embedding based estimation framework offers a promising approach for effectively and efficiently addressing the posterior PDF approximation problem.

B. Kernel Mean Embedding Based Estimation Framework

Unlike the Bayesian estimation framework that directly handles the complicated posterior PDF in the state space, the kernel mean embedding based estimation framework focuses on estimating the posterior kernel mean based on the linear estimation in a higher-dimensional RKHS, and then obtaining an approximation of the posterior PDF through a process known as preimage recovery based on the estimate.

For model (4), the posterior kernel mean is defined as

$$\mu_{x|y} = \int_{\mathbb{R}^{n_x}} \phi(\mathbf{x})p(\mathbf{x}|\mathbf{y})\mathrm{d}\mathbf{x},$$

where $\phi(\cdot)$ is the feature map defined on the state space $\mathbb{R}^{n_x}$, and its corresponding RKHS is denoted by $\mathcal{H}_x$. As far as we know, the existing posterior kernel mean estimates in the literature generally take the following form:

$$\hat{\mu}_{x|y} = A\psi(\mathbf{y}) + \beta,$$

where $\psi(\mathbf{y})$ represents a feature map defined on the measurement space $\mathbb{R}^{n_y}$, and the corresponding RKHS is denoted by $\mathcal{H}_y$. Different methods employ different choices of $A : \mathcal{H}_y \longrightarrow \mathcal{H}_x$ and $\beta \in \mathcal{H}_x$ based on linear estimation theory [22], [31], [37]. Since it is difficult to derive analytical expressions for A and β, they are approximated via sampling, with equations (1) and (3) being employed. As a result, the posterior kernel mean estimate has the following empirical expression:

$$\hat{\mu}^e_{x|y} = \sum_{i=1}^{N} w_i\phi(\mathbf{x}_i), \tag{6}$$

where $\mathbf{x}_i$ is a sample point with its weight w_i. Then, based on $\hat{\mu}^e_{x|y}$, the preimage, defined as the function $q(\cdot)$ satisfying

$$\hat{\mu}^e_{x|y} = \int_{\mathbb{R}^{n_x}} \phi(\mathbf{x})q(\mathbf{x})\mathrm{d}\mathbf{x},$$

should be recovered as the approximation of the posterior PDF $p(\mathbf{x}|\mathbf{y})$.

C. Limitations and Motivations

Although remarkable advancements have been made, the existing kernel mean embedding based methods still exhibit certain limitations:

1) *Empirical approximation error:* Since the posterior kernel mean estimate relies on empirical sample, its accuracy is sensitive to the size and quality of the sample. As the filtering process proceeds, approximation errors accumulate, potentially leading to performance degradation over time.
2) *Time-consuming preimage recovery:* Recovering the preimage from posterior kernel mean estimate remains challenging, as it typically requires computationally intensive numerical optimization to obtain an approximation of the posterior density, thereby increasing the overall computational cost.

3) *Invalid posterior kernel mean estimate:* The posterior kernel mean estimate obtained through linear estimation may not admit a preimage that is a PDF, one reason being that the weights w_i , $1 \leq i \leq N$, in (6) may be negative, as pointed out in [37]. This phenomenon contributes to the difficulty in approximating the posterior PDF.

To overcome these limitations, this paper proposes the following key innovations. First, the AKKR based on the Gaussian kernel is proposed. Second, the analytical preimage is derived. Third, an efficient regularization method is developed for the posterior kernel mean estimate. Finally, building upon these advancements, the AKKR based posterior PDF approximation algorithm and AKKF are provided.

Conventions: Due to the simplicity and characteristic property of the Gaussian kernel, the rest of this paper is discussed within the RKHS $\mathcal{H}$ generated by kernel $k(\mathbf{z}_1, \mathbf{z}_2) = \mathcal{N}(\mathbf{z}_1; \mathbf{z}_2, \Sigma)$, where $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{n_x}$ and Σ is a positive definite matrix. Correspondingly, the feature map is $\phi(\mathbf{z}) = \mathcal{N}(\cdot; \mathbf{z}, \Sigma)$.

IV. ANALYTICAL KERNEL KALMAN RULE

A. Measurement Model Approximation in RKHS

The complexity of the posterior PDF is largely influenced by the nonlinearity of the measurement function $h(\cdot)$. Therefore, obtaining a suitable approximation of $h(\cdot)$ is a critical step in simplifying the problem. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ denote a set of points, which may optionally be drawn from the prior PDF $p(\mathbf{x})$. The set X represents the region of interest where we aim to approximate $h(\cdot)$ accurately. Denote

$$\begin{aligned} \Phi &= [\phi(\mathbf{x}_1), \phi(\mathbf{x}_2), \ldots, \phi(\mathbf{x}_N)], \\ Y &= [h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_N)]. \end{aligned} \tag{7}$$

Then, as shown in [22], $h(\cdot)$ can be linearly approximated in RKHS as follows:

$$\tilde{h}(\mathbf{x}) = H\phi(\mathbf{x}),$$

where the linear operator $H : \mathcal{H} \to \mathbb{R}^{n_y}$ has the following expression:

$$H = YG^{-1}\Phi^{\mathrm{T}}, \tag{8}$$

with the (i, j)-th element of the Gram matrix G given by

$$G_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j). \tag{9}$$

Thus, the measurement model (4) is equivalent to

$$\mathbf{y} = H\phi(\mathbf{x}) + h_\delta(\mathbf{x}) + \mathbf{v}, \tag{10}$$

where $h_\delta(\mathbf{x}) = h(\mathbf{x}) - H\phi(\mathbf{x})$ is the approximation error of $h(\cdot)$. Since $h_\delta(\mathbf{x})$ may be nonlinear with respect to $\phi(\mathbf{x})$, to simplify and compensate for $h_\delta(\mathbf{x})$, the equation (10) is linearly converted as follows:

$$\mathbf{y} = H\phi(\mathbf{x}) + \tilde{\mathbf{v}}, \tag{11}$$

where $\tilde{\mathbf{v}}$ is considered to be independent of $\mathbf{x}$, and has zero mean and covariance matrix

$$\tilde{R} = R + R_\delta, \tag{12}$$

with R_δ given by

$$R_\delta = \frac{1}{m}\sum_{i=1}^{m} h_\delta(\tilde{\mathbf{x}}_i) h_\delta^{\mathrm{T}}(\tilde{\mathbf{x}}_i),$$

where the sample points $\{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_m\}$ are independent and identically distributed according to the prior PDF $p(\mathbf{x})$.

Remark 1:
It is worth emphasizing that the proposed measurement model representation in RKHS given by (11) is more reasonable than that in [22], since (11) accounts for the approximation error of $h(\cdot)$. Due to the strict positive definiteness of the Gaussian kernel, the measurement function $h(\cdot)$ is exactly fitted by $H\phi(\mathbf{x})$ at the evaluated point set X. Thus, X cannot be used to compute R_δ anymore.

Remark 2:
Theoretically, the measurement function $h(\cdot)$ can be approximated to any accuracy in $\mathcal{H}$ under appropriate criteria (see, e.g., [38, pp. 158]) . Specifically, for $r \in [1, \infty)$ with $\int_{\mathbb{R}^{n_x}} |h(\mathbf{x})|^r p(\mathbf{x}) \mathrm{d}\mathbf{x} < \infty$ and for any given $\epsilon > 0$, there exists a linear operator $H_\epsilon : \mathcal{H} \to \mathbb{R}^{n_y}$ such that

$$\int_{\mathbb{R}^{n_x}} |H_\epsilon \phi(\mathbf{x}) - h(\mathbf{x})|^r p(\mathbf{x}) \mathrm{d}\mathbf{x} < \epsilon.$$

B. Analytical Posterior Kernel Mean Estimate

Since (11) has a linear form with respect to $\phi(\mathbf{x})$, by using the linear minimum mean squared error estimation, the posterior kernel mean estimate is given as follows:

$$\hat{\mu} = \rho + \mathcal{P}H^{\mathrm{T}}(H\mathcal{P}H^{\mathrm{T}} + \tilde{R})^{-1}(\mathbf{y} - H\rho), \tag{13}$$

where ρ and $\mathcal{P}$ are the prior kernel mean and covariance operator, respectively.

We next aim to derive an analytical expression of the posterior kernel mean estimate $\hat{\mu}$. To this end, we should first derive the analytical expressions of the prior kernel mean ρ and the covariance operator $\mathcal{P}$.

Define $\mathcal{P}(\mathbf{z}_1, \mathbf{z}_2)$ based on (2) as follows:

$$\begin{aligned} \mathcal{P}(\mathbf{z}_1, \mathbf{z}_2) &\coloneqq \langle \phi(\mathbf{z}_1), \mathcal{P}\phi(\mathbf{z}_2) \rangle_{\mathcal{H}} \\ &= \int_{\mathcal{R}^{n_x}} k(\mathbf{z}_1, \mathbf{x}) k(\mathbf{z}_2, \mathbf{x}) \mathrm{d}P(\mathbf{x}) - \rho(\mathbf{z}_1)\rho(\mathbf{z}_2). \end{aligned}$$

Then, we have the following theorem.

Theorem 1:
The prior kernel mean and covariance operator can be analytically expressed as

$$\begin{aligned} \rho(\mathbf{z}) &= \mathcal{N}(\mathbf{z}; \mathbf{m}, \Sigma + P), \\ \mathcal{P}(\mathbf{z}_1, \mathbf{z}_2) &= \mathcal{N}(\mathbf{z}_1; \mathbf{z}_2, 2\Sigma)\mathcal{N}\big((\mathbf{z}_1 + \mathbf{z}_2)/2; \mathbf{m}, P + \Sigma/2\big) \\ &\quad - \mathcal{N}(\mathbf{z}_1; \mathbf{m}, \Sigma + P)\mathcal{N}(\mathbf{z}_2; \mathbf{m}, \Sigma + P), \end{aligned}$$

with $\mathbf{z}, \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^{n_x}$.

To provide a proof of Theorem 1, we first present the following lemma on the product of Gaussian PDFs, which will be frequently used later.

Lemma 1:
The product of any two Gaussian PDFs satisfies the following identity:

$$\mathcal{N}(\mathbf{x};\mathbf{a}_1,\Sigma_1)\mathcal{N}(\mathbf{x};\mathbf{a}_2,\Sigma_2)=c\cdot\mathcal{N}(\mathbf{x};\mathbf{a}_3,\Sigma_3), \tag{14}$$

where

$$\begin{aligned}&c=\mathcal{N}(\mathbf{a}_1;\mathbf{a}_2,\Sigma_1+\Sigma_2),\\ &\mathbf{a}_3=\Sigma_3(\Sigma_1^{-1}\mathbf{a}_1+\Sigma_2^{-1}\mathbf{a}_2),\Sigma_3=(\Sigma_1^{-1}+\Sigma_2^{-1})^{-1}.\end{aligned}$$

Proof:
By some calculations, we have

$$\begin{aligned}&\mathcal{N}(\mathbf{x};\mathbf{a}_1,\Sigma_1)\mathcal{N}(\mathbf{x};\mathbf{a}_2,\Sigma_2)\\ &=\tilde{c}\exp\left\{-\frac{1}{2}\Big(\|\mathbf{x}-\mathbf{a}_1\|^2_{\Sigma_1^{-1}}+\|\mathbf{x}-\mathbf{a}_2\|^2_{\Sigma_2^{-1}}\Big)\right\}\\ &=\tilde{c}\exp\left\{-\frac{1}{2}\Big(\|\mathbf{x}\|^2_{\Sigma_1^{-1}+\Sigma_2^{-1}}-2(\Sigma_1^{-1}\mathbf{a}_1+\Sigma_2^{-1}\mathbf{a}_2)^{\mathrm{T}}\mathbf{x}\right.\\ &\qquad\left.+\|\mathbf{a}_1\|^2_{\Sigma_1^{-1}}+\|\mathbf{a}_2\|^2_{\Sigma_2^{-1}}\Big\}\right.\\ &=\tilde{c}\exp\left\{-\frac{1}{2}\Big(\|\mathbf{x}\|^2_{\Sigma_3^{-1}}-2\mathbf{a}_3^{\mathrm{T}}\Sigma_3^{-1}\mathbf{x}+\|\mathbf{a}_3\|^2_{\Sigma_3^{-1}}\right.\\ &\qquad\left.-\|\mathbf{a}_3\|^2_{\Sigma_3^{-1}}+\|\mathbf{a}_1\|^2_{\Sigma_1^{-1}}+\|\mathbf{a}_2\|^2_{\Sigma_2^{-1}}\Big\}\right.\\ &=\tilde{c}\exp\left\{-\frac{1}{2}\Big(\|\mathbf{x}-\mathbf{a}_3\|^2_{\Sigma_3^{-1}}-\|\mathbf{a}_3\|^2_{\Sigma_3^{-1}}+\|\mathbf{a}_1\|^2_{\Sigma_1^{-1}}\right.\\ &\qquad\left.+\|\mathbf{a}_2\|^2_{\Sigma_2^{-1}}\Big\}\right.\\ &=c\cdot\mathcal{N}(\mathbf{x};\mathbf{a}_3,\Sigma_3),\end{aligned}$$

where $\tilde{c}$ and c are constants with respect to $\mathbf{x}$, and from [39] we have

$$\begin{aligned}c&=\int_{\mathbb{R}^{n_x}}\mathcal{N}(\mathbf{x};\mathbf{a}_1,\Sigma_1)\mathcal{N}(\mathbf{x};\mathbf{a}_2,\Sigma_2)\mathrm{d}\mathbf{x}\\ &=\mathcal{N}(\mathbf{a}_1;\mathbf{a}_2,\Sigma_1+\Sigma_2).\end{aligned}$$

This completes the proof. □

Based on Lemma 1, we next provide a proof of Theorem 1.

Proof of Theorem 1:
According to Lemma 1 and the definition of the kernel mean, we have

$$\begin{aligned}\rho(\mathbf{z})&=\int_{\mathbb{R}^{n_x}}k(\mathbf{x},\mathbf{z})p(\mathbf{x})\mathrm{d}\mathbf{x}\\ &=\int_{\mathbb{R}^{n_x}}\mathcal{N}(\mathbf{x};\mathbf{z},\Sigma)\mathcal{N}(\mathbf{x};\mathbf{m},P)\mathrm{d}\mathbf{x}\\ &=\mathcal{N}(\mathbf{z};\mathbf{m},\Sigma+P).\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\mathcal{P}(\mathbf{z}_1,\mathbf{z}_2)\\ &=\int_{\mathbb{R}^{n_x}}k(\mathbf{x},\mathbf{z}_1)k(\mathbf{x},\mathbf{z}_2)\mathrm{d}P(\mathbf{x})-\rho(\mathbf{z}_1)\rho(\mathbf{z}_2)\\ &=\int_{\mathbb{R}^{n_x}}\mathcal{N}(\mathbf{x};\mathbf{z}_1,\Sigma)\mathcal{N}(\mathbf{x};\mathbf{z}_2,\Sigma)\mathcal{N}(\mathbf{x};\mathbf{m},P)\mathrm{d}\mathbf{x}\\ &\quad-\rho(\mathbf{z}_1)\rho(\mathbf{z}_2)\\ &=\int_{\mathbb{R}^{n_x}}\mathcal{N}\Big(\mathbf{x};(\mathbf{z}_1+\mathbf{z}_2)/2,\Sigma/2\Big)\mathcal{N}(\mathbf{x};\mathbf{m},P)\mathrm{d}\mathbf{x}\\ &\quad\cdot\mathcal{N}(\mathbf{z}_1;\mathbf{z}_2,2\Sigma)-\rho(\mathbf{z}_1)\rho(\mathbf{z}_2)\\ &=\mathcal{N}(\mathbf{z}_1;\mathbf{z}_2,2\Sigma)\mathcal{N}\big((\mathbf{z}_1+\mathbf{z}_2)/2;\mathbf{m},P+\Sigma/2\big)\\ &\quad-\mathcal{N}(\mathbf{z}_1;\mathbf{m},\Sigma+P)\mathcal{N}(\mathbf{z}_2;\mathbf{m},\Sigma+P).\end{aligned}$$

Note that (14) is repeatedly used in the derivation above. □

Based on Theorem 1, we now provide the analytical expression of the posterior kernel mean estimate $\hat{\mu}$ in the following theorem.

Theorem 2:
The analytical expression of the posterior kernel mean estimate (13) is given as

$$\hat{\mu}(\mathbf{z})=\rho(\mathbf{z})+\sum_{i=1}^{N}w_i\mathcal{P}(\mathbf{z},\mathbf{x}_i),\ \mathbf{z}\in\mathbb{R}^{n_x}, \tag{15}$$

where w_i is the i-th element of the vector

$$\mathbf{w}=G^{-1}Y^{\mathrm{T}}(YG^{-1}KG^{-1}Y^{\mathrm{T}}+\tilde{R})^{-1}(\mathbf{y}-YG^{-1}\rho_X) \tag{16}$$

with $\rho_X=[\rho(\mathbf{x}_1),\rho(\mathbf{x}_2),\ldots,\rho(\mathbf{x}_N)]^{\mathrm{T}}$, and $K\in\mathbb{R}^{N\times N}$ with (i,j)-th element

$$K_{i,j}\coloneqq\mathcal{P}(\mathbf{x}_i,\mathbf{x}_j),\quad 1\le i,j\le N. \tag{17}$$

Proof:
Substituting (8) into (13) yields

$$\begin{aligned}\hat{\mu}(\mathbf{z})&=\phi(\mathbf{z})^{\mathrm{T}}\mathcal{P}\Phi G^{-1}Y^{\mathrm{T}}(YG^{-1}\Phi^{\mathrm{T}}\mathcal{P}\Phi G^{-1}Y^{\mathrm{T}}+R)^{-1}\\ &\quad\cdot(\mathbf{y}-YG^{-1}\Phi^{\mathrm{T}}\rho)+\langle\phi(\mathbf{z}),\rho\rangle_{\mathcal{H}}\\ &=\rho(\mathbf{z})+\phi(\mathbf{z})^{\mathrm{T}}\mathcal{P}\Phi G^{-1}Y^{\mathrm{T}}\\ &\quad\cdot(YG^{-1}KG^{-1}Y^{\mathrm{T}}+\tilde{R})^{-1}(\mathbf{y}-YG^{-1}\rho_N)\\ &=\rho(\mathbf{z})+\phi(\mathbf{z})^{\mathrm{T}}\mathcal{P}\Phi\mathbf{w}\\ &=\rho(\mathbf{z})+\sum_{i=1}^{N}w_i\mathcal{P}(\mathbf{z},\mathbf{x}_i),\end{aligned}$$

which completes the proof. □

Theorem 2 addresses the first limitation outlined in III-C. Equation (15) is called the AKKR for posterior estimation under measurement model (4). To further highlight the distinctive features of the AKKR, its relationships with the existing three representative methods are given as follows:

- *Relationship with Bayes' rule:* The posterior kernel mean estimate $\hat{\mu}$ given by the AKKR is obtained through an additive correction to the prior kernel mean ρ, exhibiting a linear update structure. This structure is

fundamentally different from the classical Bayes' rule, in which the posterior distribution is proportional to the product of the prior and the likelihood. The linearity of the update not only avoids the complexity of multiplicative normalization but also provides a more transparent and computationally efficient way to incorporate new information into the prior.

- *Relationship with Kalman-type estimate:* The structure of (15) is analogous to that of the Kalman-type estimate in which the posterior state estimate is given by the prior mean plus a gain-weighted correction. This resemblance underlies the naming of (15) as the AKKR. However, as clearly illustrated in VII-A, a key difference lies in that the Kalman-type estimate yields a posterior mean estimate only, whereas the AKKR provides an estimate of the entire posterior distribution through the posterior kernel mean estimate, as discussed in Section V.
- *Relationship with the existing kernel mean embedding based estimation methods:* It is evident that (15) differs from the empirical expression (6), in which both the prior and the likelihood are entirely approximated by sample. As shown later, the AKKR enables the derivation of an analytical preimage, which in turn facilitates the efficient construction of a continuous approximation of the posterior PDF, as opposed to one based on discrete samples.

V. PREIMAGE, REGULARIZATION AND ALGORITHM

A. Analytical Preimage

To obtain an approximation of the posterior PDF in the state space, in this subsection, we focus on the preimage recovery from the posterior kernel mean estimate given by the AKKR. Specifically, we aim to derive an analytical form of the function $q(\cdot)$ that satisfies the following equation:

$$\int_{\mathbb{R}^{n_x}} k(\mathbf{z},\mathbf{x})q(\mathbf{x})\mathrm{d}\mathbf{x} = \hat{\mu}(\mathbf{z}),$$

where $q(\mathbf{x})$ serves as an approximation of the posterior PDF. To this end, we first provide the following key lemma.

Lemma 2:
The following equation holds:

$$\mathcal{P}(\mathbf{z},\mathbf{x}_i) = \int_{\mathbb{R}^{n_x}} k(\mathbf{z},\mathbf{x})g_i(\mathbf{x})\mathrm{d}\mathbf{x},$$

where

$$g_i(\mathbf{x}) = \big(\mathcal{N}\big(\mathbf{x};\mathbf{f}_i,\tilde{P}\big) - p(\mathbf{x})\big)\rho(\mathbf{x}_i),$$
$$\mathbf{f}_i = \tilde{P}\big(\Sigma^{-1}\mathbf{x}_i + P^{-1}\mathbf{m}\big),\ \ \tilde{P} = \big(\Sigma^{-1}+P^{-1}\big)^{-1}.$$

Proof:
It follows from Theorem 1 that

$$\begin{aligned}\mathcal{P}(\mathbf{z},\mathbf{x}_i) &= \mathcal{N}(\mathbf{z};\mathbf{x}_i,2\Sigma)\mathcal{N}\big((\mathbf{z}+\mathbf{x}_i)/2;\mathbf{m},P+\Sigma/2\big)\\ &\quad - \mathcal{N}(\mathbf{z};\mathbf{m},\Sigma+P)\mathcal{N}(\mathbf{x}_i;\mathbf{m},\Sigma+P)\\ &= \mathcal{N}(\mathbf{x}_i;\mathbf{m},\Sigma+P)\mathcal{N}(\mathbf{z};\mathbf{f}_i,P_f)\\ &\quad - \mathcal{N}(\mathbf{x}_i;\mathbf{m},P+\Sigma)\mathcal{N}(\mathbf{z};\mathbf{m},P+\Sigma)\\ &= \mathcal{N}(\mathbf{x}_i;\mathbf{m},\Sigma+P)\big(\mathcal{N}(\mathbf{z};\mathbf{f}_i,\Sigma+\tilde{P})\\ &\quad - \mathcal{N}(\mathbf{z};\mathbf{m},P+\Sigma)\big)\\ &= \int_{\mathbb{R}^{n_x}} k(\mathbf{z},\mathbf{x})g_i(\mathbf{x})\mathrm{d}\mathbf{x},\end{aligned}$$

which completes the proof. □

Based on Lemma 2, we then present the analytical preimage of the posterior kernel mean estimate in the following theorem.

Theorem 3:
The preimage of $\hat{\mu}$ exists uniquely and is given by

$$q(\mathbf{x}) = \sum_{i=1}^{N}\tilde{w}_i\mathcal{N}\big(\mathbf{x};\mathbf{f}_i,\tilde{P}\big) + \left(1-\sum_{i=1}^{N}\tilde{w}_i\right)p(\mathbf{x}), \quad (18)$$

where $\tilde{w}_i$ is the i-th element of the vector

$$\tilde{\mathbf{w}} \coloneqq \mathbf{w}\odot\rho_X. \quad (19)$$

Proof:
It follows from Theorem 2 and Lemma 2 that

$$\begin{aligned}\hat{\mu}(\mathbf{z}) &= \rho(\mathbf{z}) + \sum_{i=1}^{N} w_i\mathcal{P}(\mathbf{z},\mathbf{x}_i)\\ &= \rho(\mathbf{z}) + \sum_{i=1}^{N} w_i\int_{\mathbb{R}^{n_x}} k(\mathbf{z},\mathbf{x})g_i(\mathbf{x})\mathrm{d}\mathbf{x}\\ &= \int_{\mathbb{R}^{n_x}} k(\mathbf{z},\mathbf{x})\Big(\mathcal{N}(\mathbf{x};\mathbf{m},P) + \sum_{i=1}^{N} w_i g_i(\mathbf{x})\Big)\mathrm{d}\mathbf{x}\\ &= \int_{\mathbb{R}^{n_x}} k(\mathbf{z},\mathbf{x})q(\mathbf{x})\mathrm{d}\mathbf{x}.\end{aligned}$$

The uniqueness follows from the fact that the Gaussian kernel is characteristic. □

Theorem 3 addresses the second limitation discussed in Section III-C, providing a foundation for efficiently approximating the posterior PDF. Moreover, it reveals that the preimage inherently possesses a Gaussian mixture structure.

While $q(\cdot)$ integrates to 1, it does not necessarily satisfy the non-negativity condition $q(\mathbf{x}) \geq 0$, $\forall \mathbf{x}\in\mathbb{R}^{n_x}$, as the weights $\tilde{w}_i$, $i = 1,2,\ldots,N$ may be negative. This means that, the preimage $q(\cdot)$ of the posterior kernel mean estimate given by (13) may not be a PDF. This issue arises because the linear estimation does not inherently preserve the non-negativity, even if $\phi(\mathbf{x})$ is non-negative. Therefore, regularization of the posterior kernel mean estimate is required.

B. Convex Regularization of Posterior Kernel Mean Estimate

To design an effective regularzation of the posterior kernel mean estimate, we project $\hat{\mu}$ onto the set $\mathcal{K}$ of kernel means of all PDFs. The projection $\hat{\mu}^*$ serves as a regularization for $\hat{\mu}$ and is defined as

$$\hat{\mu}^* = \arg\min_{\nu \in \mathcal{K}} \|\nu - \hat{\mu}\|_{\mathcal{H}}^2.$$

Remark 3:
Note that $\|\nu - \hat{\mu}\|_{\mathcal{H}}$ corresponds to the commonly used maximum mean discrepancy between ν and $\hat{\mu}$ (see, e.g., [27] for details). Therefore, the projection $\hat{\mu}^*$ provides the closest kernel mean of distribution in the sense of maximum mean discrepancy. This guarantees that $\hat{\mu}^*$ preserves the maximum similarity to $\hat{\mu}$ while being a valid posterior kernel mean estimate.

However, due to the infinite-dimensional nature of $\mathcal{K}$, the problem becomes inherently infinite dimensional, posing substantial challenges in obtaining a solution. Guided by the inherent structure of $\hat{\mu}$ given by (15), we instead project $\hat{\mu}$ onto a finite dimensional subspace defined as

$$\tilde{\mathcal{K}} = \Big\{ \sum_{i=1}^{N+1} a_i \nu_i; \sum_{i=1}^{N+1} a_i = 1, a_i \geq 0, i = 1, 2, \ldots, N+1 \Big\},$$

where ν_i represents $\mathcal{N}(\mathbf{x}; \mathbf{f}_i, \tilde{P} + \Sigma)$ for $i = 1, 2, \ldots, N$ and ν_{N+1} represents $\mathcal{N}(\mathbf{x}; \mathbf{m}, P + \Sigma)$. Then, the following finite-dimensional optimization problem is formulated:

$$\hat{\mu}^* = \arg\min_{\nu \in \tilde{\mathcal{K}}} \|\nu - \hat{\mu}\|_{\mathcal{H}}^2, \tag{20}$$

whose optimal solution is provided in the following theorem.

Theorem 4:
The optimal solution of problem (20) is

$$\hat{\mu}^* = \sum_{i=1}^{N} a_i^* \mathcal{N}(\mathbf{x}; \mathbf{f}_i, \tilde{P} + \Sigma) + a_{N+1}^* \rho(\mathbf{x}),$$

with $\mathbf{a}^* = [a_1^*, a_2^*, \ldots, a_{N+1}^*]^{\mathrm{T}}$ being the optimal solution of the following convex quadratic programming problem:

$$\begin{aligned} \min_{\mathbf{a} \in \mathbb{R}^{N+1}} \ & \mathbf{a}^{\mathrm{T}} J \mathbf{a} - 2 \underline{\mathbf{w}}^{\mathrm{T}} J \mathbf{a} \\ \text{s.t.} \ & \sum_{i=1}^{N+1} a_i = 1, \\ & a_i \geq 0, \ 1 \leq i \leq N+1, \end{aligned} \tag{21}$$

where $\underline{\mathbf{w}} = [\tilde{\mathbf{w}}^{\mathrm{T}}, 1 - \sum_{i=1}^{N} \tilde{\mathbf{w}}_i]^{\mathrm{T}}$ and the semi-positive definite matrix J with elements

$$J_{i,j} = \begin{cases} \mathcal{N}(\mathbf{f}_i; \mathbf{f}_j, 2\tilde{P} + \Sigma) & \text{if } 1 \leq i \leq j \leq N, \\ \mathcal{N}(\mathbf{f}_i; \mathbf{m}, \tilde{P} + P + \Sigma) & \text{if } 1 \leq i \leq N, j = N+1, \\ \mathcal{N}(\mathbf{m}; \mathbf{m}, 2P + \Sigma) & \text{if } i = j = N+1. \end{cases}$$

Proof:
Incorporating the constraint $\nu \in \tilde{\mathcal{K}}$, the objective function in (20) becomes

$$\begin{aligned} \|\nu - \hat{\mu}\|_{\mathcal{H}}^2 &= \Big\| \sum_{i=1}^{N+1} (a_i - \underline{w}_i) \nu_i \Big\|_{\mathcal{H}}^2 \\ &= \sum_{i=1}^{N+1} \sum_{i=j}^{N+1} (a_i - \underline{w}_i)(a_j - \underline{w}_j) \langle \nu_i, \nu_j \rangle_{\mathcal{H}} \\ &= \|\mathbf{a} - \underline{\mathbf{w}}\|_J^2, \end{aligned}$$

where $\underline{w}_i$ denotes the i-th element of $\underline{\mathbf{w}}$ and $J \in \mathbb{R}^{(N+1)\times(N+1)}$ is the matrix with its (i,j)-th element $J_{i,j} = \langle \nu_i, \nu_j \rangle_{\mathcal{H}}$ for $1 \leq i, j \leq N+1$. Ignoring constant terms, the objective function can be expressed as $\mathbf{a}^{\mathrm{T}} J \mathbf{a} - 2 \underline{\mathbf{w}}^{\mathrm{T}} J \mathbf{a}$.

Next, we show that (21) is convex quadratic programming problem by verifying that J is semi-positive definite. The symmetry of J is evident. For any $b_i \in \mathbb{R}$ with $i = 1, 2, \ldots, N+1$, we have

$$\begin{aligned} \sum_{i=1}^{N+1} \sum_{j=1}^{N+1} b_i b_j J_{i,j} &= \Big\langle \sum_{i=1}^{N+1} b_i \nu_i, \sum_{j=1}^{N+1} b_j \nu_j \Big\rangle_{\mathcal{H}} \\ &= \Big\| \sum_{i=1}^{N+1} b_i \nu_i \Big\|_{\mathcal{H}}^2 \geq 0. \end{aligned}$$

Thus, J is semi-positive definite.

Finally, we given the expressions of elements of J. If $1 \leq i, j, \leq N$, it follows from Lemma 1 that

$$\begin{aligned} \langle \nu_i, \nu_j \rangle &= \int k(\mathbf{x}_1, \mathbf{x}_2) \mathcal{N}(\mathbf{x}_1; \mathbf{f}_i, \tilde{P}) \mathcal{N}(\mathbf{x}_2; \mathbf{f}_j, \tilde{P}) \mathrm{d}\mathbf{x}_1 \mathrm{d}\mathbf{x}_2 \\ &= \int \mathcal{N}(\mathbf{x}_2; \mathbf{f}_i, \Sigma + \tilde{P}) \mathcal{N}(\mathbf{x}_2; \mathbf{f}_j, \tilde{P}) \mathrm{d}\mathbf{x}_2 \\ &= \mathcal{N}(\mathbf{f}_i; \mathbf{f}_j, 2\tilde{P} + \Sigma). \end{aligned}$$

Since the calculations for the remaining two cases follow similar arguments, they are omitted here. □

Remark 4:
For the convex quadratic optimization problem (21), there are several mature solvers available that provide efficient algorithms, such as the CVX (see, e.g., [40]).

Theorem 4 guarantees the validity of the estimate and prevents the accumulation of inconsistencies during filtering, thereby addressing the third limitation identified in III-C.

Based on Theorem 4, we provide the preimage of $\hat{\mu}^*$ presented in (20), as stated in the following theorem.

Theorem 5:
The unique preimage of $\hat{\mu}^*$ is given by

$$q^*(\mathbf{x}) = \sum_{i=1}^{N} a_i^* \mathcal{N}(\mathbf{x}; \mathbf{f}_i, \tilde{P}) + a_{N+1}^* p(\mathbf{x}). \tag{22}$$

Proof:
Using Theorem (1) and the characteristic property of Gaussian kernel, it is straightforward to derive (22). □

Remark 5:
It is worth noting that although q^* consists of $N+1$ Gaussian components, N of them share a common covariance matrix. Therefore, in subsequent statistical inference processes, such as during the filtering prediction step in Section VI, where the covariance matrix needs to be decomposed, it does not incur significant computational cost.

Theorem 5 provides a posterior PDF approximation in the form of Gaussian mixture, which offers several advantages over discrete sample-based methods such as Monte Carlo methods and some existing kernel mean embedding approaches. These advantages are summarized as follows:

- *Suitability for statistical inference:* Gaussian mixture distributions inherit many desirable properties of Gaussian PDFs, such as efficient and accurate integral approximations. These properties make them well-suited for statistical inference tasks, including moment computation and nonlinear transformations. For example, this property is leveraged in the prediction step of the proposed filter in VI.
- *Avoidance of particle degeneracy:* Discrete sample based methods rely on resampling to mitigate particle degeneracy, which can lead to sample impoverishment (loss of diversity). Gaussian mixtures do not require resampling, thereby avoiding this bottleneck. This feature makes them particularly suitable for filtering, as illustrated by in VII-B and VII-C.
- *Reduced computational cost:* Gaussian mixtures have greater representational capability compared with discrete sample when the number of mixture components equals that of sample points. Thus, to approximate the same posterior PDF, Gaussian mixtures require significantly fewer components to achieve much higher approximation accuracy. This substantially reduces computational complexity, as demonstrated in VII-C.

C. AKKR Based Posterior PDF Approximation Algorithm

In the minimum mean squared error (MSE) sense, the optimal estimate of $\mathbf{x}$ and its MSE matrix are given by the mean and covariance matrix of the posterior PDF $p(\mathbf{x}|\mathbf{y})$, respectively. Thus, according to Theorem 5, the state estimate and its MSE matrix obtained by AKKR are given as

$$\hat{\mathbf{x}} = \sum_{i=1}^{N} a_i^* \mathbf{f}_i + a_{N+1}^* \mathbf{m}, \tag{23}$$

$$\hat{P} = \sum_{i=1}^{N} a_i^* (\tilde{P} + \mathbf{f}_i \mathbf{f}_i^{\mathrm{T}}) + a_{N+1}^* (P + \mathbf{m}\mathbf{m}^{\mathrm{T}}) - \hat{\mathbf{x}}\hat{\mathbf{x}}^{\mathrm{T}}. \tag{24}$$

The proposed AKKR based posterior PDF approximation algorithm is summarized in Algorithm 1.

Algorithm 1 AKKR based posterior PDF approximation algorithm

Input: h, R, $\mathbf{y}$, $\mathbf{m}$, P, N.
1: Generate two independent and identically distributed samples from $\mathcal{N}(\mathbf{m}, P)$: $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\tilde{\mathbf{x}}_j\}_{j=1}^m$.
2: Calculate $\tilde{R}$ using (12) and $\tilde{\mathbf{w}}$ using (19), where $\mathbf{w}$, Y, G and K are given by (16), (7), (9) and (17), respectively.
3: **if** $\underline{\mathbf{w}} = [\tilde{\mathbf{w}}^{\mathrm{T}}, 1 - \sum_{i=1}^N \tilde{\mathbf{w}}_i]^{\mathrm{T}} \succeq 0$ **then**
4: The posterior PDF $q^*(\mathbf{x})$ is given by (18).
5: **else**
6: Solve the convex quadratic programming problem (21), and then obtain the posterior PDF $q^*(\mathbf{x})$ given by (22).
7: **end if**
8: Calculate the posterior state estimate $\hat{\mathbf{x}}$ using (23) and MSE matrix $\hat{P}$ using (24), respectively.
Output: $q^*(\mathbf{x})$, $\hat{\mathbf{x}}$, $\hat{P}$.

Remark 6:
It is worth noting that, since a Gaussian mixture preserves the analytical moment properties of the Gaussian distribution, the AKKR can be readily extended to cases where the prior PDF is a Gaussian mixture. Consequently, Algorithm 1 can be extended to scenarios where the prior PDF $p(\mathbf{x})$ is a Gaussian mixture without further technical difficulties, significantly broadening its applicability.

Remark 7:
For numerical stability in algorithm implementation and appropriate selection of the parameter Σ in kernel function, we consider the following variable substitutions:

$$\breve{\mathbf{x}} = A\mathbf{x},\ \breve{\mathbf{y}} = B\mathbf{y} + \mathbf{d},\ \breve{\mathbf{v}} = B\mathbf{v} + \mathbf{d},$$
$$\breve{h}(\cdot) = Bh\big(A^{-1}(\cdot)\big) + \mathbf{d},$$

where A, B and $\mathbf{d}$ are chosen such that $\breve{\mathbf{x}}$ and $\breve{h}(\breve{\mathbf{x}})$ are approximately zero-mean and have identity covariance matrices. Under these substitutions, the measurement model (4) can be equivalently rewritten as

$$\breve{\mathbf{y}} = \breve{h}(\breve{\mathbf{x}}) + \breve{\mathbf{v}}.$$

In this case, we set the parameter matrix Σ to

$$\Sigma = \Big(\frac{4}{n_x + 2}\Big)^{\frac{2}{n_x+4}} N^{\frac{-2}{n_x+4}} I,$$

and then employ Algorithm 1 to compute the posterior PDF approximation of $\breve{\mathbf{x}}$:

$$\tilde{q}^*(\breve{\mathbf{x}}) = \sum_{i=1}^{N+1} \tilde{a}_i^* \mathcal{N}(\breve{\mathbf{x}}; \mathbf{m}_i, P_i).$$

Finally, using the linear equation $\breve{\mathbf{x}} = A\mathbf{x}$, the posterior PDF approximation of $\mathbf{x}$ is obtained as follows

$$q^*(\mathbf{x}) = \sum_{i=1}^{N+1} \tilde{a}_i^* \mathcal{N}(\mathbf{x}; A^{-1}\mathbf{m}_i, A^{-1} P_i A^{-\mathrm{T}}).$$

VI. ANALYTICAL KERNAL KALMAN FILTER

Consider the following nonlinear stochastic dynamic system:

$$\begin{aligned} \mathbf{x}_{k+1} &= f_k(\mathbf{x}_k) + \mathbf{w}_k, \\ \mathbf{y}_k &= h_k(\mathbf{x}_k) + \mathbf{v}_k, \end{aligned} \tag{25}$$

where k is the time index, $\mathbf{x}_k \in \mathbb{R}^{n_x}$ and $\mathbf{y}_k \in \mathbb{R}^{n_y}$ are the state and the measurement, respectively, the state transition function $f_k : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ and the measurement function $h_k : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ are both known nonlinear functions, $\{\mathbf{w}_{k-1}\}$ and $\{\mathbf{v}_k\}$ are mutually independent Gaussian white noise sequences with mean zero and covariance matrices $Q_{k-1} \succeq 0$ and $R_k \succ 0$, respectively.

For each time step k, the proposed AKKF consists of two steps: prediction and update, as specified as follows.

1) Prediction. Suppose that the posterior PDF at time step $k-1$ is given as

$$p(\mathbf{x}_{k-1}|Y_{k-1}) = \sum_{i=1}^{N_{k-1}} \alpha_{i,k-1}\mathcal{N}(\mathbf{x}_{k-1}; \mathbf{m}_{i,k-1}, P_{i,k-1}).$$

Then, a commonly used method for obtaining the estimate of $\mathbf{x}_k$ and the corresponding MSE matrix is the following cubature rule [9]:

$$\mathbf{x}_{k|k-1} = \frac{1}{2n_x}\sum_{i=1}^{N_{k-1}}\sum_{j=1}^{2n_x} \alpha_{i,k-1} f_{k-1}(\mathbf{x}_{i,j,k-1}), \tag{26}$$

$$\begin{aligned} P_{k|k-1} = \frac{1}{2n_x}\sum_{i=1}^{N_{k-1}}\sum_{j=1}^{2n_x} \alpha_{i,k-1} f_{k-1}(\mathbf{x}_{i,j,k-1}) \\ \cdot f_{k-1}(\mathbf{x}_{i,j,k-1})^{\mathrm{T}} + Q_{k-1} - \mathbf{x}_{k|k-1}\mathbf{x}_{k|k-1}^{\mathrm{T}}, \end{aligned} \tag{27}$$

where $\mathbf{x}_{i,j,k-1}$ is the j-th column of $\mathbf{m}_{i,k-1}\mathbf{1}^{\mathrm{T}} + [S_{i,k-1}, -S_{i,k-1}]$ with $S_{i,k-1} = \mathrm{chol}(n_x P_{i,k-1})$, for $i = 1, 2, \ldots, N$. Then, the prior PDF of state at time step k $p(\mathbf{x}_k|Y_{k-1})$ is approximated by $\mathcal{N}(\mathbf{x}_k; \mathbf{x}_{k|k-1}, P_{k|k-1})$.

2) Update. By replacing the prior PDF $\mathcal{N}(\mathbf{m}, P)$ in (4) with $\mathcal{N}(\mathbf{x}_k; \mathbf{x}_{k|k-1}, P_{k|k-1})$ and h in (4) with h_k, and then adopting Algorithm 1, we can obtain the approximated posterior PDF $p(\mathbf{x}_k|Y_k)$, state estimate $\mathbf{x}_{k|k}$ and its MSE matrix $P_{k|k}$.

One cycle of the proposed AKKF is summarized in Algorithm 2.

Algorithm 2 AKKF

Input: $p(\mathbf{x}_k|Y_{k-1})$, f_{k-1}, h_k, $\mathbf{y}_k$, Q_{k-1}, R_k, N.

I. Prediction

1: Calculate the state estimate $\mathbf{x}_{k|k-1}$ and its MSE matrix $P_{k|k-1}$ using (26) and (27), respectively.

II. Update

2: Implement Algorithm 1, where the inputs are h_k, R_k, $\mathbf{y}_k$, $\mathbf{x}_{k|k-1}$, $P_{k|k-1}$, N, to obtain the posterior PDF $p(\mathbf{x}_k|Y_k)$, posterior state estimate $\mathbf{x}_{k|k}$ and its MSE matrix $P_{k|k}$.

Output: $p(\mathbf{x}_k|Y_k)$, $\mathbf{x}_{k|k}$, $P_{k|k}$.

Remark 8:
In the prediction step, the Gaussian approximation is adopted for simplicity. In fact, according to Remark 6, it can be extended to the Gaussian mixture approximation in the prediction step to improve state estimation accuracy.

VII. EXAMPLES

To demonstrate the effectiveness of Algorithm 1, an example for complex posterior distribution approximation is provided in VII-A. To demonstrate the effectiveness of Algorithm 2, two examples, including a univariate non-stationary growth model and a target tracking scenario, are provided in VII-B and VII-C, respectively.

Some better-known nonlinear filters are used for comparison, including the cubature Kalman filter (CKF) [9], and the sampling importance resampling filter, which is a particle filter (PF) [41]. Moreover, the recent kernel embedding based method in [22] is also adopted for comparison. All the filtering algorithms were implemented using Octave on a device with an Intel Core i7 2.60 GHz processor.

A. Complex Posterior Distribution Approximation

Consider the following measurement model:

$$y = h(x) + v,$$

where $x, y \in \mathbb{R}$, v is a Gaussian noise with zero mean and variance $\sigma_v = 2$, and the prior distribution of x is $\mathcal{N}(0, 1)$. In this example, we assume that a realization of the measurement y is 3. Then, the posterior PDF of x is

$$p(x|y) \propto \exp\{-(h(x) - 3)^2/4 - x^2/2\}.$$

To demonstrate the effectiveness of Algorithm 1 under different cases, we present results for different choices of h with varying degrees of nonlinearity. Besides, we set $N = 40$ points in Algorithm 1.

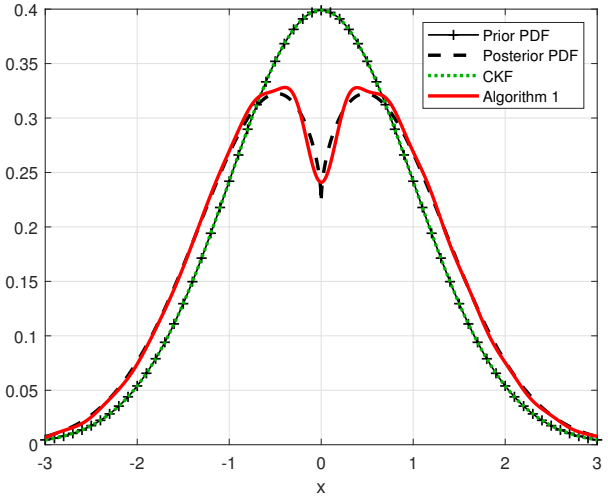

Fig. 1. $h(x) = 0.5\sqrt{|x|}$

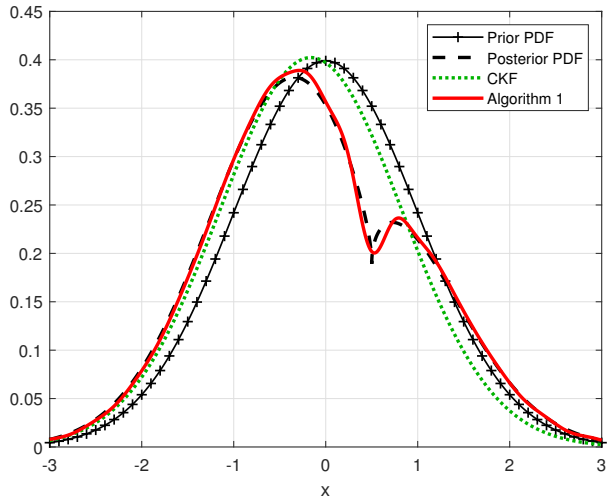

Fig. 2. $h(x) = 0.5\sqrt{|x-0.2|}$

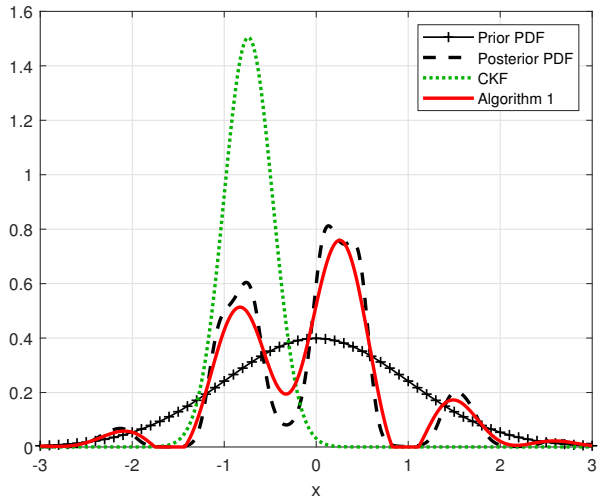

Fig. 4. $h(x) = \cos(2x+0.5) + 2\sin(5x) + 1$

In Figs. 1 and 2, we set $h = 0.5\sqrt{|x+c|}$ with $c = 0$ and $c = -0.2$, respectively. Since CKF estimates the first two moments of the posterior distribution only, its estimate results are represented as Gaussian PDFs. Although CKF and the proposed AKKR both originate from the linear minimum mean squared error estimation, the latter estimates the posterior kernel mean in RKHS rather than the posterior mean in the state space. As a result, Algorithm 1 captures more statistical information about the posterior distribution, leading to more accurate posterior PDF approximation. As demonstrated by the Figs. 1 and 2, Algorithm 1 presents superior performance in approximating bimodal posterior PDFs.

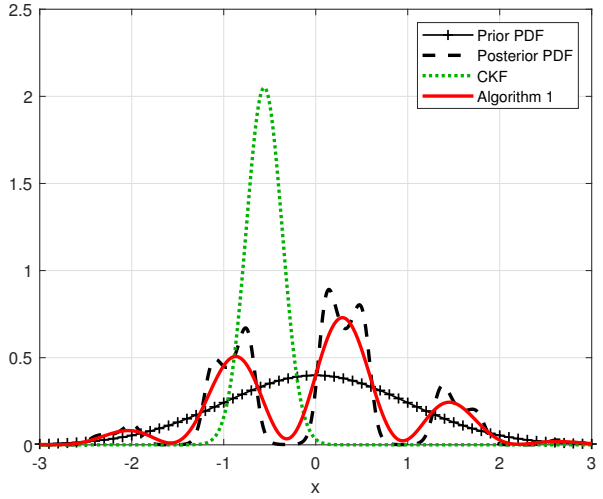

Fig. 3. $h(x) = 3\sin(5x) + 1$

In Figs. 3 and 4, we set $h = a\cos(2x+0.5) + b\sin(5x) + 1$ with (a, b) being $(0, 3)$ and $(1, 2)$, respectively, to demonstrate the capability of Algorithm 1 in handling highly nonlinear systems. We can see that CKF approximates one of the modes of the posterior PDF, while AKKR closely matches the posterior PDF across the entire domain. This highlights the superior performance of AKKR in handling highly nonlinear systems. Additionally, it is clearly observed in Fig. 4 that the posterior PDF approximation provided by Algorithm 1 is nearly zero around $x = -1.5$ and $x = 1$. In fact, before applying the convex regularization in V-B, these values were negative. This observation supports the statement in V-A that the preimage of the posterior kernel mean estimate is not necessarily a valid PDF, and demonstrates the effectiveness of the proposed regularization as well.

B. Univariate Non-Stationary Growth Model

The univariate non-stationary growth model has been widely studied in the literature (see, e.g., [12], [41]–[44]) and serves as a suitable test case for the proposed filter due to its strong nonlinearity. The nonlinear stochastic dynamic system, whose state transition follows the univariate non-stationary growth model, is given as follows:

$$x_k = \frac{x_{k-1}}{2} + \frac{25x_{k-1}}{1+x_{k-1}^2} + 8\cos\big(1.2(k-1)\big) + w_{k-1},$$

$$y_k = \frac{x_k^2}{20} + v_k,$$

where w_k and v_k are Gaussian noises with zero mean and variances $\sigma_w^2 = 10$ and $\sigma_v^2 = 0.1$, respectively. The initial state follows the Gaussian distribution with zero mean and variance 1.

In this example, the PF uses 50 particles, and both the method in [22] and the proposed AKKF employ 5 points for measurement function approximation, ensuring a fair comparison under comparable computational cost. A PF with 500 particles is included as a performance benchmark.

Fig. 5 presents the root mean squared errors (RMSEs) of the compared filters over 500 Monte Carlo runs. The proposed AKKF achieves the best estimation performance among the efficient methods, reducing the average RMSE

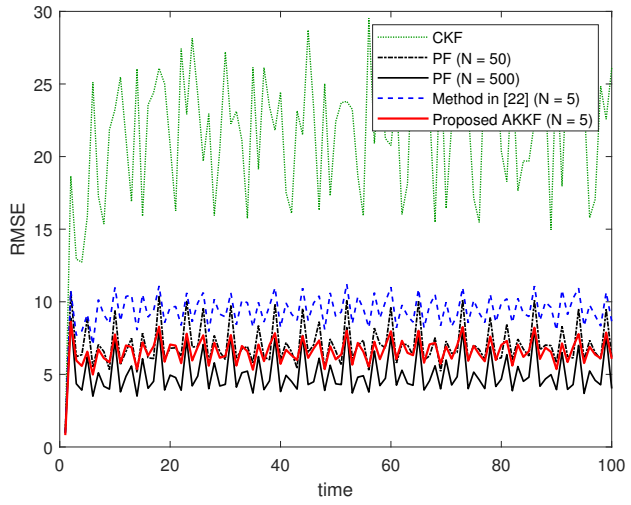

Fig. 5. RMSEs of compared filters

by 6.78% compared to PF with 50 particles. The PF with 500 particles yields the highest estimation accuracy due to the increased number of particles, and serves as a performance benchmark. However, its computational cost is approximately 12 times that of the proposed AKKF. Both the method in [22] and the proposed AKKF outperform CKF, demonstrating the advantage of kernel mean embedding based approaches. The relatively lower accuracy of CKF may stem from its limited ability to approximate multimodal posterior distributions, as it estimates only the first two moments, whereas the other three methods are based on distribution approximation.

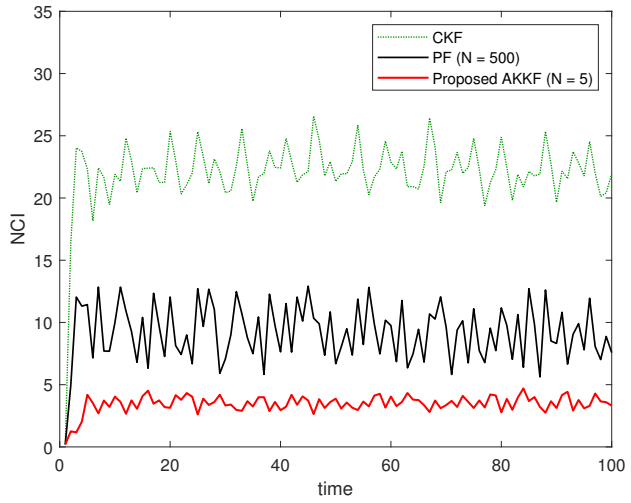

Fig. 6. NCIs of CKF and proposed AKKF.

As analyzed convincingly in [45], the noncredibility index (NCI) is used to assess the reliability of the filter's estimate by comparing the filter's MSE matrix $P_{k|k}$ with the true one; the closer the NCI is to zero, the more accurate the filter's MSE matrix $P_{k|k}$ is. Fig. 6 presents the NCIs of CKF, PF with 500 particles, and the proposed AKKF over 500 Monte Carlo runs. Among these, the proposed AKKF exhibits the most accurate MSE matrices. It is worth noting that the NCIs of PF with 5 particles and the method in [22] are not shown here because they are infinite in this example. This phenomenon arises from the degeneracy of these two methods, where the probability mass collapses onto a single sample point, causing the estimated MSE matrix to be nearly zero and resulting in infinite NCIs. This also highlights the ability of the proposed AKKF to effectively mitigate the degeneracy problem.

TABLE II
Average RMSEs of Proposed AKKF

	N			
(σ_w^2, σ_v^2)	3	5	10	15
$(1, 5)$	7.1758	6.7677	6.1674	6.0926
$(5, 1)$	8.0276	7.3599	6.5313	6.4657
$(5, 5)$	8.0985	7.4326	7.1014	6.9579
$(1, 10)$	7.2012	6.7808	6.6662	6.4030
$(10, 1)$	8.2994	7.6594	7.0404	6.8023
$(10, 10)$	8.3482	7.6797	7.3076	7.2105

Table II presents the average RMSEs of the proposed AKKF over 500 Monte Carlo runs under different parameter settings. With (σ_w^2, σ_v^2) fixed, it can be observed that the estimation accuracy improves as the number of points N increases. Besides, when N remain unchanged, the estimation accuracy is inversely proportional to the values of σ_w^2 and σ_v^2, which aligns with intuition.

C. Target Tracking

As in [12], [22], we consider a target moving in a plane with a nearly constant velocity. The state transition model is given by

$$\mathbf{x}_k = F_{k-1}\mathbf{x}_{k-1} + G_{k-1}\mathbf{w}_{k-1},$$

where $\mathbf{x}_k = [x_k, \dot{x}_k, y_k, \dot{y}_k]^{\mathrm{T}}$, $[x_k, y_k]^{\mathrm{T}}$ and $[\dot{x}_k, \dot{y}_k]^{\mathrm{T}}$ are the position and velocity of the target, respectively, and

$$F_{k-1} = \text{block-diag}(F, F), G_{k-1} = \text{block-diag}(G, G),$$
$$F = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}, G = \begin{bmatrix} \Delta t^2/2 \\ \Delta t \end{bmatrix},$$

with the sampling period $\Delta t = 0.2$s, the process noise follows the Gaussian distribution $\mathcal{N}(0, Q_{k-1})$ with $Q_{k-1} = GG^T(\mathrm{m}^2/\mathrm{s}^2)^2$. The initial state is generated from the Gaussian distribution $\mathcal{N}(\bar{\mathbf{x}}_0, P_0)$ with $\bar{\mathbf{x}}_0 = [1000\mathrm{m}, -40\mathrm{m/s}, 1000\mathrm{m}, -40\mathrm{m/s}]^{\mathrm{T}}$ and $P_0 = \mathrm{diag}([10^5\mathrm{m}^2, 100\mathrm{m}^2/\mathrm{s}^2, 10^5\mathrm{m}^2, 100\mathrm{m}^2/\mathrm{s}^2]^{\mathrm{T}})$.

The measurement model is

$$\mathbf{y}_k = \begin{bmatrix} \sqrt{x_k^2 + y_k^2} \\ \frac{x_k\dot{x}_k + y_k\dot{y}_k}{\sqrt{x_k^2+y_k^2}} \end{bmatrix} + \mathbf{v}_k,$$

where the measurement noise $\{\mathbf{v}_k\}$ is the zero-mean Gaussian white noise sequence with covariance matrix

$R_k = \mathrm{diag}([100\mathrm{m}^2, 10\mathrm{m}^2/\mathrm{s}^2]^{\mathrm{T}})$. We set 4000, 10000, and 20000 particles in the PF and $N = 20, 30, 40$ points in the proposed AKKF.

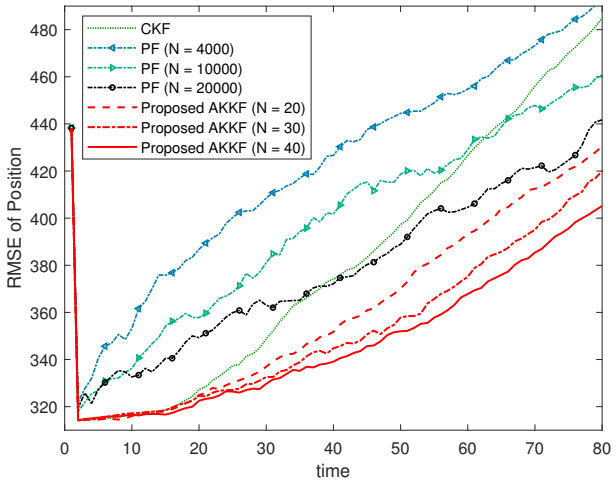

Fig. 7. RMSEs of position estimates (m).

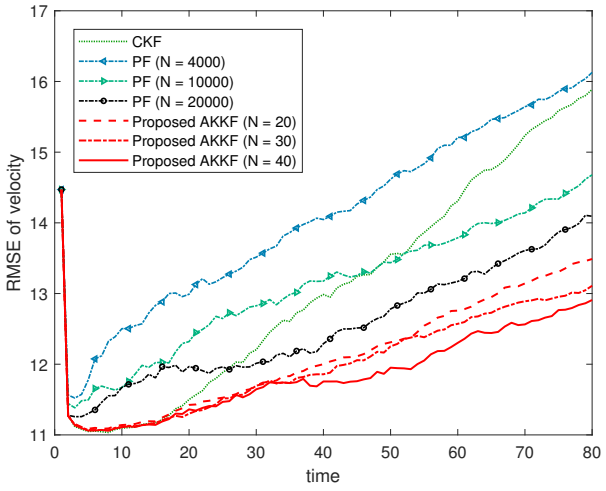

Fig. 8. RMSEs of velocity estimates (m/s).

The RMSEs of position and velocity estimates over 80 time steps and 500 Monte Carlo runs are shown in Figs. 7 and 8. It can be seen that the proposed AKKF has superior estimation accuracy among all the methods and exhibits an increasing trend in estimation accuracy as the number of sample points N increases. In addition, the RMSEs for all the filtering methods increase over time. This phenomenon arises from that the measurements only provide the range and range rate of the target relative to the sensor, which results in limited information.

Fig. 9 presents the NCIs of the filters over 500 Monte Carlo runs. As expected, the proposed AKKF significantly outperforms CKF and PF in terms of reliability, with its NCI remaining stable over time. Furthermore, the performance of the proposed AKKF improves as the number N of points increases.

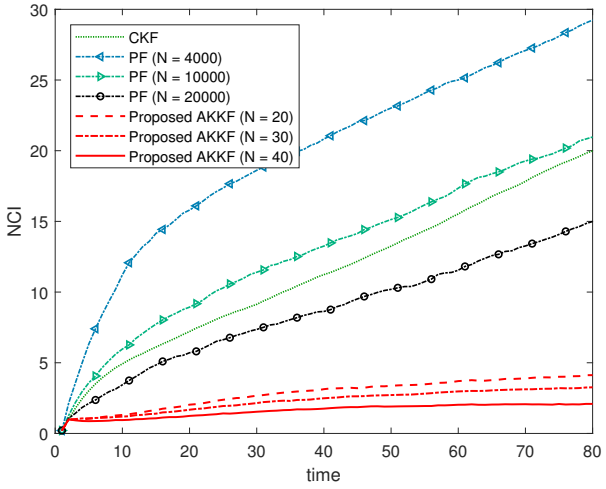

Fig. 9. NCIs of the compared filters.

TABLE III
Average Relative Computation Time

Method	Time
CKF	1.00
Proposed AKKF ($N = 20$)	39.25
Proposed AKKF ($N = 30$)	52.36
Proposed AKKF ($N = 40$)	95.71
PF ($N = 4000$)	127.74
PF ($N = 10000$)	354.64
PF ($N = 20000$)	762.65

Table III reports the relative computation time of the filters averaged over 500 Monte Carlo runs. It can be seen that CKF has high computational efficiency, as it requires simple matrix operations only. As the number of sample points N increases, the proposed AKKF requires more computation time, mainly due to the necessity of solving a convex quadratic programming problem of size $N + 1$ at specific time steps. Nevertheless, the proposed AKKF remains more efficient than PF. Additionally, combining Figs. 7–9, it is evident that the proposed AKKF provides superior estimation accuracy and reliability with lower computational cost, achieving a balance between estimation accuracy and computational cost.

VIII. CONCLUSION

This paper has developed the AKKR for nonlinear filtering in a model-driven setting, eliminating the need for empirical approximations through deriving analytical expression of the posterior kernel mean estimate. Then, the analytical preimage with a Gaussian mixture structure has been obtained, enabling a convex quadratic programming based regularization for valid and efficient posterior PDF approximation. Building on these results, the AKKR based posterior PDF approximation and AKKF have

been established. Three representative examples have confirmed the effectiveness and accuracy of the proposed algorithms across a range of nonlinear scenarios.

REFERENCES

[1] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms, and Software*. New York, NY, USA: John Wiley & Sons, 2001.

[2] T. Badar, S. Särkkä, Z. Zhao, and A. Visala, "Rao–Blackwellized particle filter using noise adaptive Kalman filter for fully mixing state-space models," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 5, pp. 6972–6982, 2024.

[3] M. S. Grewal, A. P. Andrews, and C. G. Bartone, *Global Positioning Systems, Inertial Navigation, and Integration*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2013.

[4] D. Ma, T. Ma, Y. Li, Y. Ling, and Y. Ben, "A long-term underwater robust terrain contour-aided navigation method with a single beam echo sounder," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 5, pp. 5801–5811, 2024.

[5] N. Meskin and K. Khorasani, *Fault Detection and Isolation: Multi-Vehicle Unmanned Systems*. New York, NY, USA: Springer, 2011.

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

[7] P. Setoodeh, S. Habibi, and S. Haykin, *Nonlinear filters: Theory and Applications*. Hoboken, NJ, USA: John Wiley & Sons, 2022.

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

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

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

[11] L. Guo, S. Hu, J. Zhou, and X. Rong Li, "Nonlinear filtering based on third-degree approximation via sequential homotopy," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5334–5348, 2022.

[12] J. Lan and X. Rong Li, "Multiple conversions of measurements for nonlinear estimation," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4956–4970, 2017.

[13] Y. Tao and S. S.-T. Yau, "Outlier-robust iterative extended Kalman filtering," *IEEE Signal Processing Letters*, vol. 30, pp. 743–747, 2023.

[14] A. F. Garcia-Fernandez, 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] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, pp. 1805–1819, 2005.

[16] Y. Lin, L. Miao, and Z. Zhou, "An improved MCMC-based particle filter for GPS-aided SINS in-motion initial alignment," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7895–7905, 2020.

[17] M. Tian, Z. Chen, H. Wang, and L. Liu, "An intelligent particle filter for infrared dim small target detection and tracking," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 6, pp. 5318–5333, 2022.

[18] L. Dai and F. Daum, "On the design of stochastic particle flow filters," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 3, pp. 2439–2450, 2023.

[19] X. Chen and Y. Li, "Normalizing flow-based differentiable particle filters," *IEEE Transactions on Signal Processing*, vol. 73, pp. 493–507, 2025.

[20] A. H. Chughtai, A. Majal, M. Tahir, and M. Uppal, "Variational-based nonlinear Bayesian filtering with biased observations," *IEEE Transactions on Signal Processing*, vol. 70, pp. 5295–5307, 2022.

[21] 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.

[22] L. Guo, J. Wang, Y. Zhao, and J.-F. Zhang, "Consensus-based distributed nonlinear filtering with kernel mean embedding," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 61, no. 2, pp. 4973–4987, 2025.

[23] M. L. Psiaki, "Gaussian mixture nonlinear filtering with resampling for mixand narrowing," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5499–5512, 2016.

[24] L. Yu, T. Yang, and A. B. Chan, "Density-preserving hierarchical EM algorithm: Simplifying Gaussian mixture models for approximate inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1323–1337, 2019.

[25] P. H. Leong, S. Arulampalam, T. A. Lamahewa, and T. D. Abhayapala, "Gaussian-sum cubature Kalman filter with improved robustness for bearings-only tracking," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 513–517, 2014.

[26] S. Hu, L. Guo, M. Tang, and Y. Rong, "Split-and-merge based simultaneous input and state filter for nonlinear dynamic systems," in *2022 China Automation Congress (CAC)*, Xiamen, China, Nov. 2022, pp. 5472–5476.

[27] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1–2, pp. 1–141, 2017.

[28] L. Song, K. Fukumizu, and A. Gretton, "Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 98–111, 2013.

[29] K. Fukumizu, L. Song, and A. Gretton, "Kernel Bayes' rule: Bayesian inference with positive definite kernels," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3753–3783, 2013.

[30] M. Kanagawa, Y. Nishiyama, A. Gretton, and K. Fukumizu, "Filtering with state-observation examples via kernel Monte Carlo filter," *Neural Computation*, vol. 28, no. 2, pp. 382–444, 2016.

[31] M. Sun, M. E. Davies, I. K. Proudler, and J. R. Hopgood, "Adaptive kernel Kalman filter," *IEEE Transactions on Signal Processing*, vol. 71, pp. 713–726, 2023.

[32] G. H. W. Gebhardt, A. Kupcsik, and G. Neumann, "The kernel Kalman rule: Efficient nonparametric inference by recursive least-squares and subspace projections," *Machine Learning*, vol. 108, no. 12, pp. 2113–2157, 2019.

[33] L. Song, X. Zhang, A. Smola, A. Gretton, and B. Schölkopf, "Tailoring density estimation via reproducing kernel moment matching," in *Proceedings of the 25th Annual International Conference on Machine Learning*, New York, NY, USA, Jul.05–09, 2008, pp. 992–999.

[34] M. Welling, "Herding dynamical weights to learn," in *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, QC, CA, Jun. 14–18, 2009, pp. 1121–1128.

[35] F. Bach, S. Lacoste-Julien, and G. Obozinski, "On the equivalence between herding and conditional gradient algorithms," in *Proceedings of the 29th Annual International Conference on Machine Learning*, Edinburgh, UK, Jun. 26–01, 2012, pp. 1355–1362.

[36] S. Lacoste-Julien, F. Lindsten, and F. Bach, "Sequential kernel herding: Frank-Wolfe optimization for particle filtering," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, San Diego, CA, USA, May 09–12, 2015, pp. 544–552.

[37] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Canada, Jun. 14–18, 2009, pp. 961–968.

[38] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.
[39] H. Zhu, C. Han, and Y. Lin, "A reduced Gaussian mixture representation based on sparse modeling," in *15th International Conference on Information Fusion*, Singapore, Jul. 09–12, 2012, pp. 684–691.
[40] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," https://cvxr.com/cvx, Mar. 2014.
[41] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.
[42] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proceedings F Radar and Signal Processing*, vol. 140, no. 2, pp. 107–113, 1993.
[43] J. Kotecha and P. Djuric, "Gaussian sum particle filtering," *IEEE Transactions on Signal Processing*, vol. 51, no. 10, pp. 2602–2612, 2003.
[44] 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.
[45] X. Rong Li and Z. Zhao, "Measuring estimator's credibility: Noncredibility index," in *9th International Conference on Information Fusion*, Florence, Italy, Jul. 10–13, 2006, pp. 1–8.