51st IEEE Conference on Decision and Control
December 10-13, 2012. Maui, Hawaii, USA

A Probabilistic Approach to Optimal Estimation Part I: Problem Formulation and Methodology

Fabrizio Dabbene, Mario Sznaier, and Roberto Tempo

Abstract— **The classical approach to system identification is based on statistical assumptions about the measurement error and provides estimates that have stochastic nature. Worst-case identification, on the other hand, only assumes the knowledge of deterministic error bounds and provides guaranteed estimates.**

The focal point of this paper is to provide a rapproachement between these two paradigms and propose a novel probabilistic framework for system identification. The main idea in this line of research is to "discard" sets of measure at most ϵ, where ϵ is a probabilistic accuracy, from the set of deterministic estimates. Therefore, we are decreasing the so-called worst-case radius of information at the expense of a given probabilistic "risk."

The main results of the paper establish rigorous theoretical properties of a trade-off curve, called *optimal violation function*, which shows how the radius of information decreases as a function of the accuracy. In the companion paper [8], we develop algorithms (randomized and deterministic) which exploit these theoretical results for efficiently computing the optimal violation function.

Keywords: System identification, optimal algorithms, uncertain systems

I. Introduction and Preliminaries

In the last decades, several authors focused their attention on the so-called set-membership identification which has the objective to compute so-called optimal estimators, as well as hard bounds on the estimation error, see for instance [19], and [15]. Set-membership identification has been formulated, see e.g. [18], [26], [14], [23], in the general framework of information-based complexity (IBC), which is a theory developed in the computer science area for studying the complexity of problems approximately solved due to the presence of partial and/or contaminated information; see [29] and [30]. Classical applications of IBC include distributed computations, clock synchronization and computer vision.

In the worst-case setting, the noise is a deterministic variable bounded within a set of given radius. The objective is to derive optimal algorithms which minimize (with respect to the noise) the maximal distance between the true-but-unknown system parameters and their estimates. The drawback of the hard bound approach is that these bounds on the estimation errors may be too large in many instances, in particular when the goal is to use system identification in the context of closed-loop control. On the other hand, the mainstream stochastic approach to system identification, see [16] and the special issues [17], [25], assumes that the available observations are contaminated by *random* noise, and has the goal to derive soft bounds on the estimation errors.

The focal point of this paper is to study a *rapprochement* between these two settings, see [21], [13], [22], [5], [11] for earlier work in this direction. That is, the measurement noise is confined within a given set (as in the worst-case setting), but it is also a random variable with given probability distribution (so that statistical information is used). We recall that this rapprochement has been also studied in the context of control design in the presence of uncertainty, see [27], [4], [3], and [2] which provide a rigorous methodology for deriving controllers satisfying the desired performance specifications with high level of probability.

The specific problem formulation we consider in this paper is the probabilistic setting of IBC. The objective is to compute the so-called *probabilistic radius of information*, and the related probabilistic optimal estimate, when the noise is uniformly distributed. We remark that, contrary to the statistical setting which mainly concentrates on asymptotic results, the probabilistic radius introduced in this paper provides a quantification of the estimation error based on a finite number of observations. In this sense, this approach has close relations with the works [6], [7], where noise-free non-asymptotic confidence sets for the estimates are derived. Furthermore, the paper is also related to the work [28], where a probabilistic density function over the consistency set is considered.

We now provide a brief overview of information-based complexity, see II for formal definitions (this section also contains an illustrative example dealing with system parameter identification). We are interested in computing an optimal approximation of $\mathcal{S}x \in Z$ where $\mathcal{S}$ is a given linear mapping $\mathcal{S} : X \to Z \subseteq \mathbb{R}^s$, and $x \in X \subseteq \mathbb{R}^n$; x and $\mathcal{S}$ are called, respectively, problem element and solution operator (that is, $\mathcal{S}x$ represents a linear combination of the unknown parameters of the system to be identified). The element x is not exactly known. Rather, only approximate information $y = \mathcal{I}x + q$ is available, where $\mathcal{I}$, the so-called information operator, is linear, and the noise q is confined within a norm-bounded set $\mathcal{Q} \subset \mathbb{R}^m$. An approximation to $\mathcal{S}x$ is provided by an algorithm (or estimator) $\mathcal{A}$, generally nonlinear, acting on the information y. Optimal algorithms minimize the maximal distance between the true-but-unknown solution $\mathcal{S}x$ and the estimated solution $\mathcal{A}(y)$ for the worst-case noise $q \in \mathcal{Q}$. The error of an optimal algorithm is called the worst-case radius of information. Section II also contains

F. Dabbene and R. Tempo are with the CNR-IEIIT Institute, Politecnico di Torino, Italy (e-mail: fabrizio.dabbene@polito.it, roberto.tempo@polito.it).

M. Sznaier is with Northeastern University, Boston, MA 02115, USA (e-mail: msznaier@ece.neu.edu).

978-1-4673-2066-5/12/$31.00 ©2012 IEEE

the formal definition of the consistency set $\mathcal{I}^{-1}(y)$ which plays a major role in this paper. Roughly speaking, this is the set of all parameters x which are compatible with the given data y, the model $y = \mathcal{I}x + q$ and the noise $q \in \mathcal{Q}$.

In Section III we introduce the probabilistic setting where we "discard" sets of (probabilistic) measure at most ϵ from the consistency set with the objective to decrease the worst-case radius. Therefore, we obtain a new error which represents the probabilistic radius of information. This approach may be very useful, for example, for system identification in the presence of outliers [1], where "bad measurements" may be discarded.

In Section IV we present the main results of the paper for uniformly distributed noise. Theorem 1 shows that the induced measure (through the inverse of the operator information $\mathcal{I}$) over the consistency set $\mathcal{I}^{-1}(y)$ is uniform and the induced measure of the set $\mathcal{S}\mathcal{I}^{-1}(y)$ (which is the transformation of the consistency set $\mathcal{I}^{-1}(y)$ through the solution operator $\mathcal{S}$) is log-concave. Theorem 2 proves crucial properties, from the computational point of view, of the so-called optimal violation function $v_o(r)$, which shows how the risk ϵ decreases as a function of the radius r. In particular, this result shows that $v_o(r)$ is non-increasing, and for fixed $r > 0$, it can be obtained as the maximization of a specially constructed unimodal function.

The companion paper [8] presents (deterministic and randomized) algorithms for computing the optimal violation function. In this paper, the properties of the violation function are exploited to derive deterministic and randomized relaxations by means of low complexity algorithms. In the same paper, the effectiveness of the proposed methodology is demonstrated using an application example of parameter estimation of an FIR system.

A. Notation

We write $\|\cdot\|$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ to denote the ℓ_p, ℓ_2 and ℓ_∞ norms, respectively. The ℓ_p norm-ball of center ξ_c and radius r is denoted by

$$\mathcal{B}(\xi_c, r) \doteq \{\xi \,|\, \|\xi - \xi_c\| \leq r\},$$

and we write $\mathcal{B}(r) \doteq \mathcal{B}(0, r)$. We denote by $\mathcal{B}_2(\xi_c, r)$ and by $\mathcal{B}_\infty(\xi_c, r)$ the ℓ_2 and ℓ_∞ norm-balls, respectively. We use the notation $x \sim p_A$ to indicate that the random vector x has probability density function (pdf) $p_A(x)$ with support set A. The *uniform* density $\mathcal{U}_A$ over the set $A \subset \mathbb{R}^n$ is defined as

$$\mathcal{U}_A(x) \doteq \begin{cases} 1/\mathrm{vol}\,[A] & \text{if } x \in A; \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{vol}\,[A]$ represents the Lebesgue measure (volume) of the set A, see [12] for details regarding volumes of sets. The uniform density $\mathcal{U}_A$ generates a uniform measure $\mu_{\mathcal{U}(A)}$ such that, for any measurable set B, $\mu_{\mathcal{U}(A)}(B) = \mathrm{vol}\,[B \cap A]\,/\mathrm{vol}\,[A]$. The $n \times n$ identity matrix is indicated by I_n. A set H is said to be *centrally symmetric* with center $\bar{x}$ if $x \in H$ implies that its reflection with respect to $\bar{x}$ also belongs to H, i.e. $(2\bar{x} - x) \in H$.

II. Information-Based Complexity

This section provides a formal overview of the information-based complexity definitions used in the paper and an introductory example illustrating the main concepts in a system identification context. An introduction to the IBC framework is given in [30] and in [23]; see the monograph [29] for an advanced treatment of the topic.

Let X be a linear normed n-dimensional space over the real field, which represents the set of (unknown) problem elements $x \in X$. Define a linear operator $\mathcal{I}$, called *information operator*, which maps X into a linear normed m-dimensional space Y

$$\mathcal{I} : X \to Y.$$

In general, exact information about the problem element $x \in X$ is not available and only perturbed information, or *data*, $y \in Y$ is given. That is, we have

$$y = \mathcal{I}x + q \tag{1}$$

where q represents *uncertainty* which may be deterministic or random. In either case, we assume that $q \in \mathcal{Q}$, where $\mathcal{Q} \subseteq \mathbb{R}^m$ is a bounding set. Due to the presence of uncertainty q, the problem element $x \in X$ may not be easily recovered from the data $y \in Y$. Then, we introduce a linear operator $\mathcal{S}$, called a *solution operator*, which maps X into Z

$$\mathcal{S} : X \to Z$$

where Z is a linear normed s-dimensional space over the real field, where $s \leq n$. Given $\mathcal{S}$, our aim is to estimate an element $\mathcal{S}x \in Z$ knowing the corrupted information $y \in Y$ about the problem element $x \in X$.

An *algorithm* $\mathcal{A}$ is a mapping (in general nonlinear) from Y into Z, i.e.

$$\mathcal{A} : Y \to Z.$$

An algorithm provides an approximation $\mathcal{A}(y)$ of $\mathcal{S}x$ using the available information $y \in Y$ of $x \in X$. The outcome of such an algorithm is called an *estimator* $z = \mathcal{A}(y)$.

We now introduce a set which plays a key role in the subsequent definitions of radius of information and optimal algorithm. Given data $y \in Y$, we define the *consistency set* as follows

$$\mathcal{I}^{-1}(y) \doteq \{x \in X \,|\, \text{ there exists } q \in \mathcal{Q} : y = \mathcal{I}x + q\} \tag{2}$$

which represents the set of all problem elements $x \in X$ compatible with (i.e. not invalidated by) $\mathcal{I}x$, uncertainty q and bounding set $\mathcal{Q}$. Note that, under the sufficient information assumption stated next, the set $\mathcal{I}^{-1}(y)$ is guaranteed to be bounded. For the sake of simplicity, we assume that the three sets X, Y, Z are equipped by the same ℓ_p norm, and that the set $\mathcal{Q}$ is an ℓ_p norm-ball of radius ρ, that is $\mathcal{Q} \equiv \mathcal{B}(\rho)$. Note that in this case the set $\mathcal{I}^{-1}(y)$ can be written as

$$\mathcal{I}^{-1}(y) = \{x \in X \,|\, \|\mathcal{I}x - y\| \leq \rho\}. \tag{3}$$

The following assumption regarding the operators $\mathcal{I}$ and $\mathcal{S}$ is now introduced.

Assumption 1 (Sufficient information and feasibility): We assume that the information operator $\mathcal{I}$ is a one-to-one mapping, i.e. $m \geq n$ and $\operatorname{rank} \mathcal{I} = n$. Similarly, $n \geq s$ and $\mathcal{S}$ is full row rank. Moreover, we assume that the set $\mathcal{I}^{-1}(y)$ has non-empty interior.

Note that, in a system identification context, the assumption on $\mathcal{I}$ and on the consistency set represent necessary conditions for identifiability of the problem element $x \in X$. The assumption of full-rank $\mathcal{S}$ is equivalent to assuming that the elements of the vector $z = \mathcal{S}x$ are linearly independent (otherwise, one could always estimate a linearly independent set and use it to reconstruct the rest of the vector z). We now provide an illustrative example showing their role in the context of system identification.

Example 1 (System parameter identification): Consider a parameter identification problem which has the objective to identify a linear system from noisy measurements. In this case, the problem elements are represented by the trajectory $\xi = \xi(t, x)$ of a dynamic system, parametrized by some unknown parameter vector $x \in X$. The system trajectory may for instance be represented as follows

$$\xi(t, x) = \sum_{i=1}^{n} x_i \psi_i(t) = \Psi^\top(t) x,$$

with given basis functions $\psi_i(t)$, and $\Psi^\top(t) \doteq [\psi_1(t) \quad \cdots \quad \psi_n(t)]$. We suppose that m noisy measurements of $\xi(t, x)$ are available for $t_1 < t_2 < \cdots < t_m$, that is

$$y = \mathcal{I}x + q = [\Psi(t_1) \quad \cdots \quad \Psi(t_m)]^\top x + q. \tag{4}$$

In this context, we usually assume unknown but bounded errors, such that $|q_i| \leq \rho$, $i = 1, \ldots, m$, that is $\mathcal{Q} = \mathcal{B}_\infty(\rho)$. Then, the aim is to obtain a parameter estimate using the measured data y. Hence, the solution operator is given by the identity,

$$\mathcal{S}x = x$$

and $Z \equiv X$. The consistency set is sometimes referred to as feasible parameters set, and is given as follows

$$\mathcal{I}^{-1}(y) = \left\{ x \in X : \|y - [\Psi(t_1) \quad \cdots \quad \Psi(t_m)]^\top x\|_\infty \leq \rho \right\}. \tag{5}$$

In the case of time series prediction, we are interested on predicting s future values of the function $\xi(t, x)$ based on m past measurements, and the solution operator takes the form

$$z = \mathcal{S}x = \xi(t_{m+1}, x) = \Psi^\top(t_{m+1}) x$$

for a one-step prediction, and by

$$\begin{aligned} z = \mathcal{S}x &= \{\xi(t_{m+1}, x), \quad \cdots \quad , \xi(t_{m+s}, x)\} \\ &= [\Psi(t_{m+1}) \quad \cdots \quad \Psi(t_{m+s})]^\top x, \end{aligned}$$

for a s steps prediction. ◇

Next, we define approximation errors and optimal algorithms when q is deterministic or random.

A. Worst-Case Setting

Given data $y \in Y$, we define the worst-case error $r^{\mathrm{wc}}(\mathcal{A}, y)$ of the algorithm $\mathcal{A}$ as

$$r^{\mathrm{wc}}(\mathcal{A}, y) \doteq \max_{x \in \mathcal{I}^{-1}(y)} \|\mathcal{S}x - \mathcal{A}(y)\|. \tag{6}$$

This error is based on the available information $y \in Y$ about the problem element $x \in X$ and it measures the approximation error between Sx and $A(y)$. An algorithm $\mathcal{A}_{\mathrm{o}}^{\mathrm{wc}}$ is called *worst-case optimal* if it minimizes $r^{\mathrm{wc}}(\mathcal{A}, y)$ for any $y \in Y$. That is, given data $y \in Y$, we have

$$r_{\mathrm{o}}^{\mathrm{wc}}(y) \doteq r^{\mathrm{wc}}(\mathcal{A}_{\mathrm{o}}^{\mathrm{wc}}, y) \doteq \inf_{\mathcal{A}} r^{\mathrm{wc}}(\mathcal{A}, y). \tag{7}$$

The minimal error $r_{\mathrm{o}}^{\mathrm{wc}}(y)$ is called the *worst-case radius of information*[1].

This optimality criterion is meaningful in estimation problems as it ensures the smallest approximation error between the actual (unknown) solution $\mathcal{S}x$ and its estimate $\mathcal{A}(y)$ for the worst element $x \in \mathcal{I}^{-1}(y)$ for any given data $y \in Y$. Obviously, a worst-case optimal estimator is given by $z_{\mathrm{o}}^{\mathrm{wc}} = \mathcal{A}_{\mathrm{o}}^{\mathrm{wc}}(y)$.

We notice that optimal algorithms map data y into the ℓ_p–*Chebychev center* of the set $\mathcal{S}\mathcal{I}^{-1}(y)$, where the Chebychev center $z_{\mathrm{c}}(H)$ of a set $H \subseteq Z$ is defined as

$$\max_{h \in H} \|h - z_{\mathrm{c}}(H)\| \doteq \inf_{z \in Z} \max_{h \in H} \|h - z\| \doteq r_{\mathrm{c}}(H).$$

Optimal algorithms are often called *central algorithms* and $z_{\mathrm{c}}(\mathcal{S}\mathcal{I}^{-1}(y)) = z_{\mathrm{o}}^{\mathrm{wc}}$. We remark that, in general, the Chebychev center of a set $H \subset Z$ may not be unique and not necessarily belongs to H, for example, when H is not convex or it is a discrete set.

Remark 1 (Interpretation of the Chebychev center): Note that, by construction, for any given set H (not necessarily convex, nor connected), the ℓ_p–Chebychev center $z_{\mathrm{c}}(H)$ of H and its radius $r_{\mathrm{c}}(H)$ are given by the center and radius of the smallest ℓ_p norm-ball enclosing the set H. That is, we can compute $z_{\mathrm{c}}(H)$ and $r_{\mathrm{c}}(H)$ solving the optimization problem

$$\min_{z, r} r \quad \text{subject to } \mathcal{B}(z, r) \supseteq H. \tag{8}$$

Note that, as remarked above, the optimal ball $\mathcal{B}(z, r)$ need not be unique. It follows immediately that if the set H is centrally symmetric with center $\bar{z}$, then $\bar{z}$ is a Chebychev center of H. ◇

The computation of the worst-case radius of information $r_{\mathrm{o}}^{\mathrm{wc}}(y)$ and of the derivation of optimal algorithms $\mathcal{A}_{\mathrm{o}}^{\mathrm{wc}}$ have been the focal point of several papers in a system identification setting, see e.g. [18].

[1] In the IBC context, this error is usually referred to as "local" radius of information, to distinguish from the so-called "global" radius, see [29] for further details.

III. Probabilistic Setting with Random Uncertainty

In this section, we introduce a probabilistic counterpart of the worst-case setting previously defined, that is we define optimal algorithms $\mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}$ and the probabilistic radius $r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon)$ for the so-called probabilistic setting when the uncertainty q is random and $\epsilon \in (0,\, 1)$ is a given parameter called *accuracy*. Roughly speaking, in this setting the error of an algorithm is measured in a worst-case sense, but we "discard" a set of measure at most ϵ from the consistency set $\mathcal{S}\mathcal{I}^{-1}(y)$. Hence, the probabilistic radius of information may be interpreted as the smallest radius of a ball discarding a set whose measure is at most ϵ. Therefore, we are decreasing the worst-case radius of information at the expense of a probabilistic "risk" ϵ. In a system identification context, reducing the radius of information is clearly a highly desirable property. Using this probabilistic notion, we can compute a trade-off function which shows how the radius of information decreases as a function of the parameter ϵ.

We now state a formal assumption regarding the random uncertainty q.

Assumption 2 (Random measurement uncertainty): We assume that the uncertainty q is a real random[2] vector with given probability density $p_{\mathcal{Q}}(q)$ and support set $\mathcal{Q} = \mathcal{B}(\rho)$. We denote by $\mu_{\mathcal{Q}}$ the probability measure generated by $p_{\mathcal{Q}}(q)$ over the set $\mathcal{Q}$.

Remark 2 (Induced measure over $\mathcal{I}^{-1}(y)$): We note that the probability measure over the set $\mathcal{Q}$ *induces*, by means of equation (1), a probability measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ over the set $\mathcal{I}^{-1}(y)$. Indeed, for any measurable set $B \subseteq X$, we consider the probability measure $\mu_{\mathcal{Q}}$ as follows: $\tilde{\mu}_{\mathcal{I}^{-1}}(B) = \mu_{\mathcal{Q}}(q \in \mathcal{Q} \,|\, \exists x \in B \cap \mathcal{I}^{-1}(y) \,|\, \mathcal{I}x + q = y)$. This probability measure is such that points outside the consistency set $\mathcal{I}^{-1}(y)$ have measure zero, and $\tilde{\mu}_{\mathcal{I}^{-1}}\big(\mathcal{I}^{-1}(y)\big) = 1$, that is this induced measure is concentrated over $\mathcal{I}^{-1}(y)$. This induced measure is formally defined in [29, Chapter 6], where it is shown that it is indeed a *conditional measure*. Similarly, we denote by $\tilde{p}_{\mathcal{I}^{-1}}$ the induced probability density, having support over $\mathcal{I}^{-1}(y)$. We remark that Theorem 1 in Section IV studies the induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}(\cdot)$ over the set $\mathcal{I}^{-1}(y)$ when q is uniformly distributed within $\mathcal{Q}$, showing that this measure is still uniform. In turn, the induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ is mapped through the linear operator $\mathcal{S}$ into a measure over $\mathcal{S}\mathcal{I}^{-1}(y)$, which we denote as $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}}$. Similarly, the induced density is denoted as $\tilde{p}_{\mathcal{S}\mathcal{I}^{-1}}$. In Theorem 1 in Section IV we show that the induced measure $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}}$ is log-concave in the case of uniform density over $\mathcal{Q}$. $\diamond$

Given corrupted information $y \in Y$ and accuracy level $\epsilon \in (0, 1)$, we define the probabilistic error (to level ϵ) $r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon)$ of the algorithm $\mathcal{A}$ as

$$r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon) \doteq \inf_{\mathcal{X}_\epsilon \text{ such that } \tilde{\mu}_{\mathcal{I}^{-1}}(\mathcal{X}_\epsilon) \leq \epsilon} \; \max_{x \,\in\, \mathcal{I}^{-1}(y) \setminus \mathcal{X}_\epsilon} \|\mathcal{S}x - \mathcal{A}(y)\| \tag{9}$$

where the notation $\mathcal{I}^{-1}(y) \setminus \mathcal{X}_\epsilon$ indicates the set-theoretic difference between $\mathcal{I}^{-1}(y)$ and $\mathcal{X}_\epsilon$,

$$\mathcal{I}^{-1}(y) \setminus \mathcal{X}_\epsilon \doteq \left\{ x \in \mathcal{I}^{-1}(y) \,|\, x \notin \mathcal{X}_\epsilon \right\}.$$

Clearly, $r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon) \leq r^{\mathrm{wc}}(\mathcal{A}, y)$ for any algorithm $\mathcal{A}$, data $y \in Y$ and accuracy level $\epsilon \in (0, 1)$, which implies a reduction of the approximation error in a probabilistic setting.

An algorithm $\mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}$ is called *probabilistic optimal* (to level ϵ) if it minimizes the error $r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon)$ for any $y \in Y$ and $\epsilon \in (0, 1)$. That is, given data $y \in Y$ and accuracy level $\epsilon \in (0, 1)$, we have

$$r_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon) \doteq r^{\mathrm{pr}}(\mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}, y, \epsilon) = \inf_{\mathcal{A}} r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon). \tag{10}$$

The minimal error $r_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon)$ is called the *probabilistic radius of information* (to level ϵ) and the corresponding optimal estimator is given by

$$z_{\mathrm{o}}^{\mathrm{pr}}(\epsilon) \doteq \mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon). \tag{11}$$

The problem we study in the next section is the computation of $r_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon)$ and the derivation of probabilistic optimal algorithms $\mathcal{A}_{\mathrm{o}}^{\mathrm{pr}}$. To this end, as in [29], we reformulate equation (9) in terms of a chance-constrained optimization problem [20]

$$r^{\mathrm{pr}}(\mathcal{A}, y, \epsilon) = \min \left\{ r \,|\, v(r, \mathcal{A}) \leq \epsilon \right\},$$

where the violation function for given algorithm $\mathcal{A}$ and radius r is defined as

$$v(r, \mathcal{A}) \doteq \tilde{\mu}_{\mathcal{I}^{-1}} \left\{ x \in \mathcal{I}^{-1}(y) \,|\, \|\mathcal{S}x - \mathcal{A}(y)\| > r \right\}.$$

Then, this formulation leads immediately to

$$r_{\mathrm{o}}^{\mathrm{pr}}(y, \epsilon) = \min \left\{ r \,|\, v_{\mathrm{o}}(r) \leq \epsilon \right\}, \tag{12}$$

where the *optimal violation function* for a given radius r is given by

$$v_{\mathrm{o}}(r) \doteq \inf_{\mathcal{A}} \tilde{\mu}_{\mathcal{I}^{-1}} \left\{ x \in \mathcal{I}^{-1}(y) : \|\mathcal{S}x - \mathcal{A}(y)\| > r \right\}. \tag{13}$$

IV. Random Uncertainty Uniformly Distributed

In this section, which contains the main technical results of the paper, we study the case when q is uniformly distributed over the ball $\mathcal{Q} \equiv \mathcal{B}(\rho)$, i.e. $q \sim \mathcal{U}_{\mathcal{Q}}$ and $\mu_{\mathcal{Q}} \equiv \mu_{\mathcal{U}(\mathcal{Q})}$. First, we address a preliminary technical question: *If $\mu_{\mathcal{Q}}$ is the uniform measure over $\mathcal{Q}$, what is the induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ over the set $\mathcal{I}^{-1}(y)$ defined in equation (2)?* The next result shows that this distribution is indeed still uniform. Furthermore, we prove that the induced measure on $\mathcal{S}\mathcal{I}^{-1}(y)$ is log-concave.

Remark 3 (Log-concave measures): We recall that a measure $\mu(\cdot)$ is log-concave if, for any compact subsets A, B and $\lambda \in [0,\, 1]$, it holds

$$\mu(\lambda A + (1 - \lambda)B) \geq \mu(A)^{\lambda} \mu(B)^{1-\lambda}$$

[2]For simplicity, in this assumption we consider the case when the density of q exists (that is the distribution is differentiable).

where $\lambda A + (1-\lambda)B$ denotes the Minkowski sum[3] of the two sets λA and $(1-\lambda)B$. Note that the Brunn-Minkowski inequality [24] asserts that the uniform measure over convex sets is log-concave. Furthermore, any Gaussian measure is log-concave. $\diamond$

Theorem 1 (Measures over $\mathcal{I}^{-1}(y)$ and $\mathcal{S}\mathcal{I}^{-1}(y)$): Let $q \sim \mathcal{U}(\mathcal{Q})$ with $\mathcal{Q} \equiv \mathcal{B}(\rho)$, then, for any $y \in Y$ it holds:

(i) The induced measure $\tilde{\mu}_{\mathcal{I}^{-1}}$ is uniform over $\mathcal{I}^{-1}(y)$, that is $\tilde{\mu}_{\mathcal{I}^{-1}} \equiv \mu_{\mathcal{U}(\mathcal{I}^{-1}(y))}$;

(ii) The induced measure $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}}$ over $\mathcal{S}\mathcal{I}^{-1}(y)$ is log-concave. Moreover, if $\mathcal{S} \in \mathbb{R}^{n,n}$, then this measure is uniform, that is $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}} \equiv \mu_{\mathcal{U}(\mathcal{S}\mathcal{I}^{-1}(y))}$.

The proof of this result is available in the paper [9].

The result in this theorem can be immediately extended to the more general case when $\mathcal{Q}$ is a compact set. We now introduce an assumption regarding the solution operator $\mathcal{S}$.

Assumption 3 (Regularized solution operator): In the sequel, we assume that the solution operator is regularized, so that $\mathcal{S} = \left[\bar{\mathcal{S}}\ 0_{s,n-s}\right]$, with $\bar{\mathcal{S}} \in \mathbb{R}^{s,s}$.

Remark 4 (On Assumption 3): Note that the assumption is made without loss of generality. Indeed, for any full row rank $\mathcal{S} \in \mathbb{R}^{s,n}$, we introduce the change of variables $T = [T_1\ T_2]$, where T_1 is an orthonormal basis of the column space of $\mathcal{S}^\top$ and T_2 is an orthonormal basis of the null space of $\mathcal{S}$ (in Matlab notation, we write $T_1 = \texttt{orth}(\mathcal{S}^\top)$ and $T_2 = \texttt{null}(\mathcal{S})$). Then, T is orthogonal by definition, and it follows

$$\begin{aligned} z &= \mathcal{S}x = \mathcal{S}TT^\top x = \mathcal{S}\left[T_1\ T_2\right]T^\top x \\ &= \left[\mathcal{S}T_1\ \mathcal{S}T_2\right]T^\top x = \left[\bar{\mathcal{S}}\ 0_{s,n-s}\right]\tilde{x} = \widetilde{\mathcal{S}}\tilde{x}, \end{aligned}$$

where we introduced the new problem element $\tilde{x} \doteq T^\top x$ and the new solution operator $\widetilde{\mathcal{S}} \doteq \mathcal{S}T$. Note that, with this change of variables, equation (1) is rewritten as

$$y = \widetilde{\mathcal{I}}\tilde{x} + q,$$

by introducing the transformed information operator $\widetilde{\mathcal{I}} \doteq \mathcal{I}T$. We observe that any algorithm $\mathcal{A}$, being a mapping from Y to Z, is invariant to this change of variable. It is immediate to conclude that the new problem defined in the variable $\tilde{x}$ and operators $\widetilde{\mathcal{I}}$ and $\widetilde{\mathcal{S}}$ satisfies Assumption 3. $\diamond$

Instrumental to the next developments, we first introduce the degenerate cone (cylinder) in the element space X, with given center $z_{\mathrm{c}} \in Z$ and radius r, as follows

$$\mathcal{C}(z_{\mathrm{c}}, r) \doteq \{x \in \mathbb{R}^n \,|\, \|\mathcal{S}x - z_{\mathrm{c}}\| \leq r\} \subset X. \qquad (14)$$

Note that this set is the inverse image through $\mathcal{S}$ of the norm-ball $\mathcal{B}(z_{\mathrm{c}}, r) \subset Z$. Moreover, due to Assumption 3, the cylinder $\mathcal{C}(z_{\mathrm{c}}, r)$ is parallel to the coordinate axes, that is any element x of the cylinder can be written as

$$x \in \mathcal{C}(z_{\mathrm{c}}, r) \quad \Leftrightarrow \quad x = \begin{bmatrix} \bar{\mathcal{S}}^{-1}\zeta \\ \xi \end{bmatrix},$$

with $\zeta \in \mathcal{B}(z_{\mathrm{c}}, r) \subset \mathbb{R}^s$ and $\xi \in \mathbb{R}^{n-s}$. Hence, for the case $s < n$, the cylinder is unbounded, while for $s = n$ it is simply a linear transformation through $\mathcal{S}^{-1}$ of an ℓ_p norm-ball. Next, for given center $z_{\mathrm{c}} \in Z$ and radius $r > 0$, we define the intersection set between the cylinder $\mathcal{C}(z_{\mathrm{c}}, r)$ and the consistency set $\mathcal{I}^{-1}(y)$

$$\Phi(z_{\mathrm{c}}, r) \doteq \mathcal{I}^{-1}(y) \cap \mathcal{C}(z_{\mathrm{c}}, r) \subset X \qquad (15)$$

and its volume

$$\phi(z_{\mathrm{c}}, r) \doteq \mathrm{vol}\left[\Phi(z_{\mathrm{c}}, r)\right]. \qquad (16)$$

Finally, we define the set $\mathcal{H}(r)$ of all centers $z_{\mathrm{c}} \in \mathbb{R}^s$ for which the intersection set $\Phi(z_{\mathrm{c}}, r)$ is non-empty, i.e.

$$\mathcal{H}(r) \doteq \{z_{\mathrm{c}} \in \mathbb{R}^s \,|\, \Phi(z_{\mathrm{c}}, r) \neq \emptyset\}. \qquad (17)$$

Note that, even if the cylinder $\mathcal{C}(z_{\mathrm{c}}, r)$ is in general unbounded, the set $\Phi(z_{\mathrm{c}}, r)$ is bounded whenever $z_{\mathrm{c}} \in \mathcal{H}(r)$, since $\mathcal{I}^{-1}(y)$ is bounded.

We are now ready to state the main theorem of this section, that provides useful properties from the computational point of view of the optimal violation function defined in (13).

Theorem 2: Let $q \sim \mathcal{U}(\mathcal{Q})$ with $\mathcal{Q} \equiv \mathcal{B}(\rho)$, and $\mathcal{S} = \left[\bar{\mathcal{S}}\ 0_{s,n-s}\right]$, with $\bar{\mathcal{S}} \in \mathbb{R}^{s,s}$. Then, the following statements hold

(i) For given $r > 0$, the optimal violation function $v_{\mathrm{o}}(r)$ is given by

$$v_{\mathrm{o}}(r) = 1 - \frac{\phi_{\mathrm{o}}(r)}{\mathrm{vol}\left[\mathcal{I}^{-1}(y)\right]}, \qquad (18)$$

where $\phi_{\mathrm{o}}(r)$ is the solution of the optimization problem

$$\phi_{\mathrm{o}}(r) \doteq \max_{z_{\mathrm{c}} \in \mathcal{H}(r)} \phi(z_{\mathrm{c}}, r) \qquad (19)$$

with $\phi(z_{\mathrm{c}}, r)$ and $\mathcal{H}(r)$ defined in (16) and (17), respectively.

(ii) For given $r > 0$, the function (16) is continuous semi-strictly quasi-concave[4] in $z_{\mathrm{c}} \in \mathcal{H}(r)$;

(iii) The function $v_{\mathrm{o}}(r)$ is right-continuous and non-increasing for $r > 0$.

The proof of this result is available in the paper [9].

Remark 5 (Unimodality of the function $\phi(z_{\mathrm{c}}, r)$): Point (ii) in Theorem 2 is crucial from the computational viewpoint. Indeed, as remarked for instance in [10], a semi-strictly quasi-concave function cannot have local maxima. Roughly speaking, this means that the function $\phi(\cdot, r)$ is unimodal, and therefore any local maximal solution of problem (19) is also a global maximum. Note that from

[3]The Minkowski sum of two sets A and B is obtained adding every element of A to every element of B, i.e. $A+B = \{a+b \,|\, a \in A, b \in B\}$.

[4]A function f defined on a convex set $A \in \mathbb{R}^n$ is semi-strictly quasi-concave if $f(y) < f(\lambda x + (1-\lambda)y)$ holds for any $x, y \in A$ such that $f(x) > f(y)$ and $\lambda \in (0,\, 1)$.

the Brunn-Minkowski inequality it follows that, if there are multiple points $z_{\rm o}(i)$ where $\phi(\cdot)$ achieves its global maximum, then the sets $\Phi(z_{\rm o}^{(i)}, r)$ are all homothetic, see [24]. Further, from the definition of $\Phi(\cdot, r)$, this implies that $\Phi(z_{\rm o}^{(i)}, r) = \Phi(z_{\rm o}^{(j)}, r) + z_{\rm o}^{(i)} - z_{\rm o}^{(j)}$. $\diamond$

Remark 6 (Probabilistic radius): Theorem 2 provides a way of computing the optimal probabilistic radius of information $r_{\rm o}^{\rm pr}(y, \epsilon)$. Indeed, for given $r > 0$, the probabilistic radius of information (to level ϵ) is given by the solution of the following one-dimensional "inversion" problem

$$r_{\rm o}^{\rm pr}(y, \epsilon) = \min \{ r \,|\, v_{\rm o}(r) \leq \epsilon \} . \qquad (20)$$

Note that point (iii) in Theorem 2 guarantees that such solution always exists for $\epsilon \in (0, 1)$, and it is unique. The corresponding optimal estimate is then given by

$$z_{\rm o}^{\rm pr}(\epsilon) = \mathcal{A}_{\rm o}^{\rm pr}(y, \epsilon) = z_{\rm o}(r_{\rm o}^{\rm pr}(y, \epsilon)),$$

where we denoted by $z_{\rm o}(r)$ a solution of the optimization problem (19).

$\diamond$

Theorem 2 shows that the problem we are considering is indeed a well-posed one, since it has a unique solution (even though not a unique minimizer in general). However, its solution requires the computation of the volume of the intersection set $\Phi(z_{\rm c}, r)$, which is in general a very hard task. A notable exception in which the probabilistic optimal estimate is immediately computed for q uniformly distributed in $\mathcal{Q}$ is the special case when the consistency set $\mathcal{I}^{-1}(y)$ is centrally symmetric with center $\bar{x}$. Indeed, in this case it can be seen that $\mathcal{S}\mathcal{I}^{-1}(y)$ is also a centrally symmetric around $\bar{z} = \mathcal{S}\bar{x}$, and so is the density $\tilde{\mu}_{\mathcal{S}\mathcal{I}^{-1}}$. Hence, the optimal probabilistic estimate coincides with the center $\bar{z}$, since it follows from symmetry that the probability measure of the intersection of $\mathcal{S}\mathcal{I}^{-1}(y)$ with an ℓ_p norm-ball is maximized when the two sets are concentric. Moreover, this estimate coincides with the classical worst-case (central) estimate, which in turn coincides with the classical least squares estimates.

V. Conclusions

This paper deals with the rapproachement between the stochastic and worst-case settings for system identification. The problem is formulated within the probabilistic setting of information-based complexity, and it is focused on the approach to "discard" sets of small measure from the set of deterministic estimates. The paper establishes rigorous optimality properties of a trade-off curve, called *violation function*, which shows how the radius of information decreases as a function of the accuracy.

References

[1] E.W. Bai, H. Cho, R. Tempo, and Y.Y. Ye. Optimization with few violated constraints for linear bounded error parameter estimation. *IEEE Transactions on Automatic Control*, 47:1067–1077, 2002.

[2] G. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

[3] G. Calafiore, F. Dabbene, and R. Tempo. Research on probabilistic methods for control system design. *Automatica*, 47:1279–1293, 2011.

[4] G. Calafiore and F. Dabbene (Eds.). *Probabilistic and Randomized Methods for Design under Uncertainty*. Springer-Verlag, London, 2006.

[5] M.C. Campi, G.C. Calafiore, and S. Garatti. Interval predictor models: Identification and reliability. *Automatica*, 45(2):382–392, 1990.

[6] M.C. Campi and E. Weyer. Guaranteed non-asymptotic confidence regions in system identification. *Automatica*, 41:1751–1764, 2005.

[7] M.C. Campi and E. Weyer. Non-asymptotic confidence sets for the parameters of linear transfer functions. *IEEE Transactions on Automatic Control*, 55:2708–2720, 2010.

[8] F. Dabbene, M. Sznaier, and R. Tempo. A probabilistic approach to optimal estimation - Part II: Algorithms and applications. In *Proceedings IEEE Conference on Decision and Control*, 2012.

[9] F. Dabbene, M. Sznaier, and R. Tempo. Probabilistic optimal estimation and filtering under uncertainty. *arXiv:1203.1429v2*, 2012.

[10] A. Danilidis and Y. Garcia Ramos. Some remarks on the class of continuous (semi-)strictly quasi convex functions. *Journal of Optimization Theory and Applications*, 133:37–48, 2007.

[11] M. Gevers, X. Bombois, B. Codrons, G. Scorletti, and B.D.O. Anderson. Model validation for control and controller validation in a prediction error identification framework - Part I : Theory. *Automatica*, 39(3):403–415, 2003.

[12] P.R. Halmos. *Measure Theory*. Springer-Verlag, New York, 1950.

[13] H.D. Hanebeck, J. Horn, and G. Schmidt. On combining statistical and set-theoretic estimation. *Automatica*, 35:1101–1109, 1999.

[14] A.J. Helmicki, C.A. Jacobson, and C.N. Nett. Control-oriented system identification: A worst-case/deterministic approach in H_∞. *IEEE Transactions on Automatic Control*, 36:1163–1176, 1991.

[15] H. Hjalmarsson. From experiment design to closed loop control. *Automatica*, 41(3):393–438, 2005.

[16] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, 1999.

[17] L. Ljung and A. Vicino. Special issue on 'system identification' - editorial. *IEEE Transactions on Automatic Control*, 50(10):787–803, 2005.

[18] M. Milanese and R. Tempo. Optimal algorithms theory for robust estimation and prediction. *IEEE Transactions on Automatic Control*, 30:730–738, 1985.

[19] M. Milanese and R. Tempo. Optimal estimation theory fordynamic systems with set membership uncertainty: an overview. *Automatica*, 27(6):997–1009, 1991.

[20] A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *Journal of Optimization Theory and Applications*, 17:969–996, 2006.

[21] B.M. Ninness and G.C. Goodwin. Rapprochement between bounded-error and stochastic estimation theory. *International Journal of Adaptive Control and Signal Processing*, 9:107–132, 1995.

[22] W. Reinelt, A. Garulli, and L. Ljung. Comparing different approaches to model error modeling in robust identication. *Automatica*, 38(5), 2002.

[23] R.S. Sánchez-Peña and M. Sznaier. *Robust Systems: Theory and Applications*. John Wiley, New York, 1998.

[24] R. Schneider. *Convex bodies: the Brunn-Minkowski theory*. Cambridge University Press, 1993.

[25] T. Söderström, P.M.J. Van den Hof, B. Wahlberg, and S. Weiland. Special issue on 'data-based modelling and system identification' - editorial. *Automatica*, 41(3), 2005.

[26] R. Tempo. Robust estimation and filtering in the presence of bounded noise. *IEEE Transactions on Automatic Control*, 33:864–867, 1988.

[27] R. Tempo, G. Calafiore, and F. Dabbene. *Randomized Algorithms for Analysis and Control of Uncertain Systems*. Communications and Control Engineering Series. Springer-Verlag, London, 2005.

[28] F. Tjarnstrom and A. Garulli. A mixed probabilistic/bounded-error approach to parameter estimation in the presence of amplitude bounded white noise. In *Proceedings of the IEEE Conference on Decision and Control*, pages 3422–3427, 2002.

[29] J.F. Traub, G.W. Wasilkowski, and H. Woźniakowski. *Information-Based Complexity*. Academic Press, New York, 1988.

[30] J.F. Traub and A.G. Werschulz. *Complexity and Information*. Cambridge University Press, Cambridge, 1998.