The Generalised Moment-based Filter

A. Benavoli

***Abstract*— Can we solve the filtering problem from the only knowledge of few moments of the noise terms? In this paper, by exploiting set of distributions based filtering, we solve this problem without introducing additional assumptions on the distributions of the noises (e.g., Gaussianity) or on the final form of the estimator (e.g., linear estimator). Given the moments (e.g., mean and variance) of random variable X, it is possible to define the set of all distributions that are compatible with the moments information. This set can be equivalently characterized by its extreme distributions: a family of mixtures of Dirac's deltas. The lower and upper expectation of any function g of X are obtained in correspondence of these extremes and can be computed by solving a linear programming problem. The filtering problem can then be solved by running iteratively this linear programming problem. In this paper, we discuss theoretical properties of this filter, we show the connection with set-membership estimation and its practical applications.**

***Index Terms*— Generalised moments, set of distributions, robustness, set-membership estimation, Kalman filter.**

I. Introduction

In the linear non-Gaussian case, it is well-known that the Kalman filter (KF) is the best linear minimum variance estimator. If the distributions of the non-Gaussian noises are unknown (we only know their mean and variance), there exist few alternative approaches to state estimation apart from KF (Monte Carlo methods cannot be used in this case since they assume that the distributions of the noises are known).

In this unknown-distribution setting, quantifying the uncertainty/reliability of the KF estimate is a big issue, since the estimation error $\hat{x}_k - x_k$ has an unknown distribution. Thus, the best one can hope for is to give bounds of the estimation error using for instance the Chebyshev inequality. This method, however, has several limitations: (i) it can only be applied to determine lower/upper bounds for the probability of intervals of type $|\hat{x}_k - x_k| \gtreqless \delta\sigma_k$; (ii) it can produce confidence regions that are too large. To alleviate the conservativeness of Chebyshev inequality, Spall [1] has proposed to compute confidence regions by using instead the Kantorovich inequality. In [2], Maryak et al. derive narrower probability bounds for he estimation error by using central limit theorem type arguments. However, these approaches can only be used to compute confidence regions for the KF estimate in non-Gaussian settings, while cannot be used to compute bounds for the expectation of other functions of interest of the state. Furthermore, they cannot be extended to the nonlinear case.

In this paper, we address these issues by using set of distributions based filtering [3] considering a particular class of distributions: the set of distributions which have the same first m *generalised moments*. Given real-valued functions f_i for $i = 0, 1, \ldots, m$ of a random variable X we call the real numbers $\mu_i = E[f_i]$ generalised moments (here $E[\cdot]$ denotes expectation). The knowledge of m (generalised) moments μ_i is not enough to uniquely specify the distribution of a variable, the generalised moment problem then consists to find all the distributions compatible with the moments information [4].

The moment problem has received great attention in the control theory community since, for instance, the solution to the optimal control problem for a linear plant can be obtained by solving a special kind of moment problem [5, Ch. 7]. Other applications of the moment problem to system and control theory are presented in [6], [7]. Here, Byrnes and Lindquist solve the moment problem by selecting among the admissible distributions the one which minimizes a Kullback-Leibler divergence based cost. In the moment problem, it is common to impose some cost criterion (e.g., maximum entropy, KL divergence etc.) [8], [9], [10] to select one of the admissible distributions. In this way the moment problem reduces to maximise/minimise a cost subject to moment constraints.

It is well known that given the mean and variance of a real-valued variable X, the maximum entropy distribution is the Gaussian distribution with same mean and variance. However, we will show that the most critical distributions (the ones that have the most extreme behaviour) compatible with the mean/variance information are trimodal mixtures of Dirac's deltas. Thus, maximum entropy is not a robust criterion since it chooses a unimodal distribution (Gaussian) when the true distributions may be multimodal.[1]

In this paper, we follow a different approach. Instead of choosing a single distribution (imposing some criterion), we address the problem by dealing with all the distributions compatible with the moment constraints. Thus, we determine lower and upper bounds of the expectations of all the functions of interest in estimation. To obtain this goal, we exploit the following results:
(i) given m generalised moments (e.g., mean, variance etc.) of random variable X, it is possible to define the set of all distributions that are compatible with the moments information;
(ii) this set of distributions is closed and convex and, thus, can be equivalently characterized by an upper (or lower) expectation model, that can be computed by solving a (infinite) linear programming problem;
(iii) by assuming that m generalised moments of initial state, process and measurement noise are known and considering the corresponding upper (or lower) expectation models, the filtering problem can then be solved by propagating in time the lower and upper expectations of any real-valued function

Author's address: Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland. e-mail: alessio@idsia.ch. Acknowledgements: this work has been partially supported by the Swiss NSF grant n. 200020-137680/1.

[1] For example, the MMSE or MAP point estimate computed assuming a unimodal-symmetric distribution can be in a region of low probability if the true distribution is multimodal with the same mean and variance.

of interest g by using the approach proposed in [3].
The obtained Generalised Moment Based Filter has several interesting properties. It reduces to set-membership estimation when only the supports of the noises are known. It can be used to determine uncertainty bounds for the Kalman filter estimate when the distributions of the process or measurement noises are unknown. It can compute the set of all Bayesian optimal estimates which are compatible with the moment information. It can be applied to both linear/nonlinear systems [11, Sec. 8].

II. Imprecise information, only moments known

Consider a variable X taking values x in the possibility space $\mathcal{X}$ (e.g., a finite set or a subset of the euclidean space $\mathbb{R}^n$) and let $\mathcal{B}$ be the Borel σ-algebra of measurable subsets of $\mathcal{X}$. Consider a sequence of $m+1$ real-valued functions f_i for $i = 0, 1, \ldots, m$ defined on $\mathcal{X}$ and measurable with respect to $\mathcal{B}$. We call f_i for $i = 0, 1, \ldots, m$ *generalised moment functions* (gmfs). Assume that the only information about X is expressed in terms of expectations of these gmfs, i.e., we know that

$$\mu_i = E_P[f_i] = \textstyle\int_{\mathcal{X}} f_i(x) dP(x) \text{ for } i = 0, 1, \ldots, m, \quad (1)$$

where dP denotes a Borel measure on $\mathcal{X}$, $\mu_i \in \mathbb{R}$ are finite and known. In the rest of the paper, we assume that $\mu_0 = 1$ and $f_0 = 1$ and, thus, that P is a probability measure, i.e., $E_P[1] = \int_{\mathcal{X}} dP(x) = 1$ (normalization) and, that, $\boldsymbol{\mu} = [1, \mu_1, \mu_2, \ldots, \mu_m]^T$ is a feasible moment vector, i.e., a moment vector of some distribution.

Notice that the knowledge of the expectation of $m+1$ moment functions is not enough to uniquely specify the probability measure P, so we can consider the set of all probabilities which are compatible with this information:

$$\mathcal{P}(\boldsymbol{\mu}) = \left\{ P : \quad \textstyle\int_{\mathcal{X}} \mathbf{f}(x) dP(x) = \boldsymbol{\mu} \ \right\}, \quad (2)$$

where $\mathbf{f} = [1, f_1, \ldots, f_m]^T$ is the vector of gmfs.

Given a real-valued objective function g defined on $\mathcal{X}$ and integrable with respect all probabilities in $\mathcal{P}(\boldsymbol{\mu})$, our goal is to compute

$$\underline{E}[g] = \inf_{P \in \mathcal{P}(\boldsymbol{\mu})} E_P[g], \quad \overline{E}[g] = \sup_{P \in \mathcal{P}(\boldsymbol{\mu})} E_P[g], \quad (3)$$

i.e., the lower and upper bounds on the expectation of g over the set of probability measures $\mathcal{P}(\boldsymbol{\mu})$ whose vector of gmfs match $\boldsymbol{\mu}$. A particular case of interest of the above optimization problem is when $\mathcal{X} \subseteq \mathbb{R}$ and $f_i = X^i$ and, thus, the constraints (1) become:

$$E[1] = 1, \ E[X] = \mu_1, \ldots, E[X^m] = \mu_m. \quad (4)$$

By considering (4), we are assuming that the only knowledge about X is represented by the first $m+1$ raw moments. For $m = 2$, this means that we only know the space of possibilities $\mathcal{X}$, the mean μ_1 and the variance $\mu_2 - \mu_1^2$ of X.

If instead of the first $m+1$ raw moments we know for instance the first q-quantiles $r_1, \ldots, r_q$ of X, then we can model this knowledge through (1) by setting $m = q+1$, $f_i = I_{\{X \leq r_i\}}$ and $\mu_i = i/m$ for $i = 1, 2, \ldots, m$, where I_Ω is the indicator function of the set Ω. Thus, the formulation (1) is very general and allows to model a large variety of situations.

A. Optimization problem

Consider the problem (3) with the constraints (1). Hereafter, we focus on $\overline{E}[g]$ (upper bound) only, but all results hold true for the lower bound as well, since $\underline{E}[g] = -\overline{E}[-g]$. Observe that, in the optimization problem (3): (i) the optimization variables are the amount of non-negative mass assigned to each point x in $\mathcal{X}$, (ii) the objective $E_P[g]$ and the constraints $E_P[\mathbf{f}] = \boldsymbol{\mu}$ are linear functions of the optimization variables. Therefore, if $\mathcal{X}$ is finite then (3) is a conventional linear program, while if $\mathcal{X}$ is infinite then (3) is a semi-infinite linear program (i.e., infinite number of decision variables but finite number of constraints).

Hence, since (3) is a (infinite) linear program, from the fundamental theorem of linear program we know that in the search for an optimal solution we can focus only on basic solutions (extreme points). Karr [12] has in fact proved that the set of probability measures $\mathcal{P}(\boldsymbol{\mu})$ which are feasible for the semi-infinite linear program problem (3) is convex and compact with respect to the weak* topology. As a result, $\mathcal{P}(\boldsymbol{\mu})$ can be expressed as the convex hull of its extreme points and these extreme points are probability measures that have at most $m+1$ distinct points of support in $\mathcal{X}$ (e.g., on $\mathbb{R}$ they are mixtures of $m+1$ Dirac's deltas), see [13, Lemma 3.1]. A consequence of this is that the integral $E_P[g] = \int_{\mathcal{X}} g(x) dP(x)$ with respect to the probability $P \in \mathcal{P}(\boldsymbol{\mu})$ becomes a sum over $m+1$ points when calculated on the extreme solution which gives the upper expectation (the same holds for the integrals in the constraints in (1)). The aim of the optimization is thus to find centres and weights of this $m+1$ mixture of Diracs'.

Since (3) is a (infinite) linear program, we can define its dual problem. Because (3) has $m+1$ constraints, then the dual has $m+1$ optimization variables $\mathbf{z} = [z_0, z_1, \ldots, z_m]$ with $z_i \in \mathbb{R}$ [13, Sec. 3] and it is equal to:

$$\overline{E}[g] = \inf_{\mathbf{z}} \mathbf{z}^T \boldsymbol{\mu}, \quad s.t. \ \ \mathbf{z}^T \mathbf{f}(x) \geq g(x), \quad \forall x \in \mathcal{X}. \quad (5)$$

Observe that when $\mathcal{X}$ is infinite, the dual is a semi-infinite linear program, since the number of constraints is infinite.

The dual can also be rewritten in the equivalent minimax formulation [14, Sec. 3.1.3]:

$$\overline{E}[g] = \quad \inf_{\mathbf{z}} \sup_{x \in \mathcal{X}} g(x) - \mathbf{z}^T (\mathbf{f}(x) - \boldsymbol{\mu}). \quad (6)$$

When $\mathcal{X}$ is finite, (5) becomes a linear program and, thus, can easily be solved. In case $\mathcal{X}$ is infinite, a way to solve it is first to discretise $\mathcal{X}$ and then, use standard linear programming. We will discuss other approaches in Section VI. Let us now consider some examples.

B. The case of $m = 0$

In this case, the only constraint is $\int_{\mathcal{X}} dP(x) = 1$. Assume for instance that $\mathcal{X} = [-d, d] \subset \mathbb{R}$ with $d > 0$, then in this case the only knowledge about the value x of X is that $x \in [-d, d]$ or, equivalently, $|x| \leq d$. We can model this norm bounded constraint by the following probabilistic constraint $P(x \in \mathcal{X}) = \int_{\mathcal{X}} dP(dx) = 1$. This means that the only information about the probability measure of X is its support $\mathcal{X}$. We can then consider the set of all probability measures

whose support is $\mathcal{X}$ and use this set to compute lower and upper bounds for the expectation of real-valued functions of interest g of X. In this case, it holds that

$$\overline{E}[g] = \sup_{P\in\mathcal{P}(\boldsymbol{\mu})} \int_{\mathcal{X}} g(x)dP(x) = \sup_{x\in\mathcal{X}} g(x), \tag{7}$$

and $\underline{E}[g] = \inf_{x\in\mathcal{X}} g(x)$. This follows by the fact that in this case ($m = 0$) the set of extreme points of the closed convex set of probabilities $\mathcal{P}(\boldsymbol{\mu})$ is the set of all $m + 1 = 1$ mixtures of Dirac's deltas in $\mathcal{X}$. Hence, it is clear that for any function g it holds $\overline{E}[g] = \sup_{x\in\mathcal{X}} g(x)$ (it is enough to take $p(x) = \delta_{x_0}(x)$ with $x_0 = \arg\sup_{x\in\mathcal{X}} g(x)$). For instance in case $g = X$ and $\mathcal{X} = [-d, d]$, one gets that $\underline{E}[X] = -d$, $\overline{E}[X] = d$, which are respectively the lower and upper mean of X.

C. The case of $m = 2$, $f_1 = X$ and $f_2 = X^2$

In this case, it is assumed that the only knowledge about X is represented by the space of possibilities and the first two raw moments (equivalently, mean and variance) of X. Assume that $\mathcal{X} = \mathbb{R}$, $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}^+$ and $\mu_2 - \mu_1^2 \geq 0$ (this is the positivity constraint for the variance of X). The constraints on $\boldsymbol{\mu} = [1, \mu_1, \mu_2]^T$ have been imposed to guarantee that $\boldsymbol{\mu}$ is a feasible moment sequence (i.e, there exists at least a probability distribution with moment vector $\boldsymbol{\mu}$). In this case, we know that the set of extreme points of $\mathcal{P}(\boldsymbol{\mu})$ are at most mixtures of $m + 1 = 3$ Dirac's deltas. Our aim is to compute lower and upper bounds for the expectation of real-valued functions g of X. Since the first two raw moments are known, it is clear that $\underline{E}[X] = \overline{E}[X] = \mu_1$ and $\underline{E}[(X - \mu_1)^2] = \overline{E}[(X - \mu_1)^2] = \mu_2 - \mu_1^2$.

Consider instead the following function $g = I_{\{|X-\mu_1|\leq\gamma\sigma\}}$ with $\gamma > 0$, where $|X - \mu_1| \leq \gamma\sigma$ is the standard $\gamma\sigma$ credible interval: centred on the mean μ_1 and with standard deviation $\sigma = \sqrt{\mu_2 - \mu_1^2}$. Then, it holds that

$$\underline{E}[I_{\{|X-\mu_1|\leq\gamma\sigma\}}] = 1 - \tfrac{1}{\gamma^2}, \tag{8}$$

and the lower expectation in (8) is obtained by a trimodal mixture of Dirac's deltas, see [11, Sec. II]. Notice that the lower expectation of $I_{\{|X-\mu_1|\leq\gamma\sigma\}}$ corresponds to the worst-case (equality) in the Chebyshev inequality $P(|X - \mu_1| \leq \gamma\sigma) \geq 1 - \frac{1}{\gamma^2}$, i.e., the probability of the set $|X - \mu_1| \leq \gamma\sigma$ is equal to $1 - \frac{1}{\gamma^2}$. Thus we have obtained the Chebyshev inequality.

D. Multivariate moment problem

In all the above examples we have considered univariate moment problems, i.e., $\mathcal{X} \subseteq \mathbb{R}$. However, we have previously assumed that X can take values in a generic possibility space $\mathcal{X}$ including thus the multivariate case $\mathcal{X} \subseteq \mathbb{R}^n$.

Consider then a multivariate variable $X = [X_1, \ldots, X_n]^T$ on $\mathcal{X} \subseteq \mathbb{R}^n$ and assume that the first α raw moments of X are known, i.e., $E[X_1^{\alpha_1} X_2^{\alpha_2} \cdots X_n^{\alpha_n}] = \mu_{\alpha_1\alpha_2\cdots\alpha_n}$, for any non-negative integer α_i such that $\alpha_1 + \alpha_2 + \cdots + \alpha_n \leq \alpha$ [15]. The above multivariate moment problem can be expressed in the form (1) by choosing $f_i(x) = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$. Thus, given a real-valued function g of X, the goal is always to determine lower and upper bounds for $E_P[g]$ with $P \in \mathcal{P}(\boldsymbol{\mu})$. For instance, in case $X = [X_1, X_2]^T$, we have 6 constraints $E[1] = 1$, $E[X_1] = \mu_{10}$, $E[X_2] = \mu_{01}$, $E[X_1X_2] = \mu_{11}$, $E[X_1^2] = \mu_{20}$, $E[X_2^2] = \mu_{02}$, which is equivalent to assume the knowledge of support, mean and covariance matrix of X.

III. Filtering

The objective in this paper is to solve the filtering problem assuming that the only information available on the distributions of initial state X_0, state dynamics $X_k|X_{k-1}$ and measurement equation $Y_k|X_k$ is represented by the expectation of generalised moment functions, i.e.,

$$E_{X_0}[\boldsymbol{f}(X_0)] = \boldsymbol{\mu}_0^x, \quad E_{X_k}[\boldsymbol{f}(X_k)|x_{k-1}] = \boldsymbol{\mu}_k^x(x_{k-1}), \tag{9}$$

$$E_{Y_k}[\boldsymbol{f}(Y_k)|x_k] = \boldsymbol{\mu}_k^y(x_k), \tag{10}$$

where the moment vectors $\boldsymbol{\mu}_k^x(x_{k-1})$ and $\boldsymbol{\mu}_k^y(x_k)$ depend on the conditioning variables x_{k-1} and x_{k-1} and $\boldsymbol{f}(X_k)$, $\boldsymbol{f}(Y_k)$ are generalized moment functions. Note also that, here we are assuming that the distributions of the noises are unknown and non stationary, i.e., they can vary with time, while the (time evolution of the) moments are known and given by (9)–(10). In Section II, we have shown that the only knowledge of a finite number of moments of a variable X does not allow to specify a single distribution. Instead of arbitrarily choosing a distribution compatible with the given moments (e.g., maximum entropy distribution), we solve the filtering problem by considering all the distributions in the set $\mathcal{P}(\boldsymbol{\mu})$ that satisfy (9)–(10).

To obtain this goal we do not simply apply the standard Bayesian filtering to the set of extreme points (e.g., Dirac's deltas) of $\mathcal{P}(\boldsymbol{\mu})$. In fact, because of the prediction step, the number of extreme points of the set of posterior densities, which characterizes our imprecise information on the state X_t, increases exponentially with time, see for instance [11, Sec. V]. So we cannot characterize our information on the state given all the past measurements by means of the posterior set of distributions.[2] In [3], it has been shown that an efficient solution of the moment based filtering problem can be computed by propagating in time the lower and upper expectation of functions of interest g of X_k (e.g., mean, credible interval, etc.). To summarise this result, we introduce the short notation $X^\ell = \{X_0, \ldots, X_\ell\}$ and $Y^\ell = \{Y_1, \ldots, Y_\ell\}$.

Theorem 1. *Assume that our information on the initial state, state dynamics and measurement equation is represented by the upper expectation models $\overline{E}_{X_0}$, $\overline{E}_{X_k}[\cdot|X_{k-1}]$ and, respectively, $\overline{E}_{Y_k}[\cdot|X_k]$, which are assumed to be known for $k = 1, \ldots, t$. Furthermore, assume that, for each $k = 1, \ldots, t$, X^{k-2} and Y^{k-1} are epistemically irrelevant to X_k given X_{k-1} and that X^{k-1} and Y^{k-1} are irrelevant to Y_k given X_k, meaning that*

$$\overline{E}_{X_k}[h_1|x^{k-1}, y^{k-1}] = \overline{E}_{X_k}[h_1|x_{k-1}], \tag{11}$$

$$\overline{E}_{Y_k}[h_2|x^k, y^{k-1}] = \overline{E}_{Y_k}[h_2|x_k], \tag{12}$$

[2] This is the same problem that arises in set-membership estimation in the multivariate case, i.e., the number of vertices (extreme points) that describes exactly the membership-set of the state at a given time t increases exponentially with time.

for any bounded scalar functions $h_1 : \mathcal{X}^k \times \mathcal{Y}^{k-1} \to \mathbb{R}$ *and given* x^{k-1}, y^{k-1}, $h_2 : \mathcal{X}^k \times \mathcal{Y}^k \to \mathbb{R}$ *and given* x^k, y^{k-1}. *Then, assuming that* $\overline{E}_{X^t,Y^t}[\prod_{k=1}^t I_{\{\tilde{y}_k\}}] > 0$ *and given the sequence of measurements* $\tilde{y}^t = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_t\}$, *the posterior upper expectation* $\overline{E}_{X_t}[g|\tilde{y}^t]$ *for any bounded scalar function* $g : \mathcal{X}_t \to \mathbb{R}$ *is equal to the unique value* $\nu \in \mathbb{R}$ *that solves the following optimization problem:*

$$\inf \nu \quad s.t. \quad \overline{E}_{X^t,Y^t}\Big[(g-\nu)\prod_{k=1}^t I_{\{\tilde{y}_k\}}\Big] \leq 0, \tag{13}$$

where the above joint upper expectation is given by:

$$\overline{E}_{X_0}\Big[\overline{E}_{X_1}\Big[\overline{E}_{Y_1}\Big[\ldots\overline{E}_{X_t}\Big[\overline{E}_{Y_t}\big[(g-\nu)\prod_{k=1}^t I_{\{\tilde{y}_k\}}|X_t\big]\Big|X_{t-1}\Big] \ldots\Big|X_1\Big]\Big|X_0\Big]\Big]. \tag{14}$$

■

The proof and explanations can be found in [3, Th. 2].[3] Notice that, the irrelevance conditions (11)–(12) are just the generalization to (lower) upper expectation models of the Markov conditions of independence assumed in Bayesian filtering. Furthermore, the (lower) upper expectation model obtained by solving (13) is in fact equal to the lower envelope of the posterior expectations that we obtain by applying Bayes' rule to the set of probability measures associated to the joint model (14). It can in fact be verified that Theorem 1 reduces to the solution of Bayesian filtering for density functions when upper expectations are replaced by expectations and the indicators $I_{\{\tilde{y}_k\}}$ by $I_{B(\tilde{y}_k,\gamma)}$, where $B(\tilde{y}_k,\gamma)$ is ball of radius γ centred at $\tilde{y}_k$, and then taking the limit for $\gamma \to 0$ [3, Corol. 1].[4] Bayesian filtering represents the most informative case, in the next section we will show that set-membership estimation represents the least informative case.

A. Set-membership estimation

Consider the following scalar system

$$\begin{cases} x_k &= ax_{k-1} + w_{k-1}, \\ y_k &= cx_k + v_k, \end{cases} \tag{15}$$

where $a, c \in \mathbb{R}$ are known, and assume that the only information on the initial state and disturbances is represented by the following norm bounded constraints

$$|x_0| \leq \rho_0, \quad |w_k| \leq \rho_w, \quad |v_k| \leq \rho_v, \tag{16}$$

for some given scalar $\rho_0, \rho_w, \rho_v > 0$. This means that the only available information on initial state and disturbances is represented by their spaces of possibilities (i.e., the sets where they take values). From (15)–(16), it in fact follows that, given x_{k-1}, the space of possibilities of X_k is $\mathcal{X}_k(x_{k-1}) = \{x_k \in \mathbb{R} : \ |x_k - ax_{k-1}| \leq \rho_w\}$ and, given x_k, the space of possibilities of Y_k is $\mathcal{Y}_k(x_k) = \{y_k \in \mathbb{R} : \ |y_k - cx_k| \leq \rho_v\}$.

[3]The only difference w.r.t. [3, Th. 2] is that here for conditioning we apply the so called Regular extension (13). Applying regular extension is equivalent to compute the lower envelope of the posterior expectations that we obtain by applying Bayes' rule to the set of probability measures (associated to the joint model) that assign positive mass to the observations [14, Appendix J].

[4]Bayes' rule for density functions is defined as the limit of Bayes' rule for probability mass function (discrete observations) when the discretisation step goes to zero, see for instance [14, Sec. 6.10].

The notation $\mathcal{X}_k(x_{k-1})$ and $Y_k(x_k)$ is used to highlight the dependence of the possibility spaces of the conditional models on x_{k-1} and, respectively, x_k. In Section II we have shown that (15)–(16) is a particular case of a moment problem, i.e., the $m = 0$ moment problem, in which we have a single constraint (on the support of the variable) and the upper (lower) expectation of any function g is simply equal to the supremum (infimum) of g in $\mathcal{X}$. Thus, the upper expectation models corresponding to the (15)–(16) are:

$$\begin{aligned} \overline{E}_{X_0}[g] &= \sup_{|x_0|\leq\rho_0} g(x_0), \\ \overline{E}_{X_k}[g|x_{k-1}] &= \sup_{|x_k - ax_{k-1}|\leq\rho_w} g(x_k), \\ \overline{E}_{Y_k}[h|x_k] &= \sup_{|y_k - cx_k|\leq\rho_v} h(y_k), \end{aligned} \tag{17}$$

for any given x_{k-1}, x_k and real-valued functions of the state (g) and of the measurement (h).

Assume that $t = 1$ (we can generalize the following derivations to the case $t > 1$) and consider (14) in case initial state, measurement model and state transition are described by (17). Then, (14) is equal to:

$$\sup_{\{x_0:|x_0|\leq\rho_0\}} \sup_{\{x_1:|x_1-ax_0|\leq\rho_w\}} \sup_{\{y_1:\ |y_1-cx_1|\leq\rho_v\}} (g(x_1)-\nu) I_{\{\tilde{y}_1\}}(y_1). \tag{18}$$

If $\tilde{y}_1$ is an observation compatible with the constraints (15)–(16), one has $\sup_{\{x_0:|x_0|\leq\rho_0\}} \sup_{\{x_1:|x_1-ax_0|\leq\rho_w\}} \sup_{\{y_1:\ |y_1-cx_1|\leq\rho_v\}} I_{\{\tilde{y}_1\}}(y_1) > 0$ thus we can apply Theorem 1. The joint upper expectation (18) can be rewritten as

$$\begin{aligned} &\sup_{\{x_0:|x_0|\leq\rho_0\}} \sup_{\{x_1:|x_1-ax_0|\leq\rho_w\}} \Big[(g(x_1)-\nu)^+ \sup_{\{y_1:\ |y_1-cx_1|\leq\rho_v\}} I_{\{\tilde{y}_1\}}(y_1) \\ &+(g(x_1)-\nu)^- \inf_{\{y_1:\ |y_1-cx_1|\leq\rho_v\}} I_{\{\tilde{y}_1\}}(y_1)\Big], \end{aligned} \tag{19}$$

where $(g-\nu)^+ = \sup(0, g-\nu)$ and $(g-\nu)^- = \inf(0, g-\nu)$ are respectively the positive and negative part of $(g-\nu)$. It holds that $\inf_{\{y_1:\ |y_1-cx_1|\leq\rho_v\}} I_{\{\tilde{y}_1\}}(y_1) = 0$, and $\sup_{\{y_1:\ |y_1-cx_1|\leq\rho_v\}} I_{\{\tilde{y}_1\}}(y_1) = I_{\{x_1:\ |\tilde{y}_1-cx_1|\leq\rho_v\}}$. Thus, (19) can be rewritten as

$$\sup_{\{x_0:|x_0|\leq\rho_0\}} \sup_{\{x_1:|x_1-ax_0|\leq\rho_w, |\tilde{y}_1-cx_1|\leq\rho_v\}} (g(x_1)-\nu)^+. \tag{20}$$

Since $(g(x_1)-\nu)^+$ is always non-negative, the infimum value of ν such that (20) is non-positive is

$$\nu = \sup_{\{x_0:|x_0|\leq\rho_0\}} \sup_{\{x_1:|x_1-ax_0|\leq\rho_w\}\cap\{x_1:|\tilde{y}_1-cx_1|\leq\rho_v\}} g(x_1), \tag{21}$$

which, by (13), is the solution of the filtering problem $\nu = \overline{E}[g|\tilde{y}_1]$, i.e., the upper posterior expectation of the function g of X_1 given the observation $\tilde{y}_1$. Observe that the hypothesis $\overline{E}_{X,Y_1}[I_{\{\tilde{y}_1\}}] > 0$ of Theorem 1 ensures that the intersection in the second supremum in (21) is non empty. Consider for instance the case $g = X_1$ and assume that $a = c = 1$, then from (21) one has that $\nu = \overline{E}_{X_1}[X_1|\tilde{y}_1] = \min(\rho_0 + \rho_w, \tilde{y}_1 + \rho_v)$, and, thus, $\underline{E}_{X_1}[X_1|\tilde{y}_1] = -\overline{E}_{X_1}[-X_1|\tilde{y}_1] = \max(-\rho_0 - \rho_w, \tilde{y}_1 - \rho_v)$ gives the upper posterior mean of X_1.

It can be noticed that the solution ν of (18) coincides with the set-membership estimate of X_1 and the posterior expectation of any function of X_k at any time k can be simply obtained by applying interval analysis. Actually, in set-membership estimation one aims to compute the posterior support of X_1 and not the lower and upper posterior means. It can easily be shown that the interval $[\underline{E}_{X_1}[X_1|\tilde{y}_1], \overline{E}_{X_1}[X_1|\tilde{y}_1]]$ coincides with the posterior support.

B. Markov's moment problem

Assume that, besides m-moments, we know lower and upper bounds of the density function. For instance, for the measurement density function, this means

$$l(y_k|x_k) \leq p(y_k|x_k) \leq u(y_k|x_k),$$

where the two functions $0 < l(y_k|x_k), u(y_k|x_k)$ are assumed to be known real-valued bounded continuous functions for any $x_k \in \mathcal{X}_k$. This is the Markov's moment problem [5, Ch. 7] in which, besides the knowledge of the moments, an additional condition is imposed on the distribution $dP(x)$ in (1): the distribution is required to have a density $p(x)$ which is lower and upper bounded by known functions. If we assume a bounded density model for the measurement equation, then it can be shown [16] that for any real-valued function h of Y_k the upper expectation $\overline{E}_{Y_k}\big[h\big|x_k\big]$ can be obtained by solving:

$$\begin{aligned}\inf_{\mathbf{z}} \mathbf{z}^T\boldsymbol{\mu}(x_k) + &\int_{\mathcal{Y}_k} (h(y_k) - \mathbf{z}^T\mathbf{f}(y_k))^+ u(y_k|x_k)dy_k \\ &+ \int_{\mathcal{Y}_k} (h(y_k) - \mathbf{z}^T\mathbf{f}(y_k))^- l(y_k|x_k)dy_k.\end{aligned} \tag{22}$$

Provided the upper expectation model in (22) is well-defined, the extreme densities which give the lower/upper expectations are piecewise densities that can only assume values in the set $\{l(y_k|x_k), u(y_k|x_k)\}$ for any $y_k \in \mathcal{Y}_k$ and $x_k \in \mathcal{X}_k$ and that have at most $m+1$ points of discontinuity in $\mathcal{Y}_k$. Observe that, when $l(y_k|x_k) = u(y_k|x_k) = p(y_k|x_k)$, we are back to a standard PDF model and (22) reduces to $E_{Y_k}\big[h\big|x_k\big] = \int_{\mathcal{Y}_k} h(y_k)p(y_k|x_k)dy_k$.

In the following, we assume the model (22) for the measurement equation with $l(y_k|x_k) > 0$ and, thus, the measurement model is a set of PDFs defined by gmfs and lower and upper bounds on the density.

Observe that, when $l(y_k|x_k) = 0$ one has that $\underline{E}_{Y_k}\big[I_{B(\tilde{y}_k,\gamma)}\big|x_k\big] = 0$ for any $x_k \in \mathcal{X}_k$ and small γ, and in this case it can be shown that the moment based filter gives the same posterior inferences as set-membership estimation even in the case, besides the supports, additional information on the moments is available for initial state and noises. The intuitive explanation for this behaviour is that $\underline{E}_{Y_k}\big[I_{B(\tilde{y}_k,\gamma)}\big|x_k\big] = 0$ implies that the set of conditional probabilities that characterizes the observation model includes distributions that are zero on a neighborhood of the observation $\tilde{y}_k$. This means the event "the observation y_k falls in a neighborhood of $\tilde{y}_k$" has zero probability. Thus, Bayes' rule is not applicable (the likelihood is zero). Applying (13) in this case, it is equivalent to apply Bayes' rule only to the likelihoods that assign positive mass (but arbitrarily close to zero) to a neighborhood of $\tilde{y}_k$. If we do that, it results that the posterior inferences coincide with those obtained via set-membership estimation. This behaviour can be avoided if $l(y_k|x_k) > 0$ for any y_k, x_k. This happens for instance when the density function of the measurement model is known and positive (e.g., Gaussian) or when lower and upper bounds for the density are given with $l(y_k|x_k) > 0$.

IV. An algorithm for the GMF filter problem

Hereafter, we describe an algorithm that allows to solve (13) in case the upper expectation models for initial state, state dynamics and measurement equation are given by (9)–(10) with the additional constraints $l(y_k|x_k) \leq p(y_k|x_k) \leq u(y_k|x_k)$ for the measurement model. To solve (5), we discretise the support $\mathcal{X}_k$ for $k = 0, 1, \ldots, t$ so that (5) becomes a linear program.

1) For each $k = 0, \ldots, t$ discretise X_k by generating n points (equally spaced) in $\mathcal{X}_k$.
2) Set a value of ν and set $g(\cdot, t, \nu) = g(\cdot) - \nu$.
3) Do the following backward propagation for $k = t, \ldots, 1$: For each discretised value x_{k-1}^j of X_{k-1}, solve

$$\begin{aligned}&g(x_{k-1}^j, k-1, \nu) = \inf_{\mathbf{z}} \mathbf{z}^T\boldsymbol{\mu}_x^k(x_{k-1}^j) \\ &s.t.\ \mathbf{z}^T\mathbf{f}(x_k^i) - \overline{E}_{Y_k}[g(x_k^i, k, \nu)I_{\{B(\tilde{y}_k,\gamma)\}}|x_k] \geq 0, \quad \forall x_k^i \in \mathcal{X}_k,\end{aligned}$$

where $\boldsymbol{\mu}_x^k(x_{k-1}^j)$ are the known moments of $\mathbf{f}(x_k)$ given x_{k-1}^j.
4) Solve:

$$\begin{aligned}res = \quad &\inf_{\mathbf{z}} \mathbf{z}^T\boldsymbol{\mu}_x^0 \\ s.t. \quad &\mathbf{z}^T\mathbf{f}(x_0^i) - g(x_0^i, 0, \nu) \geq 0, \quad \forall x_0^i \in \mathcal{X}_0,\end{aligned}$$

where $\boldsymbol{\mu}_0$ are the known moments of $\mathbf{f}(x_0)$.
5) Repeat steps 2–4 until the infimum value of ν such that $res \leq 0$ is achieved.

The solution of step 5 gives $\nu = \overline{E}_{X_k}[g|\tilde{y}^k]$. Observe that also to compute $\overline{E}_{Y_k}[g(x_k^i, k, \nu)I_{\{B(\tilde{y}_k,\gamma)\}}|x_k]$ we need to solve an optimization problem like (22). [5]

V. Numerical simulations

In [11], we have shown the application of the generalised moment based filter (GMBF) to both a linear and nonlinear scalar systems in which the measurement noise is Gaussian and only the mean and variance are known for initial state and process noise. Here, we modify the assumptions on the noises to obtain a percentiles based filter.

Consider the one-dimensional model (15). We assume to know: (i) the support $\mathcal{X}_0$, the 25th percentile $E[I_{\{X_0 \leq -0.3\}}] = 0.25$, the 50th percentile $E[I_{\{X_0 \leq 0\}}] = 0.5$ and the 75th percentile $E[I_{\{X_0 \leq 0.3\}}] = 0.75$; (ii) a similar model is assumed for the process noise. These are the percentiles of a Cauchy distribution with zero location parameter and 0.3 scale parameter $\mathcal{C}(0, 0.3)$. Conversely, the measurement noise is assumed to be Gaussian distributed $v_k \sim \mathcal{N}(0, r)$ with $r > 0$ (i.e., $l(y_k|x_k) = u(y_k|x_k) = \mathcal{N}(y_k, x_k, r)$). From the knowledge of the percentiles of the process noise

[5] We have implemented the above algorithm in matlab-tomlab by solving the dual problems by using `cplex` algorithm and the infimum on ν problem using `npsol`.

and (15), we can derive: $E[I_{\{X_k \leq -0.3+ax_{k-1}\}}|x_{k-1}] = 0.25$, $E[I_{\{X_k \leq ax_{k-1}\}}|x_{k-1}] = 0.5$ and $E[I_{\{X_k \leq 0.3+ax_{k-1}\}}|x_{k-1}] = 0.75$. We can the employ the algorithm of Section IV with $\mathcal{X}_k = \mathcal{X} = [-50, 50]$ and discretisation step 0.1 to compute the 95% robust credible interval, i.e., the smaller interval which has lower probability 0.95 of including the true state. For comparison, we report the estimate of a KF that assumes $x_0 \sim \mathcal{N}(0, q)$, $w_k \sim \mathcal{N}(0, q)$ and $v_k \sim \mathcal{N}(0, r)$ where the variance $q = (0.4447)^2$ ensures that the zero-mean Gaussian distribution has the same 25th, 50th and 75th percentiles of $\mathcal{C}(0, 0.3)$.

A trajectory of 8 timesteps has been considered based on the assumption $x_0, w_k \sim \mathcal{C}(0, 0.3)$ (the true model is stationary but the GMBF is not assuming stationarity). For performance comparison, we have computed the optimal Bayesian estimate (OPF), i.e., the posterior mean, obtained by a particle filter (2500 particles) based on the true unknown distributions of x_0, w_k (i.e., the Cauchy distribution). The results are shown in Figure 1. It can be noticed that the KF estimate is wrong at time 7. The trajectory of the system and the OPF are not included in the 95% KF-CI (KF-Chebyshev Inequality) based interval, while they are included in the GMBF credible interval (because the likelihood model is Gaussian with a small variance, KF recovers very fast in 1-2 time steps). These violations happen in several trajectories: the average (in the 230 MC runs) coverage of the 95% KF-CI based interval is around 85%, while the 95% credible interval of GMBF has a coverage of 98%. The flat-tails of the Cauchy distribution makes the KF to be not robust.

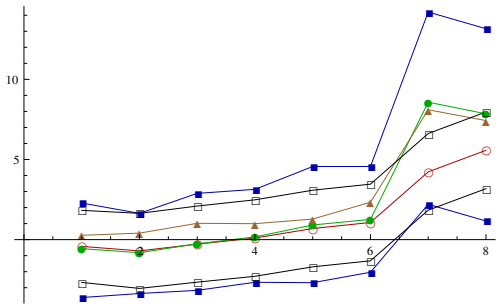

Fig. 1. Trajectory (brown triangles), KF estimate (red empty circles), OPF estimate (green full circles), GMBF lower and upper limits of the credible interval (blue full squares) and KF Chebyshev inequality based interval (black empty squares) for a single MC run.

VI. Conclusions

In this paper, we have solved the filtering problem in the case only few generalised moments of initial state and noise terms are known. We have also shown that this filter reduces to set-membership estimation when only the supports of the noises are known.

A main issue for future work is how to efficiently extend this approach to the multivariate case. A global discretisation of the space $\mathcal{X}$ is not efficient in high dimensions. These are some ideas that can be exploited.

First, we can assume that the noises are bounded and that we have some further information expressed as gmfs and bounds on the PDF. Since the noises are bounded we can use set-membership estimation to compute some polytopic outer approximation of the support [17] and, thus, apply the discretisation to this set. In this way, we perform a local discretisation which is much more efficient. The additional information on the noises (mean, variance) can be used to compute a 95% credible region that can be much smaller than the 100% credible region (support) computed by set-membership estimation without caring about the additional information on the noises. Second, we can solve the problem without discretisation by using the minimax formulation in (6) and nonlinear optimization approaches. Third, for piecewise polynomial functions f, g, the inner supremum in (6) can be solved efficiently [13] by using linear matrix inequalities. Thus, we could use a polynomial approximation instead of discretisation to practically solve the semi-infinite linear program.

References

[1] J. Spall, "The Kantorovich inequality for error analysis of the Kalman filter with unknown noise distributions," *Automatica J. IFAC*, vol. 31, pp. 1513–1517, 1995.

[2] J. Maryak, J. Spall, and B. Heydon, "Use of the Kalman filter for inference in state-space models with unknown noise distributions," *IEEE Transactions on Automatic Control*, vol. 49, no. 1, pp. 87 – 90, 2004.

[3] A. Benavoli, M. Zaffalon, and E. Miranda, "Robust filtering through coherent lower previsions," *Automatic Control, IEEE Transactions on*, vol. 56, pp. 1567 –1581, July 2011.

[4] J. Shohat and J. Tamarkin, *The problem of moments*. American Mathematical Society, 1950.

[5] M. Krein and A. Nudelmán, *The Markov moment problem and extremal problems*, vol. 50. Amer Mathematical Society, 1977.

[6] C. Byrnes and A. Lindquist, "A convex optimization approach to generalized moment problems, control and modeling of complex systems," in *Cybernetics in the 21st Century: Festschrift in Honor of Hidenori Kimura on the Occasion of his 60th*, pp. 3–21, 2003.

[7] C. Byrnes and A. Lindquist, "Important moments in systems, control and optimization," in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pp. 91 –96, dec. 2009.

[8] A. Ferrante, M. Pavon, and F. Ramponi, "Hellinger Versus Kullback-Leibler Multivariable Spectrum Approximation," *Automatic Control, IEEE Transactions on*, vol. 53, pp. 954 –967, may 2008.

[9] T. Georgiou and A. Lindquist, "Kullback-Leibler approximation of spectral density functions," *Information Theory, IEEE Transactions on*, vol. 49, pp. 2910 – 2917, nov. 2003.

[10] M. Pavon and A. Ferrante, "On the Georgiou-Lindquist approach to constrained Kullback-Leibler approximation of spectral densities," *Automatic Control, IEEE Transactions on*, vol. 51, pp. 639 – 644, april 2006.

[11] A. Benavoli and B. Noack, "Pushing kalman's idea to the extremes," in *Proc. 15th Int. Conf. Information Fusion*, (Singapore), pp. 1202–1209, 2012.

[12] A. Karr, "Extreme points of certain sets of probability measures, with applications," *Mathematics of Operations Research*, vol. 8, no. 1, pp. 74–85, 1983.

[13] A. Shapiro, "On duality theory of conic linear problems," in *in Semi-Infinite Programming Recent Advances*, pp. 135–165, 2001.

[14] P. Walley, *Statistical Reasoning with Imprecise Probabilities*. New York: Chapman and Hall, 1991.

[15] D. Bertsimas and I. Popescu, "Optimal inequalities in probability theory: A convex optimization approach," in *SIAM Journal on Optimization*, pp. 780–804, 2005.

[16] J. Smith, "Generalized Chebychev inequalities: theory and applications in decision analysis," *Operations Research*, pp. 807–825, 1995.

[17] L. Chisci, A. Garulli, and G. Zappa, "Recursive state bounding by parallelotopes," *Automatica*, vol. 32:7, pp. 1049–1055, 1996.