Pushing Kalman's Idea to the Extremes

Alessio Benavoli
IDSIA, Manno, Switzerland,
alessio@idsia.ch

Benjamin Noack
ISAS, Karlsruhe, Germany,
noack@kit.edu

***Abstract*— The paper focuses on the fundamental idea of Kalman's seminal paper: how to solve the filtering problem from the only knowledge of the first two 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 noise terms (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 of distributions can be equivalently characterized by its extreme distributions which is 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.**

Keywords: moments, Chebyshev bounds, imprecise probability.

I. Introduction

The attractiveness of the Kalman filter (KF) lies in the fact that it is simple (only the mean and variance of the state must be propagated through time) and optimal in several different senses: (i) in the Gaussian case it is the minimum variance estimator; (ii) in the non-Gaussian case it is the best linear minimum variance estimator (LMVE). 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. In fact, in this case the estimation error $\hat{x}_k - x_k$ has an unknown distribution. Thus, the best one can hope for is to give bound of the estimation error. The Chebyshev inequality can be used for this purpose: $P(|\hat{x}_k - x_k| > \delta\sigma_k) \leq \frac{1}{\delta^2}$, where σ_k^2 is the variance of $\hat{x}_k$. This method, however, has several limitations:

1) it can only be applied to determine lower/upper bounds for the probability of intervals of type $|\hat{x}_k - x_k| \gtreqless \delta\sigma_k$ but, in general, it cannot be extended to compute lower/upper bounds for the expectation of other functions of interest of the state (e.g., mean, some utility function to be used for control, a risk measure [1] etc.).
2) it can produce confidence regions that are too large (i.e., too conservative) for practical applications in which some further information on the distributions of the noises (besides mean and variance) is available.

To alleviate the conservativeness of Chebyshev inequality, Spall in [2] has proposed to compute confidence regions by using the Kantorovich inequality[1] and by assuming the unknown noises distributions to be from the class of log-concave distributions. In another paper, Maryak and Spall derive the same kind of bound by instead assuming that the unknown noises distributions are symmetric and unimodal. However, these two 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 the nonlinear case, KF cannot be applied and its nonlinear version, i.e., the extended Kalman filter (EKF), does not have any optimality property in general. Thus, Chebyshev-like inequalities applied to the EKF estimate/variance do not have any guarantee of reliability.

The objective of this paper is to solve these issues. The proposed solution is based on the following points:

- 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 of distributions is closed and convex and, thus, can be equivalently characterized by its extreme distributions which is a family of Dirac's deltas.
- The lower and upper expectation of any function g of X are obtained in correspondence of these extreme distributions and can be computed by solving a linear programming problem.
- The filtering problem can then be solved by propagating in time the lower and upper expectation of the functions of interest g by using an approach similar to the one proposed in [3].

II. Imprecise information, only moments known

Consider a real-valued X and assume that the only probabilistic information about the value x of X is:

$$E[I_{\{\Omega\}}] = 1,\ E[X] = \mu_1, \cdots, E[X^m] = \mu_m. \qquad (1)$$

where $I_{\{\Omega\}}$ is the indicator function of the set Ω, which is the space of possibilities of X, and μ_i for $i = 1, \ldots, m$ are the first m non-central moments of X. This means that the only knowledge about the variable X is represented by the first m non-central moments. Hereafter, we assume that $\Omega = \mathbb{R}$, i.e.,

[1] For a random variable X such that $0 < m \leq X \leq M$, the Kantorovich inequality states that a.s. $1 \leq E(X)E(X^{-1}) \leq (m+M)^2/4mM$.

we are considering a univariate filtering problem. Notice that m moments are not enough to uniquely specify the distribution of X, so we can consider the set of all Probability Density Functions (PDF) which are compatible with this information:

$$\mathcal{P} = \left\{ p(\cdot) \geq 0 : \begin{array}{rcl} \int p(x)dx & = & 1 \\ \int xp(x)dx & = & \mu_1 \\ & \vdots & \\ \int x^m p(x)dx & = & \mu_m \end{array} \right\} \quad (2)$$

Since we have assumed that $\mathcal{X} = \mathbb{R}$, the above integral are on $\mathbb{R}$. The constraints $p(\cdot) > 0$ and $\int p(x)dx = 1$ ensure that $p(\cdot)$ is a well-defined PDF.

In state estimation we are interested on computing the expectation of functions g of X, for instance the mean $g = X$, the variance $g = (X - E[X])^2$, the credible (Bayesian confidence) interval $g = I_{\{\mathcal{X}\}}$, where $\mathcal{X}$ is a some subset of $\mathbb{R}$. In case our information on X does not allow to specify a single precise distribution for X we cannot compute expectations. However, we can use the set of PDFs $\mathcal{P}$ compatible with our information on X to compute lower and upper bounds for the expectations of any real-valued function of interest g of X.

In practice, we aim to solve the following problem

$$\overline{E}[g] = \max_{p \in \mathcal{P}} E_p[g] = \max_{p \in \mathcal{P}} \int g(x)p(x)dx \quad (3)$$

where $E_p[g]$ denotes the expectation w.r.t. the PDF $p \in \mathcal{P}$ and $\overline{E}[g]$ the upper expectation of g w.r.t. $\mathcal{P}$. Hereafter, we focus on maximization problems (upper bound), but all results hold true for the lower-bound problem as well, since $\underline{E}[g] = -\overline{E}[-g]$. Karr [4] has proved that the set of probability measures $\mathcal{P}$ which are feasible for the infinite-dimensional problem (3) is convex and compact with respect to the weak* topology. As a result, $\mathcal{P}$ can be expressed as the convex hull of its extreme points. Furthermore, an optimal solution of (3) is obtained at an extreme point of $\mathcal{P}$ and these extreme points are probability measures that have at most $m+1$ distinct points of support in $\mathbb{R}$ [5]–[7], that is they are mixtures of Dirac's deltas with at most $m+1$ components. A consequence of this is that the integral in (3) becomes a sum over $m+1$ points when calculated on the mixture of Dirac's deltas which gives the upper expectation.

Observe that, (3) is an infinite linear optimization problem (the cost and the constraints are linear in the unknown p) which explains why the optimal solution is obtained at an extreme point of $\mathcal{P}$. Since (3) is a linear program, we can define its dual problem. Because the maximum is given by a mixture of Dirac's deltas, the dual problem can be written as the following linear semi-infinite program [7, Sec. 3]:

$$\begin{array}{rll} \overline{E}[g] = & \min\limits_{\tilde{z}} z^T q & \\ s.t. & z^T f(x) - g(x) \geq 0, & \forall x \in \Omega, \end{array} \quad (4)$$

where $q = [1, \mu_1, \mu_2, \ldots, \mu_m]^T$ and $f(x) = [1, x, x^2, \ldots, x^m]^T$. Observe that q must be a *feasible moment sequence* otherwise the problem does not have solution [5], e.g, in case $m = 2$ it must hold that $\mu_2 - \mu_1^2 > 0$ (this is the positivity constraint for the variance of X). A semi-infinite program can be seen as a special case of a bilevel program, i.e.,

$$\begin{array}{rll} \overline{E}[g] = & \min\limits_{z} z^T q & \\ s.t. & \min\limits_{x \in \Omega} z^T f(x) - g(x) \geq 0, & \end{array} \quad (5)$$

or, equivalently,

$$\begin{array}{rll} \overline{E}[g] = & \min\limits_{x \in \Omega} \min\limits_{z} z^T q & \\ s.t. & z^T f(x) - g(x) \geq 0, & \end{array} \quad (6)$$

which is a *minmin* problem. Notice that the function $z^T f(x) - g(x)$ can be nonlinear and, thus, (6) may be difficult. However, if the support set Ω is finite, (4) is a simple linear program and, thus, an approximate solution of (4) can be obtained by discretizing X and bounding the support Ω.

Let us now consider two examples of (3). Assume that $m = 2$ and $\mu_2 - \mu_1^2 > 0$ and consider the case $g = I_{\{X \leq x\}}$, where $I_{\{X \leq x\}}$ is the indicator function of the set $\{u \in \mathbb{R} : -\infty \leq u \leq x\}$ and, thus, $E[I_{\{X \leq x\}}]$ gives the Cumulative Distribution Function (CDF) of X, i.e., $F(x)$. It can be proved that the solution of (3) or, equivalently, (5) is given by the Chebyshev-Markov inequality:

$$\begin{array}{l} \overline{E}[I_{\{X \leq x\}}] = \overline{F}(x) = \left\{ \begin{array}{ll} \frac{\mu_2 - \mu_1^2}{\mu_2 - \mu_1^2 + (\mu_1 - x)^2}, & x \leq \mu_1, \\ 1, & x > \mu_1, \end{array} \right. \\ \underline{E}[I_{\{X \leq x\}}] = \underline{F}(x) = \left\{ \begin{array}{ll} 0, & x \leq \mu_1, \\ 1 - \frac{\mu_2 - \mu_1^2}{\mu_2 - \mu_1^2 + (\mu_1 - x)^2}, & x > \mu_1. \end{array} \right. \end{array} \quad (7)$$

Notice in fact that for $x = \mu_1 + \delta\sigma$ for some $\delta > 0$ and $\sigma^2 = \mu_2 - \mu_1^2$, one gets $\underline{F}(\mu_1 + \delta\sigma) = 1 - \frac{1}{1+\delta^2}$. This is the lower bound (worst-case) of the one-tailed version of Chebyshev inequality, i.e., $P(X - \mu_1 \leq \delta\sigma) \geq 1 - 1/(1+\delta^2)$. Fig. 1 shows the lower and upper CDF in the standard case, i.e., $\mu_1 = 0$ and $\sigma^2 = \mu_2 - \mu_1^2 = 1$. The following theorem

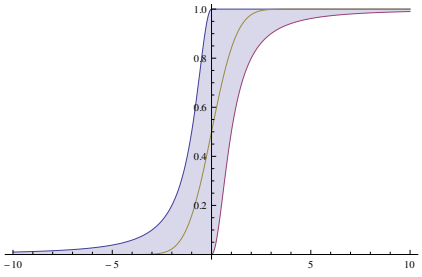

Fig. 1. Lower and upper bounds for the CDF of X in the standard case. The central plot is the CDF of the standard Normal distribution.

gives the densities in $\mathcal{P}$ that attain the bounds (7).

Theorem 1. *The set of extreme densities which attain the bounds (7) is:*

$$\begin{array}{rcl} \mathcal{P} & = & \big\{p : \ p(x) = w_1(m_2)\delta(x - m_1(m_2)) \\ & + & (1 - w_1(m_2))\delta(x - m_2), \ m_2 \in \mathbb{R} \backslash \{\mu_1\}\big\}, \end{array} \quad (8)$$

where $\delta(x - x_0)$ denotes the Dirac's delta centred at x_0,

$$w_1(m_2) = \frac{(\mu_1 - m_2)^2}{\mu_2 - \mu_1^2 + (\mu_1 - m_2)^2}, \quad (9)$$

$$m_1(m_2) = \frac{\mu_1(\mu_2 - \mu_1^2 + (\mu_1 - m_2)^2) - m_2(\mu_2 - \mu_1^2)}{(\mu_1 - m_2)^2}, \tag{10}$$

and m_2 is free to vary in $\mathbb{R}\backslash\{\mu_1\}$. Observe that w_1 and m_1 depends on the value of m_2, hence the notation $w_1(m_2)$ and $m_1(m_2)$. ■

Proof: *For lake of space, we just give a sketch of the proof. Consider $p(x) = w_1\delta(x - m_1) + (1 - w_1)\delta(x - m_2)$ with $w_1 \in (0, 1)$ and $m_1, m_2 \in \mathbb{R}$. In order to satisfy the constraints (1), we must choose w_1 and m_1, m_2 such that*

$$w_1 m_1 + (1 - w_1)m_2 = \mu_1, \quad w_1 m_1^2 + (1 - w_1)m_2^2 = \mu_2. \tag{11}$$

From simple algebraic manipulations it is possible to derive (9)–(10). Then minimizing and maximizing w.r.t. m_2 the expectation $E[I_{\{X\leq x\}}]$ one obtains (7). ■

These extreme PDFs are bimodal mixtures of Dirac's deltas. Consider now the following function $g = I_{\{|X-\mu_1|\leq\delta\sigma\}}$ with $\delta > 0$. Observe that $|X - \mu_1| \leq \delta\sigma$ is the standard $\delta\sigma$ credible interval: centred on the mean μ_1 and with standard deviation $\sigma = \sqrt{\mu_2 - \mu_1^2}$. It can be proved that the lower expectation of g obtained by solving (3) (by exploiting the fact that $\underline{E}[g] = -\overline{E}[-g]$) is

$$\underline{E}[I_{\{|X-\mu_1|\leq\delta\sigma\}}] = 1 - \frac{1}{\delta^2}. \tag{12}$$

Notice that the lower expectation of $I_{\{|X-\mu_1|\leq\delta\sigma\}}$ corresponds to the worst-case (equality) in the Chebyshev inequality $P(|X - \mu_1| \leq \delta\sigma) \geq 1 - \frac{1}{\delta^2}$, i.e., the probability of the set $|X - \mu_1| \leq \delta\sigma$ is exactly equal to $1 - \frac{1}{\delta^2}$.

Theorem 2. *The lower expectation in (12) is obtained by the following trimodal mixture of Dirac's deltas:*

$$\begin{aligned}\mathcal{P} &= \big\{p:\ p(x) = w(m)\delta(x - \mu_1 - m) \\ &+ \ w(m)\delta(x - m + \mu_1) + (1 - 2w(m))\delta(x - \mu_1)\big\},\end{aligned} \tag{13}$$

where $w(m) = (\mu_2 - \mu_1^2)/2m^2$ and $m^2 > \mu_2 - \mu_1^2$. Observe that besides the constraint on the first and second moment, the mixture also satisfies the symmetry constraint:

$$\int_{-\infty}^{\mu_1} p(x)dx = \int_{\mu_1}^{\infty} p(x)dx.$$

■

Proof: *It can be verified that (13) satisfies the symmetry and moments constraints. By minimizing w.r.t. m the expectation $E[I_{\{|X-\mu_1|\leq\delta\sigma\}}]$ computed w.r.t. the trimodal mixture, one obtains (12).* ■

Fig. 2 shows two members of the bimodal and trimodal family of mixtures of Dirac's deltas. In general, it holds the following.

Theorem 3. *Given any real-valued bounded function g of X, the upper posterior expectation of g is obtained by a mixture of Dirac's deltas with at most $m+1$ components, where m is the number of moment constraints.* ■

The proof can be found in [7, Lemma 3.1] and [8, Th.1].

III. Objective of the paper

The objective in this paper is to solve the filtering problem assuming that the only information available on the distributions of initial state and state dynamics is represented by the first two moments:

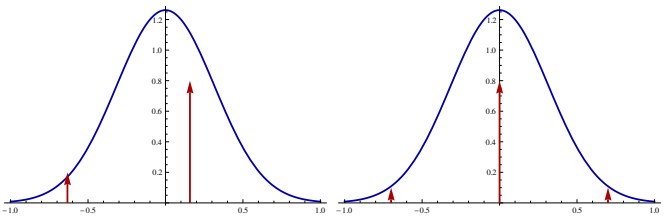

Fig. 2. Bimodal and Trimodal family of mixtures of Dirac's deltas (in red) for the case $\mu_1 = 0$, $\mu_2 = 0.1$, $w(m) = 0.8$ (trimodal) and $w_1(m_2) = 0.8$ (bimodal). The Gaussian (in blue) $\mathcal{N}(0, 0.1)$ has been reported for comparison.

$$E[X_0] = \hat{x}_0, \qquad E[X_0^2] = \hat{x}_0^2 + p_0, \tag{14}$$

$$E[X_{t+1}|X_t] = aX_t, \quad E[X_{t+1}^2|X_t] = (aX_t)^2 + q, \tag{15}$$

where $a, \hat{x}_0 \in \mathbb{R}$, $q, p_0 > 0$ are the parameter of the dynamic system. The first row represents the information on the initial state and the second on the state dynamics (for $t > 0$). The measurement equation is described by the Gaussian PDF $\mathcal{N}(y_t; x_t, r)$ as in the Kalman filter case.[2] In the next section, we discuss how to compute the updating and prediction steps from the only knowledge of (14)–(15) and without any assumption on the distributions of the noises or on the final form of the state estimator (e.g., linear form).

IV. Updating

Consider the set of densities $\mathcal{P}$ in (2) for X in the case $m = 2$, i.e., only the first two moments are known. Then consider the measurement model $\ell(y|x) = \mathcal{N}(y; x, r)$. The goal is to compute the upper posterior expectation of the function g' of X given the observation y, i.e., $\overline{E}_X[g'|y]$.

Theorem 4. *Assuming that*

$$\underline{E}_X[\ell(y|X)] = \min_{p\in\mathcal{P}} \int \ell(y|x)p(x)dx > 0, \tag{16}$$

the upper posterior expectation $\overline{E}[g'|y]$ of a bounded real-valued[3] function g' of X is the solution $\nu \in \mathbb{R}$ of

$$0 = \overline{E}_X[(g' - \nu)\ell(y|X)] = \max_{p\in\mathcal{P}} \int (g'(x) - \nu)\ell(y|x)p(x)dx. \tag{17}$$

■

Proof: Consider (17) and observe that:

$$\begin{aligned}0 &= \max_{p\in\mathcal{P}} \int (g'(x) - \nu)\ell(y|x)p(x)dx \\ &= \max_{p\in\mathcal{P}} \left[\frac{\int g'(x)\ell(y|x)p(x)dx}{\int \ell(y|x)p(x)dx} - \nu\right] \int \ell(y|x)p(x)dx,\end{aligned}$$

[2] We plan to relax also this assumption in future work.

[3] The boundedness of g' ensures that the expectations are always well-defined (integrable). This condition can be relaxed provided that the function g' is integrable. In the sequel, since we bound Ω before discretizing X, the boundedness condition is always met for g' equal to mean, variance etc.

where the last equality makes sense because the denominator is positive (16). Being $\int \ell(y|x)p(x)dx > 0$, it does not affect the optimization. Thus, by solving the equation w.r.t. μ, one gets:

$$\nu = \max_{p\in\mathcal{P}} \frac{\int g'(x)\ell(y|x)p(x)dx}{\int \ell(y|x)p(x)dx}$$

which is by definition $\overline{E}[g'|y]$. ■

Equation (17) is a particular case of the so called Generalized Bayes Rule (GBR) [9, Ch. 6]. It generalizes Bayes Rule to the case of set of distributions. From (17) it can be noticed that to compute the upper expectation we need: (i) to solve (3) or its dual (4) with $g(x) = (g'(x) - \nu)\ell(y|x)$ and, then, (ii) to find the value ν which solves the equation (17). An approximate solution of this problem can be obtained by first bounding Ω and then discretizing X on the bounded domain. In this way (4) becomes a linear programming problem that can be efficiently solved and the solution ν of the GBR can be obtained by applying any root-finding method (e.g., bisection).

In Section II, we have seen that, in the case that only two moments are known, the distribution that gives the prior upper expectation of any function g is at most a trimodal mixture of Dirac's deltas. Is this also true a-posteriori?

Theorem 5. *Given a real-valued bounded function g' of X, the upper posterior expectation of g' given the observation y is:*

$$\overline{E}[g'|y] = \max_{w_1,w_2,m_1,m_2,m_3} \int g'(x)p(x|y)dx, \qquad (18)$$

with

$$p(x|y) = \frac{\mathcal{N}(y;x,r)p(x)}{\int \mathcal{N}(y;x,r)p(x)dx}$$
$$= w_1^*\delta(x-m_1) + w_2^*\delta(x-m_2) + w_3^*\delta(x-m_3), \qquad (19)$$
$$w_i^* = \frac{w_i\mathcal{N}(y;m_i,r)}{\sum_{j=1}^3 w_j\mathcal{N}(y;m_j,r)}, \quad i=1,2,3, \qquad (20)$$
$$p(x) = w_1\delta(x-m_1) + w_2\delta(x-m_2) + w_3\delta(x-m_3),$$
$$w_3 = 1 - w_1 - w_2,$$
$$w_1m_1 + w_2m_2 + w_3m_3 = \mu_1,$$
$$w_1m_1^2 + w_2m_2^2 + w_3m_3^2 = \mu_2,$$

$w_i > 0$ *and* $m_i \in \mathbb{R}$ *for* $i = 1,2,3$. ■

Proof: Given the a-priori trimodal mixture of Dirac's deltas $p(x)$ and the likelihood $\mathcal{N}(y;x,r)$, the expression (19) is a direct consequence of the application of Bayes rule and properties of the Dirac's delta. The fact that also the upper posterior expectation of g is obtained by one of prior extreme PDFs follows by the lower envelope theorem [9, Sec. 6.4.2]. ■

From Theorem 5 it follows that the effect of updating is just that of changing the weights of the mixture of Dirac's deltas. Fig. 3 shows the lower and upper posterior expectation computed based on (17) for $g' = X$ (mean)[4] in the case $\mu_1 = 0$, $\mu_2 = r = 1$. It can be noticed that after updating the mean of X cannot be precisely known, i.e., we only know it belongs to an interval (a-priori it was precise as stated in Theorem 1 and in particular in (11)). This is due to the effect of the uncertainty in the measurement process combined with the imprecision in the knowledge of the prior distribution of X (a-priori we only know the first two moments of X).

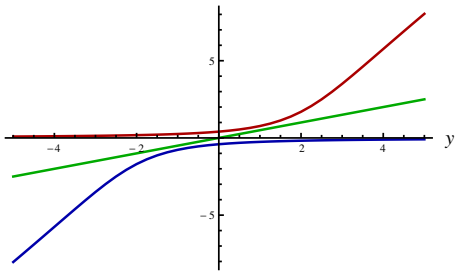

Fig. 3. Lower (blue) and upper (red) mean and KF estimate $y/2$ (green) as a function of the observation $y \in [-5,5]$ for the case $\mu_1 = 0$, $\mu_2 = r = 1$.

After updating we are only able to say that the mean of X belongs to the interval $[\underline{E}[X|y], \overline{E}[X|y]]$, that is the area between the red and blue line in Fig. 3. The KF estimate, which is the LMVE, is also reported for comparison. What is the meaning of the interval $[\underline{E}[X|y], \overline{E}[X|y]]$? In [10, Th.3], it has been proved that the interval of posterior means includes all the optimal Bayesian estimates (in the minimum-mean square error sense) that are compatible with our prior information expressed via a set distributions (in our case the set of all distributions with mean μ_1 and variance $\mu_2 - \mu_1^2$). In other words, since we cannot specify a unique prior distributions from the only knowledge of mean and variance, we cannot compute a unique optimal Bayesian estimate. In this case we are only able to say that the optimal estimate belongs to the interval $[\underline{E}[X|y], \overline{E}[X|y]]$.

Fig. 4 shows the lower and upper posterior CDF of X for the case $\mu_1 = 0$, $\mu_2 = r = 1$ and $y \in \{0, 1.5, 3\}$ and the CDF of the KF estimate (Normal CDF with mean $y/2$ and variance $1/2$). The dashed lines, which represent the a-priori CDF bounds in Fig. 1, have been reported for comparison. It can be noticed that, as one could expect, the CDF lower and upper bound gets closer when the observation y is in agreement with the prior information $E(X) = 0$ and $E(X^2) = 1$. At the increasing of y the bounds enlarges to account for conflict between data and prior.

It is also interesting to compare the 95% posterior credible interval computed using Chebyshev inequality (CI) (obtained from the posterior mean $\hat{x}$ and variance σ^2 of KF, i.e., $P(|X - \hat{x}| \leq \delta\sigma) \geq 1 - \frac{1}{\delta^2}$ with $1 - \frac{1}{\delta^2} = 0.95$), and the minimum length 95% posterior credible interval (MMPI) centred at the KF estimate $\hat{x}$, i.e., $|X - \hat{x}| \leq \hat{\eta}$, and computed as follows:

$$\hat{\eta} = \arg\min_{\eta} \underline{E}[I_{\{|X-\hat{x}|\leq\eta\}}] \geq 0.95$$

where $\underline{E}[I_{\{|X-\hat{x}|\leq\eta\}}] = -\overline{E}[-I_{\{|X-\hat{x}|\leq\eta\}}]$ and $\overline{E}$ is computed based on (17) with $g' = -I_{\{|X-\hat{x}|\leq\eta\}}$. This the lowest interval that has probability of at least 0.95 of including the true value x of the variable X.[5]

[4]The mean is not a real-valued bounded function but it is integrable and thus (17) still holds.

[5]This is the equivalent of the Bayesian Highest Posterior Density credible set.

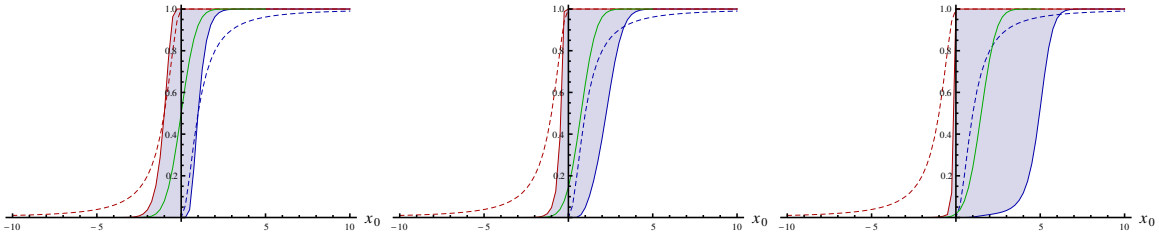

Fig. 4. Lower and upper bounds for the posterior CDF (bounds of the filled area) of X in the case $\mu_1 = 0$, $\mu_2 = r = 1$ and $y \in \{0, 1.5, 3\}$ (from left to right). The central plot is the CDF of the KF estimate (Normal CDF with mean $y/2$ and variance $1/2$). The dashed lines refer to the a-priori CDF bounds.

Fig. 5 shows these two intervals computed for different values of the observation y, $\mu_1 = 0$, $\mu_2 = r = 1$ (observe that the KF estimate/variance are $\hat{x} = y/2$ and $\sigma^2 = 1/2$ in this case). It can be noticed that CI based interval has a constant size, while MMPI size increases with y (remember from Section II that they coincided a-priori). The centres of both intervals moves according to $\hat{x} = y/2$. When $y = 0$ the true value of x is with probability 0.95 between $[-3.16, 3.16]$ based on CI and between $[-2, 2]$ based on MMPI. Thus, when $y = 0$ also if we consider the worst-case prior distribution compatible with the two moments information, we are sure that the true value x is inside $[-2, 2]$ with at least 0.95 probability. This is what MMPI says, while CI heavily overestimates the size of the 0.95 probability interval in this case. MMPI is smaller and, thus, less conservative for all value of the observations $y < 2$. MMPI is able to correctly compute the posterior interval by combining the prior information $\mu_1 = 0$ and $\mu_2 = 1$ and likelihood model $\mathcal{N}(y; x, r)$ through Bayes' rule (GBR), while CI based interval is computed by simply using the two posterior moments computed by KF (which is only the LMVE in this setting). In practice the posterior interval based on CI is assuming the model (13) with mean equal to the KF estimate $\mu_1 = y/2$ and variance equal to the KF variance $\mu_2 = \sigma^2 = 1/2$, while MMPI uses the posterior model in Theorem 5 obtained by correctly updating the prior information using Bayes' rule. At the increasing of y (for $y > 2$) MMPI size increases to account for the conflict between prior information and data as in Fig. 4. Thus, the approach proposed in this paper allows to compute the correct posterior interval for any point estimate (e.g., KF estimate in this example) and provides a tighter bound than CI.

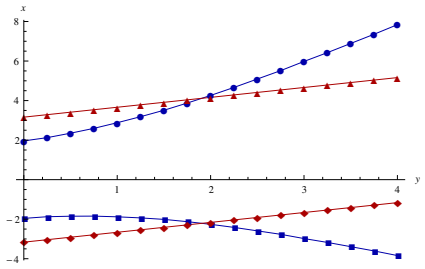

Fig. 5. 95% Chebyshev inequality from KF estimate/variance (red) and minimum length 95% posterior credible interval (blue).

V. Prediction

From the relationships in (15), it follows that $X_{t+1} = aX_t + v$ where v is a zero-mean noise with variance q. This implies that $E[X_{t+1}] = aE[X_t]$ and $E[X_{t+1}^2] = a^2 E[X_t^2] + q$.

Theorem 6. *If the variable X_t has precise first two moments, i.e., $\underline{E}[X_t] = \overline{E}[X_t] = E[X_t]$ and $\underline{E}[X_t^2] = \overline{E}[X_t^2] = E[X_t^2]$ then, given a real-valued bounded function g' of X_{t+1}, the upper expectation of g' can be obtained by solving (3) with $\mu_1 = aE[X_t]$ and $\mu_2 = a^2 E[X_t^2] + q$.* ■

Proof: This result follows directly from (3) with $\mu_1 = aE[X_t]$ and $\mu_2 = a^2 E[X_t^2] + q$. ■

Consider for instance the case $t = 0$. Since $E[X_0] = \hat{x}_0$ and $E[X_0^2] = \hat{x}_0^2 + p_0$, it follows that

$$E[X_1] = a\hat{x}_0, \quad E[X_1^2] = a^2\hat{x}_0^2 + a^2 p_0 + q.$$

These are the equations that describe the propagation of mean and variance through a linear system.

Assume now we observe y_1 and we update the probabilistic information on X_1 as described in Section IV. After updating, the posterior mean $E[X_1|y_1]$ is not precisely known. In fact, because of the prior imprecision in the probabilistic knowledge of X_1 (only the first two moments of X_1 are known), we can only determine lower and upper bounds of $E[X_1|y_1]$ as described in Section IV. Hence, the simple result of Theorem 6 cannot be used. In this case, we must follow a slightly more tortuous path.

Theorem 7. *Consider the upper expectation $\overline{E}_{X_1}$ obtained from (3) with $\mu_1 = a\hat{x}_0$ and $\mu_2 = (a\hat{x}_0)^2 + a^2 p_0 + q$ and the conditional upper expectation $\overline{E}_{X_2}[\cdot|x_1]$ obtained from (3) with $\mu_1 = ax_1$ and $\mu_2 = (ax_1)^2 + q$ for each value x_1 of the conditional variable X_1. Assuming that*

$$\underline{E}_{X_1}\left[\underline{E}_{X_2}\left[\ell(y_1|X_1)\middle|X_1\right]\right] > 0 \tag{21}$$

the upper posterior expectation of a bounded real-valued function g'' of X_2 given the observation y_1 at time 1 can be found by finding the unique solution $\nu \in \mathbb{R}$ of:

$$0 = \overline{E}_{X_1}\left[\overline{E}_{X_2}\left[(g'' - \nu)\ell(y_1|X_1)\middle|X_1\right]\right]. \tag{22}$$

■

Proof:

$$\begin{aligned} 0 &= \overline{E}_{X_1}\left[\overline{E}_{X_2}\left[(g'' - \nu)\ell(y_1|X_1)\middle|X_1\right]\right] \\ &= \overline{E}_{X_1}\left[\ell(y_1|X_1)\overline{E}_{X_2}\left[(g'' - \nu)\middle|X_1\right]\right] \\ &= \overline{E}_{X_1}\left[\ell(y_1|X_1)\left(\overline{E}_{X_2}\left[g''\middle|X_1\right] - \nu\right)\right] \end{aligned}$$

Thus, from Theorem (4) with $g' = \overline{E}_{X_2}[g''|X_1]$, it follows that

$$\nu = \overline{E}_{X_2}[g''|y_1] = \overline{E}_{X_1}[\overline{E}_{X_2}[g''|X_1]|y_1] \tag{23}$$

which by definition is the upper posterior expectation of the function g'' of X_2 given y_1. ■

From (23) it follows that to compute the prediction we must solve two nested optimization problems. Since we know that the PDF $p \in \mathcal{P}$ which gives the upper expectation in each one of the above optimization problem is a mixture of Dirac's deltas with at most $m+1$ components then the upper expectation (23) will be given by a mixture of of Dirac's deltas with at most $(m+1) \times (m+1) = (m+1)^2$ components.

This means that the number of components of the mixture of Dirac's deltas that gives the lower and upper posterior expectation increases exponentially with time (increases as $(m+1)^t$). Therefore, differently from the Bayesian solution to state estimation, the propagation of the posterior model through time is infeasible in practice because the number of extreme distributions that characterizes the posterior set of distributions increases exponentially with time, see [11].

VI. Efficient solution

Section V has shown that an exact recursive solution of the filtering problem is infeasible in practice. In fact, because of the prediction step, the number of extreme points (Dirac's deltas) of the set of posterior densities, which characterizes our imprecise information on the state X_t, increases exponentially with time. So we cannot characterize our information on the state given all the past measurements by means of the posterior set of distributions.

However, an efficient solution of the filtering problem, whose complexity increases only linearly in time, can still be computed by following a different path. Instead of aiming to compute the set of posterior distributions (i.e., by propagating in time the extreme points of this set), one can only propagate the lower and upper expectation of functions of interest g of X_t (e.g., mean, credible interval, etc.). In [3], by following this approach, a solution of the filtering problem has been derived which essentially consists of propagating in time both the lower and upper state expectations over the set of assumed probability distributions. This general solution has a structure that resembles the standard Bayesian solution to state estimation and, in fact, it reduces to it in the case the sets of probability distributions for initial state, measurement equation and state dynamics collapse to single distributions and, thus, the lower and upper expectation functionals coincide (no imprecision). Here we report the main result [3, Th. 2] in the particular case where the likelihood model is precise, i.e., the expectation of any function h of the measurement variable Y_k is $\underline{E}_{Y_k}[h|x_k] = \overline{E}_{Y_k}[h|x_k] = E_{Y_k}[h|X_k] = \int h(y_k)\mathcal{N}(y_k; x_k, r)dy_k$, which is the case considered in this paper (see section III)

Theorem 8. *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, $E_{Y_k}[h|X_k] = \int h(y_k)\mathcal{N}(y_k; x_k, r)dy_k$, 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*

$$\overline{E}_{X_k}[h_1|x^{k-1}y^{k-1}] = \overline{E}_{X_k}[h_1|x_{k-1}], \tag{24}$$

$$E_{Y_k}[h_2|x^k, y^{k-1}] = E_{Y_k}[h_2|x_k], \tag{25}$$

for any bounded scalar functions $h_1 : \mathcal{X}^k \times \mathcal{Y}^{k-1} \to \mathbb{R}$, $h_2 : \mathcal{X}^k \times \mathcal{Y}^k \to \mathbb{R}$ and given x^k, y^{k-1}. Then, given the sequence of measurements $y^t = \{y_1, y_2, \ldots, y_t\}$, the posterior upper expectation $\overline{E}_{X_t}[g|y^t]$ for any bounded scalar function $g : \mathcal{X} \to \mathbb{R}$ is equal to the unique value $\nu \in \mathbb{R}$ that satisfies the following equation:

$$0 = \overline{E}_{X^t}\Big[(g-\nu)\textstyle\prod_{i=1}^n \mathcal{N}(y_k; X_k, r)\Big], \tag{26}$$

where the above joint upper expectation is given by:

$$\overline{E}_{X_0}\Big[\overline{E}_{X_1}\Big[\ldots\overline{E}_{X_t}\Big[(g-\nu)\prod_{i=1}^{n}\mathcal{N}(y_k; X_k, r)\Big|X_{t-1}\Big]\ldots\Big|X_0\Big]\Big]. \tag{27}$$

■

The proof of Theorem 8 can be found in [3, Th. 2].[6] Equation (26) is obtained by applying the GBR to the joint upper expectation of all the state sequence, that means applying the results of Theorems 4 and 7 to the joint upper expectation of all the state sequence. Intuitively, the result follows straightforwardly from the Bayesian solution of the filtering problem by replacing standard expectations with upper expectations. Following this analogy, the conditions in (24)–(25) can be understood as the generalization of the Markov conditions of independence assumed in Bayesian filtering.

It is worth to point out that to compute $\overline{E}_{X_t}[g|\tilde{y}^t]$ in the imprecise case, we cannot in general derive a recursive solution as in the Bayesian case, see also [3, Sec. 4] for more comments about this point. In other words, to compute $\overline{E}_{X_t}[g|\tilde{y}^t]$ at any time t it is necessary to go through the joint and to find the value of ν which solves (27). This means that the computational complexity to compute $\overline{E}_{X_t}[g|\tilde{y}^t]$ increases with time but only linearly [12]. Hereafter, we describe an algorithm that allows to solve (27) in case the upper expectation models for initial state and state dynamics are given by (3) with $m = 2$ and the two moments information given in (14)–(15). To solve (4), we first bound the support Ω and then discretize X so that (4) becomes a linear program problem.

1) For each $k = 0, \ldots, t$ discretize X_k by generating n points equally spaced in $\Omega' = [x_{min}, x_{max}]$.
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 discretized value x_{k-1}^j of X_{k-1}, solve

$$g(x_{k-1}^j, k-1, \nu) = \max_z z^T q$$
$$s.t.\ z^T f(x_k^i) - g(x_k^i, k, \nu)\mathcal{N}(y_k; x_k^i, r) \geq 0, \quad \forall x_k^i \in \Omega',$$

[6] Observe that the proof in [3] has been obtained by assuming that the observation variables are discretized. Intuitively, we can see Theorem 8 as the limit of this result when the size of the discretization interval goes to zero.

where $q = [1, \mu_1, \mu_2]^T$, $f(x) = [1, x, x^2]^T$, $\mu_1 = ax_{k-1}^j$ and $\mu_2 = (ax_{k-1}^j)^2 + q$.

4) Solve:

$$\begin{aligned} res = \quad & \max_{z} z^T q \\ s.t. \quad & z^T f(x_0^i) - g(x_0^i, 0, \nu) \geq 0, \quad \forall x_0^i \in \Omega', \end{aligned}$$

where $q = [1, \mu_1, \mu_2]^T$, $f(x) = [1, x, x^2]^T$, $\mu_1 = \hat{x}_0$ and $\mu_2 = \hat{x}_0^2 + p_0$.

5) If $res = 0$ return ν, otherwise change ν based on some root-finding algorithm (e.g., bisection) and back to 2).

In the sequent simulations, we have set $x_{min} = -15$, $x_{max} = 15$ and $n = 350$.[7] We have implemented the above algorithm in matlab-tomlab by solving the maximization problem by using `cplex` algorithm and the root-finding problem using `fminbnd`.

VII. Extension to the nonlinear case

In the previous sections we have assumed that the measurement model and the state dynamic in (15) are linear and time invariant. However, our model does not require these assumptions, all the derivations above can be extended to the time variant nonlinear case by simply replacing (15) with

$$E[X_{t+1}|X_t] = f(X_t, t), \quad E[X_{t+1}^2|X_t] = (f(X_t, t))^2 + q, \tag{28}$$

and the measurement model $\mathcal{N}(y_t; x_t, r)$ with $\mathcal{N}(y_t; h(x_t, t), r)$, where $f(X_t, t)$ and $h(X_t, t)$ are nonlinear functions of the state.

VIII. Numerical example

We have performed Monte Carlo simulations in order to show the basic features of the proposed approach: hereafter called moment-based filter (MBF). These simulations compare the performance of the MBF with the KF and the optimal unknown Bayesian filter (i.e., the filter that knows the true distribution of initial state and process noise).

A. Linear case

The one-dimensional model

$$x(t+1) = ax(t) + w(t), \quad y(t) = x(t) + v(t),$$

has been considered for $t > 0$ with initial state $x(0) = \hat{x}_0 + w(0)$, where $a = 0.7$, $v(t) \sim N(0, r)$, $w(0)$ is the noise on the initial state and $w(t)$ for $t > 0$ is the process noise. We assume that the modeller only knows that $w(0)$ and $w(t)$ for $t > 0$ have zero mean and variance p_0 and, respectively, q. The system model is thus, non-Gaussian (the true distributions of $w(t)$ for $t \geq 0$ are unknown) non-stationary but with time-invariant first two moments. In particular, a trajectory of 8 timesteps and a Monte Carlo size of 230 runs have been considered and the following distribution for $w(t)$ considered: the true state at time t has been generated by the mixture of Dirac given in Theorem 1 with $\mu_1 = ax(t-1)$ and $\mu_2 = (ax(t-1))^2 + q$ for $t > 1$ and $\mu_1 = \hat{x}_0$ and $\mu_2 = \hat{x}_0^2 + p_0$ at $t = 0$ and in both cases selecting the value of m_2 so to satisfy the constraint $w_1(m_2) = 0.8$, see left plot in Fig. 2. Note that, in all simulations, both the MBF and the KF were designed without assuming the knowledge of the true distributions of the noises $w(t)$ for $t \geq 0$. However, for performance comparison, we have computed also the optimal Bayesian estimate (OPF) obtained by a particle filter (2500 particles) based on the true unknown distributions of $w(t)$ (i.e., the bimodal mixture of Diracs).

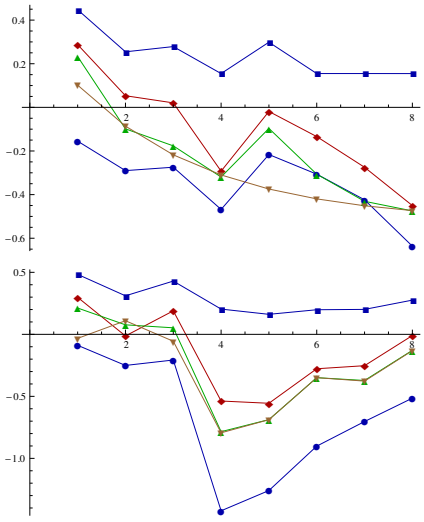

Fig. 6. Trajectory (brown), KF estimate (red), OPF estimate (green), MBF lower and upper posterior means (blue).

Two instances of trajectories generated with this model are shown in Fig. 6 together with the KF and OPF estimates and the lower and upper posterior means computed based on MBF. It can be noticed that the interval delimited by the lower and upper posterior means includes the KF and OPF estimates at each time instant. This happens in all the simulations. In fact, by definition, the interval $[\underline{E}[X_t|y^t], \overline{E}[X_t|y^t]]$ includes all the Bayesian estimates that can be obtained from the set of distributions compatible with the two moments information. Observe that, both the two distributions shown in the left plot in Fig. 2 (Gaussian and Diracs' mixture) belong to this set and, thus, the KF and OPF estimates must be contained in this interval. Conversely, the true trajectory does not have to be contained in this interval. $[\underline{E}[X_t|y^t], \overline{E}[X_t|y^t]]$ is not a credible interval but it is the interval that includes all the optimal Bayesian estimates. Notice also that, in the unknown distribution setting we cannot compute the optimal Bayesian estimator and, in this case, KF is only the LMVE. However, we can use MBF to compute bounds for the optimal Bayes estimate. These bounds are exactly the lower and upper posterior means.

The 95% posterior credible interval based on MBF and Chebyshev-inequality (CI) interval computed using the KF posterior mean and variance are shown in Fig. 7 for the second case in Fig. 6. As discussed in Section IV it can be noticed

[7] However, noticed that to increase the accuracy of the solution we can iteratively refine the discretization step in a neighbour of the previous computed solution.

that MBF based bound is in general tighter than CI for the KF estimate.

From Fig. 6, it seems that there is some correlation between the KF estimate and the lower and upper means. We have computed the statistical correlation between: (a) the difference $\overline{E}[X_t|y^t] - \underline{E}[X_t|y^t]$ and KF innovation; (b) the difference $\overline{E}[X_t|y^t] - \underline{E}[X_t|y^t]$ and the absolute distance between KF and OPF estimates. The correlation averaged over the 230 MC runs is 0.74 for (a) and 0.21 for (b). This means that in the MBF filter the imprecision increases at the increasing of KF inadequacy (innovation) and, thus, MBF is an indicator of KF inadequacy and that the MBF filter is also an indicator of the distance between the LMVE and the optimal unknown Bayes estimate.

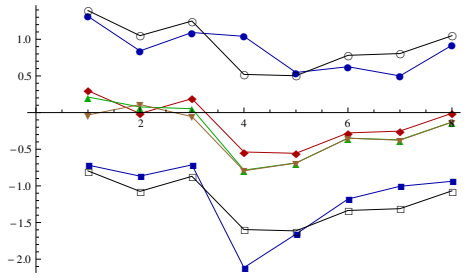

Fig. 7. Trajectory (brown), KF estimate (red), OPF estimate (green), MBF lower and upper limits of the credible interval (blue) and Chebyshev inequality based interval (black).

B. Nonlinear case

The following nonlinear model

$$\begin{cases} x(t+1) &= 0.5x(t) + 2\frac{x(t)}{1+x(t)^2} + w(t), \\ y(t) &= \frac{x(t)^2}{20} + v(t), \end{cases}$$

has been considered with the same noise models as in the linear case but with $r = 0.01$. This model is often considered as benchmark for nonlinear filtering, for the fact that the likelihood model is always symmetric w.r.t zero (and bimodal when y_t is positive). In this nonlinear setting there is not an optimal linear estimator and since we do not know the true distribution of the noises we cannot apply particle filtering but we can apply the Extended KF (EKF) which has not any guarantee of optimality in this case. Fig. 8 shows the 95% posterior credible interval based on MBF and Chebyshev-inequality (CI) interval computed using the EKF posterior mean and variance for one of the several instances where EKF almost diverges. In this case, the posterior credible interval for MBF has not been centred on the EKF estimate (it is unreliable in the nonlinear case) but on the middle point of the interval delimited by the lower and upper posterior mean. From the figure it can be noticed that the true trajectory and the OPF estimate are close and both contained in the 95% posterior credible interval based on MBF. Conversely, the EKF estimate is far from the true trajectory and the Chebyshev-inequality interval computed using the EKF posterior mean and variance diverges in the first instants and then does not include the true trajectory.

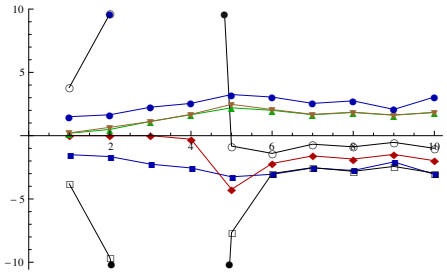

Fig. 8. Trajectory (brown), EKF estimate (red), OPF estimate (green), MBF lower and upper limits of the credible interval (blue) and Chebyshev inequality based interval (black).

IX. Conclusions

By exploiting filtering based on set of distributions we have solved the univariate filtering problem when only the first two moments of the noise terms are known and without introducing additional assumptions on the distributions of the noise terms (e.g., Gaussianity) or on the final form of the estimator (e.g., linear estimator). As future work we plan to extend this work to the multivariate case also considering the case in which more than two moments are known.

Acknowledgements

Alessio Benavoli acknowledges financial support from the Swiss NSF grant n. 200020-137680/1

References

[1] R. Kaas and M. Goovaerts, "Application of the problem of moments to various insurance problems in non-life," in *Proc. NATO ASI Insurance and Risk Theory*, (Maratea (It)), pp. 79–118, Reidel Publishing Company, 1986.

[2] J. Spall, "The Kantorovich inequality for error analysis of the Kalman filter with unknown noise distributions," *Automatica J. IFAC*, vol. 10, pp. 1513–1517, 1995.

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

[5] S. Karlin and W. Studden, *Tchebycheff systems: with applications in analysis and statistics*, vol. 376. Interscience Publishers New York, 1966.

[6] N. Akhiezer and M. Krein, *Some questions in the theory of moments*, vol. 2. Amer Mathematical Society, 1962.

[7] A. Shapiro, "On duality theory of conic linear problems, chapter 7," in *in Semi-Infinite Programming Recent Advances*, pp. 135–165, 2001.

[8] W. Rogosinski, "Moments of non-negative mass," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 245, no. 1240, p. 1, 1958.

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

[10] A. Benavoli and M. Zaffalon, "Density-ratio robustness in dynamic state estimation," *IDSIA, Technical Report 02 - 12*, 2012. http://www.idsia.ch/idsiareport/IDSIA-02-12.pdf.

[11] B. Noack, V. Klumpp, D. Brunn, and U. D. Hanebeck, "Nonlinear bayesian estimation with convex sets of probability densities," in *Proceedings of the 11th International Conference on Information Fusion (Fusion 2008)*, (Cologne, Germany), pp. 1–8, July 2008.

[12] G. De Cooman, F. Hermans, A. Antonucci, and M. Zaffalon, "Epistemic irrelevance in credal nets: the case of imprecise markov trees," *International Journal of Approximate Reasoning*, vol. 51, no. 9, pp. 1029–1052, 2010.