Non-Gaussian Bayesian Filtering by Density Parametrization Using Power Moments

Guangyu Wu [a], Anders Lindquist [b]

[a] *Department of Automation, Shanghai Jiao Tong University, Shanghai, China*

[b] *Department of Automation and School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China*

Abstract

Non-Gaussian Bayesian filtering is a core problem in stochastic filtering. The difficulty of the problem lies in parameterizing the state estimates. However the existing methods are not able to treat it well. We propose to use power moments to obtain a parameterization. Unlike the existing parametric estimation methods, our proposed algorithm does not require prior knowledge about the state to be estimated, e.g. the number of modes and the feasible classes of function. Moreover, the proposed algorithm is not required to store massive parameters during filtering as the existing nonparametric Bayesian filters, e.g. the particle filter. The parameters of the proposed parametrization can also be determined by a convex optimization scheme with moments constraints, to which the solution is proved to exist and be unique. A necessary and sufficient condition for all the power moments of the density estimate to exist and be finite is provided. The errors of power moments are analyzed for the density estimate being either light-tailed or heavy-tailed. Error upper bounds of the density estimate for the one-step prediction are proposed. Simulation results on different types of density functions of the state are given, including the heavy-tailed densities, to validate the proposed algorithm.

Key words: Bayesian methods; filtering; power moments.

1 Introduction

Stochastic filtering theory has been a fundamental topic in several areas including controls and signal processing for years and is applied in the various engineering and scientific areas, including communications, machine learning, neuroscience, economics, finance, political science, and many others. Pioneered by Norbert Wiener [38] and Andrey N. Kolmogorov [20,29] in the 1940s, and promoted by Bode and Shannon [4], Zadeh and Ragazzini [39] and others, a milestone of stochastic filtering theory was achieved by Rudolf E. Kalman [18, 26] in the 1960s. The Kalman filter (KF) consists of an iterative measurement-time update process. In the time-update step, the one-step ahead prediction of state is calculated; in the measurement-update step, the correction to the estimate of state according to the current observation is calculated. Moreover, the Kalman filter is indeed a time-variant Wiener filter [2].

Even though the Kalman filter was originally derived with the orthogonal projection method under the LQG procedures, it has a decent Bayesian interpretation and can be derived within a Bayesian framework. Indeed, the Kalman filter can be regarded as a class of Bayesian filters of which the probability density function of the states and noises are Gaussian. Without exaggeration, the research on Bayesian filtering is inspired by the development of Kalman filtering. One of the first explorations of iterative Bayesian estimation is found in Ho and Lee's paper [16], where the principle and procedure of Bayesian filtering are specified. Sprangins [31] discussed the iterative application of Bayes rule to sequential parameter estimation. Lin and Yau [24] and Chien and Fu [9] discussed Bayesian approach to optimization of adaptive systems.

In practical situations, the state of the system and the noises do not 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 does not apply. Moreover, when the probability density of the state is multi-modal, it is not feasible to estimate it with a Gaussian distribution. In the scenarios where the Kalman filter does not apply, we consider the Bayesian filter naturally due to its relaxation on the type of the

Email addresses: chinarustin@sjtu.edu.cn (Guangyu Wu), alq@kth.se (Anders Lindquist).

densities.

However, there is always a tradeoff in stochastic filtering. For the Kalman filter, the densities being Gaussian makes it feasible to obtain an analytic form of solution of the convolution in the time update step. Bayesian filter does not require the densities to be Gaussian, but the solvability of the convolution (integration) is not guaranteed. In previous research results, several numerical methods have been proposed to obtain an analytic solution to the integration in the time update step, including Gaussian/Laplace approximation [25], iterative quadrature [11,22,37], Gaussian sum approximation [1,30] and state-space calculus [15]. These are the parametric methods for parameterizing the probability density function in a continuous form. However, the flexibility of those methods is too limited, which makes it difficult to apply the methods to a wide class of density functions. Moreover, quantitative approximation performance analyses of the methods, e.g. an error upper bound of estimation, have not been proposed yet, which severely decreases the value of these methods in practical use.

Meanwhile, there are also several methods which characterize the density in a discrete form of the function, including mulitgrid method and point-mass approximation [5,7] and Monte Carlo sampling approximation [10, 14]. These nonparametric methods impose no prior constraints on the density functions which seem to enjoy the maximum flexibility, however the tradeoff is also very severe: quite a number 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 always not able to obtain its value of probability. At the same time, as to guarantee the computation efficiency of these algorithms in filtering, resampling is widely used in the filters to avoid depletion of particles with small probability values. 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.

Let us return to the methods parameterizing the density function in a continuous form. The Kalman filter estimates the first two orders of moments, but it is natural to consider using the moments of higher orders for filtering. Unfortunately the classes of density functions, which the filters are able to treat, are still very limited in previous papers [32].

In this paper, we first formulate the non-Gaussian Bayesian filtering problem in Section 2. We propose to use the higher order moments to characterize the density function. The density surrogate is also defined. A construction of the density surrogate, i.e., parametrization of the density function is proved to exist and proposed in Section 3. In doing this, we follow the procedure of [12], where a large class of trigonometric moment problems are solved by minimizing a Kulback-Leibler type criterion. Indeed, this theory can be modified to the power moment problem, which is what is needed here. The parameters of the model can be determined by a convex optimization scheme and the map from the parameters to the power moments is proved to be a homeomorphism. It ensures that the gradient-based optimization algorithms can be applied to determining the parameters of the density surrogate. The parametrization is in terms of a prior density θ, and in Section 4, we give a sufficient and necessary condition on θ for the density surrogate $\hat{\rho}$ to have all power moments to exist and be finite. With the prior distribution selected as a sub-Gaussian distribution, the estimated moments of the density surrogate are proved to be asymptotically unbiased from the true ones when using the density surrogate with the highest order of the moments used tending to infinity. By selecting a sufficient large n, we have that the estimated moments are approximately the true ones, i.e., using the density surrogate will not bring significant cumulative errors to the moment estimation of the subsequent filtering steps. To our knowledge, the asymptotic unbiasedness of the statistics of the estimated density function has not previously been proved for the Bayesian filters of which the state is a continuous function, and cumulative errors are always hard to predict. Moreover, an upper bound of the approximation error in the sense of total variation distance is proposed, which has not previously been done for the Bayesian filters with no prior constraints on the classes of the densities. Error upper bounds of the density estimate for the probability of subsets of the real line are also proposed. Simulation results of different classes of density functions, including heavy-tailed ones, are given in Section 5 to validate the performance of our filter.

2 Problem formulation

In this paper, following [15], we consider the non-Gaussian filtering problem for the first order system

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

$t = 0, 1, 2, \ldots$. The state x_t is a random variable defined on $\mathbb{R}$, and f_t, h_t are assumed to be known real numbers. The system noise η_t is a random variable defined on $\mathbb{R}$, which can be either continuous or discrete. When η_t is continuous, the probability density function is assumed to be non-Gaussian. The observation noise ϵ_t is assumed to be a Lebesgue integrable function. The noises are assumed to be independent from each other, and their densities are denoted as ρ_{η_t} and ρ_{ϵ_t}.

We adopt the Bayesian filter as used in [15]. Denoting

the collection of observations $y_t, y_{t-1}, \cdots, y_0$ by $\mathcal{Y}_t$, the conditional densities of the measurement and time updates are given by the following

Measurement update: For $t = 0$,

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

for $t \geq 1$,

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

Time update: For $t \geq 0$,

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

As are derived in (2), (3) and (4), $\rho_{x_t|\mathcal{Y}_t}(x)$ and $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ are evaluated at x. In the following part of this paper, we write $\rho_{x_t|\mathcal{Y}_t}, \rho_{x_{t+1}|\mathcal{Y}_t}$ for simplicity, if there is no ambiguity. Even though the densities are all non-Gaussian, the measurement update (3) is a multiplication of two densities, therefore can be performed easily. But it is not always possible to obtain an explicit form of the one-step prediction in (4) when the densities are not Gaussian [8]. Now the problem becomes the approximation of $\rho_{x_{t+1}|\mathcal{Y}_t}$. However, we notice that the power moments of $\rho_{x_{t+1}|\mathcal{Y}_t}$: i.e.,

$$\sigma_{k,t+1} := \int_{\mathbb{R}} x^k \rho_{x_{t+1}|\mathcal{Y}_t} dx = \mathbb{E}\left(x_{t+1}^k|\mathcal{Y}_t\right), \tag{5}$$

are easy to obtain. In fact, by (1),

$$\begin{aligned}\sigma_{k,t+1} =& \mathbb{E}\left((f_t x_t + \eta_t)^k|\mathcal{Y}_t\right) \\ =& \mathbb{E}\left(\sum_{j=0}^{k}\binom{k}{j} f_t^j x_t^j \cdot \eta_t^{k-j}|\mathcal{Y}_t\right) \\ =& \sum_{j=0}^{k}\binom{k}{j}\mathbb{E}\left(f_t^j x_t^j \cdot \eta_t^{k-j}|\mathcal{Y}_t\right).\end{aligned}$$

Therefore, since x_t and η_t are independent,

$$\sigma_{k,t+1} = \sum_{j=0}^{k}\binom{k}{j} f_t^j \mathbb{E}\left(x_t^j|\mathcal{Y}_t\right)\mathbb{E}\left(\eta_t^{k-j}\right) \tag{6}$$

for $k = 1, \cdots, 2n$. Inspired by the method of moments, we propose to use the truncated power moments to estimate $\rho_{x_{t+1}|\mathcal{Y}_t}$.

Previous research has been focusing on density approximation. In the Kalman filter (and its extended forms, e.g. extended Kamlan filter and unscented Kalman filter), approximation of the one-step prediction is a parametric estimation problem, which is done by estimating the first and second order moments. On the other hand, the particle filter is proposed as a non-parametric algorithm to tackle this problem. However, there is no convenient way of error analysis and its performance suffers from sample depletion.

In this paper, we propose a filter which not only admits treating the non-Gaussian state estimation problem but also provides an analytic error analysis to measure the performance of filtering. To this end, we define the probabilities with the identical truncated power moment sequence.

Definition 2.1. A probability density function, whose first $2n$ power moments coincide with those of the probability density ρ, is called an order-$2n$ density surrogate of ρ and is denoted by ρ^{2n}.

Denoting the density prediction as $\hat{\rho}$, we propose to perform each iteration of Bayesian filtering with the density surrogate as in Algorithm 1.

Algorithm 1 Bayesian filtering with density surrogate at time t.

Input: System parameters : f_t, h_t; non-Gaussian densities : η_t, ϵ_t; prediction at time $t-1$: $\rho_{x_0}(x)$ or $\hat{\rho}_{x_t|\mathcal{Y}_{t-1}}(x)$;

Output: Prediction at time t : $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}(x)$.

1: Calculate $\hat{\rho}_{x_t|\mathcal{Y}_t}$ by (2) or (3);
2: Calculate σ_t by (6);
3: Determine an order-$2n$ density surrogate $\rho^{2n}_{x_{t+1}|\mathcal{Y}_t}$, of which the truncated moment sequence is σ_t. The density estimate for the one-step prediction at time $t+1$ is then chosen as the density surrogate, i.e., $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t} = \rho^{2n}_{x_{t+1}|\mathcal{Y}_t}$.

Now the problem comes down to constructing an order-$2n$ density surrogate. Since the domain of ρ is $\mathbb{R}$, the problem becomes a Hamburger moment problem [27]. In the next section, we will give a formal definition to the Hamburger moment problem we will treat and give a solution to the problem, i.e. a parametrization of the density surrogate.

3 Parametrization of the density surrogate using power moments

As seen from (5), determining a density $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$ given the power moments is a (truncated) Hamburger moment

problem, For simplicity we omit the subscript t in all terms.

Definition 3.1. A sequence

$$\bar{\sigma}_{2n} = (\sigma_0, \sigma_1, \ldots, \sigma_{2n}) \tag{7}$$

is a feasible $2n$-sequence, if there is a random variable X with a probability density $\rho(x)$ defined on $\mathbb{R}$, whose moments are given by (6), that is,

$$\sigma_k = \mathbb{E}\{X^k\} = \int_{\mathbb{R}} x^k \rho(x) dx, \quad k = 0, 1, \ldots, 2n. \tag{8}$$

We say that any such random variable X has a $\bar{\sigma}_{2n}$-feasible distribution. As to ensure the existence of ρ, the Hankel matrix

$$\Sigma := \begin{bmatrix} \sigma_0 & \sigma_1 & \ldots & \sigma_n \\ \sigma_1 & \sigma_2 & \ldots & \sigma_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_n & \sigma_{n+1} & \ldots & \sigma_{2n} \end{bmatrix} \tag{9}$$

needs to be positive definite, i.e., the sequence $\bar{\sigma}_{2n}$ needs to be a positive one.

For Bayesian filtering, we need to propagate the density estimates throughout the filtering process, which makes it necessary to derive an analytic form of $\rho(x)$ with finitely many parameters. It is called the parametrization of the density function, which is essentially a dimension reduction problem. We emphasize that the solution to the problem is not unique and that there are in general infinitely many solutions. Next we proceed to describe these.

Observe that the moment conditions

$$\sigma_k = \int_{\mathbb{R}} x^k \rho(x) dx, \quad k = 0, 1, \ldots, 2n \tag{10}$$

can be written in the matrix form

$$\int_{\mathbb{R}} G(x) \rho(x) G^T(x) dx = \Sigma, \tag{11}$$

where

$$G(x) = \begin{bmatrix} 1 & x & \cdots & x^{n-1} & x^n \end{bmatrix}^T,$$

and Σ is the Hankel matrix of the form (9), of which the entries $\sigma_k, k = 0, \cdots, 2n$ are calculated by (6). Consequently, we have an order $2n$ moment estimation problem as defined in Definition 3.1. Indeed, it is an Hankel matrix representation of the Hamburger moment problem.

We note that Σ is a Hankel matrix hence symmetric. By (11) we have that, for each $a \in \mathbb{R}_{n+1}/\{0\}$, $a'\Sigma a = \int a(x)^2 \rho(x) dx$, where $a(x) = a'G(x)$. Since $a(x)^2$ is positive except possibly in a set of measure zero, $a'\Sigma a > 0$ for all a, and by Definition 4.1.11 in [17], Σ is positive definite. By Corollary 9.2 in [27], the positive definiteness of Σ also shows that there exists at least one solution to the corresponding truncated moment problem.

Let $\mathcal{P}$ be the space of probability density functions on the real line with support there, and let $\mathcal{P}_{2n}$ be the subset of all $\rho \in \mathcal{P}$ which have at least $2n$ finite moments (in addition to σ_0, which of course is 1). We note that the class of $\rho \in \mathcal{P}$ satisfying (11) is nonempty since Σ is positive definite ($\Sigma \succ 0$). In fact, Σ is in the range of the linear integral operator

$$\Gamma : \rho \mapsto \Sigma = \int_{\mathbb{R}} G(x) \rho(x) G^T(x) dx, \tag{12}$$

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

Let θ be an arbitrary prior density in $\mathcal{P}$ and consider the Kullback-Leibler (KL) (pseudo) distance

$$\mathbb{KL}(\theta \| \rho) = \int_{\mathbb{R}} \theta(x) \log \frac{\theta(x)}{\rho(x)} dx \tag{13}$$

between θ and ρ. Although not symmetric in its arguments, the KL distance is jointly convex. It is widely used in density estimation [13, 23, 35].

In [12], the KL distance was used as a distance measure between spectral densities. In this section, following the line of thought of [12], we introduce a parametrization of $\rho \in \mathcal{P}_{2n}$ satisfying (10), which is induced by the KL distance. The results are very similar to those in [12], but since here we are dealing with a power moment problem rather than a trigonometric moment problem as in [12], important details of the proofs are different, so we need to proceed with care.

Theorem 3.2. *Let Γ be defined by (12), and let*

$$\mathcal{L}_+ := \left\{ \Lambda \in \text{range}(\Gamma) \mid G(x)^T \Lambda G(x) > 0, x \in \mathbb{R} \right\}.$$

Given any $\theta \in \mathcal{P}$ and any $\Sigma \succ 0$, there is a unique $\rho \in \mathcal{P}_{2n}$ that minimizes (13) *subject to $\Gamma(\rho) = \Sigma$, i.e., subject to* (11)*, namely*

$$\hat{\rho} = \frac{\theta}{G^T \hat{\Lambda} G}, \tag{14}$$

where $\hat{\Lambda}$ is the unique solution to the problem of minimizing

$$\mathbb{J}_\theta(\Lambda) := \text{tr}(\Lambda\Sigma) - \int_{\mathbb{R}} \theta(x) \log \left[G(x)^T \Lambda G(x) \right] dx \tag{15}$$

over all $\Lambda \in \mathcal{L}_+$. Here $\mathrm{tr}(M)$ denotes the trace of the matrix M.

Proof. First form the Lagrangian

$$L(\rho, \Lambda) = \mathbb{KL}(\theta \| \rho) + \mathrm{tr}(\Lambda(\Gamma(\rho) - \Sigma)),$$

where $\Lambda \in \mathrm{range}(\Gamma)$ is the matrix-valued Lagrange multiplier, and consider the problem of maximizing the dual functional

$$\Lambda \mapsto \inf_{\rho \in \mathcal{P}_{2n}} L(\rho, \Lambda). \tag{16}$$

Clearly $\rho \mapsto L(\rho, \Lambda)$ is strictly convex, so to be able to determine the right member of (16), we must find a $\rho \in \mathcal{P}_{2n}$, for which the directional derivative $\delta L(\rho, \Lambda; \delta\rho) = 0$ for all relevant $\delta\rho$. This will further restrict the choice of Λ. Setting

$$q(x) := G(x)^T \Lambda G(x), \tag{17}$$

we have

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

with the directional derivative

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

which has to be zero at a minimum for all variations $\delta\rho$. Clearly this can be achieved only if $q(x) = \theta(x)/\rho(x)$ for all $x \in \mathbb{R}$. To complete the proof of Theorem 3.2, we need some more preliminary results, but let us first make an important observation.

Remark. The parametrization of the density surrogate by the Hankel matrix restricts the highest order of the terms of the denominator to be even, i.e., $2n$. Indeed, it is the necessary condition for a polynomial to be always positive everywhere on $\mathbb{R}$. A polynomial for which the order of the highest order term is odd always has a real zero, and the value of the polynomial changes sign at that point. It makes constructing the density surrogate problematic.

In particular, this requires the condition $q(x) > 0$ for all $x \in \mathbb{R}$, so by (11) and (17), we obtain the constraint $\Lambda \in \mathcal{L}_+$, which is proved in the following lemma.

Lemma 3.3. *$\Lambda \in \mathcal{L}_+$ only if $q(x) > 0$.*

Proof. Since $\Lambda \in \mathcal{L}_+$, we write Λ as

$$\int_{\mathbb{R}} G(y)\psi(y)G^T(y)dx = \Lambda,$$

where $\psi \in \mathcal{P}_{2n}$. Therefore we have

$$G^T(x) \int_{\mathbb{R}} G(y)\psi(y)G^T(y)dxG(x) = G^T(x)\Lambda G(x) = q(x),$$

and hence

$$q(x) = \int_{\mathbb{R}} \varphi_x(y)\psi(y)dy,$$

for each x, where

$$\varphi_x(y) = \left[G^T(x)G(y)\right]^2 \geq 0.$$

However, for each fixed x, $\varphi_x(y)$ is a polynomial such that $\varphi_x(0) = 1$, and hence $\varphi_x(y) = 0$ at most in a finite number of y. Consequently, since $\psi(y) > 0$, we have $q(x) > 0$ for all x. □

Moreover, a possible minimizer must have the form

$$\rho = \frac{\theta}{q},$$

and the dual functional must be

$$L\left(\frac{\theta}{q}, \Lambda\right) = -\mathbb{J}_\theta(\Lambda) + \int_{\mathbb{R}} \theta(x)dx,$$

where $\mathbb{J}_\theta(\Lambda)$ is given by (15). Therefore the dual problem amounts to minimizing $\mathbb{J}_\theta(\Lambda)$ over $\mathcal{L}_+$. To conclude the proof of Theorem 3.2 we need the following theorem, which will be proved in the following part of this section.

Theorem 3.4. *The functional $\mathbb{J}_\theta(\Lambda)$ has a unique minimum $\hat{\Lambda} \in \mathcal{L}_+$. Moreover*

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

By this theorem

$$\hat{\rho} = \frac{\theta}{\hat{q}}, \quad \text{where } \hat{q} = G^T\hat{\Lambda}G,$$

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

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

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 $\rho = \hat{\rho}$. This completes the proof of Theorem 3.2. □

To prove Theorem 3.4, we need to consider the dual problem to minimize $\mathbb{J}_\theta(\Lambda)$ over $\mathcal{L}_+$.

Lemma 3.5. *Any stationary point of $\mathbb{J}_\theta(\Lambda)$ must satisfy the equation*

$$\omega(\Lambda) = \Sigma, \tag{18}$$

where the map $\omega : \mathcal{L}_+ \mapsto \mathcal{S}_+$ between $\mathcal{L}_+$ and $\mathcal{S}_+ := \{\Sigma \in \text{range}(\Gamma) \mid \Sigma \succ 0\}$ is defined as

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

with q defined by (17).

Proof. From (15) and (17) we have

$$\mathbb{J}_\theta(\Lambda) = \text{tr}\{\Lambda\Sigma\} - \int_{\mathbb{R}} \theta(x)\log q(x)dx,$$

and therefore, using the fact that

$$\delta q(\Lambda;\delta\Lambda) = G^T\delta\Lambda G = \text{tr}\{\delta\Lambda G G^T\},$$

we have the directional derivative

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

which is zero for all $\delta\Lambda \in \text{range}(\Gamma)$ if and only if (18) holds. This completes the proof. □

To prove Theorem 3.4, we also 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 (18) has a unique solution, and hence that there is a unique minimum of the dual functional $\mathbb{J}_\theta$. We start with injectivity.

Lemma 3.6. *Suppose $\Lambda \in$ range(Γ). Then the map*

$$\Lambda \mapsto G^T\Lambda G \tag{19}$$

is injective.

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

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

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

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

Consequently we have $[G(x)^T G(y)]^2\psi(y) = 0$, for all $x, y \in \mathbb{R}$, which clearly implies that $\psi = 0$, and hence that $\Lambda = 0$. Consequently the map (19) is injective, as claimed. □

To prove that $\omega : \mathcal{L}_+ \to \mathcal{S}_+$ 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

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

where $q_1 = G^T\Lambda_1 G$ and $q_2 = G^T\Lambda_2 G$. In view of Lemma 3.3, this implies that $q_1 = q_2$, so by Lemma 3.6 this implies that $\Lambda_1 = \Lambda_2$, establishing that ω is injective.

Next, we shall prove that $\omega : \mathcal{L}_+ \mapsto \mathcal{S}_+$ is also surjective. To this end, 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, and hence diffeomorphic to this space. 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 $\mathcal{S}_+$. (For a more general statement, see Theorem 2.1 in [6].) Consequently it just 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 suppose it is infinite. Then, 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, which completes the proof of Theorem 3.4.

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 $\Gamma(\rho) = \Sigma$. The dual functional has the following property.

Lemma 3.7. *The dual functional $\mathbb{J}_\theta(\Lambda)$ is strictly convex.*

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

$$\delta^2\mathbb{J}_\theta(\Lambda;\delta\Lambda) = \int_{\mathbb{R}} \frac{\theta(x)}{q(x)^2}\left(G(x)^T\delta\Lambda G(x)\right)^2 dx \tag{20}$$

By (20), we have $\delta^2 \mathbb{J}_\theta \geq 0$, so it remains to show that

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

which follows directly from Lemma 3.6, replacing Λ by $\delta\Lambda$. □

This leads to the following update of Algorithm 1, which is executed for a particular choice of θ.

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

Input: System parameters: f_t, h_t; non-Gaussian densities: η_t, ϵ_t; prediction at time $t-1$: $\rho_{x_0}(x)$ or $\hat{\rho}_{x_t|\mathcal{Y}_{t-1}}(x)$;

Output: Prediction at time t: $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}(x)$.

1: Calculate $\hat{\rho}_{x_t|\mathcal{Y}_t}(x)$ by (2) or (3);
2: Calculate Σ by (6);
3: Do the optimization (15) to obtain the order-$2n$ density surrogate of (4), which is the new predictor $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}(x)$.

4 Tails and error analyses of the proposed filter

Given a prior probability density θ, Algorithm 2 provides us with a unique solution $\hat{\rho}$ to the truncated Hamburger moment problem, that is, with a unique surrogate probability density $\hat{\rho}$. In this calculation the choice of n may be crucial, as $\hat{\rho}$ may have only a finite number of moments. Indeed, we may want to consider situations when the density has a heavy tail. In this section, we establish the conditions on the prior θ for the density estimate $\hat{\rho}$ to satisfy tail specifications.

4.1 Light-tailed density surrogate and the moment error propagation

We first introduce the concept of the sub-Gaussian distributions [36], which, loosely speaking, are distributions whose tails are dominated by the tails of a Gaussian distribution, i.e., decay at least as fast as a Gaussian. More precisely, a random variable X is called sub-Gaussian if the moments of X satisfy

$$\|X\|_{L^p} = \left(\mathbb{E}|X|^p\right)^{1/p} \leq K_1\sqrt{p} \text{ for all } p \geq 1 \qquad (21)$$

or the moment generating function of X^2 is bounded at some point, namely

$$\mathbb{E}\left[\exp\left(X^2/K_2^2\right)\right] \leq 2 \qquad (22)$$

where $K_1, K_2 \in \mathbb{R}_+$ are two parameters. We denote the space of all sub-Gaussian distributions as $\mathcal{SG}$. Then we have the following theorem.

Theorem 4.1. *All power moments of the density surrogate $\hat{\rho}$ exist and are finite if and only if the prior $\theta \in \mathcal{SG}$.*

Proof. We first prove the necessity. We have

$$\mathbb{E}\left[|\hat{x}|^p\right] = \int_{\mathbb{R}} |x|^p \hat{\rho}(x)dx = \int_{\mathbb{R}} |x|^p \frac{\theta(x)}{\hat{q}(x)}dx.$$

By Lemma 3.3, we have that $\hat{q}(x) > 0$. We also note that $|x|^p, \theta(x)$ are both positive.

Since the prior $\theta \in \mathcal{SG}$, by (21) we have

$$\begin{aligned}\int_{\mathbb{R}} |x|^p \frac{\theta(x)}{\hat{q}(x)}dx &\leq \frac{1}{\min_x \hat{q}(x)} \int_{\mathbb{R}} |x|^p \theta(x)dx \\ &\leq \frac{1}{\min_x \hat{q}(x)} \left(K_1\sqrt{p}\right)^p,\end{aligned}$$

which proves that $\mathbb{E}\left[|\hat{x}|^p\right]$, $p = 1, 2, 3, \ldots$, are all finite. However we have $|\mathbb{E}\left[\hat{x}^p\right]| \leq \mathbb{E}\left[|\hat{x}|^p\right]$. Then $\mathbb{E}\left[\hat{x}^p\right]$ are also finite, and hence all moments exist and are finite.

Next we prove the sufficiency. In view of (14),

$$\begin{aligned}&\mathbb{E}_\theta\left[\exp\left(x^2/K_2^2\right)\right] \\ =&\sum_{i=0}^{\infty} \frac{1}{i!} \int_{\mathbb{R}} \left(x^2/K_2^2\right)^i \theta(x)dx \\ =&\sum_{i=0}^{\infty} \frac{1}{i!} \int_{\mathbb{R}} \left(x^2/K_2^2\right)^i G(x)^T \Lambda G(x)\hat{\rho}(x)dx\end{aligned}$$

Then, with $\hat{\Lambda}_{j,k}$ being the entries of the matrix $\hat{\Lambda}$, we have

$$\begin{aligned}&\mathbb{E}_\theta\left[\exp\left(x^2/K_2^2\right)\right] \\ =&\sum_{i=0}^{\infty}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\hat{\Lambda}_{j,k}}{i!} \int_{\mathbb{R}} \left(x^2/K_2^2\right)^i x^{j+k-2}\hat{\rho}(x)dx \\ =&\sum_{i=0}^{\infty}\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\hat{\Lambda}_{j,k}}{i!K_2^{2i}} \mathbb{E}[\hat{x}^{2i+j+k-2}].\end{aligned}$$

Since all power moments of $\hat{\rho}$ exist and are finite, it is always possible to choose a $K_2 \in \mathbb{R}_+$ such that

$$\left|\sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\hat{\Lambda}_{j,k}}{i!K_2^{2i}} \mathbb{E}[\hat{x}^{2i+j+k-2}]\right| \leq \frac{1}{2^i},\ i \geq 0,$$

and then we have

$$\mathbb{E}_\theta\left[\exp\left(x^2/K_2^2\right)\right] \leq \frac{1}{1-1/2} = 2,$$

i.e., the prior θ is sub-Gaussian by (22). This completes the proof of sufficiency. □

Error propagation through the whole filtering process is a problem in the filter design. Unlike other pdf approximation problems, the estimation is done at each time step t, which means that the approximation errors of the previous iterations may have a cumulative effect on the current estimation.

With the proposed condition on θ, we are always able to ensure the existence and boundedness of all power moments of $\hat{\rho}$. We will first analyze the error propagation of the first $2n$ power moments when $\theta \in \mathcal{SG}$. Since the approximation errors caused by the time updates could effect the measurement updates, we analyze the first $2n$ moment terms of not only $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$ but also $\hat{\rho}_{x_t|\mathcal{Y}_t}$.

Theorem 4.2. *Suppose $\hat{\rho}_{x_1|\mathcal{Y}_0}$ is a surrogate for $\rho_{x_1|\mathcal{Y}_0}$, and let $\hat{\rho}_{x_t|\mathcal{Y}_t}$ and $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$ 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 in n from those of $\rho_{x_t|\mathcal{Y}_t}$ and $\rho_{x_{t+1}|\mathcal{Y}_t}$ respectively and are approximately identical to them for a sufficiently large n, given that all power moments of x_t and the corresponding $\hat{x}_t$ exist and are finite.*

Proof. For the sake of simplicity, we omit the normalizing factor in the measurement update equations (2) and (3), which does not affect the remaining 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.,

$$\int_{\mathbb{R}} x^k \rho_{x_1|\mathcal{Y}_0} dx = \int_{\mathbb{R}} x^k \hat{\rho}_{x_1|\mathcal{Y}_0} dx, \quad k = 0, \cdots, 2n. \tag{23}$$

Then, referring to (3), we can write the moment terms of $\rho_{x_1|\mathcal{Y}_1}$ as

$$\mathbb{E}\left(x_1^k|\mathcal{Y}_1\right) = \int_{\mathbb{R}} x^k \rho_{\epsilon_1}\left(y_1 - h_1 x\right) \rho_{x_1|\mathcal{Y}_0}(x) dx$$

for $k = 0, \cdots, 2n$, and those of $\hat{\rho}_{x_1|\mathcal{Y}_1}$ as,

$$\mathbb{E}\left(\hat{x}_1^k|\mathcal{Y}_1\right) = \int_{\mathbb{R}} x^k \rho_{\epsilon_1}\left(y_1 - h_1 x\right) \hat{\rho}_{x_1|\mathcal{Y}_0}(x) dx$$

for $k = 0, \cdots, 2n$. Therefore we have,

$$\begin{aligned}&\mathbb{E}\left(x_1^k|\mathcal{Y}_1\right) - \mathbb{E}\left(\hat{x}_1^k|\mathcal{Y}_1\right)\\ &= \int_{\mathbb{R}} x^k \rho_{\epsilon_1}\left(y_1 - h_1 x\right)\left(\rho_{x_1|\mathcal{Y}_0}(x) - \hat{\rho}_{x_1|\mathcal{Y}_0}(x)\right) dx.\end{aligned} \tag{24}$$

We note that $\rho_{\epsilon_1}\left(y_1 - h_1 x\right)$ is analytic almost everywhere. Assume $\rho_{\epsilon_1}\left(y_1 - h_1 x\right)$ is analytic at point x_0, then it is feasible for us to write the Taylor series at this point. Without loss of generality, we take $x_0 = 0$, then we have

$$\rho_{\epsilon_1}\left(y_1 - h_1 x\right) = \sum_{i=0}^{+\infty} \frac{\rho_{\epsilon_1}^{(i)}\left(y_1\right)}{i!} x^i$$

Since all power moments of x_1 and $\hat{x}_1$ exist and are finite, we have

$$\begin{aligned}&\mathbb{E}\left(x_1^k|\mathcal{Y}_1\right) - \mathbb{E}\left(\hat{x}_1^k|\mathcal{Y}_1\right)\\ =&\sum_{i=0}^{+\infty} \frac{\rho_{\epsilon_1}^{(i)}\left(y_1\right)}{i!} \int_{\mathbb{R}} x^{k+i}\left(\rho_{x_1|\mathcal{Y}_0} - \hat{\rho}_{x_1|\mathcal{Y}_0}\right) dx,\end{aligned}$$

which, in view of (23), yields

$$\begin{aligned}&\mathbb{E}\left(x_1^k|\mathcal{Y}_1\right) - \mathbb{E}\left(\hat{x}_1^k|\mathcal{Y}_1\right)\\ =&\sum_{i=2n-k+1}^{+\infty} \frac{\rho_{\epsilon_1}^{(i)}\left(y_1\right)}{i!} \int_{\mathbb{R}} x^{k+i}\left(\rho_{x_1|\mathcal{Y}_0} - \hat{\rho}_{x_1|\mathcal{Y}_0}\right) dx, \quad k = 0, 1, \ldots, 2n,\end{aligned} \tag{25}$$

which tends to zero as $n \to \infty$. Thus, by properly selecting a sufficient large n, we have

$$\mathbb{E}\left(x_1^k|\mathcal{Y}_1\right) \approx \mathbb{E}\left(\hat{x}_1^k|\mathcal{Y}_1\right), \quad k = 0, \cdots, 2n,$$

Similarly we can prove

$$\mathbb{E}\left(x_t^k|\mathcal{Y}_t\right) \approx \mathbb{E}\left(\hat{x}_t^k|\mathcal{Y}_t\right), \quad k = 0, \cdots, 2n,$$

and

$$\mathbb{E}\left(x_{t+1}^k|\mathcal{Y}_t\right) \approx \mathbb{E}\left(\hat{x}_{t+1}^k|\mathcal{Y}_t\right), \quad k = 0, \cdots, 2n,$$

as claimed. □

Theorem 4.2 proves that the first $2n$ moment terms of the estimated densities with the density surrogate are approximately the true ones throughout the whole filtering process for sufficiently large n, i.e., $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$, and $\hat{\rho}_{x_t|\mathcal{Y}_t}$ are approximately order-$2n$ density surrogate of $\rho_{x_{t+1}|\mathcal{Y}_t}$ and $\rho_{x_t|\mathcal{Y}_t}$. It reveals the fact that approximation using the truncated power moments does not introduce uncontrollable cumulative errors to the first $2n$ moment terms of the estimated pdfs, given that all power moments of the true system states exist and are finite, i.e., the prior $\theta \in \mathcal{SG}$.

4.2 Heavy-tailed density surrogate and the moment errors

We have proposed the feasible class of θ to ensure the existence and boundedness of the power moments of the density surrogate $\hat{\rho}(x)$. However in some situations, one desires state estimates with heavy tails. In the following

part of this section, we analyze the error propagation of the power moments, given that the power moments are not all finite, i.e., $\theta \notin \mathcal{SG}$.

The calculation (25) is still valid, but, since the power moments of $\hat{\rho}$ are not all finite, i.e.,

$$\int_{\mathbb{R}} x^k \left(\rho_{x_{t+1}|\mathcal{Y}_t} - \hat{\rho}_{x_{t+1}|\mathcal{Y}_t}\right) dx$$

may be infinite for some k, we cannot draw the same conclusion as in the light-tailed case. However, we note that

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

and therefore

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

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

Theorem 4.3. *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 $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$.*

Theorem 4.3 reveals the fact that with a satisfactory performance of density estimation, i.e., a relatively small $\max_x \left|\rho_{x_{t+1}|\mathcal{Y}_t} - \hat{\rho}_{x_{t+1}|\mathcal{Y}_t}\right|$, the error of the estimated moments of the density is also small.

We have analyzed the error propagation of estimated power moments with light and heavy tailed density surrogates. But we note that in real applications, it is not possible for us to treat the infinite-dimensional estimation problem, i.e., to use the full power moment sequence for density estimation. Then the density estimate is not always identical to the true density for either $\theta \in \mathcal{SG}$ or $\theta \notin \mathcal{SG}$. In the next section, we propose to analyze the error upper bounds of $\hat{\rho}$ to reveal its maximum difference from the true one, given the first $2n$ terms of power moments.

4.3 Error upper bounds of the density surrogate

To our knowledge, an error upper bound for the state estimate has not been established in Bayesian filtering with non-Gaussian distributions. 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 [33, 34]. This upper bound distinguishes our proposed filter from other Bayesian filters.

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

$$\begin{aligned}
V(\hat{\rho}, \rho) &= \sup_x \left|\int_{(-\infty, x]} (\hat{\rho} - \rho) dx\right| \\
&= \sup_x \left|F_{\hat{\rho}} - F_\rho\right|
\end{aligned} \tag{26}$$

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

In [34], Shannon-entropy is used to calculate the upper bound of the total variation distance. The Shannon-entropy [28] is defined as

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

We first introduce the Shannon-entropy maximizing distribution $F_{\breve{\rho}}$, of which the moments are calculated by (6). It has the following density function [19],

$$\breve{\rho}(x) = \exp\left(-\sum_{i=0}^{2n} \lambda_i x^i\right) \tag{27}$$

where $\lambda_0, \cdots, \lambda_{2n}$ are determined so that $\breve{\rho}$ has the same moments $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{2n}$ as the density $\hat{\rho}$, i.e.,

$$\int_{\mathbb{R}} x^j \breve{\rho}(x) dx = \hat{\sigma}_j, \quad j = 0, 1, \cdots, 2n.$$

Hence

$$H[\breve{\rho}] = \sum_{i=0}^{2n} \lambda_i \int_{\mathbb{R}} x^i \breve{\rho}(x) dx = \sum_{i=0}^{2n} \lambda_i \hat{\sigma}_i$$

Then, following [34], we form the KL distance between the true density and the Shannon-entropy maximizing density, i.e., in view of (27),

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

However, if $\theta \in \mathcal{SG}$ and n is sufficiently large, $\hat{\sigma}_i$ is approximately equal to σ_i for $i = 0, 1, \ldots 2n$ ($\sigma_i \approx \hat{\sigma}_i$) by

Theorem 4.2, and hence

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

Similarly, we obtain

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

By [21, 34], 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

$$V\left(\breve{\rho},\rho\right) \leq 3\left[-1+\left\{1+\frac{4}{9}\left(H\left[\breve{\rho}\right]-H\left[\rho\right]\right)\right\}^{1/2}\right]^{1/2}$$

Then we obtain the upper bound of the error

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

In some practical situations, e.g. financial applications, error upper bounds of the probability of the state estimate within intervals, e.g. $|P(x_t \geq a) - P(\hat{x}_t \geq a)|$, $|P(a \leq x_t \leq b) - P(a \leq \hat{x}_t \leq b)|$, are desired for people to make conservative decisions. However to our knowledge, there has not been a non-Gaussian Bayesian filter which provides such kinds of tight bounds without assuming the density functions to fall within specific classes.

In Section 4, we have proved that the power moments of the density estimates are approximately the true ones, by using our proposed algorithm when the density surrogate is light-tailed. The estimation error of the moments have also been proved to be bounded and small with satisfactory density estimation performance, which can be achieved by our proposed algorithm. We note that there are a series of research results on the tight bounds of the moment problem. These results make it feasible for us to derive upper bounds for the density estimates during the filtering process by our proposed algorithm. For example, achievable upper bounds $\max P(x_t \geq a)$ and $\max P(a \leq x_t \leq b)$ given the moment constraints are proposed in [3]. By these upper bounds, we can then obtain the upper bounds of errors

$$\begin{aligned} &\left|P(x_t \geq a) - P(\hat{x}_t \geq a)\right| \\ \leq&\max\left\{\max P(x_t \geq a) - P(\hat{x}_t \geq a), P(\hat{x}_t \geq a)\right\}, \end{aligned}$$

and

$$\begin{aligned} &\left|P(a \leq x_t \leq b) - P(a \leq \hat{x}_t \leq b)\right| \\ \leq&\max\left\{\max P(a \leq x_t \leq b) - P(a \leq \hat{x}_t \leq b),\right. \\ &\left.P(a \leq \hat{x}_t \leq b)\right\}. \end{aligned}$$

In conclusion, we have performed quantitative error analyses of the state estimates. An error upper bound of the state estimate in the sense of total variation distance, together with two error upper bounds for the probability of subsets of the real line given the power moments have been proposed in this section.

5 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 the prior $\theta(x)$. For light-tailed density surrogates, $\theta(x) \in \mathcal{SG}$ can usually be chosen as a Gaussian density function. It ensures that the first $2n$ power moments of $\hat{\rho}(x)$ are finite. Therefore, the problem reduces to determining the mean and variance of the Gaussian distribution.

At each time step, the first and second order power moments, i.e., σ_1, σ_2 of the density to be estimated can be calculated by (6). In practice, we can choose $m = \sigma_1$ and $\sigma^2 > \sigma_2$ and determine the prior density $\theta(x) = \mathcal{N}(m, \sigma^2)$ for each time step. Here we note that a relatively large variance σ^2 is to better adjust to the densities with multiple modes.

Second, we considers the choice of ρ_{x_0}. In some scenarios, the true probability density of the initial state x_0 is known prior. For scenarios where the initial state x_0 is not known prior, we take an arbitrary moment sequence $\bar{\sigma}_{2n}$ which satisfies that the Hankel matrix is positive definite. Then the ρ_{x_0} can be obtained by doing the optimization (15) given the moment constraints $\bar{\sigma}_{2n}$.

We note that the density $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ does not always have an explicit function form, i.e., it is not always pos-

sible to obtain the true system states. It makes comparing the estimates of the density to the true ones infeasible. However, we note that when η_t is a discrete random variable, the density $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ can be written as

$$\rho_{x_{t+1}|\mathcal{Y}_t}(x) = \sum_{i=1}^{m} \rho_i \cdot \rho_{x_t|\mathcal{Y}_t}\left(\frac{x-\xi_i}{f_t}\right) \tag{28}$$

which is a mixture of densities and is Lebesgue integrable (analytic if $\rho_{x_{t+1}|\mathcal{Y}_t}$ is analytic). 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 parts of this section. Moreover, the average estimation error of $\hat{\rho}$ is calculated by the total variation distance $V(\hat{\rho}, \rho)$, i.e., (26) in the following simulations.

5.1 Density estimation with different number of moment terms

In this part of section, we simulate on density estimation with different number of moment terms. In Example 1, we choose the true density to be a mixture of two Gaussians where there are two modes

$$\rho(x) = \frac{0.3}{\sqrt{2\pi}} e^{\frac{(x-2)^2}{2}} + \frac{0.7}{\sqrt{2\pi}} e^{\frac{(x+2)^2}{2}}.$$

The function class of the true density is not known prior. We use power moments up to order 6 to estimate the density function. $\theta(x)$ is chosen as $\mathcal{N}(-0.8, 3^2)$. The highest order of the polynomial $q(x)$ is 6. The density estimate $\hat{\rho}(x) = \theta(x)/q(x)$, where $q(x) = 2.30 \cdot 10^{-3}x^6 + 3.02 \cdot 10^{-3}x^5 - 2.55 \cdot 10^{-2}x^4 - 6.58 \cdot 10^{-2}x^3 - 4.10 \cdot 10^{-2}x^2 + 3.58 \cdot 10^{-1}x + 1.25$. The estimated density function is given in Figure 1. We note that the two modes are well estimated by the parametrization. The estimation error is $V(\hat{\rho}, \rho) = 0.0331$.

In Example 2, we simulate the same true density as in Example 1. However, the highest order of moments used is 8 in this example. $\theta(x)$ is chosen as $\mathcal{N}(-0.8, 3^2)$. The density estimate $\hat{\rho}(x) = \theta(x)/q(x)$, where $q(x) = 3.81 \cdot 10^{-4}x^8 - 2.46 \cdot 10^{-4}x^7 - 1.37 \cdot 10^{-2}x^6 + 8.43 \cdot 10^{-3}x^5 + 1.74 \cdot 10^{-1}x^4 - 8.88 \cdot 10^{-2}x^3 - 8.64 \cdot 10^{-1}x^2 + 3.20 \cdot 10^{-1}x + 1.96$. The estimated density function is given in Figure 2. The estimation error is $V(\hat{\rho}, \rho) = 0.0208$. We note that by using power moments up to order 8, the result is better than that of order 6 in the sense of the total variation distance.

In conclusion, these two simulation results give an example that with higher order moments used, the error of density estimation is less significant, which validates the approximation in Theorem 4.2. In the following part of section, we will give simulation results on different types of density functions by our proposed algorithm.

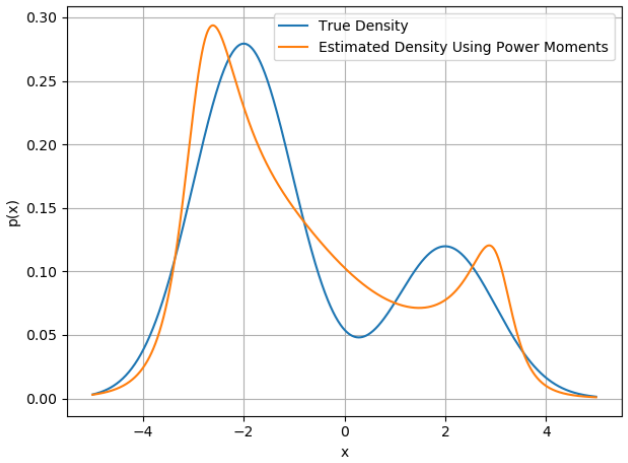

Fig. 1. Simulation result of Example 1. The blue curve represents the true density function. The orange one represents the density estimate using power moments.

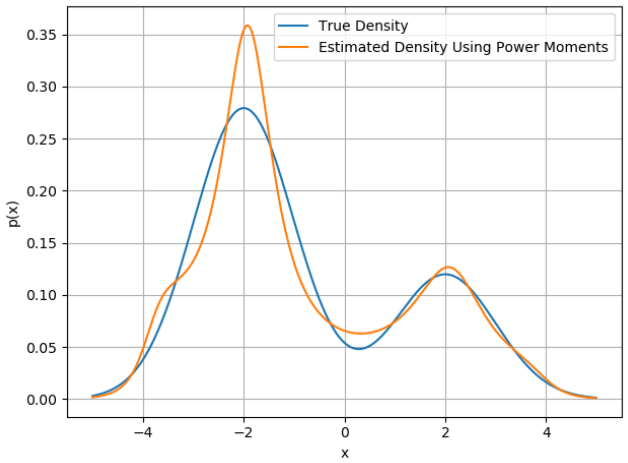

Fig. 2. Simulation result of Example 2.

5.2 Estimation of mixtures of different types of density functions

In the first two examples, we simulate a mixture of a Gaussian and a Laplacian, and the mixtures of Laplacians.

Example 3 is a bimodal density which is a mixture of a Gaussian and a Laplacian. The probability density function is

$$\rho(x) = \frac{0.5}{\sqrt{2\pi}} e^{\frac{(x-2)^2}{2}} + \frac{0.5}{2} e^{-|x+2|}.$$

$\theta(x)$ is chosen as $\mathcal{N}(0, 5^2)$. The highest order of the polynomial $q(x)$ is 4. We obtain the density estimate $\hat{\rho}(x) = \theta(x)/q(x)$, where $q(x) = 0.0203x^4 + 0.0280x^3 - 0.2252x^2 - 0.1892x + 0.9948$. The simulation result is

given in Figure 3. In this example, we note that even if the modes are of different types of distributions, the proposed density surrogate can still treat the density approximation without prior knowledge of the type of distributions. The two distinct modes are well estimated. $V(\hat{\rho}, \rho) = 0.0567$, which is a promising result.

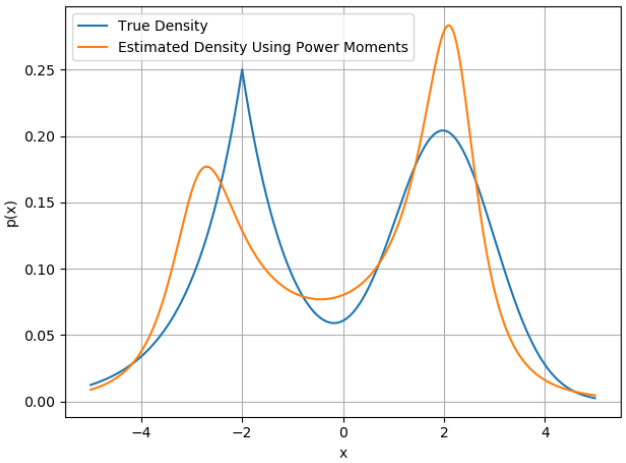

Fig. 3. Simulation result of Example 3. The blue curve represents the true density function. The orange one represents the density estimate using only power moments.

Example 4 is chosen as a bimodal density which is a mixture of two Laplacians. The probability density function is

$$\rho(x) = \frac{0.7}{2} e^{-|x-1|} + \frac{0.3}{2} e^{-|x+3|}.$$

$\theta(x)$ is chosen as $\mathcal{N}(-0.2, 7^2)$. The highest order of the polynomial $q(x)$ is 4. The density estimate $\hat{\rho}(x) = \theta(x)/q(x)$, where $q(x) = 0.0147x^4 + 0.0476x^3 - 0.0995x^2 - 0.2721x + 0.5713$. The simulation result is given in Figure 4. We note that the two modes are well characterized by the density surrogate, even when one has a relatively small probability. The estimation error is $V(\hat{\rho}, \rho) = 0.0744$.

The densities in the two examples above have two modes (peaks). In the following examples, we simulate densities with more modes to validate the performance of our proposed density surrogate.

Example 5 is chosen as a density with four modes which is a mixture of four Laplacians. The probability density function is

$$\rho(x) = \frac{0.4}{2} e^{|x|} + \frac{0.4}{2} e^{-|x-5|} + \frac{0.1}{2} e^{-|x+7|} + \frac{0.1}{2} e^{-|x-11|}.$$

$\theta(x)$ is chosen as $\mathcal{N}(0.5, 20^2)$. The highest order of the polynomial $q(x)$ is 8. The density estimate is $\hat{\rho}(x) = \theta(x)/q(x)$, where $q(x) = 4.22 \cdot 10^{-7}x^8 - 7.41 \cdot 10^{-6}x^7 - 3.48 \cdot 10^{-5}x^6 + 9.86 \cdot 10^{-4}x^5 - 5.48 \cdot 10^{-4}x^4 - 3.35 \cdot 10^{-2}x^3 + 7.55 \cdot 10^{-2}x^2 + 9.69 \cdot 10^{-2}x + 1.70 \cdot 10^{-1}$. The simulation result is given in Figure 5.

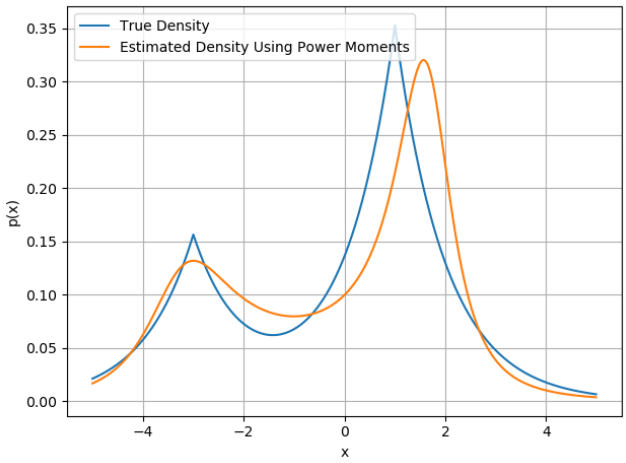

Fig. 4. Simulation result of Example 4.

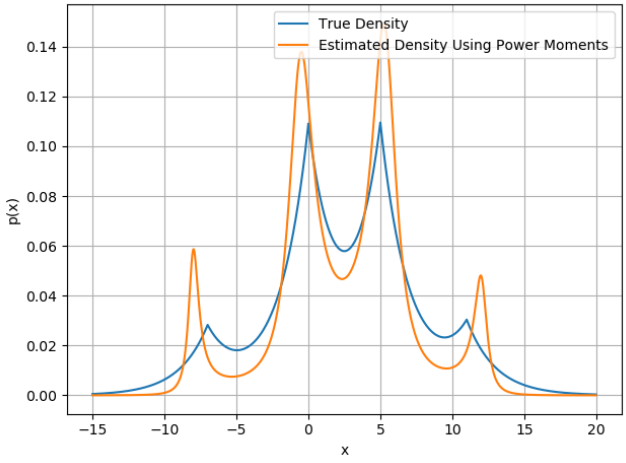

Fig. 5. Simulation result of Example 5.

The simulation result of Example 5 shows the performance of our proposed density surrogate in estimating the multi-modal density without prior knowledge on the modes or function type. The number of modes are correctly observed and the estimation error is satisfactory. The estimation error is $V(\hat{\rho}, \rho) = 0.053$.

Example 6 is chosen as a density with four modes which is a mixture of four Gaussians and a Laplacian. The probability density function is

$$\rho(x) = \frac{0.3}{\sqrt{2\pi}} e^{\frac{(x-2)^2}{2}} + \frac{0.3}{\sqrt{2\pi}} e^{\frac{(x+1)^2}{2}} + \frac{0.1}{\sqrt{2\pi}} e^{\frac{(x-6)^2}{2}} + \frac{0.1}{\sqrt{2\pi}} e^{\frac{(x+5)^2}{2}} + \frac{0.2}{2} e^{-|x-2|}.$$

It is a complicated mixture of densities. $\theta(x)$ is chosen as $\mathcal{N}(0.6, 10^2)$. The highest order of the polynomial in the denominator is 8. The density estimate is $\hat{\rho}(x) = \theta(x)/q(x)$, where $q(x) = 2.31 \cdot 10^{-5}x^8 - 8.35 \cdot 10^{-5}x^7 - 1.76 \cdot 10^{-3}x^6 + 4.83 \cdot 10^{-3}x^5 + 3.75 \cdot 10^{-2}x^4 - 6.64 \cdot 10^{-2}x^3 - 1.22 \cdot 10^{-1}x^2 + 8.95 \cdot 10^{-2}x + 3.52 \cdot 10^{-1}$. The simulation result is given in Figure 6 and $V(\hat{\rho}, \rho) = 0.096$. We note that the four modes of the state are well observed and the performance is satisfactory without prior knowledge of the density $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$. This example validates the ability of the proposed filter in estimating the complicated densities of the state.

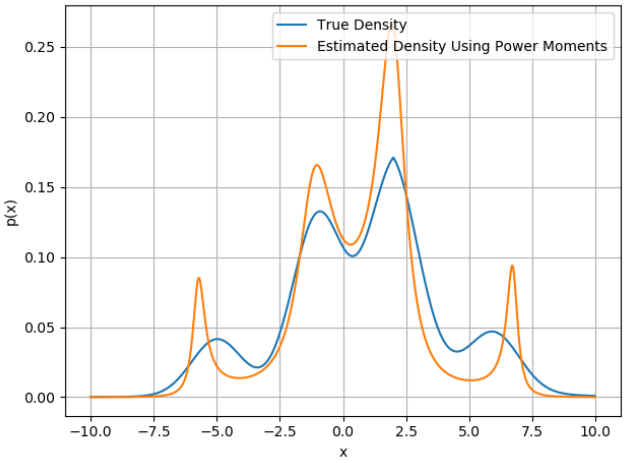

Fig. 6. Simulation result of Example 6.

Filtering problems where the state and noise distributions are heavy-tailed is a recent interest of the control community [15,40]. By using the Bayesian filter we propose, it is feasible 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 7 is chosen as a mixture of two student-t distributions. The probability density function is

$$\rho(x) = \frac{0.4 \cdot 3}{8\left(1 + \frac{(x-2)^2}{4}\right)^{\frac{5}{2}}} + \frac{0.6 \cdot 8}{3\pi\sqrt{5}\left(1 + \frac{(x+2)^2}{5}\right)^3}$$

$\theta(x)$ is chosen as $\mathcal{C}(-0.4, 5)$, where $\mathcal{C}$ denotes the Cauchy distribution. The highest order of the polynomial $q(x)$ is 4. The density estimate is $\hat{\rho}(x) = \theta(x)/q(x)$, where $q(x) = 0.0114x^4 - 0.0028x^3 - 0.1424x^2 + 0.043x + 0.7180$. The simulation result is given in Figure 7. The estimation error is $V(\hat{\rho}, \rho) = 0.032$. Example 7 validates the ability of our proposed Bayesian filter to treat the heavy-tailed filtering problem.

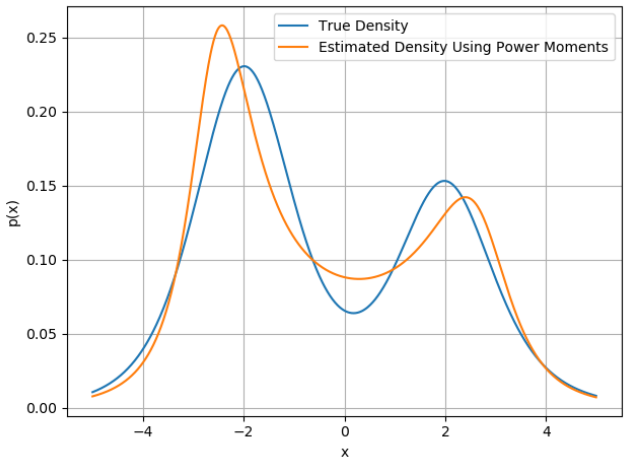

Fig. 7. Simulation result of Example 7.

6 Conclusion

In this paper, we propose the use of power moments to parameterize the state of a Bayesian filter considering the first order system. The proposed parametrization is able to characterize a much wider class of 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 formulate the density problem as a formal Hamburger moment problem. The existence of solutions to the moment problem is shown, and a Hankel matrix representation of it is proposed. The solutions of the proposed parametrization can be obtained by a convex optimization scheme. The solution to the optimization problem is proved to exist and be unique by proving that the map from the parameters to the power moments is homeomorphic. We prove that all moments of the density surrogate $\hat{\rho}$ exist and are finite if and only if θ is sub-Gaussian. Given that θ is sub-Gaussian, we prove that the estimated power moments are asymptotically unbiased and approximately the true ones throughout the filtering process. Therefore by selecting a large enough n, there are not severe cumulative errors in our proposed Bayesian filter. We also provide the upper bound of $\hat{\rho}$ when $\theta \notin \mathcal{SG}$. Error upper bounds of the state estimate are also proposed. In the simulation, we simulate mixtures of different types of density functions, including Gaussian, Laplacian 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. In future work, we plan to extend our results to the multidimensional systems. The extension is non-trivial, since the parametrization of a multivariate density function given the moment constraints is an open problem. Existence and uniqueness of solution, together with a Positivstellensätze of the parametrization, need to be proposed.

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. The kalman-bucy filter as a true time-varying wiener filter. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 119–128, 1971.

[3] Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

[4] Hendrik Wade Bode and Claude Elwood Shannon. A simplified derivation of linear least square smoothing and prediction theory. *Proceedings of the IRE*, 38(4):417–425, 1950.

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

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

[15] 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.

[16] 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.

[17] Roger A Horn and Charles R Johnson. *Matrix analysis.* Cambridge university press, 2012.

[18] RE Kalman. A new approach to liner filtering and prediction problems, transaction of asme. *Journal of Basic Engineering*, 83(1):95–108, 1961.

[19] 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.

[20] Andrei Nikolaevich Kolmogorov. *Stationary sequences in Hilbert space.* John Crerar Library, National Translations Center, 1978.

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

[22] 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.

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

[24] 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.

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

[26] Kalman RE. New results in linear filtering and prediction theory. *J. Basic Eng., ASME Trans.*, 83:95–107, 1960.

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

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

[29] AN Shiryayev. Interpolation and extrapolation of stationary random sequences. In *Selected Works of AN Kolmogorov*, pages 272–280. Springer, 1992.

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

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

[32] Krishnaswamy Srinivasan. State estimation by orthogonal expansion of probability distributions. *IEEE Transactions on Automatic Control*, 15(1):3–10, 1970.

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

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

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

[36] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[37] AH Wang and RL Klein. Optimal quadrature formula nonlinear estimators. *Information Sciences*, 16(3):169–184, 1978.

[38] Norbert Wiener et al. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, volume 8. MIT press Cambridge, MA, 1964.

[39] L. A. Zadeh and J. R. Ragazzini. An extension of wiener's theory of prediction. *Journal of Applied Physics*, 21(7):645–655, 1950.

[40] 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 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 currently pursuing the Ph.D. degree at Shanghai Jiao Tong University. His research interests are the moment problem and its applications to stochastic filtering, density steering and statistics.

Anders Lindquist received the Ph.D. degree in optimization and systems theory from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 1972, 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 the Royal Institute of Technology (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 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.