A Multivariate Non-Gaussian Bayesian Filter Using Power Moments

Guangyu Wu, *Student Member, IEEE*, and Anders Lindquist, *Life Fellow, IEEE*

***Abstract*—In this paper, we extend our results on the univariate non-Gaussian Bayesian filter using power moments [47] to the multivariate systems, which can be either linear or nonlinear. Doing this introduces several challenging problems, for example a positive parametrization of the density surrogate, which is not only a problem of filter design, but also one of the multiple dimensional Hamburger moment problem. We propose a parametrization of the density surrogate with the proofs to its existence, Positivstellensatz and uniqueness. Based on it, we analyze the errors of moments of the density estimates by the proposed density surrogate. A discussion on continuous and discrete treatments to the non-Gaussian Bayesian filtering problem is proposed to motivate the research on continuous parametrization of the system state. Simulation results are given to validate our proposed filter. To the best of our knowledge, the proposed filter is the first one implementing the multivariate Bayesian filter with the system state parameterized as a continuous function, which only requires the true states being Lebesgue integrable with first several orders of power moments being finite.**

***Index Terms*—Bayesian filter; non-Gaussian distribution; multidimensional Hamburger moment problem; density parametrization.**

I. Introduction

THe Bayesian filter provides a unified recursive approach for nonlinear filtering problems. One of the first exploration of iterative Bayesian estimation is found in Ho and Lee's paper [17], where principle and procedure of Bayesian filtering are specified. Sprangins [43] discussed the iterative application of Bayes rule to sequential parameter estimation. Lin and Yau [29] and Chien and Fu [9] discussed Bayesian approach to optimization of adaptive systems. The Bayesian filter consists of an iterative measurement-time update process, although sometimes interpreted by different names. In the time-update step, the one-step ahead prediction of state is calculated by the system equation; in the measurement-update step, the correction to the estimate of state according to the current observation is calculated by the observation equation.

However, the Bayesian filter is more a framework for nonlinear non-Gaussian filtering. The specific filters are indeed implementations of the Bayesian filtering framework. When we are confronted with a nonlinear filtering problem, it is always infeasible for us to obtain an optimal or analytic solution and be content with a suboptimal solution to the Bayesian filter [3], [25]. Mathematically speaking, the difficulty lies in the intractability of the convolution operation in the time update step. There are three main approaches to treat this problem.

We note that when the densities are assumed to be Gaussians or mixtures of Gaussians, the convolution is tractable. When all the densities involved in the filtering process are constrained to be Gaussian, the filtering process is essentially the Kalman filter [20], [21] with the system and observation equations being linear. Furthermore, its numerous variants were proposed as implementations of this approach, including the extended Kalman filter (EKF) [2], the central-difference Kalman filter (CDKF) [31], [37], the unscented Kalman filter (UKF) [19], and the quadrature Kalman filter (QKF) [18]. From the perspective of Gaussian mixtures, a smart-sensor-based method was developed to deal with Gaussian mixtures with an exponentially increasing number of Gaussians and design optimal estimation in the pioneering work [28] on UDP-like systems. With the strict constraint on the densities all being Gaussian or a mixture of them, it is feasible for the estimators above to obtain a closed form of solution to the convolution operation.

However in modern applications, the state of the system and the noises don't always follow the Gaussian distribution. In the filtering problem in econometrics, for example in the analysis of financial time series, the distributions of the noises have "heavy tails", where the Gaussian distribution doesn't apply. Modes are viewed as the central tendencies of a distribution in statistics. When the probability density of the state is multi-modal, it is obviously infeasible to estimate it with a Gaussian distribution. Confronted with the numerous cases where the first approach doesn't apply, people have been seeking for other methods to treat the intractable convolution.

In the second approach, we give up seeking an analytic solution to the convolution. Instead, we attempt to approximate the intractable convolution, where no explicit assumption on the density form is needed. We note that by assuming the densities to have a discrete form of function, we are able to write the intractable prior density as a probability mass function, which is supported on finitely many discrete points we choose on the domain $\mathbb{R}^d$. There are numerous such methods, including mulitgrid method and point-mass approximation [5], [7] and Monte Carlo sampling approximation [10], [15]. These nonparametric methods impose no prior constraints on the den-

Guangyu Wu is with Department of Automation, Shanghai Jiao Tong University, Shanghai, China. (e-mail: chinarustin@sjtu.edu.cn).

Anders Lindquist is with Department of Automation and School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China. (e-mail: alq@kth.se).

sity functions, which seem to enjoy the maximum flexibility. However the tradeoff is also very severe: quite a bunch of probability values at discrete states need to be stored and the continuity of the density is sacrificed. It means that when given an arbitrary state, we are not always able to obtain its value of probability. Meanwhile, optional resampling is widely used in the filters to avoid depletion of particles with small probability values. To design a strategy of when and how to perform the resampling operation isn't a trivial task. Moreover, the particle filter suffers from the curse of dimensionality, which requires the filter to store massive particles with the increase of dimension. In some applications, we only consider the states with significant values of probability; however the states of small values of probability are extremely important, e.g. in financial applications. In conclusion, the discrete methods for density characterization are intrinsically infeasible in tackling the problem where the states with less significant values of probability still have dominant impact on the filtering problem, such as heavy-tailed filtering [34], [48].

Meanwhile, several numerical methods have been proposed to obtain an analytic solution to the convolution in the time-update step, including Gaussian/Laplace approximation [30], iterative quadrature [11], [26], Gaussian sum approximation [1], [42] and state-space calculus [16]. The variational Bayesian filter treats the filtering problem with time-varying measurement noise parameters [36], [39], [41]. It parameterizes the posterior density as a product of a Gaussian and an Inverse-Gamma distribution with the latter one being the conjugate prior distribution for the variance of a Gaussian distribution. These are the parametric methods for parameterizing the probability density function in a continuous form. However, the flexibility of these methods is limited. Some of the methods listed above are either not able to be extended to the multivariate case, or are not able to treat the multimodal prior density. Therefore it is infeasible to apply these methods to a wide range of real applications, where there are numerous density functions which don't fall within the prescribed function classes. Moreover, quantitative approximation performance analyses, e.g. an error upper bound of estimation have not been derived for the existing methods yet. At the same time, the Bayesian filter is a recursive algorithm, which means that the estimation errors in the previous steps will have cumulative effects on the estimates of later steps. However with the estimation algorithms listed above, the cumulative errors are problematic to analyze, which severely decreases the value of these methods in practical use. A multivariate non-Gaussian Bayesian filter with the state estimation parameterized as an analytic function, where there is no constraint on the feasible classes of the prior density function, is extraordinarily desired by the researchers on stochastic filtering.

In this paper, we attempt to treat the multivariate non-Gaussian Bayesian filtering problem. We first formulate the multivariate non-Gaussian Bayesian filtering problem in Section 2. Then we propose to use the higher order moments to characterize the density function. The multivariate density surrogate is also defined. A formal definition of the truncated Hamburger moment problem is given, and the solution to it is proved to exist in Section 3. A construction of the density surrogate, i.e., parametrization of the density function is proposed in Section 4. In doing this, we follow the Kullback-Leibler optimization scheme of [12]. Then a novel Positivstellensatz is proposed to ensure the positiveness of the multivariate density surrogate. There follows the proof of the uniqueness of the solution to the optimization problem. The map from the parameters of the density surrogate to the power moments is proved to be homeomorphic, which ensures that the gradient-based optimization algorithms can be applied to determining the parameters of the density surrogate. In Section 5, given that the prior is a sub-Gaussian distribution, the estimated moments are proved to be asymptotically unbiased from the true ones using the density surrogate. And by selecting a sufficiently large order, using the density surrogate will not bring significant cumulative errors to the moment estimation of the subsequent filtering steps. Furthermore, an analytic error analysis of the multivariate Bayesian filter, i.e., an upper bound of approximation error of the multivariate prior density is proposed in Section 6, which has not yet been done for the multivariate Bayesian filters with no prior constraints on the classes of the densities. Simulation results on estimating different classes of multivariate density functions, including heavy-tailed ones, are given in Section 7. A simulation of the proposed filter on a robot localization problem is also carried out to validate the proposed filter, with a comparison to the particle filter and the unscented Kalman filter.

II. Problem formulation

In this paper, we consider the non-Gaussian multivariate filtering problem for the following system.

$$\begin{aligned} x_{t+1} &= f_t\left(x_t\right) + \eta_t \\ y_t &= h_t\left(x_t\right) + \epsilon_t \end{aligned} \tag{1}$$

$t = 0, 1, 2, \cdots$. The state x_t is a random vector defined on $\mathbb{R}^d$ endowed with its Borel sigma algebra of events. Let the system function be $f_t : \mathbb{R}^d \mapsto \mathbb{R}^d$ and the observation function be $h_t : \mathbb{R}^d \mapsto \mathbb{R}^d$.

The system noise η_t and observation noise ϵ_t are random vectors defined on $\mathbb{R}^d$, of which the distributions are denoted as $\rho_{\eta_t}(x_t)$ and $\rho_{\epsilon_t}(x_t)$ respectively. The distribution of the system noise η_t can be either a Lebesgue integrable function or a probability mass function. The distribution of the observation noise ϵ_t is assumed to be a Lebesgue integrable function. Due to the very loose constraints on the noises, we are able to treat noises of which the density functions are not smooth, such as Laplacian distributions; or even those are not continuous. The noises are assumed to be independent from each other. The probability density function of the noise random vectors is non-Gaussian if assumed continuous. We note that all the densities are multivariate.

We adopt the Bayesian filter as used in [16] and extend it to the multivariate case. Denoting the collection of observations $y_t, y_{t-1}, \cdots, y_0$ as $\mathcal{Y}_t$, the conditional densities of the measurement and time updates are given by the following

Measurement update: For $t = 0$,

$$\begin{aligned}\rho_{x_0|\mathcal{Y}_0}(x) &= \frac{\rho_{y_0|x_0}(y_0)\,\rho_{x_0}(x)}{\int_{\mathbb{R}^d}\rho_{y_0|x_0}(y_0)\,\rho_{x_0}(x)dx}\\ &= \frac{\rho_{\epsilon_0}(y_0 - h_0(x))\,\rho_{x_0}(x)}{\int_{\mathbb{R}^d}\rho_{\epsilon_0}(y_0 - h_0(x))\,\rho_{x_0}(x)dx};\end{aligned} \tag{2}$$

for $t \geq 1$,

$$\begin{aligned}\rho_{x_t|\mathcal{Y}_t}(x) &= \frac{\rho_{y_t|x_t}(y_t)\,\rho_{x_t|\mathcal{Y}_{t-1}}(x)}{\int_{\mathbb{R}^d}\rho_{y_t|x_t}(y_t)\,\rho_{x_t|\mathcal{Y}_{t-1}}(x)dx}\\ &= \frac{\rho_{\epsilon_t}(y_t - h_t(x))\,\rho_{x_t|\mathcal{Y}_{t-1}}(x)}{\int_{\mathbb{R}^d}\rho_{\epsilon_t}(y_t - h_t(x))\,\rho_{x_t|\mathcal{Y}_{t-1}}(x)dx}, x \in \mathbb{R}^d.\end{aligned} \tag{3}$$

Time update: For $t \geq 0$,

$$\begin{aligned}\rho_{x_{t+1}|\mathcal{Y}_t}(x) &= \left(\rho_{f_t(x_t)|\mathcal{Y}_t} * \rho_{\eta_t}\right)(x)\\ &= \int_{\mathbb{R}^d}\rho_{x_t|\mathcal{Y}_t}\left(f_t^{-1}(\varepsilon)\right)\rho_{\eta_t}(x-\varepsilon)d\varepsilon.\end{aligned} \tag{4}$$

$\rho_{x_{t+1}|\mathcal{Y}_t}$ is the prior of each filtering step $t+1$. In the following sections of this paper, "prior" refers to $\rho_{x_{t+1}|\mathcal{Y}_t}$, if not specified otherwise. The measurement update (3) is a multiplication of the densities, which is a straightforward calculation. However it is always infeasible to obtain an analytic form of the prior in (4) when the densities are not Gaussian.

Due to the intractability of the convolution, we naturally consider approximating $\rho_{x_{t+1}|\mathcal{Y}_t}$. Inspired by the method of moments [13], we propose to use the truncated power moments to estimate $\rho_{x_{t+1}|\mathcal{Y}_t}$. Denote the nonnegative integers as $\mathbb{N}_0$. Let $\kappa = (k_1, \cdots, k_d)$ with $k_j \in \mathbb{N}_0$. Then we are able to write the power moments in the form of linear functional

$$L^\rho(\kappa) = \int_{\mathbb{R}^d} x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}\rho(x)dx, \quad \kappa \in \mathcal{J}_{2n} \tag{5}$$

where

$$\mathcal{J}_{2n} := \{\kappa = (k_1, \cdots, k_d) \mid k_i \in \mathbb{N}_0, k_i \leq 2n, i = 1, \cdots, d\}.$$

Then we can calculate the power moments of $\rho_{x_{t+1}|\mathcal{Y}_t}$ as

$$\begin{aligned}&L^{\rho_{t+1|t}}(\kappa)\\ &= \int_{\mathbb{R}^d} x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}\cdot\rho_{x_{t+1}|\mathcal{Y}_t}(x)dx\\ &= \int_{\mathbb{R}^d} x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}\int_{\mathbb{R}^d}\rho_{x_t|\mathcal{Y}_t}\left(f_t^{-1}(\varepsilon)\right)\rho_{\eta_t}(x-\varepsilon)d\varepsilon dx\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}\rho_{x_t|\mathcal{Y}_t}\left(f_t^{-1}(\varepsilon)\right)\rho_{\eta_t}(x-\varepsilon)d\varepsilon dx\\ &= \int_{\mathbb{R}^d}\rho_{x_t|\mathcal{Y}_t}\left(f_t^{-1}(\varepsilon)\right)\int_{\mathbb{R}^d} x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}\rho_{\eta_t}(x-\varepsilon)dxd\varepsilon.\end{aligned} \tag{6}$$

Therefore we are able to obtain the power moments, even though the prior density $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ doesn't have an analytic form as a function. There have been numerous previous research results using the power moments for Bayesian filtering. For example in the well-known Kalman filter (and its extended forms such as EKF and UKF), approximation of the density is done by a parametric estimation using the first order and second order power moments. However a multivariate Bayesian filter feasible for treating the density of the state using higher order moments has not been proposed to the best of our knowledge. In this paper, our goal is to propose a multivariate Bayesian filter which not only is feasible to treat the non-Gaussian state estimation problem, but also has analytic error analyses to measure the performance of filtering.

Algorithm 1: A general framework for Bayesian filtering with density surrogate at time t.

```
Input: System parameters: $f_t, h_t$;
Non-Gaussian densities: $\eta_t, \epsilon_t$;
Prediction at time $t-1$:
$\hat{\rho}_{x_t|\mathcal{Y}_{t-1}}(x), t > 0$, or $\rho_{x_0}(x), t = 0$.
Output: Prediction at time $t$: $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}(x)$.

Step 1: Calculate $\hat{\rho}_{x_t|\mathcal{Y}_t}$ by (2) or (3);
Step 2: Calculate $L^\rho(\kappa)$ by (6) for $\kappa \in \mathcal{J}_{2n}$;
Step 3: Determine the multivariate order-$2n$ density
  surrogate $\rho^{2n}_{x_{t+1}|\mathcal{Y}_t}$, of which the truncated moments
  are $L^\rho(\kappa), \kappa \in \mathcal{J}_{2n}$. The density estimate at time
  $t+1$ is then chosen as the density surrogate, i.e.,
  $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t} = \rho^{2n}_{x_{t+1}|\mathcal{Y}_t}$.
```

To perform the density approximation in the following parts of the paper, we first define the equivalence of multivariate density functions in the sense of power moments.

Definition II.1. A probability density function, which has the identical $L^\rho(\kappa), \kappa \in \mathcal{J}_{2n}$ as ρ, is called an order-$2n$ density surrogate of ρ and denoted by ρ^{2n}.

By defining the corresponding density estimate as $\hat{\rho}$, we propose a general framework for each iteration of Bayesian filtering with the density surrogate as Algorithm 1. Now the problem amounts to constructing an order-$2n$ multivariate density surrogate. Since the domain of ρ is $\mathbb{R}^d$, the problem becomes a multidimensional Hamburger moment problem [38]. In the next section, we will give a formal definition to the multidimensional Hamburger moment problem we will treat, and show the existence of solution to it, i.e., the existence of the multivariate density surrogate.

III. The Multidimensional Hamburger moment problem and the existence of solution

In this section, we give a formal definition of the Hamburger moment problem to treat and prove the existence of solution to the multidimensional Hamburger moment problem with the power moments. First we give the definition.

Definition III.1. A sequence $\bar{\sigma} = (\sigma_\kappa, \kappa \in \mathcal{J}_{2n})$ is a feasible $2n$ sequence, if there is a random vector $X = (X_1, \cdots, X_d)^\intercal$ defined on $\mathbb{R}^d$ endowed with its Borel sigma algebra of events, whose power moments are given by (6), that is,

$$\sigma_\kappa = \sigma_{k_1,\cdots,k_d} = L^\rho(\kappa) = \int_{\mathbb{R}^d} x_1^{k_1}x_2^{k_2}\cdots x_d^{k_d}\rho(x)dx \tag{7}$$

for all $\kappa \in \mathcal{J}_{2n}$. We say that any such random vector X has a $\bar{\sigma}$-feasible distribution and denote this as $X \sim \bar{\sigma}$.

Next we show the existence of solution to the moment problem defined in Definition III.1. We note that the true

$\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ is trivially a solution to the moment problem in Definition III.1. However we require an analytic function which satisfy the moment constraints. We note that there are numerous methods for functional approximation, however the solutions they provide don't always satisfy the requirement as a state estimation of the Bayesian filter, since there are possibly infinitely many parameters in the solutions, which makes it infeasible to propagate the solutions in the filtering process. Parametrization is then the most significant problem for constructing the density surrogate, which aims to use finitely many parameters to characterize the density.

Moreover, we are provided with the truncated power moment sequence $\bar{\sigma}$ rather than the full one, which means that there might be infinitely many feasible solutions to this problem. In the following part of this section, we propose to choose proper constraints to parameterize the density surrogate which satisfies the moment conditions. We shall still emphasize here that the parametrization is not unique. Different constraints will yield different parametrizations.

In the next section, we propose to parameterize the multivariate density surrogate, i.e., to derive a unique solution to the moment problem of $\rho_{x_{t+1}|\mathcal{Y}_t}$.

IV. PARAMETRIZATION OF THE MULTIVARIATE DENSITY SURROGATE USING POWER MOMENTS

In this section, we propose to parameterize the density surrogate, i.e., to derive a unique solution to the multidimensional Hamburger moment problem of $\rho_{x_{t+1}|\mathcal{Y}_t}$. For simplicity, we omit the subscript t in all the terms of the following part of this section.

Denote the Kronecker product as $\otimes$ and we observe that the moment conditions in (7) can be written in the vector form

$$\int_{\mathbb{R}^d} F(x_1) \otimes F(x_2) \otimes \cdots \otimes F(x_d) \rho(x) dx = \breve{\Sigma}, \tag{8}$$

where

$$F(x_i) = \begin{bmatrix} 1 & x_i & \cdots & x_i^{2n-1} & x_i^{2n} \end{bmatrix}^\intercal$$

and $\breve{\Sigma}$ is written as

$$\breve{\Sigma} = \begin{bmatrix} 1 & \sigma_{0,0,\cdots,1} & \cdots & \sigma_{2n,2n,\cdots,2n} \end{bmatrix}^\intercal$$

with the power moments $\sigma_{k_1,\cdots,k_d}$ calculated by (7). We note that the dimension of $\breve{\Sigma}$ is $(2n+1)^d$. Then we have that $\breve{\Sigma}$ is in the range of the linear integral operator

$$\breve{\Gamma} : \rho \mapsto \breve{\Sigma} = \int_{\mathbb{R}^d} F(x_1) \otimes F(x_2) \otimes \cdots \otimes F(x_d) \rho(x) dx.$$

Consequently, we have an order $2n$ multidimensional moment problem as defined in Definition III.1.

A. A convex optimization scheme

We note that to directly treat the multivariate Hamburger moment problem of (8) is not feasible, since there is no prior knowledge of feasible classes of the density. If we assume the density of the state to fall within specific classes of functions, the moment problem then becomes a parametric estimation problem, where the moments are used to estimate the parameters of the parametric models given prior. However in our problem setting, the prior density of the state is only assumed to be Lebesgue integrable with first several orders of power moments being finite. No knowledge of the feasible functions of the density is required.

Let $\mathcal{P}$ be the space of probability density functions on the Euclidean space $\mathbb{R}^d$ with support there, and let $\mathcal{P}_{2n}$ be the subset of all $\rho \in \mathcal{P}$ which have finite power moments at least up to order $2n$ (in addition to $\sigma_{0,\cdots,0}$, which of course is 1).

Except for the parametric algorithms, density estimation has been done by optimization in previous results [14], [27], [46], where Kullback-Leibler distance is used as a distance measure between densities. Let θ be an arbitrary density known prior in $\mathcal{P}$ and consider the Kullback-Leibler (KL) distance

$$\mathbb{KL}(\theta \| \rho) = \int_{\mathbb{R}^d} \theta(x) \log \frac{\theta(x)}{\rho(x)} dx \tag{9}$$

between θ and ρ. Although not symmetric in its arguments, the KL distance is jointly convex and is widely used in density estimation. In this section, we extend some lines of thoughts of [12] to the multivariate case and introduce a parametrization which is induced by the KL distance. However the extension is not at all a trivial problem, of which the details will be given in this section.

Given the moment constraints (8), we first form the Lagrangian

$$L(\rho, \breve{\Lambda}) = \mathbb{KL}(\theta \| \rho) + \breve{\Lambda}(\breve{\Gamma}(\rho) - \breve{\Sigma}), \tag{10}$$

where

$$\breve{\Lambda} = \begin{bmatrix} \lambda_{0,0,\cdots,0} & \lambda_{1,0,\cdots,0} & \cdots & \lambda_{2n,2n,\cdots,2n} \end{bmatrix} \tag{11}$$

is the vector-valued Lagrange multipliers, and consider the problem of maximizing the dual functional

$$\breve{\Lambda} \mapsto \inf_{\rho \in \mathcal{P}_{2n}} L(\rho, \breve{\Lambda}). \tag{12}$$

Clearly $\rho \mapsto L(\rho, \breve{\Lambda})$ is strictly convex, so to be able to determine the right member of (12), we must find a $\rho \in \mathcal{P}_{2n}$, for which the directional derivative $\delta L(\rho, \breve{\Lambda}; \delta\rho) = 0$ for all relevant $\delta\rho$. This will further restrict the choice of $\breve{\Lambda}$. By denoting

$$F(x) = F(x_1) \otimes F(x_2) \otimes \cdots \otimes F(x_d)$$

and setting

$$q(x, \breve{\Lambda}) := \breve{\Lambda} F(x), \tag{13}$$

we have

$$L(\rho, \breve{\Lambda}) = \int_{\mathbb{R}^d} \theta(x) \log \frac{\theta(x)}{\rho(x)} dx + \int_{\mathbb{R}^d} q(x, \breve{\Lambda}) \rho(x) dx - \breve{\Lambda}\breve{\Sigma},$$

with the directional derivative

$$\delta L(\rho, \breve{\Lambda}; \delta\rho) = \int_{\mathbb{R}^d} \delta\rho(x) \left(q(x, \breve{\Lambda}) - \frac{\theta(x)}{\rho(x)} \right) dx,$$

which has to be zero at a minimum for all variations $\delta\rho$. Clearly this can be achieved only if

$$\rho(x) = \frac{\theta(x)}{q(x, \breve{\Lambda})}, \quad \forall x \in \mathbb{R}^d. \tag{14}$$

.

B. A positive parametrization for the multivariate density

We have proved that $\rho(x) = \theta(x)/q(x, \breve{\Lambda})$ maximizes the dual functional, however there is still a constraint we need to consider. We note that $\rho(x)$ and $\theta(x)$ are both probability density functions, therefore are both nonnegative on $\mathbb{R}^d$. Moreover, x is supported on $\mathbb{R}^d$ for the Hamburger moment problem, so then we need to have $\rho(x) > 0, x \in \mathbb{R}^d$.

The problem now amounts to characterizing the constraint of $\breve{\Lambda}$, under which the multivariate polynomial $q(x, \breve{\Lambda})$ is positive. However it is a challenging problem of this paper and even of the general multidimensional moment problem. The reason is that useful descriptions of strictly positive polynomials up to a fixed degree $2n$ are missing [38]. The positive definiteness of the Hankel matrix is a sufficient condition for the univariate polynomials to be positive, which is used in our previous papers [12], [47]. However it is not valid anymore for the multivariate polynomials.

In the multivarate Hamburger moment we consider, the highest order of each variable x_i is chosen as $2n$. Indeed, it is a necessary condition for a polynomial $q(x, \breve{\Lambda})$ to be strictly positive everywhere on $\mathbb{R}^d$. A polynomial of which the highest order of any variable is odd always has a real zero, and the value of the polynomial changes sign at that point. It is then not feasible to obtain a $q(x, \breve{\Lambda}) > 0$.

With highest order of each variable chosen as $2n$, we propose a new parametrization of $q(x, \breve{\Lambda})$ and give the following theorem which is a strict Positivstellensatz for $q(x, \breve{\Lambda})$. We note that by our choice of $\mathscr{J}_{2n}$, it is feasible for us to write $q(x, \breve{\Lambda})$ in a quadratic form. Denote

$$G(x_i) = \begin{bmatrix} 1 & x_i & \cdots & x_i^{n-1} & x_i^n \end{bmatrix}^\intercal$$

and

$$G(x) = G(x_1) \otimes G(x_2) \otimes \cdots \otimes G(x_d),$$

and we can write

$$q(x, \breve{\Lambda}) = q(x, \Lambda) = G^\intercal(x) \Lambda G(x). \tag{15}$$

Here we note that the dimension of $\breve{\Lambda}$ and that of Λ are different, which are $(2n+1)^d$ and $(n+1)^{2d}$ respectively. It is obvious that the latter one is always larger than the former one since $n, k \in \mathbb{N}_0 / \{0\}$. However there are only $(2n+1)^d$ moment constraints given by (7), which means that only $(2n+1)^d$ Lagrange multipliers are necessary for the dual functional. For the ease of analyses in the following sections, we write all elements of Λ by the elements of $\breve{\Lambda}$.

Denote the i_{th} element of $G(x)$ as $G(x)_i$, and the element of Λ at i_{th} row and j_{th} column as $\Lambda_{i,j}$. Let the set

$$\mathcal{R}_\kappa := \{ \Lambda_{i,j} | G(x)_i \cdot G(x)_j = x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} \}.$$

We note that all $\Lambda_{i,j} \in \mathcal{R}_\kappa$ corresponds to the same moment constraint

$$\sigma_{k_1, \cdots, k_d} = \int_{\mathbb{R}^d} x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} \rho(x) dx, \tag{16}$$

and the cardinality of each $\mathcal{R}_\kappa, 0 \leq k_0, k_1, \cdots, k_d \leq 2n$ is $(1+k_1)(1+k_2)\cdots(1+k_d)$. Therefore for each $\Lambda_{i,j} \in \mathcal{R}_\kappa$ we have

$$\Lambda_{i,j} = \frac{\lambda_{k_1, k_2, \cdots, k_d}}{(1+k_1)(1+k_2)\cdots(1+k_d)}. \tag{17}$$

By (17), we have proposed a one-to-one correspondence between $\breve{\Lambda}$ and Λ. With the proposed parametrization for $q(x, \Lambda)$, we have the following Positivstellensatz.

Theorem IV.1 (Positivstellensatz)**.** $q(x, \Lambda) > 0$ *if and only if* Λ *is positive definite.*

Proof. The necessity is obvious and we need to prove the sufficiency. Denote the eigenvalues of Λ as $m_i, i \in \mathbb{N}_0, i \leq (n+1)^d$, and the corresponding eigenvectors as v_i. We assume that there exists a $q(x, \Lambda) > 0$ of which Λ is not positive definite. By eigen decomposition, we can write Λ as

$$\Lambda = V M V^\intercal$$

where

$$V := \begin{bmatrix} v_1 & v_2 & \cdots & v_{(n+1)^d} \end{bmatrix},$$

and

$$M = \begin{bmatrix} m_1 & 0 & \cdots & 0 \\ 0 & m_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & m_{(n+1)^d}. \end{bmatrix}$$

For matrices which are not positive definite, there are at least one nonpositive eigenvalue. We assume $m_1 \leq \cdots \leq m_l \leq 0, l \leq (n+1)^d$. And assign $x_i = 0, i \in \mathbb{N}_0, i \leq (n+1)^d$. Then $G(\mathbf{0}) = [1, 0, 0, \cdots, 0]^\intercal$. So then we have

$$\begin{aligned} & q(\mathbf{0}, \Lambda) \\ = & G^\intercal(\mathbf{0}) V M V^\intercal G(\mathbf{0}) \\ = & \operatorname{tr}\left(V G^\intercal(\mathbf{0}) M G(\mathbf{0}) V^\intercal \right) \\ = & G^\intercal(\mathbf{0}) M G(\mathbf{0}) \operatorname{tr}\left(V V^\intercal \right). \end{aligned} \tag{18}$$

Since the scalar $G^\intercal(\mathbf{0}) M G(\mathbf{0}) \leq 0$ with $m_1 \leq 0$, and $\operatorname{tr}(VV^\intercal) > 0$, we have $q(\mathbf{0}, \Lambda) \leq 0$, which contradicts our assumption. Therefore when Λ has at least one eigenvalue being nonpositive, there exists at least a point $x = (0, 0, \cdots, 0)$, at which we have $q(x, \Lambda) \leq 0$. □

Remark. In the conventional research on the multidimensional truncated moment problem, we care about the existence of the positive measures rather than the parametrization of them. Therefore atomic measures [38] (probability mass functions) are always proposed as solutions to the multidimensional truncated moment problems. There have been few results on parameterizing the densities in a continuous form of function supported on $\mathbb{R}^d$. However parameterizing the density and ensuring its positiveness given the power moments are of great significance to the real applications. This theorem ensures the positiveness of $q(x, \Lambda)$ for all $d \in \mathbb{N}_0$, which contributes to the multivariate Hamburger moment problem.

Moreover, we note that the moment conditions can be written in a matrix representation

$$\int_{\mathbb{R}^d} G(x) \rho(x) G^\intercal(x) = \Sigma \tag{19}$$

where

$$\Sigma_{i,j} = \frac{\sigma_{k_1,k_2,\cdots,k_d}}{(1+k_1)(1+k_2)\cdots(1+k_d)}$$

is the element at the i_{th} row and the j_{th} column.

By Section III, we know that the class of $\rho \in \mathcal{P}$ satisfying (8) is nonempty. In fact, Σ is in the range of the linear integral operator

$$\Gamma : \rho \mapsto \Sigma = \int_{\mathbb{R}^d} G(x)\rho(x)G^{\mathsf{T}}(x)dx, \tag{20}$$

which is defined on the space $\mathcal{P}_{2n}$. Since $\mathcal{P}_{2n}$ is convex, then so is $\text{range}(\Gamma) = \Gamma\mathcal{P}_{2n}$.

In the following part of this section, we use the notations Λ and Σ instead of $\breve{\Lambda}$ and $\breve{\Sigma}$. Introducing $\breve{\Lambda}$ and the corresponding $\breve{\Sigma}$ in the previous sections is to emphasize that $(2n+1)^d$ is the minimal number of parameters to estimate. We shall always remember that all the elements of Λ and Σ can be represented by those of $\breve{\Lambda}$ and $\breve{\Sigma}$.

C. Uniqueness of the solution to the convex optimization problem

Being a multivariate probability density function, the positiveness of the parametrization in (14) is ensured by the constraint $\Lambda \succ 0$ as proved in Theorem IV.1. We then now state and prove our main results. We emphasize here that we adopt some notations from our previous paper [47] treating the univariate Bayesian filtering problem, however with different definitions.

By (14) and (15), a possible minimizer must have the form

$$\rho(x) = \frac{\theta(x)}{q(x)} = \frac{\theta(x)}{G^{\mathsf{T}}(x)\Lambda G(x)}.$$

The Lagrangian in (10) can then be written as

$$\begin{aligned} &L(\rho, \Lambda) \\ =&\mathbb{KL}(\theta\|\rho) + \text{tr}\left(\Lambda(\Gamma(\rho) - \Sigma)\right) \\ =&\int_{\mathbb{R}^d} \left(\theta \log\left(G^{\mathsf{T}}(x)\Lambda G(x)\right) + \theta\right)dx - \text{tr}\left(\Lambda\Sigma\right). \end{aligned} \tag{21}$$

Then the problem becomes minimizing the dual functional

$$\mathbb{J}_\theta(\Lambda) := \text{tr}(\Lambda\Sigma) - \int_{\mathbb{R}^d} \theta(x)\log\left(G^{\mathsf{T}}(x)\Lambda G(x)\right)dx. \tag{22}$$

Denote

$$\mathcal{L}_+ := \left\{\Lambda \in \text{range}(\Gamma) \mid \Lambda \succ 0\right\},$$

and we have the following lemma.

Lemma IV.2. *Any stationary point of* $\mathbb{J}_\theta(\Lambda)$ *must satisfy the equation*

$$\omega(\Lambda) = \Sigma, \tag{23}$$

where the map $\omega : \mathcal{L}_+ \to \mathcal{S}_+$ *between* $\mathcal{L}_+$ *and* $\mathcal{S}_+ := \{\Sigma \in \text{range}(\Gamma)\}$ *is defined as*

$$\omega : \ \Lambda \mapsto \int_{\mathbb{R}^d} G(x)\frac{\theta(x)}{q(x,\Lambda)}G^{\mathsf{T}}(x)dx$$

with q *defined by* (13).

Proof. Using the fact that

$$\delta q(\Lambda;\delta\Lambda) = G^{\mathsf{T}}\delta\Lambda G = \text{tr}\left(\delta\Lambda GG^{\mathsf{T}}\right)$$

we have the directional derivative

$$\delta\mathbb{J}_\theta(\Lambda;\delta\Lambda) = \text{tr}\left(\delta\Lambda\left[\Sigma - \int_{\mathbb{R}^d} G(x)\frac{\theta(x)}{q(x,\Lambda)}G^{\mathsf{T}}(x)dx\right]\right),$$

which is zero for all $\delta\Lambda \in \text{range}(\Gamma)$ if and only if (23) holds. This completes the proof. □

To prove the existence and uniqueness of the minimum of $\mathbb{J}_\theta$, we need to establish that the map $\omega : \mathcal{L}_+ \to \mathcal{S}_+$ is injective, establishing uniqueness, and surjective, establishing existence. In this way we prove that (23) has a unique solution, and hence that there is a unique minimum of the dual functional $\mathbb{J}_\theta$. We start with injectivity.

Lemma IV.3. *Suppose* $\Lambda \in$ *range*(Γ). *Then the map*

$$\Lambda \mapsto G^{\mathsf{T}}\Lambda G \tag{24}$$

is injective.

Proof. Since $\Lambda \in \text{range}(\Gamma)$,

$$\Lambda = \int_{\mathbb{R}^d} G(y)\psi(y)G^{\mathsf{T}}(y)dy$$

for some $\psi \in \mathcal{P}$. Suppose $G^{\mathsf{T}}\Lambda G = 0$. Then we have $\int_{\mathbb{R}^d} G^{\mathsf{T}}(x)\Lambda G(x)dx = 0$, and therefore

$$\begin{aligned} &\int_{\mathbb{R}^d} G^{\mathsf{T}}(x)\Lambda G(x)dx \\ &= \text{tr}\left(\int_{\mathbb{R}^d} G(x)^{\mathsf{T}}\int_{\mathbb{R}^d} G(y)\psi(y)G(y)^{\mathsf{T}}dy\, G(x)dx\right) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} [G(x)^T G(y)]^2\psi(y)dxdy = 0. \end{aligned}$$

Consequently we have $[G(x)^T G(y)]^2\psi(y) = 0$, for all $x, y \in \mathbb{R}^{(n+1)^d}$, which clearly implies that $\psi = 0$, and hence that $\Lambda = 0$. Consequently the map (24) is injective, as claimed. □

Theorem IV.4. *The map* $\omega : \mathcal{L}_+ \mapsto \mathcal{S}_+$ *is a homeomorphism.*

Proof. We first prove that $\omega : \mathcal{L}_+ \mapsto \mathcal{S}_+$ is injective. Suppose that $\omega(\Lambda_1) = \omega(\Lambda_2)$ for some Λ_1 and Λ_2 in $\mathcal{L}_+$. We need to show that $\Lambda_1 = \Lambda_2$. To this end, note that

$$\Delta\omega = \omega(\Lambda_1) - \omega(\Lambda_2) = \int_{\mathbb{R}^d} GG^T\frac{\theta}{q_1q_2}(q_2 - q_1)dx = 0,$$

where $q_1 = G^T\Lambda_1 G$ and $q_2 = G^T\Lambda_2 G$. Considering the element of $\Delta\omega$ at the first row and the first column, we have

$$\Delta\omega_{1,1} = \int_{\mathbb{R}^d}\frac{\theta}{q_1q_2}(q_2 - q_1)dx = 0. \tag{25}$$

Since θ is a strictly positive probability density function supported on $\mathbb{R}^d$, and $q_1, q_2 > 0$, the equality in (25) is achieved if and only if $q_1 = q_2$. Then by Lemma IV.3 this implies that $\Lambda_1 = \Lambda_2$, establishing that ω is injective.

Next, we prove that $\omega : \mathcal{L}_+ \mapsto \mathcal{S}_+$ is surjective. We first note that ω is continuous and that both sets $\mathcal{L}_+$ and $\mathcal{S}_+$ are nonempty, convex, and open subsets of the same Euclidean space $\mathbb{R}^{(2n+1)^d}$, and hence diffeomorphic to this space. We emphasize again that the dimension of the space is $(2n+1)^d$ rather than $(n+1)^{2k}$, since some of the elements in Λ and Σ are identical.

For the proof of surjectivity we shall use Corollary 2.3 in [6], by which the continuous map ω is surjective if and only if it is injective and proper, i.e., the inverse image $\omega^{-1}(K)$ is compact for any compact K in $\mathcal{S}_+$. See Theorem 2.1 in [6] for a more general statement.

Now it remains to prove that ω is proper. To this end, we first note that $\omega^{-1}(K)$ must be bounded, since, as if $\|\Lambda\| \to \infty$, $\omega(\Lambda)$ would tend to zero, which lies outside $\mathcal{L}_+$. Now, consider a Cauchy sequence in K, which of course converges to a point in K. We need to prove that the inverse image of this sequence is compact. If it is empty or finite, compactness is automatic. So we consider it to be infinite in the following proof. Since $\omega^{-1}(K)$ is bounded, there must be a subsequence (λ_k) in $\omega^{-1}(K)$ converging to a point $\lambda \in \mathcal{L}_+$. It remains to show that $\lambda \in \omega^{-1}(K)$, i.e., (λ_k) does not converge to a boundary point, which here would be $q(x) = 0$. However this does not happen since then $\det \omega(\Lambda) \to \infty$, contradicting boundedness of $\omega^{-1}(K)$. Hence ω is proper. Therefore, the map $\omega : \mathcal{L}_+ \to \mathcal{S}_+$ is proved to be homeomorphic. □

Moreover, the dual functional has the following property.

Lemma IV.5. *The dual functional $\mathbb{J}_\theta(\Lambda)$ is strictly convex.*

Proof. This is equivalent to $\delta^2 \mathbb{J}_\theta > 0$ where

$$\delta^2 \mathbb{J}_\theta(\Lambda; \delta\Lambda) = \int_{\mathbb{R}^d} \frac{\theta(x)}{q(x)^2} \left(G(x)^\intercal \delta\Lambda G(x)\right)^2 dx. \tag{26}$$

By (26), we have $\delta^2 \mathbb{J}_p \geq 0$, so it remains to show that

$$\delta^2 \mathbb{J}_p > 0, \quad \text{for all } \delta\Lambda \neq \mathbf{0},$$

which follows directly from Lemma IV.3, replacing Λ by $\delta\Lambda$. □

By Lemma IV.2, Theorem IV.4 and Lemma IV.5, we have the following theorem.

Theorem IV.6. *The functional $\mathbb{J}_\theta(\Lambda)$ has a unique minimum $\hat\Lambda \in \mathcal{L}_+$. Moreover*

$$\Gamma\left(\frac{\theta}{G^\intercal(x)\hat\Lambda G(x)}\right) = \Sigma.$$

By this theorem,

$$\hat\rho = \frac{\theta}{\hat q}, \quad \hat q = q(x, \hat\Lambda)$$

belongs to $\mathcal{P}_{2n}$ and is a stationary point of $\rho \mapsto L(\rho, \hat\Lambda)$, which is strictly convex. Consequently

$$L(\hat\rho, \hat\Lambda) \leq L(\rho, \hat\Lambda), \qquad \text{for all } \rho \in \mathcal{P}_{2n}$$

or, equivalently, since $\Gamma(\hat\rho) = \Sigma$,

$$\mathbb{KL}(\theta \| \hat\rho) \leq \mathbb{KL}(\theta \| \rho)$$

for all $\rho \in \mathcal{P}_{2n}$ satisfying the constraint $\Gamma(\rho) = \Sigma$. The above holds with equality if and only if $\rho = \hat\rho$. Finally, a complete solution is given in the following theorem for density parametrization using the multivariate power moments.

Algorithm 2: Bayesian filtering with density surrogate using power moments at time t.

```
Input: System parameters: $f_t, h_t$;
Non-Gaussian densities: $\eta_t, \epsilon_t$;
Prediction at time $t-1$:
$\hat\rho_{x_t|\mathcal{Y}_{t-1}}(x), t > 0,$ or $\rho_{x_0}(x), t = 0$.
Output: Prediction at time $t$: $\hat\rho_{x_{t+1}|\mathcal{Y}_t}(x)$.

Step 1: Calculate $\hat\rho_{x_t|\mathcal{Y}_t}$ by (2) or (3);
Step 2: Calculate $\sigma_\kappa, \kappa \in \mathcal{J}_{2n}$ of
  $\int_{\mathbb{R}^d} \hat\rho_{x_t|\mathcal{Y}_t}\left(f_t^{-1}(\varepsilon)\right) \rho_{\eta_t}(x-\varepsilon) d\varepsilon$ by (6);
Step 3: Do the optimization and obtain $\hat\Lambda$ which
  minimizes (22). $\hat\rho_{x_{t+1}|\mathcal{Y}_t}(x) = \theta(x)/(G^\intercal(x)\hat\Lambda G(x))$
  is the order-$2n$ density surrogate of
  $\int_{\mathbb{R}^d} \hat\rho_{x_t|\mathcal{Y}_t}\left(f_t^{-1}(\varepsilon)\right) \rho_{\eta_t}(x-\varepsilon) d\varepsilon$.
```

Theorem IV.7. *Let Γ be defined by (20) and Λ defined by* (17). *Given any $\theta \in \mathcal{P}$ and any Σ with σ_κ calculated by* (6), *there is a unique $\rho \in \mathcal{P}_{2n}$ that minimizes* (9) *subject to $\Gamma(\rho) = \Sigma$, i.e., subject to* (8), *namely*

$$\hat\rho = \frac{\theta}{q(x, \hat\Lambda)}, \tag{27}$$

where $\hat\Lambda$ is the unique solution to the problem of minimizing $\mathbb{J}_\theta$ in (22) *over all $\Lambda \in \mathcal{L}_+$.*

Consequently, the dual problem provides us with an approach to compute the unique $\hat\rho$ that minimizes the Kullback-Leibler distance $\mathbb{KL}(\theta\|\rho)$ subject to the constraint $\Gamma(\rho) = \Sigma$.

A parametrization using power moments for the density surrogate has been proposed in Theorem IV.7. Apply the proposed parametrization of the density surrogate to estimate $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$, we are able to design Algorithm 2 following the framework of Algorithm 1. However, the proposed filter is not the only feasible Bayesian filter by the multivariate power moments. By proposing different density parametrization strategies using power moments as Step 3 in Algorithm 1, more of such type of filters can be developed.

V. Error analyses of the density surrogate

In this section, we will first analyze the error propagation of the power moments using the density surrogate. It has been proved in [47] that all power moments of $\hat x$ exist and are finite, i.e. $\hat\rho(x)$ is light-tailed, if and only if θ is a sub-Gaussian distribution. Sub-Gaussian refers to that the tails of a distribution are dominated by those of a Gaussian distribution, i.e., decay at least as fast as a Gaussian. We will first analyze the error propagation of $\hat x$ of which all the power moments exist and are finite. We note that most Bayesian filters satisfy this condition, such as the well-known Kalman filter and the particle filter.

Theorem V.1. *Assume that all power moments of the true density $\rho_{x_{t+1}|\mathcal{Y}_t}$ and the corresponding estimate $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$ exist and are finite. Suppose $\hat{\rho}_{x_1|\mathcal{Y}_0}$ to be a surrogate for $\rho_{x_1|\mathcal{Y}_0}$, and let $\hat{\rho}_{x_t|\mathcal{Y}_t}, \hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$ be obtained from Algorithm 1 for $t = 2, 3, \cdots$. Then the power moments of $\hat{\rho}_{x_t|\mathcal{Y}_t}$ and $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$ up to order $2n$ are asymptotically unbiased and approximately those of the density surrogates $\rho_{x_t|\mathcal{Y}_t}$ and $\rho_{x_{t+1}|\mathcal{Y}_t}$ respectively with a large enough n.*

Proof. For the sake of simplicity, we omit the normalizing factor in the measurement update equations (2) and (3). It is straightforward to verify that it has no effect on the following results in this section. The first $2n$ moment terms of $\rho_{x_1|\mathcal{Y}_0}$ are identical to $\hat{\rho}_{x_1|\mathcal{Y}_0}$ after the first time update, i.e.,

$$L^{\rho_{1|0}}(\kappa) = L^{\hat{\rho}_{1|0}}(\kappa), \quad \kappa \in \mathcal{J}_{2n}. \tag{28}$$

For each $\mu \in \mathcal{J}_{2n}$, we can write the moment terms of $\rho_{x_1|\mathcal{Y}_1}$ as

$$L^{\rho_{1|1}}(\kappa) = \int_{\mathbb{R}^d} x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} \rho_{\epsilon_1}(y_1 - f_1(x)) \rho_{x_1|\mathcal{Y}_0}(x) dx$$

for all $\kappa \in \mathcal{J}_{2n}$. And those of $\hat{\rho}_{x_1|\mathcal{Y}_1}$ as,

$$L^{\hat{\rho}_{1|1}}(\kappa) = \int_{\mathbb{R}^d} x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} \rho_{\epsilon_1}(y_1 - f_1(x)) \hat{\rho}_{x_1|\mathcal{Y}_0}(x) dx$$

for all $\kappa \in \mathcal{J}_{2n}$. Therefore we have,

$$\begin{aligned} & L^{\rho_{1|1}}(\kappa) - L^{\hat{\rho}_{1|1}}(\kappa) \\ = & \int_{\mathbb{R}^d} x_1^{k_1} \cdots x_d^{k_d} \rho_{\epsilon_1}(y_1 - f_1(x)) \left(\rho_{x_1|\mathcal{Y}_0} - \hat{\rho}_{x_1|\mathcal{Y}_0} \right) dx. \end{aligned} \tag{29}$$

It is not feasible for us to calculate it directly, since we only have the moment constraints (28). We naturally consider decomposing the $\rho_{\epsilon_1}(y_1 - f_1(x))$ into a polynomial, so then (29) is able to be written in the form of a weighted sum of power moments of $\rho_{x_1|\mathcal{Y}_0} - \hat{\rho}_{x_1|\mathcal{Y}_0}$. By generalizing Exercise 13.12 in [35] to the multivariate case, it is feasible for us to write the Lebesgue integrable function $\rho_{\epsilon_1}(y_1 - h_1 x)$ as a polynomial of multiple variables, which is denoted as $\tilde{\rho}_{\epsilon_1}(y_1 - h_1(x))$. By Taylor expansion, we write it as

$$\begin{aligned} & \tilde{\rho}_{\epsilon_1}(y_1 - f_1(x)) \\ = & \tilde{\rho}_{\varepsilon_1}(y_1) + (\partial_{l_1} \tilde{\rho}_{\epsilon_1})(y_1) x_i + \frac{1}{2!} (\partial_{l_1 l_2} \tilde{\rho}_{\epsilon_1})(y_1) x_{l_1} x_{l_2} \\ + & \frac{1}{3!} (\partial_{l_1 l_2 l_3} \tilde{\rho}_{\epsilon_1})(y_1) x_{l_1} x_{l_2} x_{l_3} + \cdots \\ = & \tilde{\rho}_{\epsilon_1}(y_1) + \sum_{i=1}^{+\infty} \frac{1}{i!} (\partial_{l_{1:i}} \tilde{\rho}_{\epsilon_1})(y_1) x_{l_{1:i}} \end{aligned}$$

where

$$(\partial_{l_{1:i}} \tilde{\rho}_{\epsilon_1})(y_1) x_{l_{1:i}} := (\partial_{l_1 l_2 \cdots l_i} \tilde{\rho}_{\epsilon_1})(y_1) x_{l_1} x_{l_2} \cdots x_{l_i}.$$

We note that

$$\begin{aligned} & \mathbb{E}(x_1^\kappa | \mathcal{Y}_1) - \mathbb{E}(\hat{x}_1^\kappa | \mathcal{Y}_1) \\ = & \sum_{i=1}^{+\infty} \frac{(\partial_{l_{1:i}} \tilde{\rho}_{\epsilon_1})(y_1)}{i!} \int_{\mathbb{R}^d} x_{l_{1:k+i}} \left(\rho_{x_1|\mathcal{Y}_0} - \hat{\rho}_{x_1|\mathcal{Y}_0} \right) dx \\ + & \int_{\mathbb{R}^d} \left(\rho_{x_1|\mathcal{Y}_0} - \hat{\rho}_{x_1|\mathcal{Y}_0} \right) dx \\ = & \sum_{i=(2n)^d+1}^{+\infty} \frac{(\partial_{l_{1:i}} \tilde{\rho}_{\epsilon_1})(y_1)}{(i-k)!} \int_{\mathbb{R}^d} x_{l_{1:i}} \left(\rho_{x_1|\mathcal{Y}_0} - \hat{\rho}_{x_1|\mathcal{Y}_0} \right) dx, \end{aligned}$$

for $k = \sum_{i=1}^d k_i \leq (2n)^d$. Therefore we obtain

$$\mathbb{E}(x_1^\kappa | \mathcal{Y}_1) = \lim_{n \to +\infty} \mathbb{E}(\hat{x}_1^\kappa | \mathcal{Y}_1).$$

We also note that $\int_{\mathbb{R}^d} x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d} \rho_{\eta_t}(x - \varepsilon) dx$ in (6) is indeed a polynomial of $\varepsilon_i, i = 1, \cdots, d$, of which the highest order of ε_i in each term is $k_i \leq 2n$, $i = 1, \cdots, d$. Therefore by (6) we obtain

$$\mathbb{E}(x_2^\kappa | \mathcal{Y}_1) = \lim_{n \to +\infty} \mathbb{E}(\hat{x}_2^\kappa | \mathcal{Y}_1).$$

Similarly we have

$$\mathbb{E}(x_t^\kappa | \mathcal{Y}_t) = \lim_{n \to +\infty} \mathbb{E}(\hat{x}_t^\kappa | \mathcal{Y}_t),$$

and

$$\mathbb{E}\left(x_{t+1}^\kappa | \mathcal{Y}_t\right) = \lim_{n \to +\infty} \mathbb{E}\left(\hat{x}_{t+1}^\kappa | \mathcal{Y}_t\right),$$

which proves the asymptotic unbiasedness of the moment estimates throughout the filtering process as $n \to \infty$. By properly selecting a large enough n, we have

$$\mathbb{E}(x_t^\kappa | \mathcal{Y}_t) \approx \mathbb{E}(\hat{x}_t^\kappa | \mathcal{Y}_t), \ k_i \leq 2n, i = 1, \cdots, d,$$

and

$$\mathbb{E}\left(x_{t+1}^\kappa | \mathcal{Y}_t\right) \approx \mathbb{E}\left(\hat{x}_{t+1}^\kappa | \mathcal{Y}_t\right), \ k_i \leq 2n, i = 1, \cdots, d.$$

□

Theorem V.1 proves that the moment terms up to order $2n$ of the estimated prior densities with the density surrogate are asymptotically unbiased and approximately identical to the true ones throughout the whole filtering process. It reveals the fact that with a sub-Gaussian θ, approximation using the truncated power moments doesn't introduce significant cumulative errors to the moment terms up to order $2n$ of the estimated pdfs, with a proper choice of n.

Now we consider $\hat{\rho}$ to be heavy-tailed, which is the case for some scenarios such as financial engineering. Since the power moments of $\hat{\rho}$ are not all finite, i.e.,

$$\int_{\mathbb{R}^d} x^\kappa \left(\rho_{x_{t+1}|\mathcal{Y}_t} - \hat{\rho}_{x_{t+1}|\mathcal{Y}_t} \right) dx = \infty, \exists k_i \in \mathbb{N}_0,$$

it is not feasible for us to apply Taylor expansion to analyze the error propagation. However, we note that

$$\begin{aligned}
&\left|\mathbb{E}\left(x_{t+1}^{\kappa} | \mathcal{Y}_{t+1}\right)-\mathbb{E}\left(\hat{x}_{t+1}^{\kappa} | \mathcal{Y}_{t+1}\right)\right| \\
&=\left|\int_{\mathbb{R}^d} x^{\kappa} \rho_{\epsilon_{t+1}}\left(y_{t+1}-h_{t+1} x\right)\left(\rho_{x_{t+1} | \mathcal{Y}_t}-\hat{\rho}_{x_{t+1} | \mathcal{Y}_t}\right) d x\right| \\
&\leq \int_{\mathbb{R}^d}\left|x^{\kappa} \rho_{\epsilon_{t+1}}\left(y_{t+1}-h_{t+1} x\right)\left(\rho_{x_{t+1} | \mathcal{Y}_t}-\hat{\rho}_{x_{t+1} | \mathcal{Y}_t}\right)\right| d x \\
&=\int_{\mathbb{R}^d}|x|^{\kappa} \rho_{\epsilon_{t+1}}\left(y_{t+1}-h_{t+1} x\right)\left|\rho_{x_{t+1} | \mathcal{Y}_t}-\hat{\rho}_{x_{t+1} | \mathcal{Y}_t}\right| d x .
\end{aligned} \tag{30}$$

Since $C_k := \int_{\mathbb{R}} |x|^{\kappa} \rho_{\epsilon_{t+1}}\left(y_{t+1}-h_{t+1} x\right) dx$ is a constant unrelated to $\hat{\rho}_{x_{t+1} | \mathcal{Y}_t}$, we have

$$\begin{aligned}
&\left|\mathbb{E}\left(x_{t+1}^{\kappa} \mid \mathcal{Y}_{t+1}\right)-\mathbb{E}\left(\hat{x}_{t+1}^{\kappa} \mid \mathcal{Y}_{t+1}\right)\right| \\
&\leq C_k \max_x \left|\rho_{x_{t+1} | \mathcal{Y}_t}-\hat{\rho}_{x_{t+1} | \mathcal{Y}_t}\right| .
\end{aligned}$$

Consequently, we have proven the following theorem.

Theorem V.2. *The errors of the power moments of $\hat{\rho}_{x_{t+1} | \mathcal{Y}_{t+1}}$ are each bounded by a value which is proportional to the L_∞ norm of the error of the density surrogate $\hat{\rho}_{x_{t+1} | \mathcal{Y}_t}$.*

It reveals the fact that with a relatively small L-infinity norm of $\rho_{x_1 | \mathcal{Y}_0} - \hat{\rho}_{x_1 | \mathcal{Y}_0}$, the error of moment estimation will be tolerable.

In the previous sections, a parametrization using power moments for the multivariate density surrogate and a corresponding multivariate Bayesian filter have been proposed. The proposed Bayesian filter uses a continuous form of function to characterize the density of the state, which has not been proposed in the previous results with no constraints on feasible density functions. The continuous parametrization of the multivariate density function also makes it feasible to analyze the error of density estimate quantitatively.

To the best of our knowledge, an error upper bound for the multivariate state estimate has not been established in stochastic filtering. The reason is that a continuous form of parametrization of the system state has not been proposed. In this section, we propose an error upper bound of $\hat{\rho}(x)$ in the sense of total variation distance, which is a measure widely used in the moment problem [44], [45]. This upper bound distinguishes our proposed filter from other multivariate Bayesian filters.

The total variation distance between the density estimate $\hat{\rho}$ and the true density ρ is defined as follows:

$$V(\hat{\rho}, \rho)=\sup_x \left|\int_{(-\infty, x]}(\hat{\rho}-\rho) d x\right|=\sup_x \left|F_{\hat{\rho}}-F_\rho\right| \tag{31}$$

where $F_{\hat{\rho}}$ and F_ρ are the two distribution functions of $\hat{\rho}$ and ρ.

In [45], Shannon-entropy is used to calculate the upper bound of the total variation distance. The Shannon-entropy [40] is defined as

$$H[\rho]=-\int_{\mathbb{R}^d} \rho(x) \log \rho(x) d x .$$

Unlike the univariate case, $\rho(x)$ is a multivariate distribution in this paper. To the best of our knowledge, there has not been an existing result of a Shannon-entropy maximizing distribution $\breve{\rho}$ as for the multivariate case, However with the following theorem, it is feasible for us to obtain a Shannon-entropy maximizing distribution for the multivariate problem.

Theorem V.3 (Full chain rule [33])**.**

$$H\left(\rho(x_{1:d})\right)=\sum_{i=1}^d H\left(\rho(x_i | x_{1:i-1})\right) \leq \sum_{i=1}^d H\left(\rho(x_i)\right)$$

with equality iff $x_1, \cdots, x_n$ are mutually independent.

By Theorem V.3, it is feasible for us to turn the entropy maximizing problem of the joint distribution into maximizing the entropy of the marginal distribution of each $x_i, i = 1, \cdots, d$, and each x_i are independent from each other. So then the joint Shannon-entropy maximizing distribution $\breve{\rho}$ can be written as

$$\breve{\rho}(x_{1:d})=\arg\max_\rho H\left(\rho(x_{1:d})\right)=\prod_{i=1}^d \breve{\rho}(x_i)$$

where

$$\breve{\rho}(x_i)=\arg\max_\rho H\left(\rho(x_i)\right) .$$

The univariate Shannon-entropy maximizing distribution $\breve{\rho}$, of which the moments are calculated by (6), has the following density function [22],

$$\breve{\rho}(x_i)=\exp\left(-\sum_{j=0}^{2n} \lambda_{i,j} x_i^j\right)$$

where $\lambda_{i,0}, \cdots, \lambda_{i,2n}$ are determined by the constraints

$$\int_{\mathbb{R}^d} x_i^l \exp\left(-\sum_{j=0}^{2n} \lambda_{i,j} x_i^j\right) d x_i=\hat{\sigma}_{\kappa_{i,l}}, \tag{32}$$

and $\kappa_{i,l} := \{(k_1, \cdots, k_n) \mid k_i = l \leq 2n, k_j = 0, j \neq i\}$ for $i \leq d$.

Therefore we have

$$\begin{aligned}
& H(\breve{\rho}) \\
&=-\int_{\mathbb{R}^d} \prod_{i=1}^d \breve{\rho}(x_i) \log \left(\prod_{i=1}^d \breve{\rho}(x_i)\right) d x_{1:d} \\
&=-\int_{\mathbb{R}^d} \prod_{i=1}^d \breve{\rho}(x_i) \sum_{i=1}^d \log \breve{\rho}(x_i) d x_{1:d} \\
&=-\sum_{i=0}^d \int_{\mathbb{R}^d} \breve{\rho}(x_i) \log \breve{\rho}(x_i) d x_i \\
&=-\sum_{i=0}^d \int_{\mathbb{R}^d} \breve{\rho}(x_i) \cdot\left(-\sum_{j=0}^{2n} \lambda_{i,j} x_i^j\right) d x_i \\
&=\sum_{i=0}^d \sum_{l=0}^{2n} \lambda_{i,l} \sigma_{\kappa_{i,l}}
\end{aligned} \tag{33}$$

where the fifth equation is by (32). Referring to [45], the KL distance between the true density and the Shannon-entropy maximizing density can be written as

$$\begin{aligned}
&\mathbb{KL}\left(\rho\|\breve{\rho}\right)\\
=&\int_{\mathbb{R}^d}\rho(x)\log\frac{\rho(x)}{\breve{\rho}(x)}dx\\
=&\int_{\mathbb{R}^d}\rho(x)\log\rho(x)dx-\int_{\mathbb{R}^d}\rho(x)\log\breve{\rho}(x)dx\\
=&-H\left[\rho\right]+\sum_{i=0}^{d}\sum_{l=0}^{2n}\lambda_{i,l}\sigma_{\kappa_{i,l}}.
\end{aligned}$$

However, if θ is a sub-Gaussian and n is sufficiently large, $\hat{\sigma}_{\kappa_{i,l}}$ is approximately equal to $\sigma_{\kappa_{i,l}}$ for $i = 0, 1, \ldots 2n \left(\sigma_{\kappa_{i,l}} \approx \hat{\sigma}_{\kappa_{i,l}}\right)$ by Theorem V.1, and hence

$$\mathbb{KL}(\rho\|\breve{\rho}) \approx H[\breve{\rho}] - H[\rho].$$

Similarly, we obtain

$$\mathbb{KL}(\hat{\rho}\|\breve{\rho}) \approx H[\breve{\rho}] - H[\hat{\rho}].$$

Further by [24], [45], we have

$$\begin{aligned}
&V\left(\breve{\rho},\hat{\rho}\right)\\
\leq&3\left[-1+\left\{1+\frac{4}{9}\mathbb{KL}\left(\hat{\rho}\|\breve{\rho}\right)\right\}^{1/2}\right]^{1/2}\\
=&3\left[-1+\left\{1+\frac{4}{9}\left(H\left[\breve{\rho}\right]-H\left[\hat{\rho}\right]\right)\right\}^{1/2}\right]^{1/2}
\end{aligned}$$

and

$$\begin{aligned}
&V\left(\breve{\rho},\rho\right)\\
\leq&3\left[-1+\left\{1+\frac{4}{9}\mathbb{KL}\left(\rho\|\breve{\rho}\right)\right\}^{1/2}\right]^{1/2}\\
=&3\left[-1+\left\{1+\frac{4}{9}\left(H\left[\breve{\rho}\right]-H\left[\rho\right]\right)\right\}^{1/2}\right]^{1/2}.
\end{aligned}$$

Then we obtain the upper bound of the error

$$\begin{aligned}
&V\left(\hat{\rho},\rho\right)\\
=&\sup_x\left|F_{\hat{\rho}}\left(x\right)-F_{\rho}\left(x\right)\right|\\
\leq&\sup_x\left(\left|F_{\hat{\rho}}\left(x\right)-F_{\breve{\rho}}\left(x\right)\right|+\left|F_{\breve{\rho}}(x)-F_{\rho(x)}\right|\right)\\
\leq&\sup_x\left|F_{\hat{\rho}}\left(x\right)-F_{\breve{\rho}}\left(x\right)\right|+\sup_x\left|F_{\breve{\rho}}(x)-F_{\rho(x)}\right|\\
=&V\left(\breve{\rho},\hat{\rho}\right)+V\left(\breve{\rho},\rho\right)\\
\leq&3\left[-1+\left\{1+\frac{4}{9}\left(H\left[\breve{\rho}\right]-H\left[\hat{\rho}\right]\right)\right\}^{1/2}\right]^{1/2}\\
+&3\left[-1+\left\{1+\frac{4}{9}\left(H\left[\breve{\rho}\right]-H\left[\rho\right]\right)\right\}^{1/2}\right]^{1/2}.
\end{aligned}$$

In conclusion, we have proposed the error of the moment estimates given that the prior is either sub-Gaussian or not. And we have put forward a error upper bound of the density estimate. To our knowledge, such an error upper bound has not been proposed for the multivariate Bayesian filtering, without assuming the density function to fall within specific classes, in the previous results.

VI. Continuous vs discrete: a discussion

In the previous sections, a novel non-Gaussian Bayesian filter is proposed, of which the system state is parameterized as a continuous function. However the detailed treatments in the previous sections may have concealed the core idea of the proposed filter. In this section, we compare the proposed Bayesian filter to the existing results in a more conceptual manner, which aims to provide a bigger picture of the research on the non-Gaussian filtering and emphasize the significance of our proposed Bayesian filter.

The problem we treat doesn't restrict the non-Gaussian density to fall within specific classes of function (in our setting it is only assumed to be Lebesgue integrable of which first several orders of power moments exist and are finite), estimating the intractable prior density in the time update step is indeed an infinite-dimensional problem. The particle filter treats this estimation problem using discrete points without any assumption on the form of function of the prior density at each time step, which also turns the infinite-dimensional problem into a finite dimensional and tractable one. However characterizing the densities by discrete points requires massive particles to store the probability values of the states. The problem is even worse with the increase of dimensions, which is due to the curse of dimensionality.

Fourier decomposition was developed to map the data samples to the frequency domain, which provides an equivalence between the data samples and the spectral density. By the form of a linear integral operator, it is able to characterize the global property of data samples by a limited number of Fourier coefficients. The density estimation is then finite-dimensional and tractable. As to overcome the disadvantage of the discrete methods, there have been several previous endeavors trying to use limited number of terms in the frequency domain to characterize a wider class of distributions in stochastic filtering. For example, a univariate Bayesian filter was proposed using a state-space calculus scheme to treat the filtering problem in the frequency domain [45]. However the dimension of the state-space in each step keeps increasing which makes it infeasible for the task of filtering. Even a dimension reduction scheme was proposed (a prototype algorithm which might be problematic in real applications), it is difficult to quantitatively analyze the error of estimation introduced by dimension reduction (unlike other density estimation problems, errors in the previous steps will have cumulative effects on the following steps for stochastic filtering).

The power moments are used to characterize the intractable prior density $\rho_{x_{t+1}|\mathcal{Y}_t}$ in this paper. The power moments have the form of linear integral operators similar to the Fourier coefficients, which is then able to characterize the global property of the density to be estimated. However unlike the Fourier coefficients, the basis functions of the power moments are not orthogonal to each other and their norms are not 1, which makes it more complicated to use them for density estimation. With the proposed algorithm, it is feasible for us to treat the multivariate density estimation problem using the power moments with arbitrary number of variables. The existence, Positivstellensatz and uniqueness of the solution are all proved. These proofs serve as solid foundations of Bayesian filtering using power moments.

In conclusion, the particle filter and our proposed filter represent two approaches to treat the infinite-dimensional density estimation problem. Since the estimation problem is intrinsically infinite-dimensional, neither of them are optimal filters. However the prior density estimates of them are both asymptotically convergent to the true density, with the number of particles or the moment terms used tending to infinity. Characterizing the intractable prior density by discrete points or power moments are both able to turn the problem into a finite-dimensional and tractable one. However, most existing results on non-Gaussian Bayesian filtering are focused on the discrete representation of the density. With the results of this paper, we would like to see more attention drawn to parameterizing the density in a continuous form of function for Bayesian filtering.

VII. Simulation details and results

In the previous sections, a Bayesian filter with the density parameterized by using the power moments has been proposed. However, there are still several details to note when implementing the filter. This will be done in this section, where we will provide simulation results to validate the filter we propose.

The first problem is the choice of $\theta(x)$. Mathematically speaking, the choice is arbitrary by Theorem IV.7. However, as to achieve a faster convergence rate for optimization in practice, θ can usually be chosen as a multivariate Gaussian density function, of which the variables are independent from each other, i.e., $\theta(x)=\prod_{i=1}^{d}\theta(x_i)$. It ensures the existence the finite power moments of $\hat{\rho}(x)$ up to order $2n$. Therefore, the problem reduces to determining the mean and variance of each marginal distribution $\theta(x_i), i=1,\cdots,d$.

The first and second order power moments, i.e., $\sigma_{\kappa_{i,1}},\sigma_{\kappa_{i,2}}$ of the density to be estimated can be calculated by (5). In practice, we can choose $m_i=\sigma_{\kappa_{i,1}}$ and $\sigma_i^2>\sigma_{\kappa_{i,2}}$ and determine $\theta(x_i)=\mathcal{N}\left(m_i,\sigma_i^2\right)$. Here we note that a relatively large variance σ_i^2 is to better adjust to the densities with multiple modes.

In the following parts of this section, we will perform two types of experiments to validate the performance of our proposed filtering scheme. The core idea of the paper is to parameterize the intractable prior density by multivariate power moments up to an order. We will first simulate it to see the performance of the algorithm with different choice of the order. Then we will provide the simulation results of the proposed filter in a robot localization problem, which are compared to other prevailing methods.

A. Density estimation by multivariate power moments

In this simulation, we first estimate the prior density by multivariate power moments. We note that the prior density $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ doesn't always have an explicit form of function, i.e., it is not always feasible for us to obtain the true system states. It makes comparing the estimates of the prior density to the true ones infeasible. However, we note that when η_t is a discrete random variable, the prior density $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ can be written as

$$\rho_{x_{t+1}|\mathcal{Y}_t}(x)=\sum_{i=1}^{m}\rho_i\cdot\rho_{\eta_t}(x-\varepsilon_i) \tag{34}$$

which is a mixture of densities and has an analytic form as a function. In order to compare the density estimates to the true density for validating the performance of the proposed surrogates, we simulate the mixture of densities in the following part of this section. For the ease of visualization, the state x of the examples in this section are all chosen as two dimensional. However we note that our algorithm can treat the filtering problem with system state of arbitrary dimensions.

In the first two examples, we simulate a mixture of Gaussians. We emphasize here that our proposed density surrogate doesn't require any prior knowledge, such as how many Gaussians there are in the prior density $\rho_{x_{t+1}|\mathcal{Y}_t}$ to be estimated, or the types of functions in the mixture of densities, which distinguishes the algorithm from other existing estimation methods.

The true prior density, a weighed sum of Gaussian distributions, is denoted as

$$\rho(x)=\sum_{i=1}^{N}w_i\frac{\exp\left(-\frac{1}{2}(x-\mu_i)^\intercal{\Sigma_i}^{-1}(x-\mu_i)\right)}{\sqrt{(2\pi)^2|\Sigma_i|}}.$$

Example 1 is a mixture of four Gaussians. The weights are $w_1=w_2=w_3=w_4=0.25$. The means of the Gaussians are

$$\mu_1=[1,0]^\intercal,\mu_2=[0,1]^\intercal,\mu_3=[2,2]^\intercal,\mu_4=[-2,-2]^\intercal$$

and the covariance matrices are

$$\Sigma_1=\Sigma_2=\Sigma_3=\Sigma_4=\begin{bmatrix}1&0\\0&1\end{bmatrix}.$$

$\theta(x)$ is chosen as a multivariate Guassian distribution $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}4&0\\0&4\end{bmatrix}\right)$.

The simulation results of Example 1 is shown in the first row of Figure 1. For each example the true prior is shown in the first column. We first choose the highest order of each variable x_i in the polynomial $q(x,\Lambda)$ to be 4. The density estimate $\hat{\rho}(x)$ is shown in the second column and the third column is the absolute value of error of the density estimates, i.e., $|\rho-\hat{\rho}|$, and $\max|\rho-\hat{\rho}|=0.0175$. We then choose the highest order of each variable x_i in the polynomial $q(x,\Lambda)$ to be 6. The density estimates and the corresponding absolute value of error is given in the fourth and fifth figures of the first row. $\max|\rho-\hat{\rho}|=0.0143$.

By the simulation results of the first example, we note that our proposed parametrization is able to approximate the prior density well with power moments up to order 4 and order 6. With the increase of order, a better estimation result is obtained. This simulation also reveals the fact that the power moment terms are a more compact representation of the density as compared to the discrete data points. Only $(2\cdot2+1)^2=25$ for order 4, and $(2\cdot3+1)^2=49$ power moment terms are used for estimating the prior density, i.e., only 25 and 49 parameters are used to characterize the density for order 4 and order 6 respectively. However it is not quite possible for 25 or 49 discrete particles to represent a density with such a high accuracy.

As to verify the performance of the proposed density surrogate for estimating different mixture of Gaussian densities, we simulate on the following example.

Example 2 is chosen as a mixture of four Gaussians. The weights are $w_1 = w_2 = w_3 = w_4 = 0.25$. The means of the Gaussians are

$$\mu_1 = [1, -1]^\intercal, \mu_2 = [-1, 1]^\intercal, \mu_3 = [2, 2]^\intercal, \mu_4 = [-2, -2]^\intercal$$

and the covariance matrices are

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

$\theta(x)$ is a Gaussian distribution, which is the same as that in Example 1. The highest orders of the polynomial $q(x)$ are 4 and 6. The simulation results are given in the second row of Figure 1. We note that four modes(peaks) are well estimated by the density surrogate. The absolute value of error $\max|\rho - \hat{\rho}|$ are 0.0191 and 0.0071 respectively for order 4 and 6.

In the previous two examples, the performance of our proposed density surrogate is validated by simulations on mixtures of Gaussian distributions. In the following three examples, we simulate more complicated mixtures of densities, which have not been considered in the previous results.

Example 3 is chosen as a mixture of four multivariate generalized asymmetric Laplace distributions. A multivariate generalized asymmetric Laplace (GAL) distribution is defined as follows [23]. If the matrix Σ_i is positive-definite, the GAL distribution is truly d-dimensional and has a probability density function of the form

$$\rho_i(x) = \frac{2\exp\left(\mu_i'\Sigma_i^{-1}x\right)}{(2\pi)^{d/2}\Gamma(s)|\Sigma_i|^{1/2}}\left(\frac{Q_i(x)}{C_i(\Sigma_i, \mu_i)}\right)^{s-d/2} \cdot K_{s-d/2}(Q_i(x)C_i(\Sigma_i, \mu_i)),$$

where $K_\lambda(\cdot)$ is the modified Bessel function with index λ [32] and

$$Q_i(x) = \sqrt{x'\Sigma_i^{-1}x}, \quad C_i(\Sigma_i, \mu_i) = \sqrt{2 + \mu_i'\Sigma_i^{-1}\mu_i}.$$

We simulate $\rho(x) = \sum_{i=1}^{4} w_i \rho_i(x)$ with $w_1 = w_2 = w_3 = w_4 = 0.25$. The means of each GAL distribution $\rho_i(x)$ are respectively

$$\mu_1 = [4, 4]^\intercal, \mu_2 = [4, -4]^\intercal, \mu_3 = [-4, 4]^\intercal, \mu_4 = [-4, -4]^\intercal$$

and the covariance matrices are

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

$\theta(x)$ is chosen as a multivariate Guassian distribution $\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix}\right)$. The highest orders of the polynomial $q(x)$ are 4 and 6. The simulation results are given in the third row of Figure 1. We note that both two estimates are able to approximate the four sharp peaks with no prior knowledge of the prior density, i.e., the prior density $\theta(x)$ is chosen as a multivariate Gaussian distribution. By the simulation results, we note that with higher order power moments used, the estimated peaks tend to be closer to the true ones.

Example 4 is chosen as a mixture of four multivariate Gumbel distributions, where the modes are not clearly separated. The probability density function we consider for simulation is

$$\rho(x) = w_i \sum_{i=1}^{4} \prod_{j=1}^{2} \exp\left(-\left(x_j - \mu_{i,j} + \exp\left(-(x_j - \mu_{i,j})\right)\right)\right)$$

where $\mu_{i,j}$ denotes the j_{th} element of the mean vector μ_i. The weights are $w_1 = w_2 = w_3 = w_4 = 0.25$, and the mean vectors are

$$\mu_1 = [1, 1]^\intercal, \mu_2 = [-2, 0]^\intercal, \mu_3 = [0, -2]^\intercal, \mu_4 = [-2, -2]^\intercal.$$

$\theta(x)$ is a Gaussian distribution, which is the same as that in Example 1. The highest orders of the polynomial $q(x)$ are 4 and 6. The simulation results are given in the third row of Figure 1. With power moments up to order 4, the estimates is not able to approximate the four modes of the prior density well. With more power moments, the density estimation is satisfactory. The absolute value of error $\max|\rho - \hat{\rho}|$ are 0.0283 and 0.0118 respectively for order 4 and 6.

Heavy-tailed filtering problem is drawing more attentions in the control community in the recent years due to its applications in intelligent vehicles and underwater robots [4], [16], [48], which are working in environments with possibly more outlier observations. The proposed non-Gaussian Bayesian filter makes it feasible for us to treat this problem by choosing $\theta(x)$ as a heavy-tailed distribution. In the following example, we simulate mixtures of student-t distributions.

Example 5 is chosen as a mixture of four student-t distributions. The probability density function we consider for simulation is

$$\rho(x) = \sum_{i=1}^{4} \prod_{j=1}^{2} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{(x_j - \mu_{i,j})^2}{\nu}\right)^{-(\nu+1)/2}$$

The mean vectors are chosen as

$$\mu_1 = [1, 1]^\intercal, \mu_2 = [1, -1]^\intercal, \mu_3 = [-1, 1]^\intercal, \mu_4 = [-1, -1]^\intercal.$$

We note that the power moments up to order $\nu - 1$ exist and are finite, for $\nu \in \mathbb{N}_0$. So then we choose $\nu = 8$ to ensure that the density surrogates with moments up to order 4 and 6 both exist. $\theta(x)$ is chosen as $\mathcal{C}(0, 3)$, where $\mathcal{C}$ denotes the Cauchy distribution. The highest order of the polynomial $q(x)$ is 4 and 6. The simulation results are given in the fifth row of Figure 1. With power moments up to order 4, the estimates is not able to approximate the four modes which overlap with each other. And the tail is estimated to be narrower. However with more power moments, the overlapped modes are well approximated. And the tail is well characterized by the density surrogate. The absolute value of error $\max|\rho - \hat{\rho}|$ are 0.0283 and 0.0118 respectively for order 4 and 6.

In conclusion, the 5 examples show the performance of our proposed Bayesian filter in estimating the intractable prior density which can be either light-tailed (sub-Gaussian) or heavy-tailed. The proposed density surrogate doesn't require prior knowledge of either the number of modes or the feasible function classes of the intractable prior density, which is a clear and significant advantage for real applications.

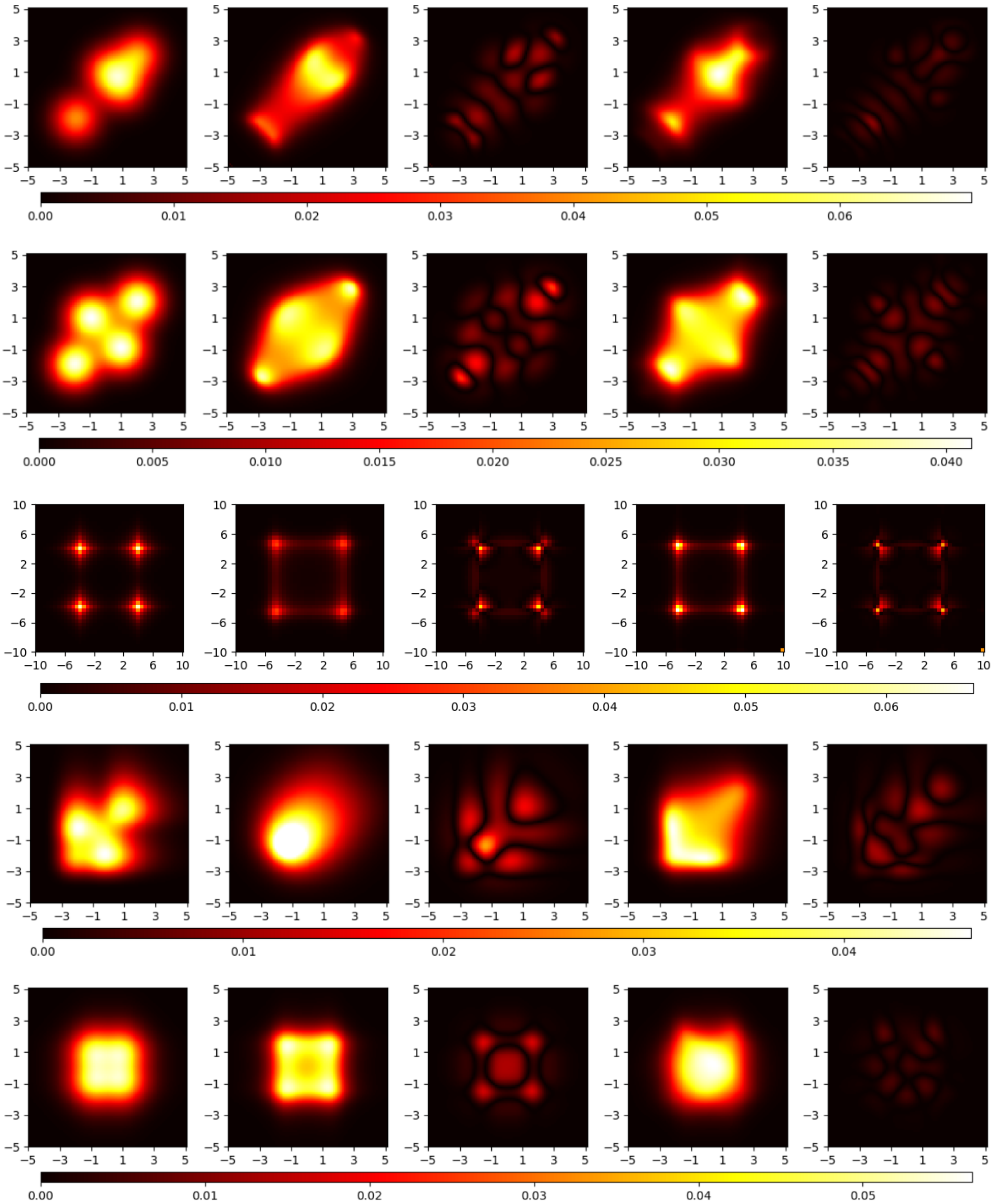

Fig. 1: Simulation results of prior density estimation. Each row is the results of an example. The first column is the true prior density. The second and fourth columns are the density estimates with order 4 and 6 density surrogates respectively. The third and fifth columns are the absolute values of error of the density estimates.

B. A robot localization problem

In this subsection, we focus on a robot localization challenge involving a sensor tasked with measuring the distances between the robot and predefined landmarks. We approximate a small area of the ground as a Euclidean plane equipped with two perpendicular coordinate axes. The robot's position with respect to these coordinates at time step k is denoted as $\{x(k), y(k)\}$. Furthermore, we denote the positions of L landmarks as $\{\check{x}_1, \check{y}_1\}, \cdots, \{\check{x}_L, \check{y}_L\}$. We note that there need to be at least two landmarks, or the location of the robot is not completely observable. In this localization task, the robot is incrementally moved one unit along the positive x and one unit along the positive y directions. The robot's controls are not perfect so it will not move exactly as commanded. Hence we need to consider noise in the particle's movements to have a reasonable chance of capturing the actual movement of the robot. In this simulation, we assume the bearing angle, of which the true value is $\pi/4$, to have an additive Gaussian noise $\mathcal{N}(0, 0.2^2)$. In addition, we assume the moving distance, of which the true value is $\sqrt{2}$, to be corrupted with an additive Gaussian noise $\mathcal{N}(0, 0.05^2)$. Besides, the distance observation of each landmark is corrupted with an additive noise. The robot localization problem is then formulated as follows. Denote the distance of the robot to each known landmark i as $z_i(k)$. The system and observation equations read

$$\begin{bmatrix} x(k+1) \\ y(k+1) \end{bmatrix} = \begin{bmatrix} x(k)+1 \\ y(k)+1 \end{bmatrix} + \begin{bmatrix} w_1(k) \\ w_2(k) \end{bmatrix}$$

and

$$\begin{bmatrix} z_1(k) \\ \vdots \\ z_L(k) \end{bmatrix} = \begin{bmatrix} \sqrt{\left(x(k)-\check{x}_1\right)^2 + \left(y(k)-\check{y}_1\right)^2} \\ \vdots \\ \sqrt{\left(x(k)-\check{x}_L\right)^2 + \left(y(k)-\check{y}_L\right)^2} \end{bmatrix} + \begin{bmatrix} v_1(k) \\ \vdots \\ v_L(k) \end{bmatrix}$$

respectively. We assume $w_1(k), w_2(k)$ to follow the Gaussian distributions $\mathcal{N}(0, 0.1^2)$ which takes the error of controlling the robot into consideration. In the previous results, the noise $v_1(k), \cdots, v_L(k)$ are always assumed to be Gaussian. However, the Gaussian distribution is not an ideal model to characterize the observation error. The reason is that when the landmark is closer, the robot is more likely to detect it and the distance observation is more accurate. However when the landmark is farther, it is more probable that the distance observation is biased more severely from the true one. In this example, we propose to use the right-skewed Gumbel distribution as an example to characterize this property, and to validate the proposed filtering scheme. A comparison of the probability density functions of the Gumbel and Gaussian distributions are given in Figure 2, where the probability density function of the Gumbel distribution is

$$p_{v_i}(x) = 4e^{-4x-e^{-4x}}. \tag{35}$$

The Gaussian distribution has identical mean and variance as the Gumbel distribution, of which the probability density function reads

$$p_{v_i}(x) = \frac{1}{\sqrt{2\pi}\cdot 0.35} e^{\frac{x^2}{2\cdot 0.35^2}}. \tag{36}$$

When $v_i > 0$, we note that it is more possible for v_i, which follows the Gumbel distribution, to be biased from 0. However, when $v_i < 0$, the probability value converges to zero more quickly with the decrease of v_i, compared with the Gaussian distribution. Therefore, the Gumbel distribution characterizes the property well.

However, the asymmetric Gumbel distribution, which acts as the model of observation noise in this localization task, causes severe problem in selecting a proper stochastic filter. To our knowledge, except for the multivariate filter based on power moments proposed (MF) in this paper, only the particle filter (PF) is feasible of performing this task. We adopt a typical sampling-importance resampling (SIR) filter [8] in this simulation. Since the observation equation is a nonlinear equation of the system state, we also simulate the problem using the unscented Kalman filter (UKF). However, the UKF is not able to treat the Gumbel observation noise. In this case, we adopt the Gaussian distribution in (36) as a substitute of (35) as the observation noise for filtering.

The initialization of the three filters are as follows. The robot starts moving from $[-6, -6]^\intercal$. The four landmarks are located at $[-1, 2]^\intercal, [5, 10]^\intercal, [12, 14]^\intercal, [18, 21]^\intercal$. The initial states $[x(0), y(0)]^\intercal$ of the MF and UKF are drawn from the Gaussian distribution $\mathcal{N}\left(\begin{bmatrix} -6 \\ -6 \end{bmatrix}, \begin{bmatrix} 2^2 & 0 \\ 0 & 2^2 \end{bmatrix}\right)$. The states of the 5000 initial particles of the PF are drawn from the Gaussian distribution $\mathcal{N}\left(\begin{bmatrix} -6 \\ -6 \end{bmatrix}, \begin{bmatrix} 5^2 & 0 \\ 0 & 5^2 \end{bmatrix}\right)$, where the variances are relatively larger to cover a wider range of possible locations. The additive noise of the distance observation follows the distribution in (35). We use multivariate power moments up to order 4 to estimate the density surrogates in the implementation of MF. A sample localization process is given in Figure 3, where the estimation results of PF and MF are given. We note that the location estimates of the MF converge to the true locations. The states of the particles of the PF also converge to the true locations. In Figure 4, the root mean square error (RMSE) curves of the MF, PF and UKF of 50 Monte-Carlo simulations are given. We note that when the state estimate converges, the RMSE errors of MF and PF are close, with the one of MF slightly smaller. However, the RMSE of the UKF is significantly larger than those of the MF and PF. It is inevitable since a Gaussian distribution is used as an approximation of the true Gumbel distribution. Moreover, the convergence of the RMSE for the UKF cannot be assured based on the curve.

From the perspective of RMSE, the MF isn't superior than the PF. However, a significant disadvantage of the Particle filter is that it needs to store massive data. For example in this simulation, we need 3 parameters for each particle, including its two-dimensional position and the weight. Hence we need 15000 parameters to characterize the density of the system state. In this simulation, the system state is only 2 dimensional. With the increase of the dimension, the parameters required increases exponentially, which we may not be able to handle. While for the MF, we only need $(4+1)^2 = 25$ parameters for this task, which is a much more compact way for the representation of the density function.

Moreover, since a convex optimization is requred in each filtering step, we need to take the time consumption into consideration. For each iteration of filtering in this example, the average processing time is 6.8 seconds on a 2.5 GHz Intel Core i7 CPU. It is a relatively long time compared to the processing time of the PF. However, it is not quite a long period of time for applications where the processing time is not very sensitive. Moreover, each optimization problem is a convex one with the existence and uniqueness of solution proved. The processing time can be decreased by designing a strategy determining the step length of the gradient methods.

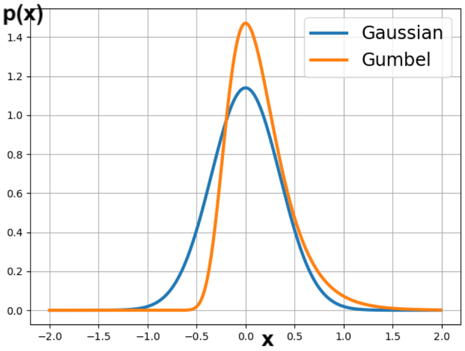

Fig. 2: Probability density functions of the Gaussian and the Gumbel distributions.

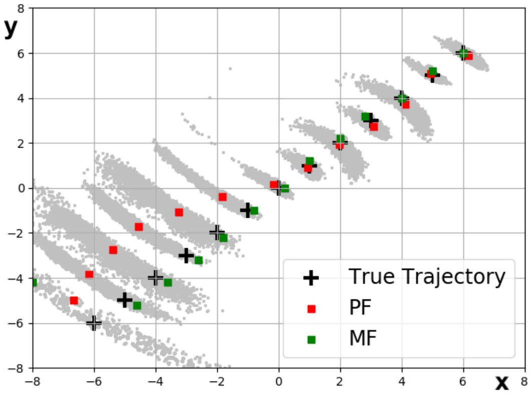

Fig. 3: A sample localization process. The black crosses represent the true trajectory of the robot. The red and green dots represent the location estimates by the particle filter and our proposed filter by multivariate power moments respectively. The gray dots represent the particles of the particle filter at each time step.

VIII. Conclusion

A multivariate non-Gaussian Bayesian filter with the state estimation parameterized as an analytic function is proposed in this paper, where the distribution of the observation noise can be a Lebesgue integrable function, and that of the system noise can be either a probability mass function or a Lebesgue integrable function. It is significant that the proposed algorithm is able to estimate the density functions without prior knowledge of the density of the state x_t, e.g. the number of modes and the feasible function class. It is not required to store massive estimates of the state at discrete points. We prove that the estimated power moments are asymptotically unbiased, approximately the true ones throughout the filtering process given a sufficiently large n. The existence of solution to the multidimensional Hamburger moment problem is established, and a novel Positivstellensatz is proposed to ensure the positiveness of the density surrogate, which also serves as a new result to the moment problem. The parameters of the proposed parametrization can be obtained by a convex optimization scheme. The solution to this problem is proved to exist and be unique by proving that the map from the parameters to the power moments is homeomorphic. Upper bounds of the state estimate are also proposed. In the simulations, we first estimate mixtures of different types of multivariate density functions using power moments, including Gaussian, Laplacian, Gumbel and student-t. The simulation results on the mixture of student-t distributions validates the ability of the proposed algorithm to treat the heavy-tailed filtering problem, which is a current key problem of stochastic filtering. We also simulate the algorithm on a robot localization problem, with a comparison to the particle filter and the unscented Kalman filter. The simulation results reveal the potential of the proposed filter in real engineering applications.

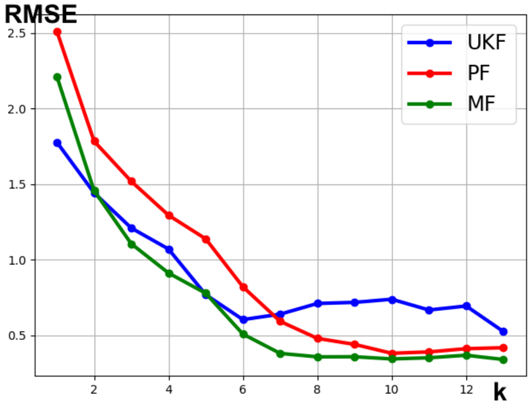

Fig. 4: RMSE as a function of time step k of 50 Monte-Carlo simulations for the unscented Kalman filter (UKF), the particle filter (PF) and the proposed filter by multivariate power moments (MF).

References

[1] Daniel Alspach and Harold Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE transactions on automatic control*, 17(4):439–448, 1972.

[2] Brian DO Anderson and John B Moore. *Optimal filtering*. Courier Corporation, 2012.

[3] Ienkaran Arasaratnam and Simon Haykin. Cubature kalman filters. *IEEE Transactions on automatic control*, 54(6):1254–1269, 2009.

[4] Mingming Bai, Yulong Huang, Yonggang Zhang, and Feng Chen. A novel heavy-tailed mixture distribution based robust kalman filter for cooperative localization. *IEEE Transactions on Industrial Informatics*, 17(5):3671–3681, 2020.

[5] Richard S Bucy and Kenneth D Senne. Digital synthesis of non-linear filters. *Automatica*, 7(3):287–298, 1971.

[6] Christopher I Byrnes and Anders Lindquist. Interior point solutions of variational problems and global inverse function theorems. *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, 17(5-6):463–481, 2007.

[7] Zhiqiang Cai, Francois Le Gland, and Huilong Zhang. *An adaptive local grid refinement method for nonlinear filtering*. PhD thesis, INRIA, 1995.

[8] Zhe Chen et al. Bayesian filtering: From kalman filters to particle filters, and beyond. *Statistics*, 182(1):1–69, 2003.

[9] Yi-Tzuu Chien and King-Sun Fu. On bayesian learning and stochastic approximation. *IEEE Transactions on Systems Science and Cybernetics*, 3(1):28–38, 1967.

[10] P Robert Christian and George Casella. Monte carlo statistical methods. *NY: Springer-Verlag*, 1999.

[11] de JFG Freitas. Bayesian methods for neural networks. *PhD, University of Cambridge, Cambridge, UK*, 1999.

[12] Tryphon T Georgiou and Anders Lindquist. Kullback-leibler approximation of spectral density functions. *IEEE Transactions on Information Theory*, 49(11):2910–2917, 2003.

[13] Alastair R Hall. *Generalized method of moments*. OUP Oxford, 2004.

[14] Peter Hall. On kullback-leibler loss and density estimation. *The Annals of Statistics*, pages 1491–1519, 1987.

[15] Johannes Edmund Handschin and David Q Mayne. Monte carlo techniques to estimate the conditional expectation in multi-stage nonlinear filtering. *International journal of control*, 9(5):547–559, 1969.

[16] Bernard Hanzon and Raimund J Ober. A state-space calculus for rational probability density functions and applications to non-gaussian filtering. *SIAM journal on control and optimization*, 40(3):724–740, 2001.

[17] YC Ho and RCKA Lee. A bayesian approach to problems in stochastic estimation and control. *IEEE transactions on automatic control*, 9(4):333–339, 1964.

[18] Kazufumi Ito and Kaiqi Xiong. Gaussian filters for nonlinear filtering problems. *IEEE transactions on automatic control*, 45(5):910–927, 2000.

[19] Simon Julier, Jeffrey Uhlmann, and Hugh F Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on automatic control*, 45(3):477–482, 2000.

[20] Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[21] Rudolf E. Kalman and Richard S. Bucy. New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(1):95–108, 1961.

[22] Jagat Narain Kapur and Hiremaglur K Kesavan. Entropy optimization principles and their applications. In *Entropy and energy dissipation in water resources*, pages 3–20. Springer, 1992.

[23] Tomasz J Kozubowski, Krzysztof Podgórski, and Igor Rychlik. Multivariate generalized laplace distribution and related random fields. *Journal of Multivariate Analysis*, 113:59–72, 2013.

[24] S. Kullback. Correction to a lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 16(5):652–652, 1970.

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

[26] Harold J Kushner and Amarjit S Budhiraja. A nonlinear filtering algorithm based on an approximation of the conditional distribution. *IEEE Transactions on Automatic Control*, 45(3):580–585, 2000.

[27] Jonathan Q Li and Andrew R Barron. Mixture density estimation. In *NIPS*, volume 12, pages 279–285, 1999.

[28] Hong Lin, Hongye Su, Peng Shi, Zhan Shu, Renquan Lu, and Zheng-Guang Wu. Optimal estimation and control for lossy network: stability, convergence, and performance. *IEEE Transactions on Automatic Control*, 62(9):4564–4579, 2017.

[29] Ta-tung Lin and Stephen S Yau. Bayesian approach to the optimization of adaptive systems. *IEEE Transactions on Systems Science and Cybernetics*, 3(2):77–85, 1967.

[30] David JC MacKay. Choice of basis for laplace approximation. *Machine learning*, 33(1):77–86, 1998.

[31] Magnus Norgaard, Niels K Poulsen, and Ole Ravn. New developments in state estimation for nonlinear systems. *Automatica*, 36(11):1627–1638, 2000.

[32] Frank WJ Olver, Daniel W Lozier, Ronald F Boisvert, Charles W Clark, and ABO Dalhuis. Digital library of mathematical functions. *National Institute of Standards and Technology from http://dlmf. nist. gov/(release date 2011-07-01), Washington, DC*, 2010.

[33] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

[34] Michael Roth, Emre Özkan, and Fredrik Gustafsson. A student's t filter for heavy tailed process and measurement noise. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5770–5774. IEEE, 2013.

[35] Walter Rudin. *Real and complex analysis*. McGraw Hill Education India, 2015.

[36] Simo Särkkä and Jouni Hartikainen. Non-linear noise adaptive kalman filtering via variational bayes. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2013.

[37] Tor Steinar Schei. A finite-difference method for linearization in nonlinear estimation algorithms. *Automatica*, 33(11):2053–2058, 1997.

[38] Konrad Schmüdgen. *The moment problem*, volume 14. Springer, 2017.

[39] İsmail Şenöz, Thijs van de Laar, Dmitry Bagaev, and Bert de Vries. Variational message passing and local constraint manipulation in factor graphs. *Entropy*, 23(7):807, 2021.

[40] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[41] Václav Šmídl and Anthony Quinn. *The variational Bayes method in signal processing*. Springer Science & Business Media, 2006.

[42] Harold W Sorenson and Daniel L Alspach. Recursive bayesian estimation using gaussian sums. *Automatica*, 7(4):465–479, 1971.

[43] J Spragins. A note on the iterative application of bayes' rule. *IEEE Transactions on Information Theory*, 11(4):544–549, 1965.

[44] Aldo Tagliani. Maximum entropy solutions and moment problem in unbounded domains. *Applied mathematics letters*, 16(4):519–524, 2003.

[45] Aldo Tagliani. A note on proximity of distributions in terms of coinciding moments. *Applied Mathematics and Computation*, 145(2-3):195–203, 2003.

[46] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[47] Guangyu Wu and Anders Lindquist. Non-Gaussian Bayesian filtering by density parametrization using power moments. *Automatica*, 153:111061, 2023.

[48] Hao Zhu, Guorui Zhang, Yongfu Li, and Henry Leung. A novel robust kalman filter with unknown non-stationary heavy-tailed noise. *Automatica*, 127:109511, 2021.

Guangyu Wu (S'22) received the B.E. degree from Northwestern Polytechnical University, Xi'an, China, in 2013, and two M.S. degrees, one in control science and engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016, and the other in electrical engineering from the University of Notre Dame, South Bend, USA, in 2018. He is pursuing the Ph.D. degree at Shanghai Jiao Tong University.

Anders Lindquist (M'77–SM'86–F'89–LF'10) received the Ph.D. degree in optimization and systems theory from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1972, and an honorary doctorate (Doctor Scientiarum Honoris Causa) from Technion (Israel Institute of Technology) in 2010, and Doctor Jubilaris from KTH in 2022.

He is currently a Zhiyuan Chair Professor at Shanghai Jiao Tong University, China, and Professor Emeritus at KTH, Stockholm, Sweden. Before that he had a full academic career in the United States, after which he was appointed to the Chair of Optimization and Systems at KTH. Dr. Lindquist is a Member of the Royal Swedish Academy of Engineering Sciences, a Foreign Member of the Chinese Academy of Sciences, a Foreign Member of the Russian Academy of Natural Sciences, a Member of Academia Europaea (Academy of Europe), an Honorary Member the Hungarian Operations Research Society, a Life Fellow of IEEE, a Fellow of SIAM, and a Fellow of IFAC. He received the 2003 George S. Axelby Outstanding Paper Award, the 2009 Reid Prize in Mathematics from SIAM, and the 2020 IEEE Control Systems Award, the IEEE field award in Systems and Control.