Available online at www.sciencedirect.com
SCIENCE DIRECT®
Systems & Control Letters 54 (2005) 1219–1232
SYSTEMS & CONTROL LETTERS
www.elsevier.com/locate/sysconle

Near optimal solutions to least-squares problems with stochastic uncertainty

Giuseppe Calafiore[a,*], Fabrizio Dabbene[b]

[a] *Dipartimento di Automatica e Informatica, Politecnico di Torino, Cso Duca degli Abruzzi 24, 10129 Torino, Italy*
[b] *IEIIT-CNR, Politecnico di Torino, Italy*

Received 11 November 2002; received in revised form 16 June 2003; accepted 19 January 2005
Available online 26 May 2005

Abstract

In this paper, we consider least-squares (LS) problems where the regression data is affected by parametric stochastic uncertainty. In this setting, we study the problem of minimizing the expected value with respect to the uncertainty of the LS residual. For general nonlinear dependence of the data on the uncertain parameters, determining an exact solution to this problem is known to be computationally prohibitive. Here, we follow a probabilistic approach, and determine a probable near optimal solution by minimizing the empirical mean of the residual. Finite sample convergence of the proposed method is assessed using statistical learning methods. In particular, we prove that if one constructs the empirical approximation of the mean using a finite number N of samples, then the minimizer of this empirical approximation is, with high probability, an ε-suboptimal solution for the original problem. Moreover, this approximate solution can be efficiently determined numerically by a standard recursive algorithm. Comparisons with gradient algorithms for stochastic optimization are also discussed in the paper and several numerical examples illustrate the proposed methodology.

Keywords: Least-squares; Uncertainty; Robustness; Learning theory; Randomized algorithms; Stochastic gradient methods

1. Introduction

In the standard least-squares (LS) framework, the objective is to determine a solution vector x^* such that the squared Euclidean norm $\|Ax-y\|^2$ of the residual of a (usually over-determined) system of linear equations is minimized. However, in many practical applications the data matrices A, y are not exactly known. This uncertainty in the data can be modeled assuming A, y to be generic, possibly nonlinear functions of a vector of uncertain real parameters

$$A(\delta) \in \mathbb{R}^{m,n}, \quad y(\delta) \in \mathbb{R}^m, \quad \delta = [\delta_1 \ \delta_2 \ \cdots \ \delta_\ell]^{\mathrm{T}},$$

where the uncertain parameter δ is assumed to belong to a given bounded set $\Delta \subset \mathbb{R}^\ell$.

* Corresponding author. Tel.: +39 11 564 7071; fax: +39 11 564 7099.
E-mail addresses: giuseppe.calafiore@polito.it (G. Calafiore), fabrizio.dabbene@polito.it (F. Dabbene).

0167-6911/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.sysconle.2005.01.006

To solve the LS problem in face of uncertainty, two main approaches are possible. In the robust, or worst-case, approach one looks for a min/max solution: let

$$f(x,\delta) \doteq \|A(\delta)x - y(\delta)\|^2, \tag{1}$$

then a robust least-squares (RLS) solution is one that minimizes the worst-case residual against the uncertainty, i.e.

$$x^*_{wc} = \arg \min_x \max_{\delta\in\Delta} f(x,\delta). \tag{2}$$

This worst-case framework is discussed for instance in the papers [2,3,12], and is closely related to Tikhonov-type regularization [16].

Alternatively, one can take a probabilistic viewpoint, and assume a stochastic nature of the uncertainty. In this case, a probability distribution $p_\delta(\delta)$ is assumed on the set Δ, and one looks for a solution minimizing the expected value of the residual

$$x^*_E = \arg \min_x E_\delta[f(x,\delta)]. \tag{3}$$

We refer to problem (3) as the least-squares with stochastic uncertainty (LSSU) problem. Unfortunately, both problems (2) and (3) are numerically hard to solve. In [3] it is shown that the deterministic problem (2) is in general NP-hard. When the uncertainty enters the data in a rational manner, it is possible to compute a suboptimal solution that minimizes an upper bound on the optimal worst-case residual, using semi-definite relaxations, see [3]. In [2,12] a solution with lower computational complexity is derived for the case of unstructured uncertainty in the data A, b. However, no exact efficient method is known for the general structured nonlinear case. Similarly, in the stochastic problem (3), even the mere evaluation of the objective function, for fixed x, can be numerically prohibitive, since it amounts to the computation of a multi-dimensional integral.

In this paper, we focus on the solution to the LSSU problem (3). Indeed, this problem falls in the general family of stochastic optimization programs, see for instance the survey [20]. Since, in general, one cannot compute exact expectations, a usual initial step in stochastic optimization is to use random sampling to construct an approximation of the original objective, and then compute a candidate solution with respect to this approximation. Known methods for stochastic programming then provide convergence results and confidence intervals for the optimal solution [4,8,9,13]. A drawback of these results is that they are of asymptotic nature and do not provide explicit bounds on the *number of samples* (which impacts on the number of iterations) needed to reach a satisfactory solution.

In this paper, we propose a new solution concept based on probabilistic levels. In particular, we show that a solution obtained by minimizing an empirical version of the mean, constructed using a finite number N of samples, results to be ε-suboptimal with high probability, for the minimization of the actual unknown expectation.

The paper is organized as follows. In Section 1.1 the notation is set and the main assumptions used throughout the paper are stated. To illustrate the LSSU framework, in Section 2 we discuss a particular case of this problem where the expected value can be explicitly computed, and observe that the LSSU problem reduces to regularized deterministic LS, which can be solved via standard methods. The general case, when the expectation cannot be computed explicitly, is discussed in Section 3. In this section, we present the Learning Theory approach to stochastic optimization, and state the main result of the paper in Theorem 2. Section 3.1 discusses a simple technique for numerical computation of the approximate solution. Section 4 discusses an alternative approach to confidence level solutions for LSSU, based on the stochastic gradient methods, recently proposed in [11]. Section 5 presents some numerical examples and comparisons. Conclusions are drawn in Section 6.

1.1. Notation and assumptions

Given a function $g(\delta) : \Delta \to \mathbb{R}$, and a probability density $p_\delta(\delta)$, the expected value operator on $g(\delta)$ is defined as

$$E_\delta[g(\delta)] = \int_{\delta\in\Delta} g(\delta)\, p_\delta(\delta)\, \mathrm{d}\delta.$$

Given N independent identically distributed (i.i.d.) samples $\delta^{(1)}, \ldots, \delta^{(N)}$ drawn according to $p_\delta(\delta)$, the

empirical expectation operator on $g(\delta)$ is defined as

$$\hat{E}_N[g(\delta)] = \frac{1}{N}\sum_{i=1}^{N} g(\delta^{(i)}).$$

Consider the function

$$\phi(x) \doteq E_\delta[f(x,\delta)], \tag{4}$$

where $f(x,\delta) = \|A(\delta)x - y(\delta)\|^2$, and let $\Delta \subset \mathbb{R}^\ell$ be a bounded set. Furthermore, denote by x^* a minimizer of $\phi(x)$, i.e.

$$x^* \doteq \arg\min_{x\in\mathbb{R}^n} \phi(x). \tag{5}$$

We assume that we know a-priori that the solution x^* is contained in a ball $\mathcal{X} \subset \mathbb{R}^n$ of center x_0 and radius $R < \infty$

$$\mathcal{X} \doteq \{x \in \mathbb{R}^n : \|x - x_0\| \leqslant R\}, \tag{6}$$

and define the achievable minimum as $\phi^* = \min_{x\in\mathcal{X}} \phi(x)$.

Let $f^*(\delta) \doteq \min_{x\in\mathcal{X}} f(x,\delta)$, and assume that the total variation of f is bounded by a constant $V > 0$, i.e.

$$f(x,\delta) - f^*(\delta) \leqslant V, \quad \forall x \in \mathcal{X},\ \forall \delta \in \Delta.$$

This implies that the total variation of the expected value is also bounded by V, i.e.

$$\phi(x) - \phi^* \leqslant V, \quad \forall x \in \mathcal{X}.$$

Notice that we only assume that there exist a constant V such that the above holds, but do not need to actually know its numerical value.

In this paper, $\mathcal{R}(X)$ denotes the linear subspace span by the columns of matrix X, and $\mathcal{N}(X)$ denotes the nullspace of X. For a square matrix P, the notation $P \succ 0$ (resp. $P \succcurlyeq 0$) means that P is symmetric and positive definite (resp. positive semidefinite).

2. Closed form solutions for affine uncertainty

To illustrate the LSSU framework, we consider in this section the special case when the uncertain parameter δ enters the data affinely. It can be easily shown that in this situation the expected value of the LS residual can be computed in closed-form. Therefore, the LSSU problem can be recast as a standard regularized LS problem. The case of generic nonlinear dependence of the data on the uncertain parameters, which is the key focus of this paper, is then treated in Section 3.

To simplify the discussion, we consider the case when only the matrix A is uncertain, i.e.

$$A(\delta) = A_0 + \sum_{i=1}^{\ell} \delta_i A_i, \quad y(\delta) = y.$$

Assume further that $p_\delta(\delta) = p_{\delta_1}(\delta_1)p_{\delta_2}(\delta_2)\cdots p_{\delta_\ell}(\delta_\ell)$ and that $E_\delta[\delta] = 0$, that is the parameters δ_i are zero-mean, independent random variables. For the sequel, only the knowledge of the covariances

$$\sigma_i^2 \doteq E_{\delta_i}[\delta_i^2], \quad i = 1,\ldots,\ell$$

is required. In fact, a standard computation leads to the following closed-form expression for the expected value of $f(x,\delta) = \|A(\delta)x - y\|^2$

$$\phi(x) = E_\delta[f(x,\delta)] = \|A_0 x - y\|^2 + x^{\mathrm{T}} Q x, \tag{7}$$

where

$$Q \doteq \sum_{i=1}^{\ell} \sigma_i^2 A_i^{\mathrm{T}} A_i. \tag{8}$$

The objective function in (7) has the form of a regularized LS objective, and a minimizing solution (which always exists) can be easily computed in closed-form as detailed in the following theorem.

Theorem 1. *Let $A(\delta) = A_0 + \sum_{i=1}^{\ell}\delta_i A_i$, where $A_i \in \mathbb{R}^{m,n}$, $i = 0,\ldots,\ell$ are given matrices, and δ_i, $i = 1,\ldots,\ell$ are independent random uncertain parameters having zero mean and given covariances σ_i^2. Let $y \in \mathbb{R}^m$ be given. Then, the minimizing solutions of*

$$\phi(x) = E_\delta[\|A(\delta)x - y\|^2]$$

are the solutions of the modified normal equations

$$(A_0^{\mathrm{T}} A_0 + Q)x = A_0^{\mathrm{T}} y, \tag{9}$$

where $Q \succcurlyeq 0$ is given in (8)*. A minimizing solution always exists. In particular, when $A_0^{\mathrm{T}} A_0 + Q \succ 0$ the solution is uniquely given by*

$$x^* = (A_0^{\mathrm{T}} A_0 + Q)^{-1} A_0^{\mathrm{T}} y.$$

Proof. Differentiating the convex quadratic objective (7) with respect to x, the first-order optimality conditions yield immediately (9). The only thing that needs to be proved is that these linear equations always admit a solution. Clearly, (9) has a solution if and only if $A_0^{\mathrm{T}} y \in \mathscr{R}(A_0^{\mathrm{T}} A_0 + Q)$, which is implied by $\mathscr{R}(A_0^{\mathrm{T}}) \subseteq \mathscr{R}(A_0^{\mathrm{T}} A_0 + Q)$. Now, since $\mathscr{R}(A_0^{\mathrm{T}}) = \mathscr{R}(A_0^{\mathrm{T}} A_0)$ (see for instance [6, Chapter 2]), solvability of (9) is implied by the condition $\mathscr{R}(A_0^{\mathrm{T}} A_0) \subseteq \mathscr{R}(A_0^{\mathrm{T}} A_0 + Q)$. In turn, this latter condition is equivalent to

$$\mathscr{N}(A_0^{\mathrm{T}} A_0 + Q) \subseteq \mathscr{N}(A_0^{\mathrm{T}} A_0).$$

This inclusion is readily proved as follows: for any $x \in \mathscr{N}(A_0^{\mathrm{T}} A_0 + Q)$, we have that

$$x^{\mathrm{T}}(A_0^{\mathrm{T}} A_0 + Q)x = x^{\mathrm{T}} A_0^{\mathrm{T}} A_0 x + x^{\mathrm{T}} Q x = 0.$$

Since both terms in the sum cannot be negative, it must hold that $x^{\mathrm{T}} A_0^{\mathrm{T}} A_0 x = x^{\mathrm{T}} Q x = 0$, which implies that $x \in \mathscr{N}(A_0^{\mathrm{T}} A_0)$, and this concludes the proof. □

We remark that this result is quite standard, and can be easily extended to the case when the independence assumption on the δ_i's is removed, and the term y is considered uncertain too, see for instance [5]. However, in the case of generic nonlinear functional dependence of A, y on the uncertainty δ, and for generic density $p_\delta(\delta)$, the expectation of the residual cannot be computed in an efficient numerical way (nor in closed-form, in general). This motivates the developments of the next section.

3. Learning Theory approach to expected value minimization

Since the minimization of the expected value $\phi(x)$ is in general numerically difficult (and indeed, as already remarked, even the evaluation of $\phi(x)$ for fixed x may be prohibitive), we proceed in two steps. First, we compute an empirical version of the mean, and then compute a minimizer of this empirical expectation.

A fundamental question at this point is whether the minimum of the empirical expectation converges in some suitable sense to the minimum of the true unknown expectation. Several *asymptotic* results of convergence are available in the stochastic optimization literature, see for instance [8,13,14]. Here however, we depart from these usual approaches, typically based on central limit arguments, and use the Learning Theory framework [17] to provide both asymptotic *and* finite sample convergence results. This approach relies on the law of uniform convergence of empirical means to their expectations. These results are summarized below.

Suppose N i.i.d. samples $\delta^{(1)}, \ldots, \delta^{(N)}$ extracted at random according to $p_\delta(\delta)$ are collected, and the *empirical mean* is computed

$$\hat{\phi}(x) \doteq \hat{E}_N[f(x, \delta)]. \tag{10}$$

The number N of uncertainty samples used to construct $\hat{\phi}(x)$ is here referred to as the *sample size* of the empirical mean. Let $\hat{x}_N$ denote a minimizer of the empirical mean:

$$\hat{x}_N \doteq \arg \min_{x \in \mathbb{R}^n} \hat{\phi}(x). \tag{11}$$

We are interested in assessing quantitatively how close $\phi(\hat{x}_N)$ is to the actual unknown minimum $\phi(x^*)$. To this end, notice first that as x varies over $\mathscr{X}$, $f(x, \cdot)$ spans a family $\mathscr{F}$ of measurable functions of δ, namely

$$\mathscr{F} \doteq \{f(x, \delta) : x \in \mathscr{X}\},$$
$$f(x, \delta) = \|A(\delta)x - y(\delta)\|^2. \tag{12}$$

A first key step is to bound (in probability) the relative deviation between the actual and the empirical mean

$$\frac{|E_\delta[f(\cdot, \delta)] - \hat{E}_N[f(\cdot, \delta)]|}{V}$$

for all $f(\cdot, \delta)$ belonging to the family $\mathscr{F}$. In other words, for given relative scale error $\varepsilon \in (0, 1)$, we require that

$$\Pr\left\{\sup_{x \in \mathscr{X}} \frac{|\phi(x) - \hat{\phi}(x)|}{V} > \varepsilon\right\} \leqslant \alpha(N) \tag{13}$$

with $\alpha(N) \to 0$ as $N \to \infty$. Notice that the uniformity of bound (13) with respect to x is crucial, since x is *not* fixed and known in advance: The uniform "closeness" of $\hat{\phi}(x)$ to $\phi(x)$ is the feature that allows us to perform the minimization on $\hat{\phi}(x)$ instead of on $\phi(x)$. Property (13) is usually referred to as the Uniform Convergence of the Empirical Mean (UCEM) property. A fundamental result of Learning Theory states

that the UCEM property holds for a function class $\mathscr{F}$ whenever a particular measure of the complexity of the class, called the P-dimension of $\mathscr{F}$ (P-DIM ($\mathscr{F}$)), is finite. Moreover, this property holds independently of the probability distribution of the data. The interested reader can refer to the monographs [17,18,15] for formal definitions and further details.

The next lemma shows that the function class (12) under consideration has indeed finite P-dimension, and explicitly provides an upper bound on P-DIM ($\mathscr{F}$).

Lemma 1 (*P-dimension of* $\mathscr{F}$). *Consider the function family* $\mathscr{F}$ *defined in* (12). *Then,*

$$\text{P-DIM}\,(\mathscr{F}) \leqslant 9n.$$

Proof. Let $M = \sup_{x\in\mathscr{X},\delta\in\Delta} f(x,\delta)$, and define the family of binary valued functions $\bar{\mathscr{F}}$, whose elements are the functions

$$\bar{f}(x,\delta,c) \doteq \begin{cases} 1 & \text{if } f(x,\delta)\geqslant c,\\ 0 & \text{otherwise} \end{cases}$$

for $c \in [0, M]$. Then, from Lemma 10.1 in [18], we have that P-DIM ($\mathscr{F}$) = VC($\bar{\mathscr{F}}$), where VC($\bar{\mathscr{F}}$) denotes the Vapnik–Chervonenkis dimension of the class $\bar{\mathscr{F}}$. Notice that the functions in $\bar{\mathscr{F}}$ are quadratic in the parameter vector $x \in \mathbb{R}^n$, therefore a bound on the VC-dimension can be derived from a result of Karpinski and Macintyre [7]:

$$\text{VC}(\bar{\mathscr{F}}) \leqslant 2n\log_2(8e) < 9n. \qquad \square$$

With the above premises, we are in position to state the key result of this paper, which provides an explicit bound on the sample size N needed to obtain a reliable estimate of the minimum of $\phi(x)$.

Theorem 2. *Let* $\alpha, \varepsilon \in (0, 1)$, *and let*

$$N \geqslant \frac{128}{\varepsilon^2}\left[\ln\frac{8}{\alpha} + 9n\left(\ln\frac{32\text{e}}{\varepsilon} + \ln\ln\frac{32\text{e}}{\varepsilon}\right)\right]. \quad (14)$$

Let x^* *be a minimizer of* $\phi(x)$ *defined in* (5), *and let* $\hat{x}_N$ *be a minimizer of the empirical mean* $\hat{\phi}(x)$. *Then, if* $\hat{x}_N \in \mathscr{X}$, *it holds with probability at least* $(1-\alpha)$ *that*

$$\frac{\phi(\hat{x}_N) - \phi(x^*)}{V} \leqslant \varepsilon, \quad (15)$$

that is, $\hat{x}_N$ *is an* ε*-suboptimal solution* (*in the relative scale*), *with high probability* $(1-\alpha)$. *A solution* $\hat{x}_N$ *such that the above holds is called an* $(1-\alpha)$*-probable* ε*-near minimizer of* $\phi(x)$, *in the relative scale* V.

Proof. Consider the function family $\mathscr{G}$ generated by the functions

$$g(x,\delta) \doteq \frac{f(x,\delta) - f^*(\delta)}{V},$$

as x varies over $\mathscr{X}$. The family $\mathscr{G}$ is a simple rescaling of $\mathscr{F}$ and maps Δ into the interval $[0, 1]$, therefore the P-dimension of $\mathscr{G}$ is the same as that of $\mathscr{F}$. Define

$$\phi_g(x) \doteq E_\delta[g(x,\delta)] = \frac{\phi(x) - K}{V} \quad (16)$$

and

$$\hat{\phi}_g(x) \doteq \hat{E}_N[g(x,\delta)] = \frac{1}{N}\sum_{i=1}^{N} g(x,\delta^{(i)}) = \frac{\hat{\phi}(x) - \hat{K}}{V}, \quad (17)$$

where

$$K \doteq E_\delta[f^*(\delta)], \quad \hat{K} \doteq \hat{E}_N[f^*(\delta)] = \frac{1}{N}\sum_{i=1}^{N} f^*(\delta^{(i)}).$$

Notice that a minimizer $\hat{x}$ of $\hat{\phi}(x)$ is also a minimizer of $\hat{\phi}_g(x)$. Then, Theorem 2 in [19] guarantees that, for $\alpha, \nu \in (0, 1)$,

$$\text{PR}\left\{\sup_{g\in\mathscr{G}}\left|E_\delta[g(\delta)] - \hat{E}_N[g(\delta)]\right| > \nu\right\} \leqslant \alpha,$$

holds irrespective of the underlying distribution of δ, provided that

$$N \geqslant \frac{32}{\nu^2}\left[\ln\frac{8}{\alpha} + \text{P-DIM}\,(\mathscr{G})\left(\ln\frac{16\text{e}}{\nu} + \ln\ln\frac{16\text{e}}{\nu}\right)\right]. \quad (18)$$

Applying this theorem with $\nu = \varepsilon/2$, and using the bound P-DIM ($\mathscr{G}$) = P-DIM ($\mathscr{F}$) $\leqslant 9n$ obtained in Lemma 1, we have that, for all $x \in \mathscr{X}$, it holds with

probability at least $(1-\alpha)$ that

$$|\phi_g(x) - \hat{\phi}_g(x)| \leqslant \frac{\varepsilon}{2}. \tag{19}$$

From (19), evaluated in $x = x^*$ it follows that

$$\phi_g(x^*) \geqslant \hat{\phi}_g(x^*) - \frac{\varepsilon}{2} \geqslant \hat{\phi}_g(\hat{x}_N) - \frac{\varepsilon}{2}, \tag{20}$$

where the last inequality follows since $\hat{x}_N$ is a minimizer of $\hat{\phi}_g$. From (19), evaluated in $x = \hat{x}_N$ it follows that

$$\hat{\phi}_g(\hat{x}_N) \geqslant \phi_g(\hat{x}_N) - \frac{\varepsilon}{2},$$

which substituted in (20), gives

$$\phi_g(x^*) \geqslant \phi_g(\hat{x}_N) - \varepsilon.$$

From the last inequality and (16) it follows that

$$\phi(\hat{x}_N) - \phi(x^*) \leqslant \varepsilon V,$$

which concludes the proof. □

Remark 1. Notice that the quality of the approximate solution $\hat{x}_N$ is expressed relative to the total variation scale V. This latter quantity is dependent on the choice of the a-priori set $\mathscr{X}$, and it is clearly non-decreasing with respect to R. This reflects the intuitive fact that the better we can a-priori localize the solution, the better is the assessment we can make on the *absolute-scale* precision to which the solution will actually be computed by the algorithm.

3.1. Numerical computation of $\hat{x}_N$

While Theorem 2 provides the theoretical properties of $\hat{x}_N$, in this section we briefly discuss a simple numerical technique to compute it.

Notice that the objective function $\hat{\phi}(x)$ has a sum-of-squares structure

$$\begin{aligned}\hat{\phi}(x) &= \frac{1}{N}\sum_{i=1}^{N} \|A(\delta^{(i)})x - y(\delta^{(i)})\|^2 \\ &= \frac{1}{N}\|\mathscr{A}x - \mathscr{Y}\|^2,\end{aligned}$$

where

$$\mathscr{A} \doteq \begin{bmatrix} A(\delta^{(1)}) \\ A(\delta^{(2)}) \\ \vdots \\ A(\delta^{(N)}) \end{bmatrix}, \quad \mathscr{Y} \doteq \begin{bmatrix} y(\delta^{(1)}) \\ y(\delta^{(2)}) \\ \vdots \\ y(\delta^{(N)}) \end{bmatrix}.$$

Therefore, an exact minimizer of $\hat{\phi}(x)$ can be readily computed as $\hat{x}_N = \mathscr{A}^\dagger \mathscr{Y}$, where $\mathscr{A}^\dagger$ is the Moore–Penrose pseudo-inverse of $\mathscr{A}$. Remark that, since $\mathscr{A}, \mathscr{Y}$ are functions of $\delta^{(i)}$, $i = 1, \ldots, N$, the resulting solution $\hat{x}_N$ is a random quantity, whose probability distribution is defined over the product space $\Delta \times \Delta \times \cdots \times \Delta$ (N times). The solution $\hat{x}_N$ can be alternatively defined as the result given at the Nth iteration by the following standard recursive form of the LS algorithm, see e.g. [6].

Algorithm 1. *Assuming that $A(\delta^{(1)})$ is full-rank, an exact minimizer $\hat{x}_N$ of the empirical mean* (10) *can be recursively computed as*

$$\begin{aligned}\hat{x}_{k+1} = \hat{x}_k &+ K_{k+1}^{-1} A^{\mathrm{T}}(\delta^{(k+1)}) \\ &\times (y(\delta^{(k+1)}) - A(\delta^{(k+1)})\hat{x}_k),\end{aligned} \tag{21}$$

where

$$K_{k+1} = K_k + A^{\mathrm{T}}(\delta^{(k+1)})A(\delta^{(k+1)}),$$

and the recursion for $k = 1, \ldots, N$ is started with $K_0 = 0$, $\hat{x}_0 = 0$.

To summarize, the solution approach that we propose is the following:

1. Given the a-priori set $\mathscr{X}$, fix the desired probabilistic levels α, ε, and determine the theoretical bound for N given in (14);
2. Compute $\hat{x}_N$. This computation needs random samples $\delta^{(i)}$, $i = 1, \ldots, N$ extracted according to $p_\delta(\delta)$ (see further comments on this point in Remark 2);
3. If $\hat{x}_N \in \mathscr{X}$, then with probability greater than $(1-\alpha)$ this solution is an ε-suboptimal minimizer for $\phi(x)$, in the relative scale V.

Remark 2. For the implementation of the proposed method, two typical situations are possible. In a first situation, we explicitly know the uncertainties distribution $p_\delta(\delta)$ and the functional dependence

$A(\delta), y(\delta)$. In this case one can generate the appropriate random samples $\delta^{(i)}$, $i = 1, \ldots, N$, using standard techniques for random sample generation (see for instance [15]). The probabilistic assessments in Theorem 2 are in this case explicitly referred to the probability measure p_δ. In other practical situations, the uncertainty δ is embedded in the data, and the corrupted data $A(\delta^{(i)}), y(\delta^{(i)})$ are directly available as observations. In this respect, we notice that the results in Theorem 2 hold irrespective of the underlying probability distribution, and hence they can be applied also in the cases where the measure p_δ exists but is unknown. In this case, $\hat{x}_N$ is computed using directly the corrupted data $A(\delta^{(i)}), y(\delta^{(i)})$ relative to the ith experiment, for $i = 1, \ldots, N$, and the results of Theorem 2 hold with respect to the unknown probability measure p_δ.

Remark 3. Notice that, if the iterative Algorithm 1 is used for the computation of $\hat{x}_N$ then, in a specific instance of the problem one may observe practical convergence in a number of iterations much smaller than N. This is in the nature of the results based on the Vapnik–Chervonenkis theory of learning, which provides theoretical bounds that hold a-priori, for any problem instance, and for all possible probability distributions of the uncertainties.Therefore, bound (14) holds always and a-priori (before even starting the estimation experiment), while practical convergence can only be assessed a-posteriori, on a specific instance of the problem. This issue is further discussed in the numerical examples section.

In the next section, we discuss an alternative approach to an approximate solution of the LSSU problem, based on stochastic gradient (SG) algorithms for stochastic optimization [10,11]. In this latter approach, a candidate solution $\hat{x}$ is computed, with the property that its associated cost is a good approximation of the optimal value of the original problem, with high probability. The learning theory approach described previously basically works in two steps: a first step where the empirical mean $\hat{\phi}(x)$ is estimated, and a successive step where a minimizer for it is computed. Contrarily, the SG method bypasses the empirical mean estimation step, and directly searches for a near optimal solution iteratively, following random gradient descent steps. The purpose of the next developments is to specialize the SG approach to the problem under study, and then use these results for comparison with those given in Theorem 2.

4. Stochastic gradient approach

The gradient of the function $f(x, \delta)$ defined in (1) is given by

$$h(x, \delta) = \partial_x f(x, \delta) = 2A^{\mathrm{T}}(\delta)(A(\delta)x - y(\delta)).$$

Assume there exist a constant $L > 0$ such that the norm of the gradient is uniformly bounded by L on $\mathscr{X} \times \Delta$. Consider the following algorithm.

Algorithm 2. *Let $N > 0$ be an a-priori fixed number of steps, and let λ_k, $k = 0, \ldots, N-1$ be a finite sequence of stepsizes, such that*

$$\lambda_k > 0, \ \lambda_k \to 0, \ \text{and} \ \sum_{k=0}^{N-1} \lambda_k \to \infty \ \text{as} \ N \to \infty.$$

Let $\delta^{(0)}, \ldots, \delta^{(N-1)}$ be i.i.d. samples drawn according to $p_\delta(\delta)$, and let $x_0 \in \mathscr{X}$ be an initial guess. Let further $\hat{x}_0 = 0$, $m_0 = 0$, and denote with $[x]_{\mathscr{X}}$ the projection of x onto $\mathscr{X}$, i.e.

$$[x]_{\mathscr{X}} = x_0 + \beta(x - x_0),$$
$$\text{where } \beta = \min\left(1, \frac{R}{\|x - x_0\|}\right).$$

Let the candidate stochastic solution $\hat{x}_N$ be obtained via the following recursion:

$$x_{k+1} = [x_k - \lambda_k h(x_k, \delta^{(k)})]_{\mathscr{X}}, \tag{22}$$

$$\hat{x}_k = \frac{m_{k-1}}{m_k}\hat{x}_{k-1} + \frac{\lambda_k}{m_k}x_k, \tag{23}$$

$$m_k = m_{k-1} + \lambda_k, \tag{24}$$

for $k = 0, \ldots, N-1$.

From a classical result on stochastic optimization of Nemirowskii and Yudin [10], we have that for the solution computed by Algorithm 2 it holds that

$$E[\phi(\hat{x}_N)] - \phi^* \leqslant \frac{R^2 + L^2\sum_{k=0}^{N-1}\lambda_k^2}{2\sum_{k=0}^{N-1}\lambda_k}. \tag{25}$$

In particular, if we choose constant stepsizes $\lambda_k = \lambda = \gamma/\sqrt{N}$, then the right-hand side of (25) becomes $(R^2 +$

$L^2\gamma^2)/(2\gamma\sqrt{N})$, which goes to zero as $O(1/\sqrt{N})$, for $N \to \infty$. If the constants R, L are known, then the optimal choice for γ is $\gamma = R/L$.

The following result, adapted from [11], gives a precise assessment of the quality of the solution obtained using the above algorithm, in terms of probabilistic levels.

Theorem 3. *Let $\alpha, \varepsilon \in (0, 1)$, and let*

$$N \geqslant \frac{1}{\alpha^2\varepsilon^2}\left(\frac{LR}{V}\right)^2. \tag{26}$$

Let x^* be a minimizer of $\phi(x)$ defined in (4)*, and let $\hat{x}_N$ be the outcome of Algorithm* 2*, with stepsizes $\lambda_k = \lambda = R/L\sqrt{N}$. Then, it holds with probability at least $(1-\alpha)$ that*

$$\frac{\phi(\hat{x}_N) - \phi(x^*)}{V} \leqslant \varepsilon, \tag{27}$$

that is, the algorithm returns a $(1-\alpha)$-probable ε-near minimizer of $\phi(x)$, in the relative scale V.

Notice that the update step (21) of Algorithm 1 and (22) of Algorithm 2 have a similar form. In particular, the recursive LS algorithm (Algorithm 1) can be interpreted as a stochastic gradient algorithm with matrix stepsizes defined by the gain matrices K_k^{-1}, as opposed to the scalar stepsizes λ_k appearing in (22). Interestingly however, the theoretical derivations follow two completely different routes, and lead to different bounds on the number N of steps required to attain the desired relative scale accuracy. In particular, bound (26) requires the knowledge of the parameters L, V, which can be hard to determine in practice, but does not depend directly on the problem dimension n. Contrary, bound (14) is independent of the L, V parameters, but depends on n, through the VC-dimension bound.

More importantly, we remark that bound (14) is almost independent of the probabilistic level α, since α appears under a logarithm, while bound (26) has a strong quadratic dependence on α. For this reason, we expect the bound (14) to be better than (26), when a high level of confidence is required.

We also remark that in [11] a modification of Algorithm 2 is also considered, which introduces a mechanism of "averaging from a pool of experts". With this modified approach, a sample bound

$$N \geqslant \frac{1}{2\varepsilon^4}\ln\frac{1}{\alpha}\left(\frac{LR}{V}\right)^2 \tag{28}$$

is obtained. However, while this modified bound improves in terms of the dependence of α, it is considerably worse in terms of the dependence on ε, which now appears with a fourth power.

5. Numerical examples

In the following sections, we illustrate the proposed approach on three numerical examples, and compare its performance with the stochastic gradient approach described in Section 4. In particular, Section 5.1 presents an example on polynomial interpolation, and Section 5.2 discusses a case with affine uncertainty. Also, an application to the problem of receding-horizon state estimation for uncertain systems is proposed in Section 5.3.

5.1. Polynomial interpolation

We consider a problem of robust polynomial interpolation borrowed from [3]. For given integers $n \geqslant 1, m$, we seek a polynomial of degree $n-1$, $p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}$ that interpolates given points (a_i, y_i), $i = 1, \ldots, m$, that is

$$p(a_i) \simeq y_i, \quad i = 1, \ldots, m.$$

If the data values (a_i, y_i) were known exactly, we would obtain a linear equation in the unknown x, with Vandermonde structure

$$\begin{bmatrix} 1 & a_1 & \cdots & a_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & a_m & \cdots & a_m^{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \simeq \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix},$$

which can be solved via standard LS. Now, we suppose that the interpolation points are not known exactly. For instance, we assume that the y_i's are known exactly, while there is interval uncertainty on the abscissae

$$a_i(\delta) = a_i + \delta_i, \quad i = 1, \ldots, m,$$

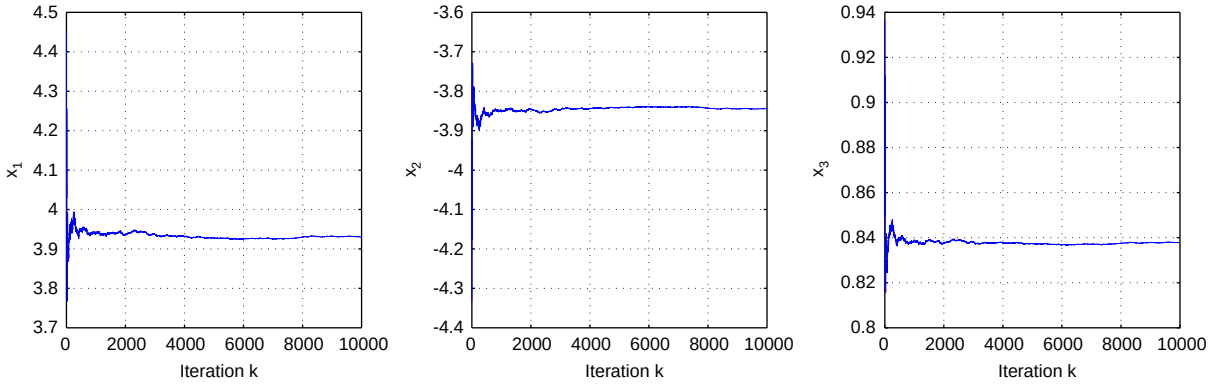

Fig. 1. Convergence of Algorithm 1 for $N = 10,000$ iterations. Solution after N iterations: $\hat{x}_N = [3.926 \ -3.840 \ 0.837]^{\mathrm{T}}$.

where δ_i are uniformly distributed in the intervals $[-\rho, \rho]$, i.e.

$$\Delta = \{\delta = [\delta_1, \ldots, \delta_m]^{\mathrm{T}} : \|\delta\|_\infty \leqslant \rho\}.$$

We therefore seek an interpolant that minimizes the average interpolation error

$$E_\delta[\|A(\delta)x - y\|^2],$$

where

$$A(\delta) = \begin{bmatrix} 1 & a_1(\delta) & \cdots & a_1^{n-1}(\delta) \\ \vdots & \vdots & & \vdots \\ 1 & a_m(\delta) & \cdots & a_m^{n-1}(\delta) \end{bmatrix}.$$

For a numerical example, we considered the data

$$(a_1, y_1) = (1, 1), \quad (a_2, y_2) = (2, -0.5),$$
$$(a_3, y_3) = (4, 2)$$

with uncertainty level $\rho = 0.2$.

The standard LS solution (obtained setting $\delta = 0$) is

$$x_{\mathrm{LS}} = \begin{bmatrix} 4.333 \\ -4.250 \\ 0.917 \end{bmatrix}.$$

We assume the a-priori search set $\mathcal{X}$ to be the ball of radius $R = 10$ centered in $x_0 = x_{\mathrm{LS}}$.

We wish to obtain a solution having relative scale error $\varepsilon = 0.1$ with high confidence $(1 - \alpha) = 0.999$, using Algorithm 1. In this case, the theoretical bound (14) would require $N \geqslant 3,115,043$ samples of the uncertainty. However, as already remarked, while this is the a-priori bound, we can expect practical convergence for much smaller sample sizes. Indeed, in the example at hand, we observe practical convergence of Algorithm 1 already for $N \simeq 10,000$, see Fig. 1.

We then compared the above results to the ones that can be obtained using the stochastic gradient approach of Algorithm 2. To this end, we first performed a preliminary step in order to obtain reasonable estimates of the parameters L, V. With the above choice of $\mathcal{X}$, we obtained the approximate bound $L/V \leqslant 0.25$. Therefore, the theoretical bound (26) would imply the (prohibitive) number of samples $N \geqslant 625,000,000$ to achieve the desired probabilistic levels. Also from a practical point of view, we observed slower convergence with respect to Algorithm 1. Moreover, the behavior of the algorithm appeared to be very sensitive to the choice of the stepsize λ.

The evolution of the estimate for $N = 100,000$, and with $\lambda = 10^{-3}$ is shown in Fig. 2.

5.2. An example with affine uncertainty

We next consider a numerical example with affine uncertainty on the matrix A. Since in this case the solution can be computed exactly as shown in Theorem 1, we can directly test the quality of the randomized solution $\hat{x}_N$ against the exact solution. Let

$$A(\delta) = A_0 + \sum_{i=1}^{3} \delta_i A_i, \quad y^{\mathrm{T}} = [0 \ 2 \ 1 \ 3]$$

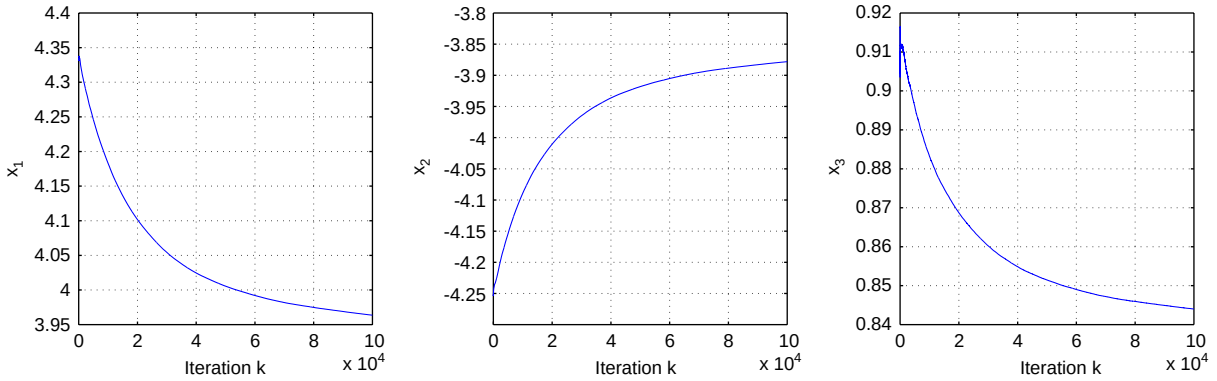

Fig. 2. Evolution of Algorithm 2 for $N = 100,000$ iterations, $\lambda = 10^{-3}$. The algorithm has not yet converged. Solution after N iterations: $\hat{x}_N = [3.961 \ -3.876 \ 0.844]^{\mathrm{T}}$.

with

$$A_0 = \begin{bmatrix} 3 & 1 & 4 \\ 0 & 1 & 1 \\ -2 & 5 & 3 \\ 1 & 4 & 5.2 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and let δ_i be Gaussian random perturbations,[1] with zero mean and standard deviations σ_1=0.067, σ_2=0.1, σ_3=0.2. In this case, the exact solution from Theorem 1 is unique and results to be

$$x^* = \begin{bmatrix} -2.352 \\ -2.076 \\ 2.481 \end{bmatrix}.$$

The standard LS solution (obtained neglecting the uncertainty terms, i.e. setting $A(\delta) = A_0$) results to be

$$x_{\mathrm{LS}} = \begin{bmatrix} -10 \\ -9.728 \\ 9.983 \end{bmatrix},$$

which is quite "far" from x^*, being $\|x_{\mathrm{LS}} - x^*\| = 13.166$. We fix the a-priori search set $\mathscr{X}$ to have center $x_0 = x_{\mathrm{LS}}$, and radius $R = 20$.

To seek a randomized solution having a-priori relative error $\varepsilon = 0.1$ with high confidence $(1 - \alpha) = 0.999$, the theoretical bound (14) would require $N \geqslant 3,115,043$ samples. Fig. 3 shows the first $N = 20,000$ iterations of Algorithm 1, which resulted in the final solution

$$\hat{x}_N = \begin{bmatrix} -2.342 \\ -2.067 \\ 2.472 \end{bmatrix}.$$

We next compared the performance of Algorithm 1 with that of Algorithm 2. For the above choice of $\mathscr{X}$, an estimated bound for the ratio L/V is $L/V \leqslant 0.11$, and therefore the theoretical bound (26) would imply $N \geqslant 484,000,000$ iterations to guarantee the desired probabilistic levels.

Numerical experiments showed that approximate convergence could be reached for $N = 400,000$, with the choice $\lambda = 10^{-2}$, yielding the solution

$$\hat{x}_N = \begin{bmatrix} -2.390 \\ -2.114 \\ 2.519 \end{bmatrix}.$$

We noticed that the SG algorithm failed to converge for larger values of λ. The evolution of the estimate is shown in Fig. 4.

[1] To be precise, *truncated* Gaussian distributions should be considered in our context, since the set Δ is assumed to be bounded. However, from a practical point of view, there is very little difference in considering a genuine zero-mean Gaussian random variable with standard deviation σ, or a truncated Gaussian bounded between, say, -4σ and 4σ.

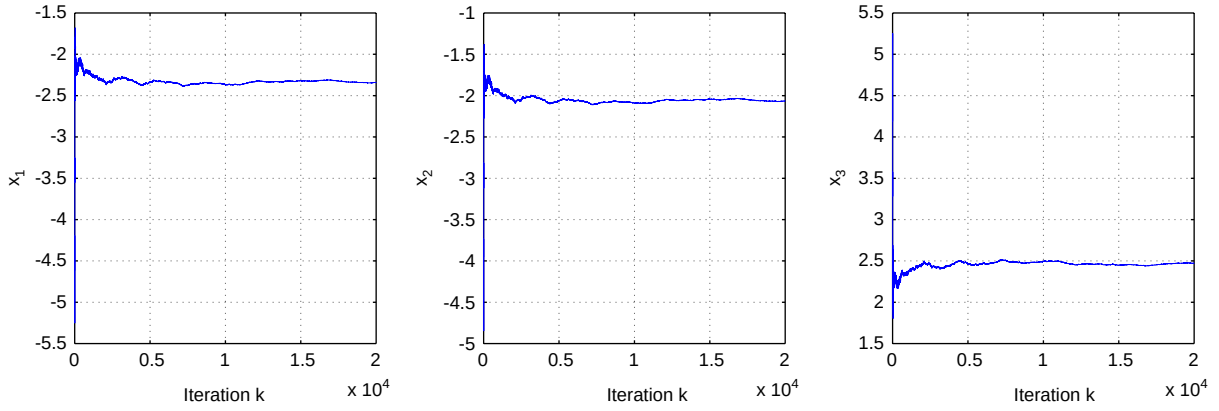

Fig. 3. Evolution of Algorithm 1 for $N = 20,000$ iterations. Solution after N iterations: $\hat{x}_N = [-2.342 \ -2.067 \ 2.472]^{\mathrm{T}}$.

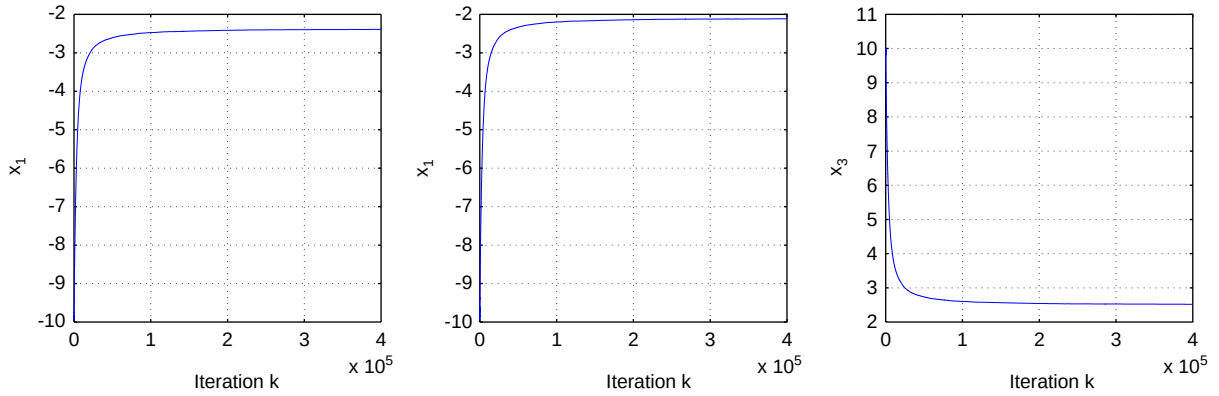

Fig. 4. Evolution of Algorithm 2 for $N = 400,000$ iterations, $\lambda = 10^{-2}$. Solution after N iterations: $\hat{x}_N = [-2.390 \ -2.114 \ 2.519]^{\mathrm{T}}$.

5.3. Receding-horizon estimation for uncertain systems

As a last example, we consider a problem of finite-memory state estimation for discrete-time uncertain linear systems. For systems *without* uncertainty, a LS solution framework for this problem has been recently proposed in [1]. The basic idea of this method is the following: Assume that at a certain time t a prediction $\bar{x}_t$ for the state $x_t \in \mathbb{R}^n$ of the linear system

$$x_{k+1} = Fx_k + \xi_k$$

is available, along with measurements $z_t, \ldots, z_{t+h}$ of the output of the system up to time $t + h$, where the assumed output model is

$$z_k = Cx_k + \eta_k,$$

and the process and measurement noises ξ_k, η_k, as well as the state x_t are assumed to have unknown statistics. The objective is to determine estimates $\hat{x}_t, \ldots, \hat{x}_{t+h}$ of the system states. In [1], the key assumption is made that these estimates should satisfy the nominal state recursions (without noise), i.e. be of the form

$$\hat{x}_{t+k} = F^k \hat{x}_t, \quad k = 0, \ldots, h. \tag{29}$$

From this assumption, it clearly follows that the only quantity that one needs to estimate is $\hat{x}_t$, since all the subsequent state estimates are then determined

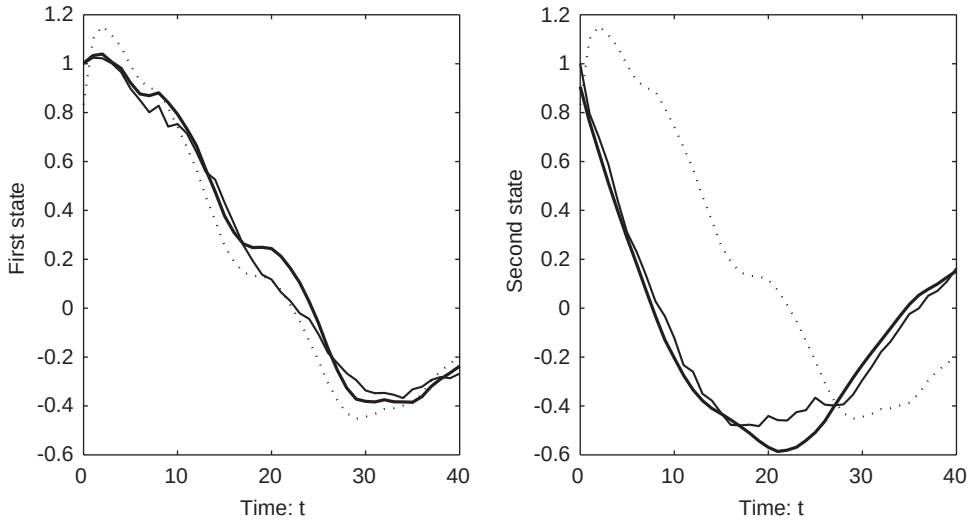

Fig. 5. Smoothing estimates on the states of the uncertain system (30)–(31), obtained by means of the randomized robust filter. Bold lines shows the state estimates obtained by the robust filter, light lines show a simulation of the actual states of the system, and dotted lines show the estimates obtained by the LS filter of [1] that neglects uncertainty. Left figure: first state; right figure: second state.

by (29). From (29), the estimated outputs are in turn given by

$$\hat{z}_{t+k} = CF^k\hat{x}_t, \quad k = 0, \ldots, h,$$

and therefore the natural criterion proposed in [1] is to minimize a least-squares error objective that takes into account the deviations of the estimates $\hat{z}_{t+k}$ from the actual measurements z_{t+k}, as well as an additional term that takes into account one's belief in the accuracy of the initial prediction $\bar{x}_t$. Collecting the output measurement in vector $Z_t \doteq [z_t^T, \ldots, z_{t+h}^T]^{\mathrm{T}}$, and the output estimates in vector $\hat{Z}_t(\hat{x}_t) \doteq [\hat{z}_t^T, \ldots, \hat{z}_{t+T}^h]^{\mathrm{T}}$, the optimization criterion is hence written as

$$J_t(\hat{x}_t) \doteq \mu^2\|\hat{x}_t - \bar{x}_t\|^2 + \|\hat{Z}_t(\hat{x}_t) - Z_t\|^2,$$

where $\mu > 0$ is a scalar weighting parameter. Determining $\hat{x}_t$ such that the above criterion is minimized is a standard LS problem.

Notice that all the above holds under the hypothesis that the model matrices F, C are perfectly known. Here, we now relax this assumption and consider the case where F, C are arbitrary functions of a vector $\delta \in \Delta \subset \mathbb{R}^\ell$ of random uncertain parameters, having probability density $p_\delta(\delta)$. The system hence becomes

$$x_{k+1} = F(\delta)x_k + \xi_k,$$
$$y_k = C(\delta)x_k + \eta_k,$$

and the objective $J_t(\hat{x}_t)$ explicitly writes

$$J_t(\hat{x}_t, \delta) = \|A(\delta)\hat{x}_t - y\|^2,$$

where we defined

$$A(\delta) \doteq \begin{bmatrix} \mu I \\ K(\delta) \end{bmatrix}; \quad y \doteq \begin{bmatrix} \mu\bar{x}_t \\ Z_t \end{bmatrix};$$

$$K(\delta) \doteq \begin{bmatrix} C(\delta) \\ C(\delta)F(\delta) \\ C(\delta)F^2(\delta) \\ \vdots \\ C(\delta)F^h(\delta) \end{bmatrix}.$$

In presence of uncertainty, a sensible estimation approach would therefore amount to determining $\hat{x}_t$ such that the expectation with respect to δ of $J_t(\hat{x}_t, \delta)$ is minimized, i.e.

$$\hat{x}_t^* = \arg\min_{x\in\mathbb{R}^n} \phi(x), \quad \phi(x) = E_\delta[J_t(x, \delta)].$$

A probable ε-near solution for this problem can be determined according to Theorem 2. Notice that, to

the best of the authors' knowledge, no efficient exact method is available for solving this problem. Notice also that even when the uncertainty enters the system matrices $F(\delta), C(\delta)$ in a simple form (such as affine), the data matrix $A(\delta)$ has a very structured and nonlinear dependence on δ. Finally, we remark that applying the estimation procedure in a sliding-window fashion we obtain a finite-memory smoothing filter, in the sense that measurements over the forward time window $t, t+1, \ldots, t+h$ are used to determine an estimate $\hat{x}_t$ of the state at the initial time instant t.

To make a simple numerical example, we modified the model presented in [1], introducing uncertainty. Let therefore

$$F(\delta) = \begin{bmatrix} 0.9950 + \delta_1 & 0.0998 + \delta_2 \\ -0.0998 - \delta_2 & 0.9950 + \delta_3 \end{bmatrix}, \tag{30}$$

$$C(\delta) = [\,1 + \delta_4 \quad 1\,] \tag{31}$$

with $\delta^{\mathrm{T}} \doteq [\delta_1, \ldots, \delta_4]$ and $\delta_1, \ldots, \delta_4$ independent and uniformly distributed in the intervals $\delta_1 \in [-0.1, 0]$, $\delta_2 \in [-0.01, 0.01]$, $\delta_3 \in [-0.1, 0]$, $\delta_4 \in [-0.1, 0.1]$. We selected estimation window $h = 10$, $\mu = 1$ and run Algorithm 1 up to $N = 10,000$ iterations, for each time instant t. Smoothed estimates have been computed over simulation time t from zero to 40. The simulation is run with initial state and initial prediction $x_0 = \bar{x}_0 = [1\ 1]^{\mathrm{T}}$, and process and measurements noises are set to independent Gaussian with standard deviations equal to 0.02 and 0.01, respectively. Fig. 5 shows the results obtained by the robust smoothing filter on this example. Notice the net improvement gained over the LS estimates of [1] which neglected uncertainty.

6. Conclusions

This paper presented a solution approach to stochastic uncertain LS problems based on minimization of the empirical mean. From the computational side, a probable near optimal solution may be efficiently determined by means of a standard recursive LS algorithm that processes at each iteration a randomly extracted instance of the uncertain data. From the theoretical side, a key departure is taken with respect to the standard asymptotic convergence arguments used in stochastic approximation, in that the convergence properties of the method are assessed for finite sample size, within the framework of statistical learning theory. As a result, the numerical complexity of computing an approximate solution can be a-priori bounded by a function of the desired accuracy ε and probabilistic level of confidence α.

The proposed method is compared with existing techniques based on stochastic gradient descent and it is shown to outperform these methods in terms of theoretical sample complexity and practical convergence, as illustrated in the numerical examples.

References

[1] A. Alessandri, M. Baglietto, G. Battistelli, Receding-horizon estimation for discrete-time linear systems, IEEE Trans. Automat. Control 48 (3) (2003) 473–478.
[2] S. Chandrasekaran, G.H. Golub, M. Gu, A.H. Sayed, Parameter estimation in the presence of bounded data uncertainties, SIAM J. Matrix Anal. Appl. 19 (1998) 235–252.
[3] L. El Ghaoui, H. Lebret, Robust solutions to least-squares problems with uncertain data, SIAM J. Matrix Anal. Appl. 18 (4) (1997) 1035–1064.
[4] J.L. Higle, S. Sen, On the convergence of algorithms with implications for stochastic and nondifferentiable optimization, Math. Oper. Res. 17 (1992) 112–131.
[5] H.A. Hindi, S.P. Boyd, Robust solutions to l_1, l_2, and l_∞ uncertain linear approximation problems using convex optimization, in: Proceedings of American Control Conference, vol. 6, 1998, pp. 3487–3491.
[6] T. Kailath, A. Sayed, B. Hassibi, Linear Estimation, Information and System Science, Prentice-Hall, Upper Saddle River, NJ, 2000.
[7] M. Karpinski, A.J. Macintyre, Polynomial bounds for VC dimension of sigmoidal neural networks, in: Proceedings of 27th ACM Symposium on Theory of Computing, 1995, pp. 200–208.
[8] A.J. King, R.T. Rockafellar, Asymptotic theory for solutions in statistical estimation and stochastic optimization, Math. Oper. Res. 18 (1993) 148–162.
[9] W.-K. Mak, D.P. Morton, R.K. Wood, Monte-Carlo bounding techniques for determining solution quality in stochastic programs, Math. Oper. Res. 24 (1999) 47–56.
[10] A. Nemirovskii, D. Yudin, Problem Complexity and Method Efficiency in Optimization, Wiley, Chichester, 1983.
[11] Yu. Nesterov, J.-Ph. Vial, Confidence level solutions for stochastic programming, Stochastic Programming E-Print Series, http://dochost.rz.hu-berlin.de/speps/, 2000.
[12] A.H. Sayed, V.H. Nascimento, F.A.M. Cipparrone, A regularized robust design criterion for uncertain data, SIAM J. Matrix Anal. Appl. 23 (4) (2002) 1120–1142.
[13] A. Shapiro, Asymptotic properties of statistical estimators in stochastic programming, Ann. Statist. 17 (1989) 841–858.

[14] A. Shapiro, Stochastic programming by Monte Carlo simulation methods, Stochastic Programming E-Print Series, http://dochost.rz.hu-berlin.de/speps/, 2000.
[15] R. Tempo, G. Calafiore, F. Dabbene, Randomized Algorithms for Analysis and Control of Uncertain Systems, Springer, London, 2004.
[16] A. Tikhonov, V. Arsenin, Solution to Ill-posed Problems, Wiley, New York, 1977.
[17] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[18] M. Vidyasagar, A Theory of Learning and Generalization, Springer, London, 1997.
[19] M. Vidyasagar, Randomized algorithms for robust controller synthesis using statistical learning theory, Automatica 37 (10) (2001) 1515–1528.
[20] R.J.-B. Wets, Stochastic programming, in: G.L. Nemhauser, A.H.G. Rinnoy Kan, M.J. Todd (Eds.), Optimization, North-Holland, Amsterdam, 1989.