A Non-Gaussian Bayesian Filter Using Power and Generalized Logarithmic 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

In this paper, we aim to propose a consistent non-Gaussian Bayesian filter of which the system state is a continuous function. The distributions of the true system states, and those of the system and observation noises, are only assumed Lebesgue integrable with no prior constraints on what function classes they fall within. This type of filter has significant merits in both theory and practice, which is able to ameliorate the curse of dimensionality for the particle filter, a popular non-Gaussian Bayesian filter of which the system state is parameterized by discrete particles and the corresponding weights. We first propose a new type of statistics, called the generalized logarithmic moments. Together with the power moments, they are used to form a density surrogate, parameterized as an analytic function, to approximate the true system state. The map from the parameters of the proposed density surrogate to both the power moments and the generalized logarithmic moments is proved to be a diffeomorphism, establishing the fact that there exists a unique density surrogate which satisfies both moment conditions. This diffeomorphism also allows us to use gradient methods to treat the convex optimization problem in determining the parameters. Last but not least, simulation results reveal the advantage of using both sets of moments for estimating mixtures of complicated types of functions. A robot localization simulation is also given, as an engineering application to validate the proposed filtering scheme.

Key words: Bayesian methods; filtering; power moments; generalized logarithmic moments.

1 Introduction

In this paper, we consider the non-Gaussian Bayesian filtering problem for the first-order system following our previous work [1]. The Bayesian filter offers a cohesive and recursive solution to the general stochastic filtering challenges. Ho and Lee's work [2] represents one of the initial attempts of iterative Bayesian estimation, delineating the principles and procedures of Bayesian filtering. Sprangins [3] delved into the iterative application of Bayes' rule for sequential parameter estimation. Lin and Yau [4], as well as Chien and Fu [5], explored the Bayesian approach for optimizing adaptive systems. The Bayesian filter consists of an iterative measurement-time update process, sometimes referred to by different terms. During the time update step, the system equation calculates a one-step ahead prediction of the state. In the measurement update step, the observation equation computes the correction to the state estimate according to the current observation.

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

The Bayesian filter usually doesn't provide an analytic state estimation, except for the cases where the system and observation equations are linear, and the distributions of the system state, the system noise and the observation noise are Gaussian. In the well-known Kalman filter (and its extended forms such as the extended Kalman filter and the unscented Kalman filter), to determine the integral in the time update step is essentially a parametric estimation problem, which is done by estimating the first and second order moments [6–11].

However, when the distributions are non-Gaussian, the problem is much more complicated. Given that the distributions of the system states and the noises are not conjugate pairs, the time update step doesn't provide an analytic probability density function of the system state, due to the intractable convolution [12]. The density function can then only be obtained by approximation. In the real-world applications, the distributions are usually non-Gaussian, which makes this approximation problem a significant one both in theory and in practice. Being an open problem, density approximation of the intractable integral has been a core problem of Bayesian filtering

for decades and is still a hot topic. Numerous numerical methods have been proposed to obtain an analytic solution to the convolution in the time update step. To name a few, there are Gaussian/Laplace approximation [13], iterative quadrature [14–16], Gaussian sum approximation [17, 18]. In these methods, the intractable integral is approximated by a single Gaussian density function or a weighted sum of Gaussians. By doing this, the convolution operation in the time update step is tractable again, which makes it feasible for us to obtain the integral in an analytic form of function. A non-Gaussian Bayesian filter based on state-space calculus is proposed in [19]. Instead of transforming the density function into a Gaussian or a mixture of Gaussians, this method considers the rational probability density functions, which are transfer functions of finite-dimensional linear systems by the realization theory. It provides explicit state space descriptions for products and convolutions of rational densities, which then provides an analytic density function of the system state in a rational form. These parametric methods assume that the prior density belongs to a specified function class. This causes the flexibility of these methods to be limited. If the density function doesn't fall exactly within the assumed class, the estimation result may be severely biased.

Since the problem we treat doesn't restrict the non-Gaussian density to fall within specific classes of functions, 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, which also turns the infinite-dimensional problem into a finite dimensional and tractable one [20–24]. However characterizing the prior 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 [12]. Moreover, analyzing the errors of the particle filters is an extremely difficult task due to the indeterministic estimates caused by the sampling strategy [25, 26], and its performance suffers a lot from sample depletion [22, 27]. A non-Gaussian Bayesian filter, of which the system state distribution is a continuous function of a limited number of parameters, possesses notable advantages and is sought after by researchers in stochastic filtering.

In our previous papers [1, 28], a non-Gaussian Bayesian filter is implemented by approximating the intractable integral using the power moments. A non-classical density surrogate for the system state, in the form of a continuous function, is proposed, and the parameters of the proposed parametrization can also be determined by a convex optimization scheme with moment constraints, to which the solution is proved to exist and be unique. By the proposed filter, the power moments of the density estimates asymptotically converge to the true moments. Furthermore, according to Theorem 4.5.5 in [29], as the number of moment terms used approaches infinity, the approximated integral approaches the true analytic density in probability, which reveals the fact that the proposed filter is consistent. In the field of stochastic filtering, exact filters can only be computed for models that are completely discrete or for discrete-time linear Gaussian models, where the Kalman filter can be applied [30]. For system models that do not have exact filters, the best filters we can design for them are consistent filters. Therefore, the consistency of the filter proposed in [1] is a clear advantage over other alternative filtering methods. We would also like to emphasize that an error upper bound in the sense of total variation distance exists for this filter. Such an error upper bound has not been proposed for other non-Gaussian Bayesian filters. With a relatively longer execution time for each filtering step, the proposed filter overcomes the disadvantages of the particle filters mentioned above. Moreover, the proposed filter can treat the filtering problem where the probability density of the true system state is an arbitrary Lebesgue-integrable function with first several orders of power moments being finite, without assuming the density to fall within specific function classes.

In this paper, inspired by [31], we propose to use logarithmic-type moments together with power moments to improve the performance of non-Gaussian Bayesian filtering. The paper is organized as follows. In Section 2, we note that the conventional logarithmic moment doesn't work in this problem, and we propose a novel generalized logarithmic moment. An algorithm framework for Bayesian filtering using both power and generalized logarithmic moments is also proposed. Then we prove that by our proposed algorithm, the generalized logarithmic moments of the density estimates are asymptotically unbiased and approximately identical to the true ones. In Section 3, together with the fact that the power moments of the state estimate is approximately identical to the true ones given a large n, we propose to use both the power and generalized logarithmic moments to parameterize the density of the state. Then in Section 4, we prove that the parameters of the proposed density surrogate can be uniquely determined in terms of the power and generalized logarithmic moments up to order $2n$, by proving the corresponding map being diffeomorphic. Three density approximation examples are performed in Section 5, including the mixtures of Gaussian, generalized logistic and Laplacian densities. The simulation results with a comparison to the parametrization using only the power moments validate the advantage of using both sets of moments for parametrization of the prior density. Moreover, the proposed filtering scheme is applied to a robot localization task and the performance is compared to prevailing methods including the Kalman filter and the particle filter.

2 Non-Gaussian Bayesian filtering and the generalized logarithmic moments

Consider the stochastic system

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

$t = 0, 1, 2, \ldots$, where 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 and the observation noise ϵ_t are assumed to be Lebesgue integrable functions. Moreover, the noises are assumed to be independent from each other, and their distributions are denoted as ρ_{η_t} and ρ_{ϵ_t}.

We use the Bayesian filtering framework in [19]. Then the conditional probability density functions of the measurement and time updates are given by

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 $t \geq 1$,

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

Time update: For $t \geq 0$,

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

Here $\mathcal{Y}_t$ denotes the collection of observations $y_t, y_{t-1}, \cdots, y_0$.

If not otherwise specified, in the sequel "prior" refers to the prior density function $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ at each time step t. We denote by $\mathcal{P}$ the space of all probability density functions supported on $\mathbb{R}$. 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 measurement update (3) takes the form of an analytic function with all ρ_{ϵ_t} and $\rho_{x_t|\mathcal{Y}_{t-1}}$ being non-Gaussian in general. However to obtain an explicit form of prior $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ in (4) when the densities are not Gaussian, is not always a feasible task. We proposed to use the power moments for approximating the prior density in [1]. For $\rho \in \mathcal{P}_{2n}$, the power moments are calculated by

$$\begin{aligned} \sigma_{k,t} =& \mathbb{E}\left(x_{t+1}^k | \mathcal{Y}_t\right) \\ =& \sum_{j=0}^{k} \binom{k}{j} f_t^j \mathbb{E}\left(x_t^j | \mathcal{Y}_t\right) \mathbb{E}\left(\eta_t^{k-j}\right), \end{aligned} \tag{5}$$

for $k = 1, \cdots, 2n$ [1].

Being linear integral operators, the power moments capture the macroscopic properties of the prior density functions. The success of the power moments in our previous work [1] naturally leads us to think about using other integral operators to characterize the prior density. Except for the power moments, other statistics have been used in previous research to improve the estimation performance. For example in [31], covariance lags (power moments) and cepstral coefficients (logarithmic moments) are both used to approximate the spectral density. In this paper, we adopt a similar idea for approximating $\rho_{x_{t+1}|\mathcal{Y}_t}$ by using both the power and the logarithmic moments.

However, we note that it is not feasible for us to directly use the logarithmic moments in the form of $\int_{\mathbb{R}} x^k \log \rho(x)dx$, which are always infinite, even for densities $\rho(x)$ with exponential decaying rate. Take $\rho(x) = \mathcal{N}(0, 1)$ for example. Then

$$\begin{aligned} &\int_{\mathbb{R}} x^k \log\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\right) dx \\ =& \int_{\mathbb{R}} -\frac{1}{2}\log(2\pi)x^k - \frac{1}{2}x^{k+2}dx \\ =& -\infty, \quad \forall k \in \mathbb{N}_0. \end{aligned}$$

Therefore, we propose generalized logarithmic moments, for which the first $2n+1$ terms exist and are finite. Thus the generalized logarithmic moments are here defined as

$$\begin{aligned} \xi_{k,t} &= \int_{\mathbb{R}} x^k \theta(x) \log \rho_{x_{t+1}|\mathcal{Y}_t}(x)dx \\ &= \int_{\mathbb{R}} x^k \theta(x) \log \int_{\mathbb{R}} \rho_{x_t|\mathcal{Y}_t}\left(\frac{\varepsilon}{f_t}\right) \rho_{\eta_t}(x - \varepsilon)d\varepsilon dx \end{aligned} \tag{6}$$

for $k = 1, \cdots, 2n$. They are called "generalized" because a reference density $\theta(x)$ needs to be determined before calculating them. We denote by $\mathcal{P}_{2n}^{\log}$ the subset of all $\theta \in \mathcal{P}$ which have finite generalized logarithmic moments to at least order $2n$, provided with $\rho \in \mathcal{P}_{2n}$. Here $\theta \in \mathcal{P}_{2n}^{\log}$ is a reference density function, of which the choice is not very limited indeed. The probability densities with exponential decaying rate, e.g. the exponential families, fall within the subset $\mathcal{P}_{2n}^{\log}$.

Now that the power and generalized logarithmic moments are defined, we give the following definition to characterize the equivalence of two densities in the sense of the two types of moments.

Definition 2.1. A probability density function, which has the first $2n$ power and generalized logarithmic moment terms identically the same as ρ (with $\theta(x)$ given prior), is called an order-$2n$ P&L density surrogate of ρ and denoted by ρ^{2n}.

We denote by $\hat{\rho}$ the prediction of density ρ and propose to substitute the intractable prior density $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ with the proposed density surrogate. Each iteration of Bayesian filtering with the density surrogate is given in Algorithm 1. At present we assume that it is feasible for us to obtain such a density surrogate given the power and generalized logarithmic moments and first investigate the error propagation through the whole filtering process with the density surrogate, which is one of the most important problems in designing a filter. Since the prior estimation is done at each time step t, which means that the approximation errors of the each previous iteration may cause a cumulative one on the current estimation. It distinguishes the problem we treat from conventional density estimation problems.

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 (5);
3: Calculate ξ_t by (6);
4: Determine the order-$2n$ P&L density surrogate $\rho^{2n}_{x_{t+1}|\mathcal{Y}_t}$, of which the truncated power moment sequence is σ_t and the truncated generalized logarithmic moment sequence is ξ_t. The prior density estimate at time $t+1$ is then chosen as the P&L density surrogate, i.e., $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t} = \rho^{2n}_{x_{t+1}|\mathcal{Y}_t}$.

We will first review the error propagation of the first $2n$ terms of the power moments in [1] and then analyze those of the first $2n$ terms of the generalized logarithmic moments. Since the approximation errors caused by the time updates could have cumulative effects on 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 2.2. *Suppose $\hat{\rho}_{x_1|\mathcal{Y}_0}$ is a P&L 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 and the generalized logarithmic moments of $\hat{\rho}_{x_t|\mathcal{Y}_t}$ and $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$ are asymptotically unbiased 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 and generalized logarithmic moments of x_t and the corresponding $\hat{x}_t$ exist and are finite.*

A complete proof of Theorem 2.2 is given in Appendix A. Theorem 2.2 reveals the fact that the first $2n$ generalized logarithmic moment terms of the estimated prior densities with the density surrogate are approximately identical to the true ones through the whole filtering process. Together with (A.1), we have that $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}$ and $\hat{\rho}_{x_t|\mathcal{Y}_t}$ are approximately order-$2n$ P&L density surrogates of $\rho_{x_{t+1}|\mathcal{Y}_t}$ and $\rho_{x_t|\mathcal{Y}_t}$. It reveals the fact that approximation using both moments doesn't introduce significant cumulative errors to the first $2n$ moment terms of the estimated pdfs, with n chosen as a relatively large integer.

The problem is now constructing an order-$2n$ P&L density surrogate. Since the domain of ρ is $\mathbb{R}$, the problem becomes a Hamburger moment problem [32] with the constraints of additive generalized logarithmic moments. In the next section, we will give a representation of this specific moment problem and propose a solution to it.

3 A parametrization of the density surrogate using power and generalized logarithmic moments

In this section, we give a formal definition of the approximation problem of the prior density and prove the existence of a solution to this problem given the power moments and the generalized logarithmic moments.

Definition 3.1. A sequence

$$(\sigma_1, \ldots, \sigma_{2n}, \xi_1, \ldots, \xi_{2n})$$

is a feasible $2n$ P&L sequence, if there is a random variable X with a probability density $\rho(x)$ defined on $\mathbb{R}$, whose moments are given by (5) and (6), that is,

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

and

$$\xi_k = \mathbb{E}^{\log}\{X^k\} = \int_{\mathbb{R}} x^k \theta(x) \log \rho(x) dx, k = 1, \cdots, 2n.$$

Moreover, we assume that $\sigma_0 = 1$ and $\xi_0 = 0$. Any such random variable X is said to have a (σ, ξ)-feasible distribution. We denote the random variable as $X \sim (\sigma, \xi)$.

Next we prove the existence of solution to the moment problem defined in Definition 3.1. We first paraphrase Exercise 13.12 in [33]. Let f be a real-valued measurable

function defined on $\mathbb{R}$. Then there exists a sequence of polynomials P_n such that

$$\lim_{n\to+\infty} P_n(x) = f(x)$$

almost everywhere.

We note that the true $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ is trivially a solution to the moment problem in Definition 3.1. However we require an analytic function which satisfy the moment constraints. In our problem setting, $\rho_{x_{t+1}|\mathcal{Y}_t}$ is Lebesgue measurable. Therefore there exists a $\lim_{n\to+\infty} P_n(x)$ which is equal to $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ almost everywhere, i.e., it is a solution to the moment problem above.

However this solution doesn't exactly satisfy the requirement of a state estimate of the Bayesian filter, since there are possibly infinitely many parameters in the solution, which makes it infeasible to propagate the solution in the filtering process. Parametrization is then the most significant problem, which aims to use finitely many parameters to characterize the density.

Meanwhile, we are provided with two truncated power and generalized logarithmic moment sequences rather than two full ones, 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 that satisfies the moment conditions. We still emphasize here that the parametrization is not unique. Different constraints will yield different parametrizations.

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

Since there are infinitely many feasible solutions to the moment problem, a criterion to determine a unique solution is necessary. Following [1, 34], we consider the Kullback-Leibler (KL) distance

$$\mathbb{KL}(\theta\|\rho) = \int_{\mathbb{R}} \theta(x) \log \frac{\theta(x)}{\rho(x)} dx \tag{7}$$

to measure the difference between θ and ρ, which is a widely used pseudo-measure in density estimation tasks [34–37]. Although it is not symmetric, which makes it not a real metric, the KL distance is jointly convex. We formulate the primal problem as minimizing

$$\int_{\mathbb{R}} \theta(x) \log \frac{\theta(x)}{\rho(x)} dx \tag{8}$$

subject to

$$\int_{\mathbb{R}} x^k \rho dx = \sigma_k, \quad k = 1, \ldots, 2n \tag{9}$$

and

$$\int_{\mathbb{R}} x^k \theta \log \rho dx = \xi_k, \quad k = 1, \ldots, 2n. \tag{10}$$

Here θ is a prespecified density function which we want the estimate of the prior density $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$ to be close to. As we have mentioned, there are infinitely many solutions to the truncated moment problem, among which some do not have satisfactory properties (e.g. smoothness) or have undesired massive modes (peaks). By minimizing the Kullback-Leibler distance between θ and the density estimate, with θ given as an analytic function, it is possible to obtain an analytic density estimate of the prior, as will be proved in the following part of this section. We are then able to propagate the prior density throughout the filtering process. By minimizing (8) subject to (9) and (10), a parametrization based on both the power moments and the generalized logarithmic moments is proposed in the following theorem.

Theorem 3.2. *Denoting the Lagrange multipliers as*

$$p = (p_1, p_2, \ldots, p_{2n}), \quad q = (q_1, \ldots, q_{2n}),$$

set

$$P(x) = 1 + p_1 x + p_2 x^2 + \cdots + p_{2n} x^{2n},$$
$$Q(x) = q_0 + q_1 x + q_2 x^2 + \cdots + q_{2n} x^{2n}.$$

Given any $\theta \in \mathcal{P}_{2n}^{\log}$ and any σ which satisfies

$$\begin{bmatrix} 1 & \sigma_1 & \cdots & \sigma_n \\ \sigma_1 & \sigma_2 & \cdots & \sigma_{n+1} \\ \vdots & & \ddots & \\ \sigma_n & \sigma_{n+1} & & \sigma_{2n} \end{bmatrix} \succ 0, \tag{11}$$

minimizing (8) *subject to* (9) *and* (10) *yields a unique solution $\rho \in \mathcal{P}_{2n}$ of the form*

$$\hat{\rho} = \frac{\hat{P}(x)}{\hat{Q}(x)}\theta, \tag{12}$$

where $\hat{P}(x)$ and $\hat{Q}(x)$ are coprime. $\hat{p}, \hat{q}$ corresponding to $\hat{P}(x)$ and $\hat{Q}(x)$ are the unique solutions to the problem of minimizing

$$\mathbb{J}(P, Q) = \sigma' q - \xi' p + \int_{\mathbb{R}} P\theta \log \frac{P\theta}{Q} dx - \int_{\mathbb{R}} P\theta dx. \tag{13}$$

over all $P(x), Q(x) > 0$.

Proof. The Lagrangian of the primal problem (8) with constraints (9) and (10) is

$$L(\rho,p,q)=\int_{\mathbb{R}}\theta\log\frac{\theta}{\rho}dx+\sum_{k=0}^{2n}q_k\left(\int_{\mathbb{R}}x^k\rho dx-\sigma_k\right)-\sum_{k=1}^{2n}p_k\left(\int_{\mathbb{R}}x^k\theta\log\rho dx-\xi_k\right).$$

Then we have

$$\begin{aligned}L(\rho,p,q)&=\int_{\mathbb{R}}\theta\log\frac{\theta}{\rho}dx-\int_{\mathbb{R}}(P-1)\theta\log\rho dx\\&\quad+\xi' p+\int_{\mathbb{R}}Q\rho dx-\sigma' q\\&=\int_{\mathbb{R}}\theta\log\theta dx-\int_{\mathbb{R}}P\theta\log\rho dx\\&\quad+\int_{\mathbb{R}}Q\rho dx+\xi' p-\sigma' q\end{aligned}$$

with the directional derivative

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

which yields the stationary point

$$\hat{\rho}=\frac{P\theta}{Q}.$$

Then

$$L(\hat{\rho},p,q)=\int_{\mathbb{R}}\theta\log\theta dx-\mathbb{J}(P,Q),$$

where

$$\mathbb{J}(P,Q)=\sigma' q-\xi' p+\int_{\mathbb{R}}P\theta\log\frac{P\theta}{Q}dx-\int_{\mathbb{R}}P\theta dx.$$

In particular,

$$\begin{aligned}\frac{\partial\mathbb{J}}{\partial p_k}&=-\xi_k+\int_{\mathbb{R}}x^k\theta\log\frac{P\theta}{Q}dx\\\frac{\partial\mathbb{J}}{\partial q_k}&=\sigma_k-\int_{\mathbb{R}}x^k\frac{P\theta}{Q}dx.\end{aligned}$$

□

Remark. We have choosen the constant term of $P(x)$ as 1 to yield a simpler form of $\hat{\rho}$ in (12). However, it can be any real number. By specifying

$$\int_{\mathbb{R}}\theta\log\frac{P\theta}{Q}dx=0 \tag{14}$$

and

$$\int_{\mathbb{R}}\frac{P\theta}{Q}dx=1, \tag{15}$$

it is determined together with q_0.

But here we note that our parametrization has a rational form. Therefore, by dividing both the numerator and denominator by p_0, the constant term of the numerator becomes 1, and the q_0/p_0 becomes the new q_0 in Theorem 3.2.

We note that to obtain a unique solution to the problem to minimize (13) by a gradient-based method, it remains to prove that the map from (p,q) to (ξ,σ) is a diffeomorphism. In the following section, we will complete the proof of Theorem 3.2 by proving precisely this.

4 The diffeomorphic map

In this section, we prove that the map $(P,Q)\mapsto(\sigma,\xi)$ is diffeomorphic, building upon some of the ideas presented in [31].

We begin by noting that σ_0 is always equal to one. We also consider q_0 as a normalizing factor to ensure that $\int_{\mathbb{R}}\rho(x)dx=1$, which is thus determined when $(p_1,p_2,\ldots,p_{2n})$ and $(q_1,q_2,\ldots,q_{2n})$ are known. Therefore, denoting by $\mathcal{S}_{2n}$ as the class of positive polynomials of order $2n$ with the term of order zero being a constant, we have $P,Q\in\mathcal{S}_{2n}$. Given a specified density function $\theta(x)\in\mathcal{P}_{2n}^{\log}$, we can represent the rational density function by $(P,Q)\in\mathcal{M}_{2n}$, where $\mathcal{M}_{2n}=\mathcal{S}_{2n}\times\mathcal{S}_{2n}$. Thus $\mathcal{M}_{2n}$ becomes a smooth, connected, real manifold of dimension $4n$ which is diffeomorphic to $\mathbb{R}^{4n}$.

Next, we define some additional spaces for analysis. We denote by $\mathcal{M}_{2n}^*$ the (dense) open subspace of $\mathcal{M}_{2n}$ consisting of pairs (P,Q) of coprime polynomials. For $P\in\mathcal{S}_{2n}$, we define $\mathcal{M}_{2n}(P)$ as the space of all points in $\mathcal{M}_{2n}$ with the polynomial P fixed. Similarly, defining $\mathcal{M}_{2n}(Q)$, $\mathcal{M}_{2n}(P)$ and $\mathcal{M}_{2n}(Q)$ become real, smooth, connected $2n$-manifolds that are diffeomorphic to $\mathcal{S}_{2n}$ and thus to $\mathbb{R}^{2n}$. Furthermore, the tangent vectors to $\mathcal{M}_{2n}$ at (P,Q) can be represented as perturbations $(P+\epsilon u,Q+\epsilon v)$, where u and v are polynomials of degree less than or equal to $2n-1$. Denoting the real vector space of polynomials of degree less than or equal to d by V_d, the tangent space to $\mathcal{M}_{2n}$ at a point (P,Q) is canonically isomorphic to $V_{2n-1}\times V_{2n-1}$. Additionally, the tangent space to the submanifold $\mathcal{M}_{2n}(P)$ at a point (P,Q) is given by

$$T_{(P,Q)}\mathcal{M}_{2n}(P)=\{(u,v)\in V_{2n-1}\times V_{2n-1}\mid u=0\}$$

Similarly, the tangent space to $\mathcal{M}_{2n}(Q)$ is given by

$$T_{(P,Q)}\mathcal{M}_{2n}(Q)=\{(u,v)\in V_{2n-1}\times V_{2n-1}\mid v=0\}$$

The $2n$-manifolds $\{\mathcal{M}_{2n}(P) \mid P \in \mathcal{S}_{2n}\}$ form the leaves of a foliation of $\mathcal{M}_{2n}$, as do the $2n$-manifolds $\{\mathcal{M}_{2n}(Q) \mid Q \in \mathcal{S}_{2n}\}$. Furthermore, these two foliations are complementary in the sense that if a leaf of one intersects a leaf of the other, the tangent spaces intersect only at $(0,0)$. This transversality property is equivalent to the fact that the polynomials (P,Q) form a local system of coordinates.

From a geometric perspective, this property implies that (P,Q) are smooth coordinates on $\mathcal{M}_{2n}$. We will use this to demonstrate that (P,Q) also form bona-fide coordinate systems. Let $g : \mathcal{M}_{2n} \to \mathbb{R}^{2n}$ be the map that sends (P,Q) to ξ, where the components of ξ are calculated using equation (6). We denote $\mathcal{C}_{2n} := g(\mathcal{M}_{2n})$. Additionally, for each $\xi \in \mathcal{C}_{2n}$, we define the subset $\mathcal{M}_{2n}(\xi) = g^{-1}(\xi)$.

We aim to show that $\mathcal{M}_{2n}(\xi)$ is a smooth submanifold of dimension $2n$. To achieve this, we need to compute the Jacobian matrix of g evaluated at tangent vectors to a point $(P,Q) \in \mathcal{M}_{2n}$. If the Jacobian matrix of g is full rank at every point $(P,Q) \in \mathcal{M}_{2n}$, meaning that the directional derivative exists in every direction at each point, then $\mathcal{M}_{2n}(\xi)$ is proved to be smooth.

We recall that the tangent vectors to $\mathcal{M}_{2n}$ at (P,Q) can be expressed as a perturbation $(P+\epsilon u, Q+\epsilon v)$, where u, v are polynomials of degree less than or equal to $2n-1$. For each component $(k = 1, \cdots, 2n)$

$$g_k(P,Q) = \int_{\mathbb{R}} x^k \theta(x) \log\left(\frac{P(x)}{Q(x)}\theta(x)\right) dx \tag{16}$$

of g, we construct the directional derivative as follows:

$$\begin{aligned} &D_{(u,v)} g_k(P,Q) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[g_k(P+\epsilon u, Q+\epsilon v) - g_k(P,Q)\right] \\ &= \int_{\mathbb{R}} \left(\frac{u}{P} - \frac{v}{Q}\right) \theta x^k dx \end{aligned} \tag{17}$$

in the direction $(u,v) \in V_{2n-1} \times V_{2n-1}$.

Next, we define the linear map $G : V_{2n-1} \mapsto \mathbb{R}^{2n}$ as follows:

$$G_\psi u = \int_{\mathbb{R}} \frac{u}{\psi}\theta \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^{2n} \end{bmatrix} dx. \tag{18}$$

Then the kernel of the Jacobian of g at (P,Q) is given by

$$\ker \mathrm{Jac}(g)|_{(u,v)} = \{(P,Q) | G_P u = G_Q v\} \tag{19}$$

Lemma 4.1. *The linear map G_ψ is a bijection.*

Proof. First, consider the case when $G_\psi u = 0$. This implies that

$$\int_{\mathbb{R}} \frac{u}{\psi}\theta x^k dx = 0 \tag{20}$$

for $k = 1, \cdots, 2n$.

From (14), we have that $g_0(\psi, Q) \equiv 0$ for any $(\psi, Q) \in \mathcal{S}_{2n}$, which means that directional derivative along any direction is equal to zero. We take the directional derivative along $(u,0)$, and we have

$$\begin{aligned} &D_{(u,0)} g_0(\psi, Q) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[g_k(\psi + \epsilon u, Q) - g_k(\psi, Q)\right] \\ &= \int_{\mathbb{R}} \frac{u}{\psi}\theta dx = 0. \end{aligned} \tag{21}$$

Since $u \in V_{2n-1}$, we write

$$u(x) = \sum_{i=0}^{2n-1} u_i x^i, u_i \in \mathbb{R}.$$

By (20) and (21), we shall write

$$\sum_{i=0}^{2n-1} u_i \int_{\mathbb{R}} \frac{u}{\psi}\theta x^i dx = \int_{\mathbb{R}} \frac{u^2}{\psi}\theta dx = 0.$$

Since θ, ψ are both positive, $u(x)$ needs to be zero. Therefore, we have established the injectivity of G_ψ. Furthermore, since the range and domain of G_ψ have the same dimension, namely $2n$, the map is also surjective. Consequently, we can conclude that G_ψ is a bijection. □

Proposition 4.2. *For each $\xi \in \mathcal{C}_{2n}$, the space $\mathcal{M}_{2n}(\xi)$ is a smooth $2n$-manifold. The tangent space $T_{(P,Q)}\mathcal{M}_{2n}(\xi)$ at (P,Q) consists of precisely all $(u,v) \in V_{2n-1} \times V_{2n-1}$ such that*

$$\int_{\mathbb{R}} \frac{u}{P}\theta x^k dx = \int_{\mathbb{R}} \frac{v}{Q}\theta x^k dx \tag{22}$$

for $k = 0, 1, \ldots, 2n$.

Proof. The tangent vectors of $\mathcal{M}_{2n}(\xi)$ at (P,Q) correspond to the vectors in the null space of the Jacobian of g at (P,Q), as indicated by equation (17). Consequently, by utilizing equation (19), we establish that (22) holds for $k = 1, 2, \ldots, 2n$. Additionally, according to Lemma 4.1, we can conclude that (20) holds for $k = 0$. Therefore, (22) also holds for $k = 0$. Furthermore, based on (19) and Lemma 4.1, the tangent space has a dimension of $2n$. Consequently, the Jacobian matrix $\mathrm{Jac}(g)|(P,Q)$ has full rank, and the remaining part of the claim follows from the implicit function theorem.

Since the rank of $\mathrm{Jac}(g)|\,(P,Q)$ is consistently $2n$, the connected components of the submanifolds $\mathcal{M}_{2n}(\xi)$ constitute the leaves of a foliation of $\mathcal{M}_{2n}$. However, we still need to demonstrate that the submanifolds $\mathcal{M}_{2n}(\xi)$ themselves are connected. The detailed proof for this is provided in Appendix C. Consequently, we can state the following proposition. □

Proposition 4.3. *The $2n$-manifolds $\{\mathcal{M}_{2n}(\xi) \mid \xi \in \mathcal{C}_{2n}\}$ are connected, hence forming the leaves of a foliation of $\mathcal{M}_{2n}$.*

From the results proved so far, we conclude the following corollary.

Corollary 4.4. *The foliations, $\{\mathcal{M}_{2n}(Q) \mid Q \in \mathcal{S}_{2n}\}$ and $\{\mathcal{M}_{2n}(\xi) \mid \xi \in \mathcal{C}_{2n}\}$, are complementary, i.e., any intersecting pair of leaves, with one leaf from each foliation, intersects transversely. And each intersecting pair of leaves intersects in at most one point.*

Proof. Setting $u = 0$ in (19), we obtain $G_Q v = 0$. Hence, by Lemma 4.1, $v = 0$ so that the foliations are transverse. If a leaf $\mathcal{P}_{2n}(P)$ intersects a leaf $\mathcal{M}_{2n}(\xi)$ at a point (P,Q), then the corresponding P is known. Then according to Appendix B, a unique ξ is determined. □

A similar statement for the foliation $\{\mathcal{M}_{2n}(Q) \mid Q \in \mathcal{S}_{2n}\}$ can be proved by the mirror image of this proof and will be omitted.

Next, let $h : \mathcal{M}_{2n} \to \mathbb{R}^{2n}$ be the map which sends (P,Q) to σ, the components of which are calculated by (5), and let $\mathcal{R}_{2n} = h\,(\mathcal{M}_{2n})$. Each σ satisfies (11), guaranteeing the existence of a solution to the moment problem.

Now for each $\sigma \in \mathcal{R}_{2n}$, we aim to demonstrate that the set

$$\mathcal{M}_{2n}\,(\sigma) = h^{-1}\,(\sigma) \tag{23}$$

forms a smooth manifold of dimension $2n$. The tangent vectors to $\mathcal{M}_{2n}$ at (P,Q) can be represented as a perturbation $(P + \epsilon u, Q + \epsilon v)$, where u, v are polynomials of degree less than or equal to $2n-1$. For each component

$$h_k\,(P,Q) = \int_{\mathbb{R}} x^k \theta(x) \frac{P(x)}{Q(x)} dx, \quad k = 1, \cdots, 2n, \tag{24}$$

the directional derivative of h at $(P,Q) \in \mathcal{M}_{2n}$ in the direction $(u,v) \in V_{2n-1} \times V_{2n-1}$ is

$$\begin{aligned} &D_{(u,v)} h_k\,(P,Q) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[h_k\,(P + \epsilon u, Q + \epsilon v) - h_k\,(P,Q)\right] \\ &= \int_{\mathbb{R}} \left(\frac{uQ}{Q^2} - \frac{vP}{Q^2}\right) \theta x^k dx \end{aligned} \tag{25}$$

Similar to (18), we define the linear map $H_\psi : V_{2n-1} \to \mathbb{R}^{2n}$ by

$$H_\psi u = \int_{\mathbb{R}} \frac{u\psi}{Q^2} \theta \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^{2n} \end{bmatrix} dx \tag{26}$$

and the kernel of the Jacobian of h at (P,Q) is given by

$$\ker \mathrm{Jac}(h)|_{(P,Q)} = \{(u,v) | H_Q u = H_P v\} \tag{27}$$

for $k = 0, 1, \ldots, 2n$.

Proposition 4.5. *For each $\sigma \in \mathcal{R}_{2n}$, the subspace $\mathcal{M}_{2n}(\sigma)$ is a smooth and connected $2n$-manifold. The tangent space $T_{(P,Q)}\mathcal{M}_{2n}(\sigma)$ consists of pairs $(u,v) \in V_{2n-1} \times V_{2n-1}$ satisfying*

$$\int_{\mathbb{R}} \frac{uQ}{Q^2} \theta x^k dx = \int_{\mathbb{R}} \frac{vP}{Q^2} \theta x^k dx \tag{28}$$

for $k = 1, \ldots, 2n$. Furthermore, the $2n$-manifolds $\mathcal{M}_{2n}(\sigma)$ constitute the leaves of a foliation of $\mathcal{M}_{2n}$.

Proof. Let us begin by demonstrating that the linear map H_ψ is a bijection. Suppose $H_\psi u = 0$. This implies

$$H_\psi u = \int_{\mathbb{R}} \frac{u\psi}{Q^2} \theta x^k dx = 0 \tag{29}$$

for $k = 0, \ldots, 2n$. From (15), we have $h_0\,(\psi, Q) = 1$ for any $(\psi, Q) \in \mathcal{S}_{2n}$. Therefore, the directional derivative along any direction is equal to zero. We take the directional derivative along $(u, 0)$, and we have

$$\begin{aligned} &D_{(u,0)} h_0\,(\psi, Q) \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left[h_k\,(\psi + \epsilon u, Q) - h_k\,(\psi, Q)\right] \\ &= \int_{\mathbb{R}} \frac{u\psi}{Q^2} \theta dx = 0 \end{aligned} \tag{30}$$

Since $u \in V_{2n-1}$, we write

$$u(x) = \sum_{i=0}^{2n-1} u_i x^i, u_i \in \mathbb{R}.$$

By (29) and (30), we shall write

$$\sum_{i=0}^{2n-1} u_i \int_{\mathbb{R}} \frac{u\psi}{Q^2} \theta x^i dx = \int_{\mathbb{R}} \frac{u^2 \psi}{Q^2} \theta dx = 0.$$

Since θ, ψ are both positive, we conclude that $u = 0$. Thus, H_ψ is injective. Moreover, since the range and domain of H_ψ have the same dimension, namely $2n$, the

map is also surjective. In conclusion, H_ψ is a bijection. Similar to Proposition 4.2, we can establish that the rank of $\mathrm{Jac}(h)|\,(P,Q)$ is full. As the rank of $\mathrm{Jac}(h)|\,(P,Q)$ is thus consistently $2n$, the connected components of the submanifolds $\mathcal{M}_{2n}(\sigma)$ form the leaves of a foliation of $\mathcal{M}_{2n}$. To complete the argument, it remains to demonstrate that the submanifolds $\mathcal{M}_{2n}(\sigma)$ are themselves connected. The proof is provided in Appendix B. □

Theorem 4.6. *For each $(P,Q) \in \mathcal{M}_{2n}(\sigma) \cap \mathcal{M}_{2n}(\xi)$, the dimension of*

$$\mathcal{D} := T_{(P,Q)}\mathcal{M}_{2n}(\sigma) \cap T_{(P,Q)}\mathcal{M}_{2n}(\xi) \tag{31}$$

equals the degree of the greatest common divisor of the polynomials $P(x)$ and $Q(x)$.

Proof. Every $(P,Q) \in \mathcal{D}$ satisfies both (22) and (28). By taking appropriate linear combinations of (22) and (28), we obtain the following equations

$$\int_{\mathbb{R}} \frac{u^2}{P^2}\theta dx = \int_{\mathbb{R}} \frac{uv}{PQ}\theta dx, \tag{32}$$

and

$$\int_{\mathbb{R}} \frac{uv}{PQ}\theta dx = \int_{\mathbb{R}} \frac{v^2}{Q^2}\theta dx. \tag{33}$$

Since $\theta(x)$ is a non-negative density function, we can define

$$f_1 := \frac{u}{P}\theta^{\frac{1}{2}} \quad \text{and} \quad f_2 := \frac{v}{Q}\theta^{\frac{1}{2}}, \tag{34}$$

which allows us to rewrite (22) and (28) as

$$\|f_1\|^2 = \langle f_1, f_2\rangle \quad \text{and} \quad \langle f_1, f_2\rangle = \|f_2\|^2 \tag{35}$$

using the inner product and norm of $L^2[-\infty,+\infty]$. Applying the parallelogram law, we have

$$\|f_1 - f_2\|^2 = \|f_1\|^2 + \|f_2\|^2 - 2\langle f_1, f_2\rangle = 0, \tag{36}$$

which implies $f_1 = f_2$. Consequently,

$$\frac{u}{v} = \frac{P}{Q} \tag{37}$$

which has no solution if P and Q are coprime. However, if P and Q have a greatest common factor of degree d, $u(x)$ and $v(x)$ can be polynomials of degree less than or equal to $2n-1$ with an arbitrary common factor of degree $d-1$, thereby defining a vector space of dimension d, as stated. □

In conclusion, Theorem 4.6 establishes the complementary property of the foliations $\{\mathcal{M}_{2n}(\sigma) \mid \sigma \in \mathcal{L}_+\}$ and $\{\mathcal{M}_{2n}(\xi) \mid \xi \in \mathcal{L}_+\}$ at every point $(P,Q) \in \mathcal{M}_{2n}^*$, where P and Q are coprime. Consequently, it follows that the kernels of $\mathrm{Jac}(g)|\,(P,Q)$ and $\mathrm{Jac}(h)|\,(P,Q)$ are complementary at any point (P,Q) in $\mathcal{M}_{2n}^*$. Remarkably, the Jacobian of the joint map $(P,Q) \mapsto (\sigma,\xi)$ achieves full rank. As a result, the mapping $(P,Q) \mapsto (\sigma,\xi)$ is a diffeomorphism, thereby completing the proof of the following theorem.

Theorem 4.7. *The power moments $\sigma_1, \sigma_2, \cdots, \sigma_{2n}$ and the generalized logarithmic moments $\xi_1, \xi_2, \cdots, \xi_{2n}$ serve as a valid smooth coordinate system within the open subset $\mathcal{M}_{2n}^*$ of $\mathcal{M}_{2n}$. This means that the mapping from $\mathcal{M}_{2n}^*$ to $\mathbb{R}^{4n}$ with components*

$$(\sigma_1, \sigma_2, \cdots, \sigma_{2n}, \xi_1, \xi_2, \cdots, \xi_{2n})$$

has an everywhere invertible Jacobian matrix.

Based on the conclusive findings presented in Theorem 4.7, we have now concluded the proof for Theorem 3.2.

Having obtained all the necessary results from the preceding sections, we can now provide a comprehensive algorithm for non-Gaussian Bayesian filtering utilizing moments. This algorithm, denoted as Algorithm 2, uses both types of moments and is built upon the foundation of Algorithm 1.

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 $\hat{\rho}_{x_t|\mathcal{Y}_t}$ by (2) or (3);
2: Calculate σ_t by (5);
3: Calculate ξ_t by (6);
4: Perform optimization, solving (13) to obtain the order-$2n$ $P\&L$ density surrogate of $\int_{\mathbb{R}} \hat{\rho}_{x_t|\mathcal{Y}_t}\left(\frac{\varepsilon}{f_t}\right)\rho_{\eta_t}(x-\varepsilon)d\varepsilon$, which represents the new prediction $\hat{\rho}_{x_{t+1}|\mathcal{Y}_t}(x)$.

In [1], we demonstrated that power moments, which are linear integral operators, contain abundant information for characterizing density functions and can transform the infinite-dimensional filtering problem into a finite-dimensional and tractable one. In this paper, we prove that other linear integral operators, which capture different types of macroscopic properties of the density to be estimated, provide additional information that can enhance the density estimate. In the next section, we will simulate our proposed non-Gaussian Bayesian filter on mixtures of different types of density functions. We will also compare its performance with the filter proposed in [1] for each numerical example, demonstrating

that the additional information carried by the generalized logarithmic moments improves the density estimation performance. Moreover, we will apply the proposed filter to a robot localization task and compare the performance to several prevailing methods.

5 Numerical examples

In this section, we provide numerical examples of our proposed non-Gaussian Bayesian filter that utilizes both the power moments and the generalized logarithmic moments.

We perform two types of numerical simulations for validating the performance of the proposed algorithm. We first simulate distribution approximation tasks. We compare this filter, denoted as DPBM (Density Parametrization using both Power Moments and Generalized Logarithmic Moments), with a Bayesian filter that only uses power moments, which we referred to as DPPM (Density Parametrization using Power Moments) in our previous paper [1].

To begin, we need to choose a reference density $\theta(x)$. For light-tailed density surrogates, the Gaussian density is a suitable choice for $\theta(x)$. With this selection, the first $2n$ power moments of $\hat{\rho}(x)$ exist and are finite. Now, we must determine the mean and variance of the Gaussian distribution.

For the DPPM proposed in [1], the power moments σ_1 and σ_2 of the reference density $\theta(x)$ can be calculated using (5). By choosing $m = \sigma_1$ and $\sigma^2 > \sigma_2$ and specifying the density $\theta(x) = \mathcal{N}(m, \sigma^2)$, we consistently achieve good estimations. The reason behind this is that a relatively large variance σ^2 helps adjust the estimate to densities with multiple peaks (modes).

For the DPBM, we can directly choose the reference density $\theta(x)$ as $\theta(x) = \mathcal{N}(\sigma_1, \sigma_2)$. Due to the additional information provided by the generalized logarithmic moments, we no longer need to choose a relatively larger variance for the prior density.

We first simulate a mixture of Gaussians with two modes,

$$\rho(x) = \frac{0.5}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} + \frac{0.5}{\sqrt{2\pi}} e^{-\frac{(x+2)^2}{2}}. \tag{38}$$

where we select $\theta(x)$ as $\mathcal{N}(0, 5^2)$. The degree of the polynomial $Q(x)$ is 4 for both $\hat{\rho}_m$ and $\hat{\rho}_l$, where $\hat{\rho}_m$ corresponds to DPPM and $\hat{\rho}_l$ to DPBM. The highest order of $P(x)$ in $\hat{\rho}_l$ is 4. By Algorithm 2 in [1], a non-Gaussian Bayesian filter using only power moments, we obtain $\hat{\rho}_m = \theta(x)/Q_m(x)$, where $Q_m(x) = 4.13 \cdot 10^{-2}x^4 + 5.40 \cdot 10^{-5}x^3 - 4.44 \cdot 10^{-1}x^2 - 3.07 \cdot 10^{-4}x + 1.40$. Also we obtain $\hat{\rho}_l = \theta(x) \cdot P_l(x)/Q_l(x)$ by Algorithm 2, where $Q_l(x) = 3.40 \cdot 10^{-2}x^4 - 4.39 \cdot 10^{-24}x^3 - 2.40 \cdot 10^{-1}x^2 - 2.88 \cdot 10^{-22}x + 1$ and $P_l(x) = 3.19 \cdot 10^{-3}x^4 - 8.39 \cdot 10^{-24}x^3 + 1.60 \cdot 10^{-1}x^2 + 4.49 \cdot 10^{-22}x + 1.06$.

The density estimates $\hat{\rho}_m$ and $\hat{\rho}_l$, along with the true density ρ, are depicted in Figure 1. The simulation results clearly demonstrate that incorporating the generalized logarithmic moments into the density surrogate enhances the accuracy of prior density estimation for non-Gaussian Bayesian filtering.

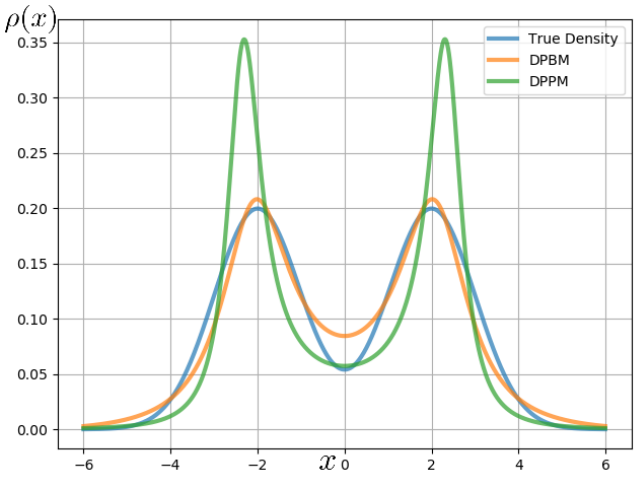

Fig. 1. Simulation results of Example 1. The blue curve represents the true prior density function. The green one represents the density estimate using only power moments. And the orange one represents the density estimate using both the power moments and the generalized logarithmic moments.

In the following example, we simulate a mixture of generalized logistic densities, which is known to be challenging to estimate accurately. Specifically, Example 2 represents a mixture of two type-I generalized logistic densities with the probability density function given by

$$\rho(x) = \frac{0.4 \cdot 2e^{-x+2}}{(1+e^{-x+2})^3} + \frac{0.6 \cdot 3e^{-x-2}}{(1+e^{-x-2})^4}.$$

We choose $\theta(x)$ as $\mathcal{N}(0.90, 5.86^2)$ as the reference density. For both $\hat{\rho}_m$ and $\hat{\rho}_l$, we use a degree-4 polynomial $Q(x)$. In $\hat{\rho}_l$, the highest order of $P(x)$ is 4. By employing the density surrogates, we obtain $\hat{\rho}_m = \frac{\theta(x)}{Q_m(x)}$, where $Q_m(x) = 1.65 \cdot 10^{-2}x^4 - 9.95 \cdot 10^{-2}x^3 + 5.27 \cdot 10^{-2}x^2 + 3.48 \cdot 10^{-1}x + 4 \cdot 10^{-1}$, and $\hat{\rho}_l = \frac{\theta(x) \cdot P_l(x)}{Q_l(x)}$, where $Q_l(x) = 1.68 \cdot 10^{-2}x^4 - 6.82 \cdot 10^{-2}x^3 - 6.75 \cdot 10^{-2}x^2 + 3.34 \cdot 10^{-1}x + 1$ and $P_l(x) = 7.10 \cdot 10^{-4}x^4 + 1.75 \cdot 10^{-3}x^3 - 6.65 \cdot 10^{-2}x^2 + 9.76 \cdot 10^{-2}x + 2.14$. The simulation results are presented in Figure 2.

It is observed that the estimate $\hat{\rho}_m$ is significantly biased compared to the true density. However, by utilizing the density surrogate that incorporates both moments, we achieve an estimate with a considerably reduced error.

This outcome is remarkable since the density surrogate only requires 10 parameters and does not rely on any knowledge of $\rho_{x_{t+1}|\mathcal{Y}_t}(x)$, such as the number of modes or the specific functional form.

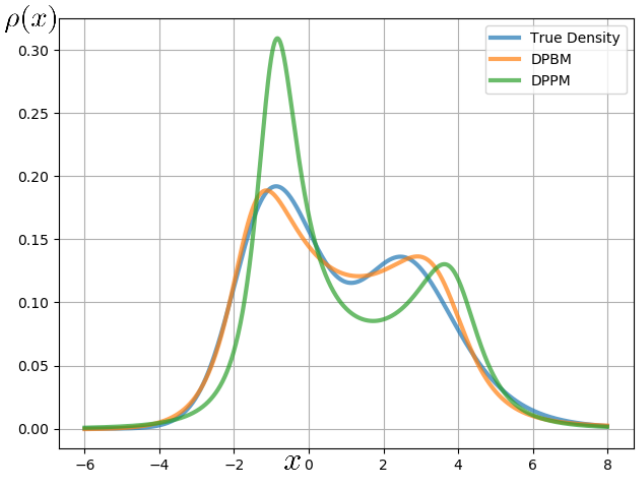

Fig. 2. Simulation results of Example 2.

In the final example, Example 3, we consider a mixture of two Laplacian densities. The probability density function is defined as follows:

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

We select $\theta(x)$ as $\mathcal{N}(-0.4, 1.5^2)$ for the reference density. The polynomial $Q(x)$ has a maximum order of 4 for both $\hat{\rho}_m$ and $\hat{\rho}_l$. In $\hat{\rho}_l$, the highest order of $P(x)$ is also 4. Utilizing the density surrogates, we obtain $\hat{\rho}_m = \frac{\theta(x)}{Q_m(x)}$, where $Q_m(x) = 5.52 \cdot 10^{-2}x^4 - 7.54 \cdot 10^{-2}x^3 - 1.69 \cdot 10^{-1}x^2 + 3.25 \cdot 10^{-1}x + 1.01$, and $\hat{\rho}_l = \frac{\theta(x) \cdot P_l(x)}{Q_l(x)}$, where $Q_l(x) = 6.78 \cdot 10^{-1}x^4 + 5.48 \cdot 10^{-2}x^3 - 1.11 \cdot x^2 + 2.39 \cdot 10^{-2}x + 1$ and $P_l(x) = 7.00 \cdot 10^{-2}x^4 + 1.03 \cdot 10^{-1}x^3 + 6.20 \cdot 10^{-1}x^2 - 3.13 \cdot 10^{-1}x + 4.67 \cdot 10^{-1}$. The simulation results are depicted in Figure 3.

In this example, it is important to highlight that the density being estimated is not smooth, and it exhibits two distinct sharp modes (peaks) that are in close proximity to each other. When utilizing only the power moments, we observe that the density estimate is unable to accurately capture the two modes, resulting in a poor approximation where only a single peak is represented. However, by incorporating the generalized logarithmic moments into the density surrogate, we significantly enhance the performance of the estimate. The resulting density approximation now successfully captures the presence of the two sharp modes and provides a much-improved representation.

In the previous numerical examples, we proved that by using the generalized logarithmic moments together

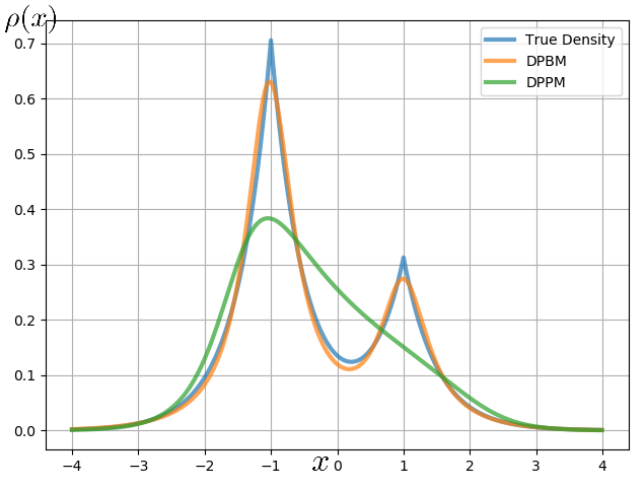

Fig. 3. Simulation results of Example 3.

with the power moments, the performance of distribution approximation is clearly improved in some situations, compared to merely using the power moments. In the following part of this section, we will give a more illustrative example, where our proposed algorithm is applied to an engineering problem and the performance is compared to several prevailing methods.

We address a robot localization challenge wherein a sensor is designated to measure the distances between the robot and predefined landmarks. We assume that the robot moves along the real line $\mathbb{R}$ with coordinate x. The robot's position with respect to coordinate x at time step k is denoted as $x(k)$. The positions of L landmarks are denoted as $\{\check{x}_1, \ldots, \check{x}_L\}$. In this localization task, the robot undergoes incremental movement, advancing one unit along the positive x direction. Imperfect controls result in deviations from the commanded movement, necessitating consideration of noise in the particle's movements to capture the actual robot movement. The moving distance, with a true value of 1, is corrupted with an additive Gaussian noise $\mathcal{N}(0, 0.03^2)$. Furthermore, the distance observation of each landmark is subject to additive noise.

The system and observation equations are

$$x(k+1) = x(k) + 1 + w(k)$$

and

$$\begin{bmatrix} z_1(k) \\ \vdots \\ z_L(k) \end{bmatrix} = \begin{bmatrix} x(k) - \check{x}_1 \\ \vdots \\ x(k) - \check{x}_L \end{bmatrix} + \begin{bmatrix} v_1(k) \\ \vdots \\ v_L(k) \end{bmatrix}$$

respectively. The sign of the signed distance $z_i(k)$ is negative when the robot is positioned to the left of landmark i and positive when it is to the right of landmark i.

We assume that the variable $w(k)$ follows a Gaussian

distribution, specifically $\mathcal{N}(0, 0.03^2)$, accounting for the error in controlling the robot. In prior results, the noises $v_1(k), \ldots, v_L(k)$ were assumed to be Gaussian to obtain a closed form of solution. In this experiment, however, we propose employing the right-skewed Gumbel distribution to validate the proposed algorithm in treating non-Gaussian filtering tasks.

The probability density functions of the Gumbel and Gaussian distributions is illustrated in Figure 4. The probability density function of the Gumbel distribution is given by

$$\rho_{v_i}(x) = 4e^{-4x-e^{-4x}}. \tag{39}$$

Meanwhile, the Gaussian distribution, with an identical mean and variance as the Gumbel distribution, has the following probability density function

$$\rho_{v_i}(x) = \frac{1}{\sqrt{2\pi}\cdot 0.35} e^{\frac{x^2}{2\cdot 0.35^2}}. \tag{40}$$

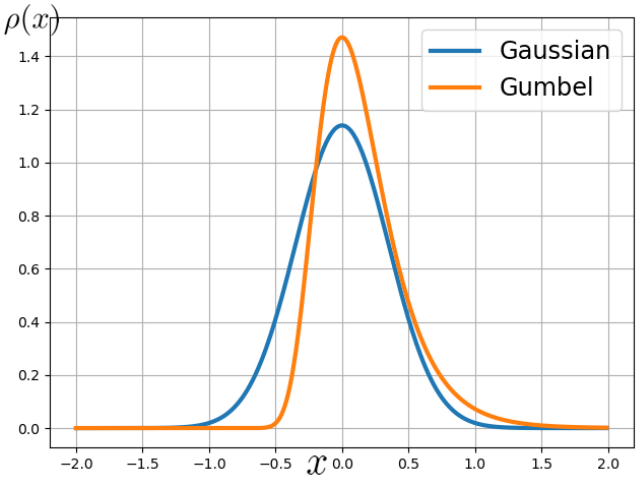

Fig. 4. Probability density functions of the Gaussian and the Gumbel distributions.

In the context of this localization task, the use of the asymmetric Gumbel distribution as the model for observation noise poses a significant challenge in selecting an appropriate stochastic filter. Except for DPBM proposed in this paper and the DPPM proposed in [1], the particle filters (PF) is the sole feasible option in prevailing methods for carrying out this task, due to the Gumbel distribution. In our simulations, we adopt a sampling-importance resampling (SIR) filter, as described in [12]. Given that the system equation and the observation equation are both linear, we also adopt the Kalman filter (KF) for this task. However, the KF faces difficulties in handling the Gumbel observation noise. Consequently, we resort to using the Gaussian distribution in (40) as a substitute for (39) to represent the observation noise during the filtering process.

The initialization procedure for the three filters is as follows: The robot initiates its movement from the position $x = -7$, and a landmark is situated at $x = 0$. The distributions of the initial states $x(0)$ for both the DPBM and KF are set to be the Gaussian distribution $\mathcal{N}(m_0, 1)$, where the mean m_0 is drawn from the Gaussian distribution $\mathcal{N}(-7, 1)$. The states of the 5000 particles in the PF are i.i.d samples drawn from the uniform distribution $U([-8, 8])$, which aims to cover a broader range of potential locations. The additive noise in the distance observation follows the Gumbel distribution in (39). The DPBM utilizes power moments and generalized logarithmic moments both up to the fourth order to estimate the density surrogates.

Figure 5 illustrates a sample robot localization process along the x-axis, showcasing estimation results by PF and DPBM. 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 DPBM filter. The gray dots represent the particles of the particle filter at each time step. We emphasize that the robot moves along the x-axis, even though the estimates and the particles of PF at different time steps are drawn slated, which aims to show the locations of the particles better. Notably, the location estimates by the DPBM converge to the true locations, while the PF particle states also converge to the correct positions. Figure 6 presents the root mean square error (RMSE) curves for 50 Monte-Carlo simulations of DPBM, PF, and KF. As the state estimates converge, the RMSE of DPBM is the smallest. The RMSE of PF is slightly larger than that of DPBM, however is quite close to it. We note that the RMSE of KF is significantly larger due to the use of a Gaussian distribution as an alternative for the true Gumbel distribution. The biased observation noise model causes obvious performance degradation of filtering.

From an RMSE standpoint, the DPBM does not significantly outperform the PF, but a notable drawback of the Particle filter is its requirement to store massive amount of data. For instance, in this simulation, the state of each particle consists of two parameters, namely its position and weight, resulting in a need for 10,000 parameters to characterize the system state density. In contrast, the DPBM only requires 9 parameters for this task, offering a more compact representation of the density function.

Additionally, considering the time consumption, the PF outperforms the DPBM in the execution time. In this example, each filtering iteration takes an average execution time of 1.08 seconds on a 2.5 GHz Intel Core i7 CPU. While this may be relatively long compared to the PF execution time, it remains manageable for applications with less sensitivity to the execution time. Moreover, the

optimization in each filtering step is convex with the solution proved to exist and be unique. It makes the execution for each filtering step to be predictable, which is a clear advantage of DPBM.

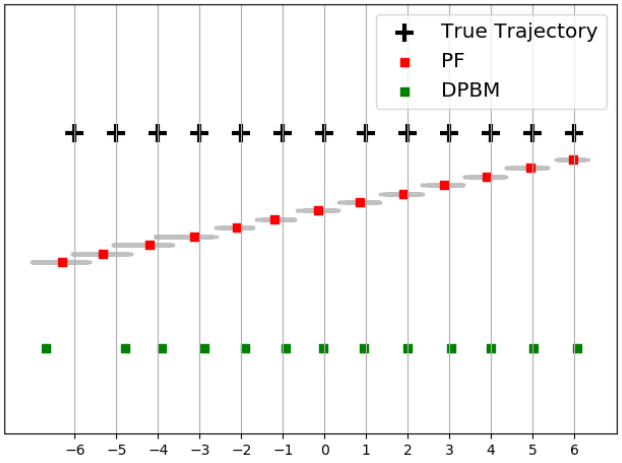

Fig. 5. A sample localization process.

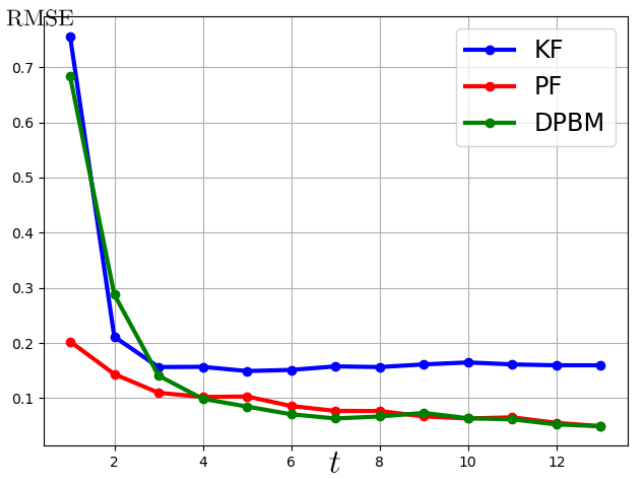

Fig. 6. RMSE as a function of time step $t = 1, \cdots, 13$ of 50 Monte-Carlo simulations for KF, PF and DPBM.

6 Conclusion

A Bayesian filter based on density parametrization using both the power moments and the generalized logarithmic moments of the densities is developed in this paper. We propose a convex optimization scheme to uniquely determine a rational density with exactly the specified power and generalized logarithmic moments, rather than estimating the parameters of a prespecified density model (such as Gaussian or Student's t) by minimizing the difference characterized by a norm, like the traditional method of moments. The map from the parameters of the proposed density surrogate to the power and generalized logarithmic moments is proved to be diffeomorphic, which reveals the fact that the parameters can be uniquely determined by the two types of moments. Furthermore, we provide the statistical property, together with numerical simulations to validate the proposed density estimator. By the results of the numerical simulations, we observe that the performance of density estimation using the proposed algorithm is quite satisfactory, which is a clear improvement as compared to that of the density surrogate using only power moments in our previous paper [1]. The presented filter is employed in a robot localization task, enabling a comprehensive performance comparison with various established filtering schemes. It is noteworthy that the Root Mean Square Error (RMSE) exhibited by the proposed filter is lower than that of the commonly used particle filter. Despite the fact that the execution time for each filtering step in the proposed algorithm is comparatively longer than that of the particle filter, the proposed algorithm effectively mitigates the necessity for the particle filter to store states of massive particles.

References

[1] G. Wu, A. Lindquist, Non-Gaussian Bayesian filtering by density parametrization using power moments, Automatica 153 (2023) 111061.

[2] Y. Ho, R. Lee, A bayesian approach to problems in stochastic estimation and control, IEEE transactions on automatic control 9 (4) (1964) 333–339.

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

[4] T.-t. Lin, S. S. Yau, Bayesian approach to the optimization of adaptive systems, IEEE Transactions on Systems Science and Cybernetics 3 (2) (1967) 77–85.

[5] Y.-T. Chien, K.-S. Fu, On bayesian learning and stochastic approximation, IEEE Transactions on Systems Science and Cybernetics 3 (1) (1967) 28–38.

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

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

[8] B. D. Anderson, J. B. Moore, The kalman-bucy filter as a true time-varying wiener filter, IEEE Transactions on Systems, Man, and Cybernetics 2 (1) (1971) 119–128.

[9] X. Kai, C. Wei, L. Liu, Robust extended Kalman filtering for nonlinear systems with stochastic uncertainties, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40 (2) (2009) 399–405.

[10] G. Chen, Approximate Kalman filtering, Vol. 2, World scientific, 1993.

[11] R. Chen, J. S. Liu, Mixture kalman filters, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62 (3) (2000) 493–508.

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

[13] D. J. MacKay, Choice of basis for laplace approximation, Machine learning 33 (1) (1998) 77–86.

[14] d. J. Freitas, Bayesian methods for neural networks, PhD, University of Cambridge, Cambridge, UK.

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

[16] A. Wang, R. Klein, Optimal quadrature formula nonlinear estimators, Information Sciences 16 (3) (1978) 169–184.

[17] H. W. Sorenson, D. L. Alspach, Recursive Bayesian estimation using Gaussian sums, Automatica 7 (4) (1971) 465–479.

[18] D. Alspach, H. Sorenson, Nonlinear bayesian estimation using gaussian sum approximations, IEEE transactions on automatic control 17 (4) (1972) 439–448.

[19] B. Hanzon, R. 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) (2001) 724–740.

[20] J. Carpenter, P. Clifford, P. Fearnhead, Improved particle filter for nonlinear problems, IEE Proceedings-Radar, Sonar and Navigation 146 (1) (1999) 2–7.

[21] A. Doucet, N. De Freitas, N. J. Gordon, et al., Sequential Monte Carlo methods in practice, Vol. 1, Springer, 2001.

[22] A. Doucet, N. J. Gordon, V. Krishnamurthy, Particle filters for state estimation of jump markov linear systems, IEEE Transactions on signal processing 49 (3) (2001) 613–624.

[23] C. Andrieu, A. Doucet, Particle filtering for partially observed Gaussian state space models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (4) (2002) 827–836.

[24] J. Vermaak, C. Andrieu, A. Doucet, S. J. Godsill, Particle methods for Bayesian modeling and enhancement of speech signals, IEEE Transactions on Speech and Audio Processing 10 (3) (2002) 173–185.

[25] D. Fox, Kld-sampling: Adaptive particle filters, Advances in neural information processing systems 14.

[26] D. Fox, Adapting the sample size in particle filters through kld-sampling, The international Journal of robotics research 22 (12) (2003) 985–1003.

[27] F. Gustafsson, Particle filter theory and practice with positioning applications, IEEE Aerospace and Electronic Systems Magazine 25 (7) (2010) 53–82.

[28] G. Wu, A. Lindquist, A multivariate non-Gaussian Bayesian filter using power moments, arXiv preprint arXiv:2211.13374.

[29] K. L. Chung, A course in probability theory, Academic press, 2001.

[30] A. Papanicolaou, Stochastic analysis seminar on filtering theory, arXiv preprint arXiv:1406.1936.

[31] C. I. Byrnes, P. Enqvist, A. Lindquist, Identifiability and well-posedness of shaping-filter parameterizations: A global analysis approach, SIAM Journal on Control and Optimization.

[32] K. Schmüdgen, The moment problem, Vol. 14, Springer, 2017.

[33] W. Rudin, Real and complex analysis, McGraw Hill Education India, 2015.

[34] T. Georgiou, A. Lindquist, Kullback-Leibler approximation of spectral density functions, IEEE Transactions on Information Theory 49 (11) (2003) 2910–2917.

[35] P. Hall, On Kullback-Leibler loss and density estimation, The Annals of Statistics (1987) 1491–1519.

[36] J. Q. Li, A. R. Barron, Mixture density estimation., in: NIPS, Vol. 12, 1999, pp. 279–285.

[37] V. Vapnik, The nature of statistical learning theory, Springer science & business media, 1999.

A Proof of Theorem 2.2

Proof. It is proved in [1] that, for a sufficiently large n,

$$\begin{aligned} \mathbb{E}\left(x_t^k|\mathcal{Y}_t\right) &\approx \mathbb{E}\left(\hat{x}_t^k|\mathcal{Y}_t\right) \\ \text{and} \quad \mathbb{E}\left(x_{t+1}^k|\mathcal{Y}_t\right) &\approx \mathbb{E}\left(\hat{x}_{t+1}^k|\mathcal{Y}_t\right) \end{aligned} \tag{A.1}$$

for $k = 1, \cdots, 2n$, for $\rho_{x_t|\mathcal{Y}_t}, \rho_{x_{t+1}|\mathcal{Y}_t} \in \mathcal{SG}$, where $\mathcal{SG}$ denotes the space of all sub-Gaussian distributions. Sub-Gaussian distributions are those whose tails are dominated by the tails of a Gaussian distribution, i.e., decay at least as fast as a Gaussian. Moreover,

$$\begin{aligned} \lim_{n\to+\infty} \mathbb{E}\left(\hat{x}_t^k|\mathcal{Y}_t\right) &= \mathbb{E}\left(x_t^k|\mathcal{Y}_t\right) \\ \text{and} \quad \lim_{n\to+\infty} \mathbb{E}\left(\hat{x}_{t+1}^k|\mathcal{Y}_t\right) &= \mathbb{E}\left(x_{t+1}^k|\mathcal{Y}_t\right) \end{aligned} \tag{A.2}$$

for $k = 1, \cdots, +\infty$. Now It remains to analyze

$$\begin{aligned} \mathbb{E}^{\log}\left(x_t^k|\mathcal{Y}_t\right) - \mathbb{E}^{\log}\left(\hat{x}_t^k|\mathcal{Y}_t\right) \\ \text{and} \quad \mathbb{E}^{\log}\left(x_{t+1}^k|\mathcal{Y}_t\right) - \mathbb{E}^{\log}\left(\hat{x}_{t+1}^k|\mathcal{Y}_t\right) \end{aligned}$$

for $k = 1, \cdots, 2n$. We note that $\mathbb{E}^{\log}\left(x_1^k|\mathcal{Y}_0\right) = \mathbb{E}^{\log}\left(\hat{x}_1^k|\mathcal{Y}_0\right)$ after the first time update, i.e.,

$$\int_{\mathbb{R}} x^k \theta(x) \left[\log\left(\rho_{x_1|\mathcal{Y}_0}\right) - \log\left(\hat{\rho}_{x_1|\mathcal{Y}_0}\right)\right] dx = 0. \tag{A.3}$$

Meanwhile, we can write the generalized logarithmic moment terms of $\rho_{x_1|\mathcal{Y}_1}$ as

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

for $k = 1, \cdots, 2n$, and those of $\hat{\rho}_{x_1|\mathcal{Y}_1}$ as,

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

for $k = 1, \cdots, 2n$. Therefore by (A.3) we have,

$$\begin{aligned} &\mathbb{E}^{\log}\left(x_1^k|\mathcal{Y}_1\right) - \mathbb{E}^{\log}\left(\hat{x}_1^k|\mathcal{Y}_1\right) \\ =&\int_{\mathbb{R}} x^k \theta(x) \left[\log\left(\rho_{x_1|\mathcal{Y}_0}\right) - \log\left(\hat{\rho}_{x_1|\mathcal{Y}_0}\right)\right] dx \\ =&0 \end{aligned}$$

for $k = 1, \cdots, 2n$. Then we have

$$\mathbb{E}^{\log}\left(x_1^k|\mathcal{Y}_1\right) = \mathbb{E}^{\log}\left(\hat{x}_1^k|\mathcal{Y}_1\right), \quad k = 1, \cdots, 2n.$$

Moreover, by (4), we have

$$\begin{aligned}
&\mathbb{E}^{\log}\left(x_2^k|\mathcal{Y}_1\right)-\mathbb{E}^{\log}\left(\hat{x}_2^k|\mathcal{Y}_1\right)\\
=&\int_{\mathbb{R}} x^k\theta(x)\log\int_{\mathbb{R}}\rho_{x_1|\mathcal{Y}_1}\left(\frac{\varepsilon}{f_1}\right)\rho_{\eta_1}(x-\varepsilon)d\varepsilon dx\\
-&\int_{\mathbb{R}} x^k\theta(x)\log\int_{\mathbb{R}}\hat{\rho}_{x_1|\mathcal{Y}_1}\left(\frac{\varepsilon}{f_1}\right)\rho_{\eta_1}(x-\varepsilon)d\varepsilon dx\\
=&\int_{\mathbb{R}} f_1x^k\theta(x)\log\int_{\mathbb{R}}\rho_{x_1|\mathcal{Y}_1}(\omega)\,\rho_{\eta_1}(x-f_1\omega)d\omega dx\\
-&\int_{\mathbb{R}} f_1x^k\theta(x)\log\int_{\mathbb{R}}\hat{\rho}_{x_1|\mathcal{Y}_1}(\omega)\,\rho_{\eta_1}(x-f_1\omega)d\omega dx
\end{aligned}$$

for $k=1,\cdots,2n$. We note that $\rho_{\eta_1}(x-f_1\omega)$ is analytic almost everywhere. Assume $\rho_{\eta_1}(x-f_1\omega)$ 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 $x_0=f_1\omega$, then we have

$$\begin{aligned}
&\rho_{\eta_1}(x-f_1\omega)\\
=&\sum_{i=0}^{+\infty}\frac{\rho_{\eta_1}^{(i)}(0)}{i!}(x-f_1\omega)^i\\
=&\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\binom{i}{j}\frac{(-f_1)^j\rho_{\eta_1}^{(i)}(0)}{i!}\omega^j x^{i-j}
\end{aligned}$$

Since all power moments and generalized logarithmic moments of x_1 and $\hat{x}_1$ exist and are finite, we have (A.4).

By (A.4), we note that $\mathbb{E}^{\log}\left(x_2^k|\mathcal{Y}_1\right)-\mathbb{E}^{\log}\left(\hat{x}_2^k|\mathcal{Y}_1\right)$ tends to zero as $n\to\infty$ by (A.1). By properly selecting a sufficient large n, we have

$$\mathbb{E}^{\log}\left(x_2^k|\mathcal{Y}_1\right)\approx\mathbb{E}^{\log}\left(\hat{x}_2^k|\mathcal{Y}_1\right),\quad k=1,\cdots,2n,$$

Similarly we can prove

$$\mathbb{E}^{\log}\left(x_t^k|\mathcal{Y}_t\right)\approx\mathbb{E}^{\log}\left(\hat{x}_t^k|\mathcal{Y}_t\right),\quad k=1,\cdots,2n,$$

and

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

as claimed. □

B Connectivity of $\mathcal{M}_{2n}(\sigma)$

It is nontrivial to prove that the set of all feasible $(p_1,\cdots,p_{2n})$ is path-connected given $\mathcal{M}_{2n}(\sigma)$.

From the view of optimization, if the feasible $(p_1,\cdots,p_{2n})$ fall into several disjoint sets, it is difficult to achieve the global optimum. In this appendix, we prove the connectivity of $\mathcal{M}_{2n}(\sigma)$.

We first prove that the map sending $(p_1,\cdots,p_{2n})$ to $\xi\in\mathcal{C}_{2n}$ is a diffeomorphism.

It is obvious that given a P, there exists a unique ξ. Now we need to prove that given a generalized logarithmic moment sequence ξ, there exists a unique P. Here we prove this by contradiction. Assume $\frac{P(x)}{Q(x)}\theta$ and $\frac{P'(x)}{Q(x)}\theta$ correspond to identical ξ where $P(x)\neq P'(x)$, i.e.

$$\int_{\mathbb{R}}\theta x^i\log\frac{P}{Q}\theta dx=\int_{\mathbb{R}}\theta x^i\log\frac{P'}{Q}\theta dx=\xi_i \tag{B.1}$$

for $i=0,1,\ddots,2n$ (specifically, ξ_0 is confined to be zero).

Therefore we have

$$\int_{\mathbb{R}}x^i\theta(x)\log\frac{P(x)\theta(x)}{P'(x)\theta(x)}dx=0,\quad i=0,1,\ldots,2n. \tag{B.2}$$

As $P(x)$ and $P'(x)$ are normalized density functions,

$$\begin{aligned}
&\sum_{i=0}^{2n}p_i\int_{\mathbb{R}}x^i\theta(x)\log\frac{P(x)\theta(x)}{P'(x)\theta(x)}dx\\
=&\int_{\mathbb{R}}\sum_{i=0}^{2n}p_ix^i\theta(x)\log\frac{P(x)\theta(x)}{P'(x)\theta(x)}dx\\
=&\,\mathbb{KL}(P\theta,P'\theta)\\
=&\,0.
\end{aligned} \tag{B.3}$$

However, $\mathbb{KL}(P\theta,P'\theta)=0$ if and only if $P\equiv P'$, i.e.

$$(p_1,\cdots,p_{2n})=(p_1',\cdots,p_{2n}'),$$

given a generalized logarithmic moment sequence ξ. This contradicts our assumption. We have that the map sending $(p_1,\cdots,p_{2n})$ to $(\xi_1,\cdots,\xi_{2n})$ is a bijection. And because the map and the inverse map are both differentiable, we have that the map is a diffeomorphism. Therefore, $\mathcal{M}_{2n}(\sigma)$ is diffeomorphic to $\mathcal{M}_{2n}(Q)$, which is again diffeomorphic to $\mathbb{R}^{2n}$ and is then path-connected.

C Connectivity of $\mathcal{M}_{2n}(\xi)$

We will prove that the $2n$-manifold $(q_1,\cdots,q_{2n})$ is path-connected given $\mathcal{M}_{2n}(\xi)$. First we prove that the map sending $(q_1,\cdots,q_{2n})$ to $\sigma\in\mathcal{R}_{2n}$ is a diffeomorphism.

It is obvious that given a pair of parameters $(q_1,\cdots,q_{2n})$, there exists a unique σ. Now we need to prove that

$$
\begin{aligned}
&\mathbb{E}^{\log}\left(x_2^k|\mathcal{Y}_1\right)-\mathbb{E}^{\log}\left(\hat{x}_2^k|\mathcal{Y}_1\right)\\
=&\int_{\mathbb{R}}x^k\theta(x)\log\int_{\mathbb{R}}\rho_{x_1|\mathcal{Y}_1}(\omega)\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\binom{i}{j}\frac{(-f_1)^j\rho_{\eta_1}^{(i)}(0)}{i!}\omega^j x^{i-j}d\omega dx\\
-&\int_{\mathbb{R}}x^k\theta(x)\log\int_{\mathbb{R}}\hat{\rho}_{x_1|\mathcal{Y}_1}(\omega)\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\binom{i}{j}\frac{(-f_1)^j\rho_{\eta_1}^{(i)}(0)}{i!}\omega^j x^{i-j}d\omega dx\\
=&\int_{\mathbb{R}}x^k\theta(x)\log\left(\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\binom{i}{j}x^{i-j}\int_{\mathbb{R}}\rho_{x_1|\mathcal{Y}_1}(\omega)\frac{(-f_1)^j\rho_{\eta_1}^{(i)}(0)}{i!}\omega^j d\omega\right)dx\\
-&\int_{\mathbb{R}}x^k\theta(x)\log\left(\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\binom{i}{j}x^{i-j}\int_{\mathbb{R}}\hat{\rho}_{x_1|\mathcal{Y}_1}(\omega)\frac{(-f_1)^j\rho_{\eta_1}^{(i)}(0)}{i!}\omega^j d\omega\right)dx\\
=&\int_{\mathbb{R}}x^k\theta(x)\log\left(\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\binom{i}{j}x^{i-j}\frac{(-f_1)^j\rho_{\eta_1}^{(i)}(0)}{i!}\mathbb{E}\left(x_1^j|\mathcal{Y}_1\right)\right)dx\\
-&\int_{\mathbb{R}}x^k\theta(x)\log\left(\sum_{i=0}^{+\infty}\sum_{j=0}^{i}\binom{i}{j}x^{i-j}\frac{(-f_1)^j\rho_{\eta_1}^{(i)}(0)}{i!}\mathbb{E}\left(\hat{x}_1^j|\mathcal{Y}_1\right)\right)dx
\end{aligned}
\tag{A.4}
$$

given a specific power moment sequence σ, there exist a unique $(q_1,\cdots,q_{2n})$. Again we prove by contradiction. Assume $\frac{P(x)}{Q(x)}$ and $\frac{P(x)}{Q'(x)}$ have the identical σ, i.e.

$$
\int_{\mathbb{R}}x^i\frac{P(x)}{Q(x)}\theta(x)dx=\int_{\mathbb{R}}x^i\frac{P(x)}{Q'(x)}\theta(x)dx=\sigma_i,\tag{C.1}
$$

for $i=0,1,\ldots,2n$.

Then we have

$$
\int_{\mathbb{R}}x^i\frac{P(x)\left(Q(x)-Q'(x)\right)}{Q(x)Q'(x)}\theta(x)dx=0,\tag{C.2}
$$

for $i=0,1,\ldots,2n$. Defining $(\tilde{q}_i)$ by

$$
Q(x)-Q'(x)=\sum_{i=0}^{2n}\tilde{q}_i x^i,\tag{C.3}
$$

we have

$$
\begin{aligned}
&\sum_{i=0}^{2n}\tilde{q}_i\int_{\mathbb{R}}x^i\frac{P(x)\left(Q(x)-Q'(x)\right)}{Q(x)Q'(x)}\theta(x)dx\\
=&\int_{\mathbb{R}}\sum_{i=0}^{2n}\tilde{q}_i x^i\frac{P(x)\left(Q(x)-Q'(x)\right)}{Q(x)Q'(x)}\theta(x)dx\\
=&0,
\end{aligned}
$$

and consequently

$$
\int_{\mathbb{R}}\sum_{i=0}^{2n}P(x)\cdot\frac{\left(\sum_{i=0}^{2n}\tilde{q}_i x^i\right)^2}{Q(x)Q'(x)}\theta(x)dx=0.\tag{C.4}
$$

Since $P(x),Q(x)$ and $Q'(x)$ are all positive, we have

$$
\sum_{i=0}^{2n}\tilde{q}_i x^i\equiv 0,\tag{C.5}
$$

i.e.

$$
\tilde{q}_i=0,\quad i=0,1,\cdots,2n\tag{C.6}
$$

This contradicts our assumption. Therefore we can conclude that the map sending $(q_1,\cdots,q_{2n})$ to $(\sigma_1,\cdots,\sigma_{2n})$ is a bijection. And because the map and the inverse map are both differentiable, we have that the map is a diffeomorphism. Since $\mathcal{M}_{2n}(\xi)$ and $\mathcal{M}_{2n}(P)$ are both differentiable, they are diffeomorphic. Then we have that $\mathcal{M}_{2n}(\xi)$ is smooth and path-connected, because $\mathcal{M}_{2n}(Q)$ also has these two properties.

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, distribution steering, system identification 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 (elected 1997), 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.