Probabilistic Optimal Estimation with Uniformly Distributed Noise

FABRIZIO DABBENE, MARIO SZNAIER, AND ROBERTO TEMPO∗

Abstract—The classical approach to system identification is based on stochastic assumptions about the measurement error, and provides estimates that have random nature. Worst-case identification, on the other hand, only assumes the knowledge of deterministic error bounds, and establishes guaranteed estimates, thus being in principle better suited for the use in control design. However, a main limitation of such deterministic bounds lies in their potential conservatism, thus leading to estimates of restricted use.

In this paper, we propose a rapprochement between the stochastic and worst-case system identification viewpoints, which is based on the probabilistic setting of information-based complexity. The main idea in this line of research is to “discard” sets of measure at most $\epsilon$, where $\epsilon$ 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.”

And the case of uniformly distributed noise, we derive new computational results, and in particular we compute a trade-off curve, called violation function, which shows how the radius of information decreases as a function of the accuracy. To this end, we construct randomized and deterministic algorithms which provide approximations of this function. We report extensive simulations showing numerical comparisons between the stochastic, worst-case and probabilistic approaches, thus demonstrating the efficacy of the methods proposed in this paper.

Keywords: Linear estimation, system identification, optimal algorithms, randomized algorithms, uncertain systems, least-squares

I. INTRODUCTION AND PRELIMINARIES

The mainstream paradigm for system identification is the classical stochastic approach, see [35] and the special issues [36], [46], which has been very successful in many applications. This approach assumes that the available observations are contaminated by random noise, and has the goal to derive soft bounds on the estimation errors. In this setting, optimality is guaranteed in a probabilistic sense and the resulting algorithms often enjoy only asymptotic convergence properties.

In the last decades, several authors focused their attention on the so-called set-membership identification which aims at the computation of hard bounds on the estimation errors, see for instance [38], and [28] for pointers to more recent developments. Set-membership identification may be embedded within the general framework of worst-case information-based complexity (IBC), see [54] and [55], so that various systems and control problems, such as time-series analysis, filtering and $H_\infty$ identification can be addressed [37], [50], [27], [43], [22]. In this 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 main drawback of this deterministic approach is that in many instances the resulting worst-case bounds could be too conservative, and therefore of limited use, in particular when the ultimate objective is to use system identification in the context of closed-loop control.

The worst-case setting is based on the “concern” that the noise may be very malicious. The computed bounds are certainly more pessimistic than the stochastic ones, but the idea is to guard against the worst-case scenario, even though it is unlikely to occur. These observations motivated a research line seeking to a rapprochement of these viewpoints, see [40], [25], [42], [13], [23]. The main idea in this line of research is to discard sets of measure at most $\epsilon$ from the set of deterministic estimates. Note that probabilistic uncertainty descriptions have been extensively studied in the context of control design, leading to a “probabilistic robust” approach [51], [11], [10], where the goal is to synthesize controllers so that the desired performance specifications are achieved with a high probability level.

The focal point of this paper is to address the rapprochement between soft and hard bounds in a rigorous fashion in an estimation context, with the goal to derive useful computational tools for linear estimation problems, see [18], [19] for preliminary results. To this end, we adopt the general abstract formulation of IBC which allows to study under the same framework the two main approaches to system identification discussed so far, and to obtain new results for the probabilistic framework. In particular, the objective is to compute (by means of randomized and deterministic algorithms) the so-called probabilistic radius of information. 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 which is based on a finite number of observations. In this sense, this approach has close relations with the works based on statistical learning theory proposed in [31], [57], [56], and with the approach in [15], [16], where distribution-free non-asymptotic confidence sets for the estimates are derived. Furthermore, the paper is also related to the work [53], where a probability density function over the consistency set is considered.

We now provide a preview of the structure and main results of the paper. Section II presents an introduction to information-based complexity and an example showing how system parameter identification and prediction may be formulated in the general IBC framework. Section III formally presents the probabilistic setting, originally introduced in [54], and illustrates it with a tutorial example regarding estimation of the parameters of a second order model corrupted by additive noise. The example is continued in other sections of the paper for illustrative purposes. In this context, the idea is to “discard” sets of (probabilistic) measure at most $\epsilon$ from the consistency set. That is, the objective is to decrease significantly the worst-case radius, thus obtaining a new error which represents the probabilistic radius of information, at the expense of a probabilistic risk $\epsilon$. This approach may be very useful, for example, for system identification in the presence of outliers [3], where “bad measurements” may be discarded. In this section, by means of a chance-constrained approach [39], we also show that the probabilistic radius is related to the minimization of the so-called optimal violation function $v_\epsilon(r)$.

Section IV deals with uniformly distributed noise and contains the main technical results of the paper. In particular, Theorem 1 shows that the conditional measure over the so-called consistency set is
uniform. Theorem 2 proves crucial properties, from the computational point of view, of the optimal violation function \( v_0(r) \). In particular, this result shows that \( v_0(r) \) is non-increasing, and for fixed \( r > 0 \), it can be obtained as the maximization of a specially constructed unimodal function. Hence, it may be easily computed by means of various optimization techniques which are discussed in the next section.

In Section V we introduce specific algorithms for computing the optimal violation function. First, we observe that the exact computation of \( v_0(r) \) requires the evaluation of the volume of polytopes. Since this problem is \( \text{NP-Hard} \) [32], we propose to use suitable probabilistic and deterministic relaxations. More precisely, first we present a randomized algorithm based upon the classical Markov Chain Monte Carlo method [49], [48], which has been studied in the context of stochastic approximation methods [17], [34]; see also [51] and [10] for further details about randomized algorithms. Secondly, we present a deterministic relaxation of \( v_0(r) \) which is based upon the solution of a semi-definite program (SDP). The performance of both algorithms is compared using the example previously introduced.

Section VI discusses normally distributed noise, and presents some connections with classical stochastic estimation. In particular, it is shown that the least-squares algorithm is “almost optimal” also in the probabilistic setting discussed in this paper. For this case, we state a bound (which is essentially tight for small-variance noise) on the probabilistic radius of information, which is given in [54] in terms of the so-called average radius of information. This bound depends on \( \epsilon \), on the noise covariance, and on the so-called information and solution operators.

Finally, in Section VII we study a numerical example of a FIR system affected by uniformly distributed noise. First, we compute deterministic and randomized relaxations of the optimal violation function. Then, by means of an extensive numerical simulation, we compare the probabilistic optimal estimate with classical least-squares and the worst-case optimal estimates.

II. INFORMATION-BASED COMPLEXITY FOR SYSTEM IDENTIFICATION

This section introduces the formal definitions used in information-based complexity and an illustrative example regarding system identification and prediction. The relevant spaces, operators and sets discussed next are shown in Figure 1.

![Fig. 1: Illustration of the information-based complexity framework.](image)

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 \( I \), called information operator, which maps \( X \) into a linear normed \( m \)-dimensional space \( Y \):

\[
I : X \rightarrow Y.
\]

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

\[
y = Ix + \eta
\]

where \( \eta \) represents additive noise (or uncertainty) which may be deterministic or random. We assume that \( \eta \in \mathcal{N} \), where \( \mathcal{N} \subseteq \mathbb{R}^m \) is a possibly unbounded set. Due to the presence of uncertainty \( \eta \), the problem element \( x \in X \) may not be easily recovered knowing data \( y \in Y \). Then, we introduce a linear operator \( S \), called a solution operator, which maps \( X \) into \( Z \):

\[
S : X \rightarrow Z
\]

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

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

\[
A : Y \rightarrow Z.
\]

An algorithm provides an approximation \( A(y) \) of \( Sz \) using the available information \( y \in Y \) of \( x \in X \). The outcome of such an algorithm is called an estimate \( z = A(y) \).

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 feasible set (also referred to as consistency set) as follows

\[
F_y = \{ x \in X : y = Ix + \eta \text{ for some } \eta \in \mathcal{N} \}
\]

which represents the set of all problem elements \( x \in X \) compatible with (i.e. not invalidated by) \( Ix \), uncertainty \( \eta \) and support set \( \mathcal{N} \).

For the sake of simplicity, we assume that the three sets \( X, Y, Z \) are equipped with the same \( \ell_p \) norm. Also, in the sequel we assume that the information operator \( I \) is a one-to-one mapping, i.e. \( m \geq n \) and rank \( I = n \). Similarly, \( n \geq s \) and \( S \) is full row rank. Moreover, we assume that the set \( F_y \) has non-empty interior; this is always satisfied e.g. whenever \( \mathcal{N} \) is the whole \( \mathbb{R}^m \). Note that, in a system identification context, the assumption on \( I \) and on the consistency set \( F_y \) are necessary conditions for identifiability of the problem element \( x \in X \). Similarly, the assumptions \( S \) full-rank and \( n \geq s \) are equivalent to assuming that the elements of the vector \( z = Sz \) 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 the role of these operators and spaces in the context of system identification; note that the IBC theoretical setting also applies to filtering problems, see for instance [50], [22].

**Example 1** (System parameter identification and prediction):

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, parameterized by some unknown parameter vector \( x \in X \). This may be represented as the following finite regression

\[
\xi(t, x) = \sum_{i=1}^{n} x_i \psi_i(t) = \Psi^T(t)x,
\]
with given basis functions $\psi_i(t)$, and $\Psi^\top(t) = [\psi_1(t) \cdots \psi_n(t)]$. We suppose that $m$ noisy measurements of $\xi(t, x)$ are available for $t_1 < t_2 < \cdots < t_m$, that is

$$y = \mathbf{I}x + \eta = [\Psi(t_1) \cdots \Psi(t_m)]^\top x + \eta. \quad (3)$$

In the set-membership approach, one usually assumes unknown but bounded errors, such that $|\eta_i| \leq \rho$, $i = 1, \ldots, m$, that is $N = \{\eta : \|\eta\|_\infty \leq \rho\}$. Then, the aim is to obtain a parameter estimate using the data $y$. Hence, the solution operator is given by the identity,

$$S_x = x$$

and $Z \equiv X$. In this case, the feasible parameters set is given as follows

$$F_y = \left\{ x \in X : \|y - [\Psi(t_1) \cdots \Psi(t_m)]^\top x\|_\infty \leq \rho \right\}. \quad (4)$$

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

$$z = S_x = \{\xi(t_{m+1}), \ldots, \xi(t_{m+s})\} = [\Psi(t_{m+1}) \cdots \Psi(t_{m+s})]^\top x.$$

Next, we define approximation errors and optimal algorithms when $\eta$ is deterministic or random. First, we briefly summarize the deterministic case which has been thoroughly analyzed in the literature, see e.g. [37]. The definitions concerning the probabilistic case are new in this context, and are introduced in Section III.

### A. Worst-Case Setting

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

$$r_{wc}(A, y) = \max_{A \in \mathcal{A}} \|S_x - A(y)\|, \quad (5)$$

where $\cdot$ denotes the $\ell_p$-norm. This error is based on the available information $y \in Y$ about the problem element $x \in X$ and measures the approximation error between $S_x$ and $A(y)$. An algorithm $\mathcal{A}_{wc}$ is called worst-case optimal if it minimizes $r_{wc}(A, y)$ for any $y \in Y$. That is, given data $y \in Y$, we have

$$r_{wc}(y) = r_{wc}(\mathcal{A}_{wc}, y) = \inf_{A} r_{wc}(A, y). \quad (6)$$

The minimal error $r_{wc}(y)$ is called the worst-case radius of information.

This optimality criterion is meaningful in estimation problems as it ensures the smallest approximation error between the actual (unknown) solution $S_x$ and its estimate $A(y)$ for the worst element $x \in F_y$ for any given data $y \in Y$. Obviously, a worst-case optimal estimate is given by $z_{wc} = A_{wc}(y)$, see Figure 1.

Note that optimal algorithm maps data into the $\ell_p$-Chebychev center of the set $S_x F_y$, where the Chebychev center $z_c(H)$ of a set $H \subseteq \mathbb{R}^n$ is defined as

$$z_c(H) = \max_{h \in H} \|h - z_c(H)\| = \inf_{z \in H} \max_{h \in H} \|h - z\|,$$

where $r_c(H)$ represents the so-called Chebychev radius. Optimal algorithms are often called central algorithms and $z_c(S_x F_y) = z_{wc}$ is the worst-case optimal estimate, frequently referred to as central estimate. We remark that, in general, the Chebychev center of a set $H \subseteq \mathbb{R}^n$ may not be unique (e.g. for $\ell_\infty$ norms), and not necessarily belongs to $H$, even if $H$ is convex.

### III. Probabilistic Setting with Random Uncertainty

In this section, we formally define a probabilistic counterpart, originally introduced in [54], of the worst-case setting previously defined. That is, we define optimal algorithms $\mathcal{A}_{wc}^N$ and the probabilistic radius $r^{pr}(A, y, \epsilon)$ for the so-called probabilistic setting when the uncertainty $\eta$ 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 $\epsilon$ from the consistency set $S_x F_y$. Hence, the probabilistic radius of information may be interpreted as the smallest radius of a ball discarding a set whose measure is at most $\epsilon$. Therefore, we are decreasing the worst-case radius of information at the expense of a probabilistic “risk” $\epsilon$. In a system identification context, reducing the radius of information is clearly a highly desirable property. Using this probabilistic notion, we compute a trade-off function which shows how the radius of information decreases as a function of the parameter $\epsilon$, as described in the tutorial Example 2 and in the numerical example presented in Section VII.

Formally, in the sequel we assume that the uncertainty $\eta$ is a real random vector with given probability measure $\mu_N$ defined on the Borel $\sigma$-algebra of $N$. The measure $\mu_N$ is assumed to be absolutely continuous with respect to the Borel measure, so that its density $\rho_N$ exists. We note that the probability measure over the set $N$ induces, by means of equation (1), a probability measure $\mu_{N \setminus X}$ over the problem element space $X$. This is a conditional measure $\mu_{X|Y}$ formally defined as follows: For any Borel measurable set $B \subseteq X$, we have

$$\mu_{X|Y}(B) = \mu_N(\eta \in Y - I(B \cap F_y) \mid \eta \in Y - I(X)) \quad (7)$$

where, for a given set $A \subseteq X$, $I(A)$ denotes the image of $A$ under the (linear) transformation $I$, that is $I(A) = \{\eta \in Z : \eta = Ix \text{ for some } x \in A\}$, and $y - I(A)$ denotes the Minkowski difference between $y$ and the set $I(A)$. This conditional measure is such that points outside the consistency set $F_y$ have measure zero, and the measure of $F_y$ is one. The interested reader is referred to [54, Appendix 2.7] for a more detailed discussion on conditional measures, see also Section IV for technical results in the case of uniform measure $\mu_{N\setminus X}$.

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

$$r^{pr}(A, y, \epsilon) = \inf_{A \in \mathcal{A}_{wc}^N} \max_{x \in F_y \setminus X} \|S_x - A(y)\| \quad (8)$$

where the notation $F_y \setminus X$ indicates the set-theoretic difference between $F_y$ and $X$. Clearly, $r^{pr}(A, y, \epsilon) \leq r_{wc}(A, y)$ for any algorithm $A$, 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 $\mathcal{A}_{wc}^N$ is called probabilistic optimal (to level $\epsilon$) if it minimizes the error $r^{pr}(A, y, \epsilon)$ 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

$$r^{pr}_{wc}(y, \epsilon) = r^{pr}(\mathcal{A}_{wc}^N, y, \epsilon) = \inf_{A} r^{pr}(A, y, \epsilon). \quad (9)$$

The minimal error $r^{pr}_{wc}(y, \epsilon)$ is called the probabilistic radius of information (to level $\epsilon$) and the corresponding optimal estimate is given by

$$z^{pr}_{wc}(\epsilon) = \mathcal{A}_{wc}^N(y, \epsilon). \quad (10)$$

The problem we study in the next section is the computation of $r^{pr}_{wc}(y, \epsilon)$ and the derivation of probabilistic optimal algorithms $\mathcal{A}_{wc}^N$. 2

1 This is the case, for instance, of non-Euclidean norms. Consider for example the tetrahedron formed by the convex hall of the points $[1 1 1]$, $[-1 -1 1]$, $[1 -1 -1]$, $[1 1 -1]$. It is easy to check that the origin is the (unique) Chebychev center in the $\ell_\infty$ norm of the set, and it lies outside of it.

2 The Minkowski sum (difference) of two sets $A$ and $B$ is obtained adding (subtracting) every element of $A$ to every element of $B$, i.e. $A \pm B = \{a \pm b : a \in A, b \in B\}$. 3
To this end, as in [54], we reformulate equation (8) in terms of a chance-constrained optimization problem [39]

\[ r^p(A, y, \epsilon) = \min \{ r : v(r, A) \leq \epsilon \}, \]

where the violation function for given algorithm \( A \) and radius \( r \) is defined as

\[ v(r, A) = \mu_{\mathcal{F}_y} \{ x \in \mathcal{F}_y : \| Sx - A(y) \| > r \}. \]

Then, this formulation leads immediately to

\[ r^\mu(y, \epsilon) = \min \{ r : v_o(r) \leq \epsilon \}, \]

where the optimal violation function for a given radius \( r \) is given by

\[ v_o(r) = \inf_{\mathcal{A}} \{ x \in \mathcal{F}_y : \| Sx - A(y) \| > r \}. \]

Roughly speaking, the function \( v_o(r) \) describes how the risk \( \epsilon \) decreases as a function of the radius \( r \). However, the computation of \( v_o(r) \) is not an easy task and requires the results proved in Section IV and the algorithms presented in Section V. To illustrate the notions introduced so far, we consider the following numerical example. The example is tutorial, and it is sufficiently simple so that all relevant sets are two dimensional and can be easily depicted.

**Example 2 (Identification of a second order model):** Our aim is to estimate the parameters of a second order FIR model

\[ y_k = x_1 u_k + x_2 u_{k-1} + \eta_k, \quad k = 1, \ldots, m \]

where the input \( u_k \) is a known input sequence. The (unknown) nominal parameters were set to \([1.25, 2.35]^{T}\) and \( m = 100 \) measurements were collected generating the input sequence \( \{ u_k \} \) according to a Gaussian distribution with zero mean value and unit variance, and the measurement uncertainty \( \eta \) as a sequence of uniformly distributed noise in the \( \ell_{\infty} \) norm, i.e., \( ||\eta|| \leq 0.5 \). Note that, in this case, the operator \( S \) is the identity, and thus \( X \equiv \mathbb{Z} \) and the sets \( \mathcal{F}_y \) and \( SF_y \) coincide. That is, the goal is to estimate \( z_i = x_i, \quad i = 1, 2 \).

First, the optimal worst-case radius defined in (6) and the corresponding optimal solution have been computed by solving four linear programs (corresponding to finding the tightest box containing the polytope \( SF_y \)). The computed worst-case optimal estimate is \( z_{wc}^\mu = [1.2499, 2.3551]^{T} \) and the worst-case radius is \( r_{wc}^\mu(y) = 0.0352 \), see Figure 2.

Subsequently, we fix the accuracy level \( \epsilon = 0.1 \), and aim at computing a probabilistic optimal radius and the corresponding optimal estimate according to definitions (9) and (10). By using the techniques discussed in Section IV, we obtained \( r_{\mu}(y, 0.1) = 0.0284 \) and \( z_{\mu}(0.1) = [1.2480, 2.3540]^{T} \), which represents about a 20% improvement compared to \( r_{wc}^\mu(y) = 0.0352 \).

**IV. RANDOM UNCERTAINTY UNIFORMLY DISTRIBUTED**

In this section, which contains the main technical results of the paper, we study the case when the uncertainty \( \eta \) is uniformly distributed over a norm bounded set \( \mathcal{N} \) and we prove that in this case the computation of the optimal violation function, and thus of the probabilistic optimal estimate, can be formulated as a concave maximization problem.

To this end, denote by \( \text{vol} [A] \) the Lebesgue measure (volume) of a set \( A \), and by \( \lambda_{A} \) the uniform measure over a set \( A \), with \( 0 < \text{vol} [A] < +\infty \), which is defined as follows: For any Borel measurable set \( B \), \( \lambda_{A} (B) \equiv \text{vol} [B \cap A] / \text{vol} [A] \).

**Assumption 1 (Uniform noise):** We assume that \( \eta \) is uniformly distributed over the \( \ell_{p} \) norm-ball \( N = B(r) = \{ \eta : ||\eta|| \leq r, \ r > 0 \} \); that is \( \mu_{\mathcal{N}} \equiv \lambda_{\mathcal{N}} \).

First, we address a preliminary technical question: If \( \mu_{\mathcal{N}} \) is the uniform measure over \( \mathcal{N} \), what is the conditional measure \( \mu_{\mathcal{F}_y} \) over the set \( \mathcal{F}_y \) defined in equation (2)? The next result shows that this distribution is indeed still uniform under the mild assumption of compactness of \( \mathcal{N} \).

**Theorem 1 (Measures over \( \mathcal{F}_y \) and \( SF_y \))** Let \( N \) be a compact set, and let \( \eta \) be uniformly distributed on \( \mathcal{N} \). Then, for any \( y \in Y \) it holds:

(i) The conditional measure \( \mu_{\mathcal{F}_y} \) is uniform over \( \mathcal{F}_y \), that is \( \mu_{\mathcal{F}_y} \equiv \lambda_{\mathcal{F}_y} \).
(ii) The conditional measure \( \mu_{SF_y} \) over \( SF_y \) is log-concave. Moreover, if \( S \in \mathbb{R}^{n,n} \), then this measure is uniform, that is \( \mu_{SF_y} \equiv \lambda_{SF_y} \).

The proof of this theorem is given in Appendix A.

**Remark 1 (Log-concave measures):** Statement (ii) of the theorem proves that the conditional measure on \( SF_y \) is log-concave. We recall that a measure \( \mu \) is log-concave if, for any compact sets \( A, B \) and \( \alpha \in [0, 1] \), the following inequality holds

\[ \mu(\alpha A + (1 - \alpha) B) \geq \mu(A)^\alpha \mu(B)^{1 - \alpha} \]

where \( \alpha A + (1 - \alpha) B \) is a Minkowski sum. Note that the Brunn-Minkowski inequality [44] asserts that the uniform measure over convex sets is log-concave. Furthermore, any Gaussian measure is log-concave.

We now introduce an assumption regarding the solution operator \( S \).

**Assumption 2 (Regularized solution operator):** In the sequel, we assume that the solution operator is regularized, so that \( S = \{ S_{0, n-1} \} \), with \( S \in \mathbb{R}^{n,n} \).

**Remark 2 (On Assumption 2):** Note that the assumption is made without loss of generality. Indeed, for any full row rank \( S \in \mathbb{R}^{n,n} \), we introduce the change of variables \( T = [T_1, T_2] \), where \( T_1 \) is...
provides useful properties from the computational point of view of $\Phi(y)$ variables, equation (1) can be rewritten as

$$z = Sx = STT^Tx = S[T_1 T_2]T^Tx = [ST_1 ST_2]T^Tx = [\tilde{S} \Phi_{\bar{z},-\bar{z}}] \tilde{x} = \tilde{S} \tilde{x},$$

where we introduced the new problem element $\tilde{x} \equiv T^Tx$ and the new solution operator $\tilde{S} \equiv ST_1$. Note that, with this change of variables, equation (1) can be rewritten as $y = \tilde{S} \tilde{x}$, which in turn coincides with the classical least squares estimates.

Instrumental to the next developments, we introduce the cylinder in the element space $X$, with given “center” $\bar{z} \in \bar{Z}$ and radius $r$, as follows

$$C(z_c, r) \equiv \{x \in \mathbb{R}^n : ||Sx - z_c|| \leq r\} = S^{-1}(B(z_c, r)) \subset X,$$

that is, $C(z_c, r)$ is the inverse image (pre-image) under the solution operator $S$ of the $\ell_p$ norm-ball $B(z_c, r) \equiv \{z : ||z - z_c|| \leq r\}$. Moreover, due to Assumption 2, the cylinder $C(z_c, r)$ is parallel to the coordinate axes, that is any element $x$ of the cylinder can be written as

$$x \in C(z_c, r) \iff x = \begin{bmatrix} \tilde{S}^{-1} \xi \\ \xi \end{bmatrix},$$

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

$$\Phi(z_c, r) \equiv \mathcal{F}_y \cap C(z_c, r) \subset X,$$

and its volume

$$\phi(z_c, r) \equiv \text{vol}[\Phi(z_c, r)].$$

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

$$\mathcal{H}(r) \equiv \{z_c \in \mathbb{R}^s : \Phi(z_c, r) \neq \emptyset\}.$$

Note that, even if the cylinder $C(z_c, r)$ is in general unbounded, the set $\Phi(z_c, r)$ is bounded whenever $z_c \in \mathcal{H}(r)$, since $\mathcal{F}_y$ is bounded due to Assumption 1.

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 (12).

**Theorem 2:** Under Assumptions 1 and 2, the following statements hold

(i) For given $r > 0$, the optimal violation function $v_o(r)$ is given by

$$v_o(r) = 1 - \frac{\phi_o(r)}{\text{vol}[\mathcal{F}_y]},$$

where $\phi_o(r)$ is the solution of the optimization problem

$$\phi_o(r) = \max_{z_c \in \mathcal{H}(r)} \phi(z_c, r),$$

with $\phi(z_c, r)$ and $\mathcal{H}(r)$ defined in (16) and (17), respectively;

(ii) For given $r > 0$, the function $\phi(z_c, r)$ is quasi-concave\(^3\) for $z_c \in \mathcal{H}(r)$, and the set $\mathcal{H}(r)$ is convex;

(iii) The function $v_o(r)$ is right-continuous and non-increasing for $r > 0$.

The proof of this result is given in Appendix B.

**Remark 3 (Unimodality of the function $\phi(z_c, r)$):** Point (ii) in Theorem 2 is crucial from the computational viewpoint. Indeed, as remarked for instance in [6], a quasi-concave function cannot have local maxima. Roughly speaking, this means that the function $\phi(z_c, r)$ is unimodal, and therefore any local maximal solution of problem (19) is also a global maximum. Note that from the Brunn-Minkowski inequality it follows that, if there are multiple points $z_c^{(i)}$ where $\phi(z_c, r)$ achieves its global maximum, then the sets $\Phi(z_c^{(i)}, r)$ are all homothetic, i.e. they can be obtained via an affine transformation, see [44].

These facts are illustrated in Figure 3, where we plot the function $\phi(z_c, r)$ for the tutorial problem considered in Example 2, for two different values of $r$. In the figure on the left, the two sets $\mathcal{F}_y$ and $C(z_c, r)$ intersect for all considered values of $z_c$, the function is unimodal, and clearly presents a unique global maximum. In the figure on the right, the radius $r$ is smaller, and there are values of $z_c$ for which $C(z_c, r)$ is completely contained in $\mathcal{F}_y$, thus leading to the “flat” region on the top. However, note that this is the only flat region, so that the function is “well-behaved” from an optimization viewpoint.

**Remark 4 (Probabilistic radius and optimal estimate):** Theorem 2 provides a way of computing the optimal probabilistic radius of information $r_{\alpha}^{\text{pr}}(y, \epsilon)$. Indeed, the probabilistic radius of information (to level $\epsilon$) is given by the solution of the following one-dimensional inverse problem

$$r_{\alpha}^{\text{pr}}(y, \epsilon) = \min \{r : v_o(r) \leq \epsilon\}. \quad (20)$$

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

$$z_{\alpha, r}(\epsilon) = A_{\alpha}^{\text{pr}}(y, \epsilon) = z_{\alpha}(r_{\alpha}^{\text{pr}}(y, \epsilon)),\quad (21)$$

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

To illustrate, continuing Example 2, we plot in Figure 4(a) the function $v_o(r)$ for $r \in (0, r_{\alpha}^{\text{pr}})$. We see that $v_o(r)$ is indeed non-increasing (actually, it is strictly decreasing), and hence the inverse problem (20) has clearly a unique solution for any $\epsilon \in (0, 1)$.

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 $\Phi(z_c, r)$, which is in general a very hard task. A notable exception in which the probabilistic optimal estimate is immediately computed for $\eta$ uniformly distributed in $N$ is the special case when the consistency set $\mathcal{F}_y$ is centrally symmetric\(^4\) with center $\bar{x}$. Indeed, in this case it can be seen that $\mathcal{S}\mathcal{F}_y$ is also centrally symmetric around $\bar{x} = S\bar{x}$, and so is the density $\mu_{\mathcal{S}\mathcal{F}_y}$. Hence, the optimal probabilistic estimate coincides with the center $\bar{x}$, since it follows from symmetry that the probability measure of the intersection of $\mathcal{S}\mathcal{F}_y$ with an $\ell_p$ 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.

\(^{3}\) A function $f$ defined on a convex set $A \subset \mathbb{R}^n$ is quasi-concave if $f(\alpha x + (1 - \alpha)y) \geq \min(f(x), f(y))$ holds for any $x, y \in A$ and $\alpha \in (0, 1)$.

\(^{4}\) 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. $(2\bar{x} - x) \in H$. 

Fig. 3: The function $\phi(z_c, r)$ for the tutorial problem considered in Example 2, for $r = 0.0284$ (a) and $r = 0.0150$ (b).

Remark 5 (Weighted $\ell_2$ norms): Note that the requirement of $\mathcal{F}_y$ being centrally symmetric is quite demanding in general, but holds naturally when (weighted) $\ell_2$ norms are considered, that is when $\eta$ is uniformly distributed in a unit ball

$$\mathcal{N} = \{\eta : \|\eta\|_W \leq \rho\}, \text{ } \|\xi\|_W = \sqrt{\xi^T W \xi}, \text{ } W = W^T \succ 0,$$

with $W \succ 0$ meaning positive definite, and $\mu_N = \lambda_N$. This framework has been also considered in the classical set-membership literature, see for instance [29], and it is well-known that in this case the set $\mathcal{F}_y$ is the ellipsoid

$$\mathcal{F}_y = \{x \in X : \|Ix - y\|_W \leq \rho\} = \{x \in X : (x - x_{ls})^T (I^T W I) (x - x_{ls}) \leq \rho^2\} \quad (21)$$

centered around the (weighted) least-squares optimal parameter estimate

$$x_{ls} = \left(I^T W I\right)^{-1} I^T W y.$$

Hence, it follows from symmetry that, for any $\epsilon \in (0, 1)$, the probabilistic optimal estimate to level $\epsilon$ is given by

$$z_{pr}^{\epsilon}(y) = A_{ls}(y) = S x_{ls}(y).$$

However, we are not aware of any closed-form expression for the corresponding probabilistic optimal radius $r_{pr}^{\epsilon}(y, \epsilon)$, see Section VI for further comments.

Remark 6 (Connections with worst-case and MLE estimates):

Since this paper considers a setup which is somehow in-between classical statistical estimation and set-membership estimation, it is of interest to discuss the differences and analogies between the various approaches. The advantages with respect to the worst-case based set-membership approach, in terms of conservatism reduction, should be evident from the discussion so far, and will be further analyzed in the numerical example of Section VII.

To better clarify the connections with classical stochastic maximum-likelihood estimation (MLE), note that in [52] it is shown that, for the case of uniform noise, the MLE estimates are not unique, and any element of $S \mathcal{F}_y$ is an MLE estimate. Hence, any approach returning estimates belonging to the consistency set is optimal in this sense. In the IBC literature, estimates with the property of belonging to the consistency set are called interpolatory, see eg. [54] for a formal definition. Interpolatory estimates enjoy interesting properties: for instance, it is easy to show that they are almost worst-case...
for Example 2, all relevant quantities have been computed exactly. However, it should be remarked that, for 
probabilistic volume oracle and a stochastic optimization approach. Hence, in this case, our approach can be seen as a tool for selecting an optimal MLE solution.

The situation is more complicated for \ell_1 or \ell_\infty norms, because in this case the central estimate is not always interpolatory. This is a consequence of the fact that the Chebychev center of a convex set may lie outside of the set for non-Euclidean norms, as discussed in Section II-A. Similarly, the probabilistic optimal estimate defined in (9) is not necessarily interpolatory. Furthermore, we note that, in this case also the classical least-squares estimate may lie outside the consistency set \( S_{F_y} \), hence it is not MLE. For instance, in Example 2 the least-squares estimate can be immediately computed as \( x^{ls} = [1.2873 \ 2.3190]^T \) and it is indeed not interpolatory.

An interesting approach, in the case of \ell_1 or \ell_\infty norms, is indeed to consider a conditional probabilistic-optimal estimate, which requires looking for the best interpolatory estimate minimizing the probabilistic radius (8). Note that, from a computational viewpoint, this is immediately obtained constraining the optimization problem (19) to \( z_c \in S_{F_y} \).

V. RANDOMIZED AND DETERMINISTIC ALGORITHMS FOR OPTIMAL VIOLATION FUNCTION APPROXIMATION

In this section, we concentrate on the solution of the optimal violation function (18) defined in Theorem 2 for fixed \( r > 0 \), and in particular on the optimization problem (19). For simplicity, we restate this problem dropping the subscript from \( z_c \).

\( \text{(P-max-int)} \colon \max_{z \in \mathcal{H}(r)} \phi(z, r), \quad \phi(z, r) = \text{vol} [F_y \cap C(z, r)] . \) \hspace{1cm} (22)

First, note that this problem is computationally very hard in general. For instance, for \ell_1 or \ell_\infty norms, the consistency set \( F_y \) is a polytope and \( C(z, r) \) is a cylinder parallel to the coordinate axes whose cross-section is a polytope. Hence, even evaluating the function \( \phi(z, r) \) appearing in (22) amounts to computing the volume of a polytope, and this problem has been shown to be NP-hard in [32]. Similarly, computing \( v_0(r) \) in (18) requires the computation of the volume of the polytope \( F_y \).

Remark 7 (Volume oracle and oracle-polynomial-time algorithm): For the case of polytopic sets, the papers [1], [21] study the problem \( \text{(P-max-int)} \) in the hypothetical setting that an oracle exists which satisfies the following property: given \( r > 0 \) and \( z \in \mathcal{H}(r) \), it returns the value of the function \( \phi(z, r) \), together with a sub-gradient of it. In this case, in [1] a strongly polynomial-time (in the number of oracle calls) algorithm is derived. Note that, even if the problem is NP hard in general, one can compute the volume of a polytope in a reasonable time for considerably complex polytopes in modest (e.g. for \( n \leq 10 \) dimensions), see [7]. In this particular case, for \ell_\infty norms, the method proposed in [21] may be used. For instance, for Example 2, all relevant quantities have been computed exactly by employing this method. However it should be remarked that, for larger dimensions, the curse of dimensionality makes the problem computationally intractable, and alternative methods need to be devised.

In the next subsections, we develop random and deterministic relaxations of problem \( \text{(P-max-int)} \) which do not suffer from these computational drawbacks.

A. Randomized algorithms for computing \( \text{(P-max-int)} \)

In this section, we propose randomized algorithms based on a probabilistic volume oracle and a stochastic optimization approach.

for approximately solving problem \( \text{(P-max-int)} \) for \( \ell_p \) norms. First of all, we compute a bounded version of the cylinder \( C(z, r) \). To this end, we note that bounds \( x^-_i, x^+_i \) on the variables \( x_i, i = s+1, \ldots, n \), can be obtained as the solution of the following \( 2(n-s) \) convex programs,

\[
\begin{align*}
 x^-_i &= \min x_i & \text{subject to } & x \in \mathcal{F}_y^i, \quad i = s+1, \ldots, n. \\
 x^+_i &= \max x_i & \text{subject to } & x \in \mathcal{F}_y^i, \quad i = s+1, \ldots, n. 
\end{align*}
\] \hspace{1cm} (23)

The problems above are convex, and for \( \ell_p \) norms can be solved for instance by (sub)gradient-based or interior point methods. In particular, problem (23) reduces to the solution to \( 2(n-s) \) linear programs for \( \ell_1 \) or \( \ell_\infty \) norms. Then, under Assumption 2, we define the cylinder

\[
\bar{C}(z, r) = \left\{ x \in \mathbb{R}^n : \| \tilde{S} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} - z \| \leq r, \quad x^-_i \leq x_i \leq x^+_i, i = s+1, \ldots, n \right\}.
\] \hspace{1cm} (24)

Note that the cylinder \( \bar{C}(z, r) \) is bounded, and has volume equal to

\[
\text{vol} [\bar{C}(z, r)] = \left( \frac{2r}{\Gamma(1/p+1)} \right)^n \frac{1}{\text{det}(\tilde{S})} \int_{1}^{n} \left( x^+_i - x^-_i \right) \text{d}V_r \] \hspace{1cm} (25)

where \( \Gamma(\cdot) \) denotes the Gamma function. By construction, we have that, for any \( r > 0 \) and \( z \in \mathcal{H}(r) \), \( \bar{C}(z, r) = F_y \cap C(z, r) \). Note that independent and identically distributed (iid) random samples inside \( \bar{C}(z, r) \) can be easily obtained from iid uniform samples in the \( \ell_p \)-norm ball, whose generation is studied in [9]. Then, a probabilistic approximation of the volume of the intersection \( \bar{C}(z, r) \) may be computed by means of the randomized oracle presented in Algorithm 1, which is based on the uniform generation of iid samples in \( \bar{C}(z, r) \).

Algorithm 1 Probabilistic Volume Oracle

1. RANDOM GENERATION

Generate \( N \) iid uniform samples \( \zeta^{(1)}, \ldots, \zeta^{(N)} \) in the \( s \)-dimensional \( \ell_p \)-ball \( B(z, r) \).

\begin{itemize}
\item For \( i = 1 \) to \( N \),
\item Generate \( s \) iid scalars \( \gamma_j \) according to the univariate Gamma density
\begin{equation}
G_{a,b}(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b}, \quad x \geq 0,
\end{equation}
\end{itemize}

with parameters \( a = 1/p, b = 1 \).

\begin{itemize}
\item Construct the vector \( \varpi \in \mathbb{R}^n \) of components \( \varpi_j = s_j^{-1/p} \), where \( s_j \) are iid random signs
\item Let \( \zeta^{(i)} = z + r \varpi \) uniformly in \( [0, 1]^n \).
\end{itemize}

End for

Generate \( N \) iid uniform samples \( \zeta^{(1)}, \ldots, \zeta^{(N)} \).

\begin{itemize}
\item For \( i = 1 \) to \( N \),
\item Generate \( \zeta^{(i)} \) uniformly in the interval \( [x^{-}_{s+j}, x^{+}_{s+j}], j = 1, \ldots, n-s \).
\end{itemize}

End for

Construct the random samples in \( \bar{C}(z, r) \) as follows

\[
\chi^{(i)} = \left[ S^{-1} \zeta^{(i)} \right], \quad i = 1, \ldots, N.
\]
Algorithm 1 Probabilistic Volume Oracle (continued)

2. CONSISTENCY TEST
- Compute the number of samples inside $F_y$ as follows
  \[
  N_y = \sum_{i=1}^{N} \mathbb{I}\left(\|z^{(i)} - y\| \leq \rho\right)
  \]
  where $\mathbb{I}(\cdot)$ denotes the indicator function, which is equal to one if the argument is true, and it is zero otherwise.

3. PROBABILISTIC ORACLE
Return an approximation of the volume $\phi(z, r)$ as follows
\[
\hat{\phi}_N(z, r) = \frac{N}{N^*} V_C
\]
where $V_C$ is defined in (25).

Note that the expected value of the random variable $\hat{\phi}_N(z, r)$ with respect to the samples $z^{(1)}, \ldots, z^{(N)} \in \mathcal{F}(z, r)$ is exactly the volume function $\phi(z, r)$ appearing in (P-max-int) that is
\[
\mathbb{E}\left[\hat{\phi}_N(z, r)\right] = \phi(z, r).
\]
This immediately follows from linearity of the expected value
\[
\mathbb{E}\left[\hat{\phi}_N(z, r)\right] = \mathbb{E}\left[1 \mathbb{N}_{1}^{N} \sum_{i=1}^{N} \mathbb{I}\left(z^{(i)} \in F_y\right) V_C\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[1 \mathbb{I}\left(z^{(i)} \in F_y\right)\right] V_C.
\]
Then, we have
\[
\mathbb{E}\left[1 \mathbb{I}\left(z^{(i)} \in F_y\right)\right] = \text{Prob}\left\{z^{(i)} \in F_y\right\} = \frac{\text{vol}\left(\Phi(z, r)\right)}{\text{vol}\left(\mathcal{C}(z, r)\right)} = \frac{\phi(z, r)}{V_C}.
\]
Hence, we reformulate the problem (P-max-int) as the following stochastic optimization problem
\[
\max_{z \in N(r)} \mathbb{E} \left[\hat{\phi}_N(z, r)\right].
\]
This problem is classical and different stochastic approximation algorithms have been proposed, see for instance \cite{34}, \cite{45} and references therein. In particular, in this paper, we use the SPSA (simultaneous perturbations stochastic approximation) algorithm, first proposed in \cite{47}, and further discussed in \cite{49}. Convergence results under different conditions are detailed in the literature, see in particular the paper \cite{26} which applies to non-differentiable functions.

Remark 8 (Scenario-based algorithms): An alternative approach based on randomized methods can be also devised employing results on the scenario optimization method introduced in \cite{8}. In particular, exploiting the results on discarded constraints, see \cite{12}, \cite{14}, an alternative algorithm can be constructed. The idea is as follows: (i) generate $N$ samples $z^{(i)} \in F_y$ according to the conditional measure $\mu_{F_y}$, ii) solve the discarded-constraint random program
\[
\min_{z, \gamma} \gamma \quad \text{s.t.} \quad \frac{1}{L} \sum_{i \in I_L} \mathbb{I}\left(\|S x^{(i)} - z\| \geq \gamma\right) \leq \epsilon
\]
where $I_L$ is a set of $L$ indices constructed discarding in a prescribed way $N - L$ indices from the set 1, 2, $\ldots$, $N$. Then, in \cite{12}, \cite{14} it is shown how to choose $N$ and the discarded set $I_L$ to guarantee, with a prescribed level of confidence, that the result of optimization problem (27) is a good approximation of the true probabilistic radius $r_{\alpha}^o(g, \epsilon)$. However, this approach entails some non-trivial technical challenges, such as the random sample generation in point (i) and the optimal discarding procedure in point (ii), whose detailed analysis goes beyond the scope of this paper.

B. A semi-definite programming relaxation to (P-max-int)

In this section, we propose a deterministic approach to (P-max-int) based on a semidefinite relaxation of the problem for $\ell_\infty$ norms (extensions to $\ell_1$ and $\ell_2$ norms are briefly discussed in Remark 9). First note that, in the case of $\ell_\infty$ norms, $N$ is a hypercube of radius $\rho$ and therefore $F_y$ is the polytope $P_X$ defined by the following linear inequalities
\[
F_y = \{x \in \mathbb{R}^n : \|Ix - y\|_\infty \leq \rho\}
\]
where 1 is a vector of ones, $1 = [1 1 \cdots 1]^T$. Since the exact computation of the volume of the intersection of two polytopic sets is in general costly and prohibitive in high dimensions, as discussed in Remark 7, we propose to maximize a suitably chosen lower bound of this volume. This lower bound can be computed as the solution of a convex optimization problem. The idea is to construct, for fixed $r > 0$, the maximal volume ellipsoid contained in the intersection $\Phi(z, r)$, which requires to solve the optimization problem
\[
\max_{z, x, P_e} \text{vol}\left[E(x, P_e)\right] \quad \text{subject to} \quad E(x, P_e) \subseteq \Phi(z, r),
\]
where the ellipsoid of center $x$ and shape matrix $P_e$ is
\[
E(x, P_e) \triangleq \{x \in \mathbb{R}^n : x = x + P_e w, \|w\|_2 \leq 1\}.
\]
The problem of deriving the maximum volume ellipsoid inscribed in a polytope is a well-studied one, and concave reformulations based on linear matrix inequalities (LMI) are possible, see for instance \cite{5}, \cite{4}. For completeness, we report this result in the next theorem.

**Theorem 3:** Let Assumptions 1 and 2 hold. Then, for given $r > 0$, a center that achieves a global optimum for problem (28) can be computed as the solution of the following semi-definite programming (SDP) problem
\[
\begin{align*}
\min_{z, \epsilon_1, \epsilon_2} \epsilon_1 & 
\text{subject to} \quad P_e \succeq 0 \quad \text{and} \\
\left[\begin{array}{c}
\rho + e_i^T(y - I x) \\
\rho + e_i^T(y - I x)
\end{array}\right] & \begin{bmatrix} I_n & P_{e_i}^T \end{bmatrix} \begin{bmatrix} I_n & P_{e_i}^T \end{bmatrix}^T \succeq 0, \\
\left[\begin{array}{c}
\rho - e_i^T(y - I x) \\
\rho - e_i^T(y - I x)
\end{array}\right] & \begin{bmatrix} I_n & -P_{e_i}^T \end{bmatrix} \begin{bmatrix} I_n & -P_{e_i}^T \end{bmatrix}^T \succeq 0,
\end{align*}
\]
where $e_i$ and $\epsilon_i$ are elements of the canonical basis of $\mathbb{R}^m$ and $\mathbb{R}^p$, respectively. Moreover, for all $r > 0$, $v_{\alpha}^o(r) \geq v_0(r)$, where we defined
\[
v_{\alpha}^o(r) \doteq 1 - \frac{\phi\left(z_{\alpha}^o(r), r\right)}{\text{vol}\left[F_y\right]},
\]
From Theorem 3, it follows that the SDP relaxation leads to a suboptimal violation function \( \psi^\text{av}(r) \). In Figure 5 we compare the optimal ball obtained via Theorem 2 and the suboptimal one obtained via SDP relaxation for the data in Example 2.

**Remark 9 (SDP relaxations for \( \ell_1 \) and \( \ell_2 \)):** An approach identical to that proposed in Theorem 3 can be developed for the \( \ell_1 \) norm, considering that also in this case the sets \( \mathcal{F}_y \) and \( \mathcal{C}(z, r) \) are a polytope and a cylinder with polytopic basis, respectively. Similarly, an algorithm can be devised for the (weighted) \( \ell_2 \) norm. In this case, the volume of an ellipsoid contained in the intersection of \( \mathcal{F}_y \) and \( \mathcal{C}(z, r) \) should be maximized, which are respectively the ellipsoid defined in (21) and a cylinder with spherical basis. It can be easily seen, see e.g. [5], that this latter problem can be rewritten as a convex SDP optimization problem.

**VI. RANDOM UNCERTAINTY: NORMALLY DISTRIBUTED AND CONNECTIONS WITH LEAST-SQUARES**

In this section, we concentrate on the case where the uncertainty \( \eta \) is normally distributed with mean value \( \bar{\eta} \) and covariance matrix \( \Sigma = \sigma^2 I > 0 \), and the set \( N \) coincides with \( \mathbb{R}^m \). This permits to draw a connection between the probabilistic setting and the classical theory of statistical estimation. Indeed, it is well known, see e.g. [30, 35] that the minimum variance unbiased estimate for the linear regression model (1) is given by the Gauss-Markov estimate

\[
x_{\text{ls}}^\text{true} = (\mathcal{I}^\top \Sigma^{-1} \mathcal{I})^{-1} \mathcal{I}^\top \Sigma^{-1} y,
\]

which coincides with the (weighted) least-squares estimate discussed in Remark 5, for \( W = \Sigma^{-1} \).

We first remark that this minimum variance problem falls into the average setting of IBC, see [54]. In particular, we recall that this setting has the objective of minimizing the expected value of the estimation error, that is, for given \( y \), the optimal average radius is defined as

\[
r_{o}^\text{av}(y) = \inf_{A} \left( \mathbb{E} \left[ ||S x - A(y)||^2 \right] \right)^{1/2},
\]

where \( \mathbb{E} [\cdot] \) denotes the expected value taken with respect to the conditional measure \( \mu_{\mathcal{F}_y} \) defined in (7), which is also Gaussian, due to well-known properties of normal measures. It follows that the optimal average estimate is immediately given by

\[
z_{o}^\text{av} = S x_{\text{av}},
\]

for any \( y \in Y \). Moreover, in [54, Chapter 6] it is proven that the optimal average radius does not depend on the measurement \( y \), and it can be computed in closed form as

\[
r_{o}^\text{av}(y) = \sqrt{\text{Trace} \left( S (\mathcal{I}^\top \Sigma^{-1} \mathcal{I})^{-1} S^\top \right)}.
\]

Concerning the probabilistic optimal estimate, we first remark that in the case of normally distributed noise, the definition of the probabilistic radius (8) still applies, observing that the consistency set \( \mathcal{F}_y \) defined in (2) in this case is given by

\[
\mathcal{F}_y = \{ x \in X : y = \mathcal{I} x + \eta; \ \eta \in \mathbb{R}^n \},
\]

and is unbounded. Hence, the “discarded” set \( \mathcal{C} \), in (8) can be also unbounded. Note that this is not an issue, since \( \mu_{\mathcal{F}_y} \) is defined over all \( \mathbb{R}^n \), so that the measure of unbounded sets is well defined.

Similarly to the worst-case and the average settings, the optimality properties of the least-square solution still hold for the probabilistic setting. Indeed, in [54, Chapter 8] it is proven that the optimal probabilistic estimate (to level \( \epsilon \)) for normal distributions is given by

\[
z_{o}^\text{pr} = S x_{\text{in}},
\]

for any \( y \in Y \). Closed-form solutions for the computation of the probabilistic radius \( r_{o}^\text{pr}(\epsilon) \) are not available, and in [54, Chapter 8] the following upper bound is given

\[
r_{o}^\text{pr}(\epsilon, y) \leq \sqrt{2 \ln \frac{\epsilon}{\epsilon}} r_{o}^\text{av}, \text{ for all } y \in Y.
\]

However, it is also observed that this bound is essentially sharp when the noise variance is sufficiently small.

**VII. NUMERICAL EXAMPLE**

As a numerical example, we consider a randomly generated instance of (1) with uniform distributed noise. In particular, \( m = 150 \) random measurements of an unknown \( n = 5 \) dimensional vector were drawn taking (we use here standard Matlab notation)

\[
\mathcal{I} = \text{round}(20 \ast \text{rand}(m, n) - 10)
\]

\[
\eta = \rho \ast (2 \ast \text{rand}(m, 1) - 1),
\]

with \( \rho = 5 \), and considering as “true” parameters \( x_{\text{true}} \) the unit vector. The solution operator was chosen as

\[
S = \begin{bmatrix}
-5 & 10 & -7 & 0 & 0 \\
3 & -4 & 7 & 0 & 0 \\
2 & 6 & 4 & 0 & 0
\end{bmatrix},
\]

leading to \( z_{\text{true}} = S x_{\text{true}} = [-2 \ 6 \ 12]^\top \). First, the optimal worst-case radius and the corresponding optimal solution have been computed by solving 10 linear programs (corresponding to finding the tightest box containing the polytope \( SF_{\mathcal{F}_y} \), see [37]). The computed worst-case optimal estimate is \( z_{o}^\text{wc} = [-1.831 \ 5.839 \ 11.883]^\top \) and the worst-case radius is \( r_{o}^\text{wc}(y) = 0.5791 \). Subsequently, in order to apply the proposed probabilistic framework, we fixed the accuracy level to \( \epsilon = 0.1 \), and computed the probabilistic optimal radius and

![Fig. 5: Optimal \( \ell_\infty \) ball for Example 2, for \( r = 0.0284 \) (in blue), and optimal ball computed as the solution of the SDP relaxation, with corresponding ellipse \( \mathcal{E}(x_E, P_E) \) (in red).](image)
the corresponding optimal estimate according to definitions (9) and (10). In this case, we were still able to use the techniques discussed in Remark 7 for computing (P-max-int) exactly. By employing a simple bisection search algorithm over \( v_0(r) \), the probabilistic radius of information was computed as \( r_{\text{pr}}^I(y, 0.1) = 0.3074 \).

The corresponding optimal probabilistic estimate is given by and \( z_{\text{pr}}^I(y, 0.1) = [-1.773 5.869 11.969]^\top \). Note that the reduction in terms of radius of information is quite significant, in the order of 50%. The intuition behind our approach is illustrated in Figure 6. As shown there, we seek for the optimal “box” discarding a set of probability measure \( \epsilon = 0.1 \). Note that, in this figure, the volume of the “discarded set” is clearly more than 10% of the total volume. The reason of this is that the probability of the discarded set is measured in the (five dimensional) space \( X \). Figure 7 shows a plot of the violation function \( v_0(r) \) computed using the different techniques discussed in this paper. It can be observed that all methods provide very consistent results.

Then, to compare, we run \( N = 10,000 \) random experiments, generating each time a difference instance of (34), and for each we computed the least-square estimate \( z^b = Sx^b \), the worst-case optimal estimate \( z^{wc}_o \), and the probabilistic optimal estimate \( z^I_{\text{pr}} \).

Figure 8 shows the (normalized) relative frequency histograms for the three estimates, while Table I reports the mean and variances of the different estimates and their corresponding errors. It can be seen that both mean and variance of the error of the proposed probabilistic estimate are much smaller than the least-squares one, and also smaller than the worst-case one.

VIII. Conclusions

This paper deals with a rapprochement 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 idea of discarding sets of small measure from the set of deterministic estimates. In the case of uniformly distributed

\[
\begin{align*}
|z^b - z^I_{\text{pr}}|_{\infty} & = 0.4764 \\
|z^{wc}_o - z^I_{\text{pr}}|_{\infty} & = 0.1864 \\
|z^b - z^I_{\text{pr}}|_{\infty} & = 0.1725 
\end{align*}
\]

The authors would like to thank Takayuki Wada for the enlightening discussions about stochastic optimization, and for clarifying the quasi-concavity properties in Theorem 2, Libor Vesely for clarifying properties of the Chebychev center of convex sets, and Marina Santacroce and Barbara Trivellato for their valuable inputs related to conditional measures and probability theory. The authors are also indebted to the anonymous reviewers for the numerous comments that allowed to improve the content of the paper.

APPENDIX A

PROOF OF THEOREM 1

Consider the transformation matrix \( T = [T_1 \ T_2] \), where \( T_1 \) is an orthonormal basis of the column space of \( Z \) and \( T_2 \) is an orthonormal basis of the null space of \( T \). Furthermore, define the linear transformation \( \eta = T\eta \). and the set \( \mathcal{N} = \{ \eta \in \mathbb{R}^m : T^\top \eta \in \mathcal{N}' \} \). Then, if the random variable \( \eta \) is uniform on \( \mathcal{N} \), the linearly transformed random variable \( \bar{\eta} \) is uniform on \( \mathcal{N} \) (see e.g. [24]).

Next, by multiplying equation (1) from the left by \( T_1^\top \) and defining \( \bar{I}_1 = T_1^\top I \) and \( \bar{y}_1 = T_1^\top y \), \( \bar{\eta}_1 = T_1^\top \eta \), \( \bar{\eta}_2 = \bar{\eta}_2 \), we get

\[
\begin{align*}
\bar{I}_1 x + \bar{\eta}_1 = \bar{y}_1 \\
\bar{\eta}_2 = \bar{y}_2,
\end{align*}
\]
Fig. 8: Relative frequencies for least-square estimates (cyan), worst-case optimal estimate (blue) and probabilistic-optimal (green). To allow comparison, the histograms have been normalized so that the area under the bars is equal to one.

since, by construction, $T_\tilde{I} = 0$. It follows from definition (2) that a point $x \in X$ belongs to $\mathcal{F}_y$ if and only if there exists $\tilde{\eta} = \begin{bmatrix} \tilde{\eta}_1 \\ \tilde{\eta}_2 \end{bmatrix} \in \mathcal{N}$ such that (35) holds, i.e. if there exist $\tilde{\eta}_1$ in the set $\mathcal{N}_1 = \{ \tilde{\eta}_1 \in \mathbb{R}^n : [\tilde{\eta}_1] \in \mathcal{N} \}$ for which $\tilde{I}_1 x + \tilde{\eta}_1 = \tilde{y}_1$. Note that the set $\mathcal{N}_1$ represents the intersection of the set $\mathcal{N}$ with the hyperplane $\{ \tilde{\eta} = [\tilde{\eta}_1 \tilde{\eta}_2] \in \mathbb{R}^m : \tilde{\eta}_2 = \tilde{y}_2 \}$. Then, (7) can be rewritten as

$$
\mu_{\mathcal{F}_y}(B) = \mu_\mathcal{N}(\tilde{\eta}_1 \in \tilde{y}_1 - \tilde{I}_1 (B \cap \mathcal{F}_y) | \tilde{\eta}_2 = \tilde{y}_2)
= \mu_\mathcal{N}(\tilde{\eta}_1 = \tilde{y}_1 - \tilde{I}_1 x \text{ for some } x \in (B \cap \mathcal{F}_y) | \tilde{\eta}_2 = \tilde{y}_2).
$$

Note that $\mu[\mathcal{N}] > 0$ due to Assumption 1, which guarantees the existence of the density. Then, since $\tilde{\eta}$ has uniform density over $\mathcal{N}$, it follows that the conditional density over the affine subset $\tilde{\eta}_1 = \tilde{y}_1 - \tilde{I}_1 x$ is also uniform. This proves statement (i).

Point (ii) follows immediately from the fact that the image of a uniform density through a linear operator $\mathbb{R}^n \to \mathbb{R}^s$ with $s \leq n$ is log-concave (see e.g. [41]). Furthermore, if the transformation is one-to-one (i.e. $S \in \mathbb{R}^{n,n}$), then it is uniform.

**APPENDIX B**

**PROOF OF THEOREM 2**

To prove point (i), we first consider equation (12). Recalling that $\mu_\mathcal{N}$ is the uniform measure over $\mathcal{N}$, we write

$$
v_n(r) = \inf \frac{\text{vol} \{ x \in \mathcal{F}_y : \| Sx - A(y) \| > r \}}{\text{vol} \mathcal{F}_y}
= \frac{1}{\text{vol} \mathcal{F}_y} \inf \text{vol} \{ x \in \mathcal{F}_y : \| Sx - z_c \| > r \}
= \frac{1}{\text{vol} \mathcal{F}_y} \inf \text{vol} \{ x \in \mathcal{F}_y : x \notin C(z_c, r) \}
= \frac{1}{\text{vol} \mathcal{F}_y} \inf \text{vol} [\mathcal{F}_y \setminus C(z_c, r)].
$$

Next, we note that this equation can be rewritten as the following maximization problem

$$
v_n(r) = 1 - \frac{1}{\text{vol} \mathcal{F}_y} \sup_{z_c} \text{vol} [\mathcal{F}_y \cap C(z_c, r)] \quad (36)
= 1 - \frac{1}{\text{vol} \mathcal{F}_y} \sup_{z_c} \phi(z_c, r). \quad (37)
$$

The statement in (i) follows immediately considering that this optimization problem can be restricted to the set $\mathcal{H}(r)$ where the intersection is non-empty. The existence of a global maximum is guaranteed because $\mathcal{H}(r)$ is compact and the function $\phi(z_c, r)$ is continuous in $z_c$.

To prove point (ii), we first show that $\mathcal{H}(r)$ is convex. Begin by noting that

$$
z_c \in \mathcal{H}(r) \iff \mathcal{S} \mathcal{F}_y \cap \mathcal{B}(z_c, r) \neq \emptyset \iff d(z_c, S \mathcal{F}_y) \leq r, \quad (38)$$
where the distance $d(z, A)$ of a point $z \in \mathbb{R}^n$ to a given set $A \subseteq \mathbb{R}^n$ is defined as $d(z, A) = \inf_{x \in A} \|z - x\|$. Since the distance function to convex sets is convex, see e.g. [6, Section 3.2.5], it follows, from convexity of $S F_y$, that $d(z, S F_y)$ is also convex. Hence, given $z_1^*, z_2^* \in \mathcal{H}(r)$, it follows that:

$$d(\alpha z_1^* + (1 - \alpha) z_2^*, S F_y) \leq \alpha d(z_1^*, S F_y) + (1 - \alpha) d(z_2^*, S F_y) \leq r$$

which implies

$$\alpha z_1^* + (1 - \alpha) z_2^* \in \mathcal{H}(r).$$

Then, note that problem (19) corresponds to maximizing the volume of the intersection $\Phi(z_c, r)$ between the two convex sets $F_y$ and $C(z_c, r)$. One of them, $F_y$, is fixed, while the other one is the set obtained translating the cylinder $C(0, r) = \{S_{1/r} z_c \mid z_c \in \mathbb{R}^n, 0 \leq z_c \leq 1\}$. Similar problems have been studied in convex analysis, see for instance [58].

In particular, the proof of continuity follows closely the proof of Lemma 4.1 in [21]. That is, consider an arbitrary direction $\xi \in \mathbb{R}^n$, and let $V_{\xi}$ be the volume of the set obtained projecting $C(0, r)$ to the hyperplane normal to $\xi$. Then, for any $\epsilon > 0$, we have that the difference between the volume of $\Phi(z_c, r)$ and $\Phi(z_c + \epsilon \xi, r)$ is bounded by $\epsilon V_{\xi}[\xi]$. Hence, $\Phi(z_c, r) - \Phi(z_c + \epsilon \xi, r)$ converges to zero for $\epsilon \to 0$, thus proving continuity.

To prove quasi-convexity, consider two points $z_1, z_2 \in \mathcal{H}(r)$ such that $\Phi(z_1, r) > \Phi(z_2, r)$. Consider then a point $z_0 = \alpha z_1 + (1 - \alpha) z_2$ where $\alpha \in (0, 1)$. From convexity of $\mathcal{H}(r)$ it follows that $z_0 \in \mathcal{H}(r)$.

Then, the following chain of inequalities holds

$$\Phi(z_0, r)^{1/n} = \Phi(\alpha z_1 + (1 - \alpha) z_2)^{1/n}$$

$$\geq \alpha \Phi(z_1, r)^{1/n} + (1 - \alpha) \Phi(z_2, r)^{1/n}$$

which follows from [58, Theorem 1, (42)] from the Brunn-Minkowski inequality for convex analysis [44] and (42) follows from the hypothesis that $\Phi(z_1, r) > \Phi(z_2, r)$. From this chain of inequalities, we have $\Phi(z_0, r) > \Phi(z_r, r)$, which implies semi-strict quasi-convexity. Combining continuity and semi-strict quasi-convexity one finally gets quasi-convexity [20].

To prove point (iii), we note that $\nu_0(r)$ is right continuous and non-increasing if and only if $\phi_0(r)$ is upper semi-continuous and non-decreasing. To show upper semi-continuity of the supremum value function $\phi_0(r)$, consider the radius $r = r_0 + \gamma$, which is nonzero since $\mathcal{H}(r)$ is assumed non-empty. Then, from point (ii) it follows that, for any $z \in \mathcal{H}(r)$, the upper level set $\mathcal{F}(z) = \{x \in \mathcal{H}(r) \mid \phi_0(r) \geq \phi_0(z, r)\}$ is strictly convex. Hence, the function $\phi_0(r, x)$ is quasi-convex, continuous and satisfies the boundedness condition defined in [33]. Then, upper semi-continuity of $\phi_0(r)$ follows from direct application of [33, Theorem 2.1]. Finally, to show that $\phi_0(r)$ is non-decreasing, take $0 < r_1 < r_2$ and denote $z_1$ and $z_2$ be the optimal solutions corresponding to $\phi_0(r_1)$ and $\phi_0(r_2)$, respectively. It follows that

$$\phi_0(z_1, r_1) \leq \phi_0(z_2, r_2) = \phi_0(r_2).$$  

since $z_0$ is the point where the maximum is attained. On the other hand, from definition (15) and $r_1 < r_2$ we have

$$\Phi(z_0, r_1) = (F_{y_1} \cap C(z_0, r_1)) \subseteq (F_{y_2} \cap C(z_0, r_2))$$

and hence

$$\phi_0(r_1) = \phi_0(z_0, r_1) \leq \phi_0(z_0, r_2).$$

Combining (46) and (47) it follows $\phi_0(r_1) \leq \phi_0(r_2)$.

REFERENCES


Roberto Tempo is currently a Director of Research of Systems and Computer Engineering at the institute IEIT, National Research Council (CNR), Politecnico di Torino, Italy. He has held visiting positions at various institutions including Chinese Academy of Sciences in Beijing, Kyoto University, The University of Tokyo, University of Illinois at Urbana-Champaign and Columbia University in New York.

Dr. Tempo’s research activities are focused on analysis and design of complex systems with uncertainty, and various applications within information technology. He is author or co-author of more than 190 research papers published in journals, books and conferences. He is a co-author of the book “Randomized Algorithms for Analysis and Control of Uncertain Systems, with Applications” Springer-Verlag, London, 2013. He is a Fellow of the IEEE and a Fellow of the IFAC. He received the “Outstanding Paper Prize Award” from the IFAC for a paper published in Automatica and of the “Distinguished Member Award” from the IEEE Control Systems Society.

In 2010 he has served the IEEE Control Systems Society as President. He is currently an Editor and Deputy Editor-in-Chief of Automatica, a Senior Editor of the IEEE Transactions on Automatic Control, a member of the Advisory Committee of the IEEE Transactions on Control of Network Systems, and an Editor at Large of the Asian Journal of Control. He was General Co-Chair of the 52nd IEEE Conference on Decision and Control, which was held in Florence, Italy, 2013.