2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Set-valued regression and cautious suboptimization: From noisy data to optimality

Jaap Eising Jorge Cortés

Abstract— **This paper deals with the problem of finding suboptimal values of an unknown function on the basis of measured data corrupted by bounded noise. As a prior, we assume that the unknown function is parameterized in terms of a number of basis functions. Inspired by the informativity approach, we view the problem as the suboptimization of the worst-case estimate of the function. The paper provides closed form solutions and convexity results for this function, which enables us to solve the problem. After this, an online implementation is investigated, where we iteratively measure the function and perform a suboptimization. This nets a procedure that is safe at each step, and which, under mild assumptions, converges to the true optimizer.**

I. Introduction

Optimization of unknown functions on the basis of measured data is a topic with many applications ranging from control to machine learning. For instance, cost or reward functions in modern model-predictive control methods are often partially unknown due to modeling difficulties. On the other hand, sampling the functions is often possible, giving us access to measurements. Accurately determining suboptimal values of such unknown functions is at best a major part of some control objective and at worst safety-critical. This motivates the problem: determine suboptimal points of the unknown function on the basis of noisy measurements. We will investigate this problem both in a *one-shot* setting, where the data are given, and in an *online* setting that allows for repeated measurements.

Literature review: Of course this is not a new problem, and there are many solutions with various setups of which we mention the ones most relevant. In order to solve the one-shot problem, the usual approach is to first employ the data to obtain a unique or, in some sense, 'best' estimate of the unknown function. For this, a number of nonparametric techniques have been developed to estimate unknown functions from data. Popular are e.g. Gaussian processes [1] and methods based on Lipschitz constants [2]–[4]. However, in this paper, we will make the assumption that the unknown function can be parameterized in terms of a number of basis functions or features. Common choices of basis functions are for instance linear, polynomial, Gaussian, or sigmoidal functions. These basis functions allow us to perform regression (see e.g. [5], [6]) on the parameters, leading to the 'best' estimate. Methods differ on what is considered this best estimate: For instance least squares (minimal Frobenius norm), ridge regression (minimal L_2 norm), sparse or Lasso regression (minimal L_1 norm). Analogously are the similar methods that have arisen in the nonlinear system-identification literature. Methods have been developed to determine models that are sparse [7], low rank [8]–[10], or both [11] within a class parameterized using a basis.

After obtaining an estimate of the unknown function, one can treat this estimate as the true unknown function and apply any well-studied optimization technique to obtain suboptimal values. Given the fact that the data is corrupted by noise, it is reasonable to require some robustness from the methods applied (see e.g. [12], [13] and the references therein). Apart from these one-shot optimization problems, we are also interested in online problems, in which we iteratively measure and optimize. This is inspired by methods such as extremum-seeking control [14]–[16], whereas our implementation is markedly different.

In contrast to this paradigm of regression (or: learning) followed by optimization, this paper can be viewed as being in line with the concept of informativity (see [17], [18]), where system properties are investigated for *all* systems compatible with the measurements. This falls within a recent surge of replacing system identification with methods on the basis of Willems' fundamental lemma [19]. While most of these works deal with linear dynamics, the paper [20] dealt with bilinear systems by embedding them into a higher dimensional linear system. Similarly, the works [21], [22] consider systems that are linear in their basis functions.

Statement of contributions: As mentioned, we take a viewpoint related to that of the informativity framework. Using the assumed basis functions, we characterize the set of all parameters compatible with a set of measurements with bounded noise. This is named *set-valued regression*. Clearly, a point can be guaranteed to be suboptimal for the unknown function only if it is for *all* functions corresponding to compatible parameters. This motivates us to introduce *cautious suboptimization*, that is, the problem of finding such values in a way that is robust against the worst-case realization of the parameters. In order to resolve this, we derive closed forms and investigate convexity of this realization. Combining these allows us to resolve the problem using any method from convex optimization. Importantly, this allows us to derive *guaranteed* upper bounds of the optimal value of the unknown functions on the basis of potentially very small data sets.

In addition to this one-shot problem, we investigate an online variant consisting of the iteration of two steps: Collecting

This work was partially funded by ONR Award N00014-23-1-2353 and SNF/FW Weave Project 200021E_20397

Jaap Eising is with the Department of Information Technology and Electrical Engineering at ETH Zürich, Switzerland, jeising@ethz.ch. Jorge Cortés is with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, cortes@ucsd.edu. The bulk of this work was done when J. Eising was affiliated with UC San Diego.

local measurements and performing cautious suboptimization. This gives rise to a procedure which given increasingly sharp upper bounds of the unknown function. Moreover, assuming that the noise is randomly generated in addition to being bounded, we prove that this procedure converges to the true optimal value.

Proofs are omitted for reasons of space and will appear elsewhere.

II. Problem statement

Let[1] $\phi_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, k$ be a collection of known *basis functions* (or *features*) and consider the set consisting of all functions $\phi^\gamma : \mathbb{R}^n \to \mathbb{R}$ linearly parameterized by $\gamma \in \mathbb{R}^k$ as $\phi^\gamma(z) = \Sigma_{i=1}^k \gamma_i \phi_i(z)$. By collecting the basis functions in a vector-function as

$$b(z) := \begin{bmatrix} \phi_1(z) & \cdots & \phi_k(z) \end{bmatrix}^\top,$$

we can write the shorthand $\phi^\gamma(z) = \gamma^\top b(z)$. Consider a function $\hat{\phi} : \mathbb{R}^n \to \mathbb{R}$ which is unknown but can be expressed as a linear combination of the features, i.e., $\hat{\phi}(z) = \phi^{\hat{\gamma}}(z)$, for some *unknown* parameter $\hat{\gamma}$.

We are interested in deducing properties of the function $\hat{\phi}$ on the basis of measurements. Suppose we sample the function for the variables z_i with $i = 1, \ldots, T$ and collect noisy measurements y_i of the true function, that is, $y_i = \hat{\phi}(z_i) + w_i$. Here, the vector w_i denotes an unknown noise, or disturbance, for each i. In order to reason with these measurements in a structured manner, we define

$$Y := \begin{bmatrix} y_1 & \cdots & y_T \end{bmatrix}, \quad W := \begin{bmatrix} w_1 & \cdots & w_T \end{bmatrix}, \qquad (1)$$

$$\Phi := \begin{bmatrix} b(z_1) & \cdots & b(z_T) \end{bmatrix} = \begin{bmatrix} \phi_1(z_1) & \ldots & \phi_1(z_T) \\ \vdots & & \vdots \\ \phi_k(z_1) & \ldots & \phi_k(z_T) \end{bmatrix}.$$

Note that $Y, W \in \mathbb{R}^{1 \times T}$, $\Phi \in \mathbb{R}^{T \times k}$, and $Y = \hat{\gamma}^\top \Phi + W$.

We consider bounded noise. In order to formalize this, let $\Pi \in \mathbb{R}^{(1+\ell)\times(1+\ell)}$ be a symmetric matrix. We partition Π as

$$\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}, \text{ with } \Pi_{11} \in \mathbb{R}, \Pi_{22} \in \mathbb{R}^{\ell \times \ell}.$$

If $\Pi_{22} < 0$, then we denote the Schur complement $\Pi | \Pi_{22} := \Pi_{11} - \Pi_{12}\Pi_{22}^{-1}\Pi_{21}$. We define the set

$$\mathcal{Z}(\Pi) := \left\{ v \in \mathbb{R}^\ell \mid \begin{bmatrix} 1 \\ v \end{bmatrix}^\top \Pi \begin{bmatrix} 1 \\ v \end{bmatrix} \geqslant 0 \right\}.$$

We make the following assumption on the noise model.

Assumption 1 (Noise model)**.** *Let* $\Pi \in \mathbb{R}^{(1+T)\times(1+T)}$ *be symmetric, such that* $\Pi_{22} < 0$, *and* $\Pi|\Pi_{22} \geqslant 0$. *The noise samples satisfy* $W^\top \in \mathcal{Z}(\Pi)$.

With Π as in Assumption 1, the set $\mathcal{Z}(\Pi)$ is nonempty, convex, and bounded. A common example of such a noise model is the case where $WW^\top \leqslant q$, for some $q \geqslant 0$, or as confidence intervals of Gaussian noise. Assuming that the noise signal satisfies Assumption 1, we can define the set of all parameters γ consistent with the measurements by:

$$\Gamma := \{\gamma \in \mathbb{R}^k \mid Y = \gamma^\top \Phi + W, W^\top \in \mathcal{Z}(\Pi)\}. \qquad (2)$$

Thus, if we define $N \in \mathbb{R}^{(1+k)\times(1+k)}$ by

$$N := \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} = \begin{bmatrix} 1 & Y \\ 0 & -\Phi \end{bmatrix} \Pi \begin{bmatrix} 1 & Y \\ 0 & -\Phi \end{bmatrix}^\top, \qquad (3)$$

it follows immediately that $\Gamma = \mathcal{Z}(N)$. Note that $\hat{\gamma} \in \Gamma$ and that we have no further information on the value of $\hat{\gamma}$. We refer to the procedure of obtaining Γ from the measurements as *set-valued regression*.

Remark II.1 (Sufficiently exciting measurements)**.** Note that the set Γ is closed. Moreover, it is bounded if and only if $N_{22} < 0$. Since $N_{22} = \Phi \Pi_{22} \Phi^\top$ and $\Pi_{22} < 0$, this holds if and only if Φ has full row rank. In turn, this requires that the basis functions are not identical and that the set of points z_i is 'rich' enough or sufficiently 'exciting', cf. [19]. •

Remark II.2 (Least-squares estimates)**.** One can check that, if $N_{22} < 0$, then $\gamma^{\text{lse}} := -N_{22}^{-1}N_{21} \in \Gamma$. Therefore $(\gamma^{\text{lse}})^\top b(z)$ is consistent with the measurements. In fact,

$$\begin{bmatrix} 1 \\ -N_{22}^{-1}N_{21} \end{bmatrix}^\top N \begin{bmatrix} 1 \\ -N_{22}^{-1}N_{21} \end{bmatrix} \geqslant \begin{bmatrix} 1 \\ \gamma \end{bmatrix}^\top N \begin{bmatrix} 1 \\ \gamma \end{bmatrix},$$

for any $\gamma^\top \in \mathcal{Z}(N)$. As such, γ^{lse} is the value for which the quadratic inequality is maximal. This leads us to refer to the function $\phi^{\text{lse}}(z; \Gamma) := (\gamma^{\text{lse}})^\top b(z) = -N_{12}N_{22}^{-1}b(z)$ as the *least-squares estimate* of $\hat{\phi}(z)$. •

We are interested in the optimization of the unknown function $\hat{\phi}$. However, based on the measurements, we cannot distinguish between the different functions ϕ^γ for $\gamma \in \Gamma$. Indeed, small changes in the parameter γ might lead to large changes in the quantitative behavior and the location of optimal values of the functions ϕ^γ. In order to be robust against such changes, we consider suboptimization problems instead: for instance, we can conclude that $\hat{\phi}(z) \leqslant \delta$ only if $\phi^\gamma(z) \leqslant \delta$ for all $\gamma \in \Gamma$. This motivates the following.

Problem 1 (Cautious optimization for set-valued regression)**.** Consider an unknown function $\hat{\phi}$, a noise model Π such that Assumption 1 holds, measurements of the true function (Y, Φ), and Γ as in (2). Then,

(a) (Verification of suboptimality): given $z \in \mathbb{R}^n$, find the smallest of $\delta \in \mathbb{R}$ for which $\phi^\gamma(z) \leqslant \delta$ for all $\gamma \in \Gamma$;
(b) (One-shot cautious suboptimization): using the solution to (a), and given a set $\mathcal{S} \subseteq \mathbb{R}^n$, find $z \in \mathcal{S}$ for which (a) yields the minimal value of δ;
(c) (Online cautious suboptimization): determine where to collect new measurements to iteratively improve the bound obtained in (b).

Note that Problem 1 can be posed instead as a question regarding properties of the measurements (Y, Φ), as in the

[1] Throughout the paper, we use the following notation. We denote by $\mathbb{N}$ and $\mathbb{R}$ the sets of nonnegative integer and real numbers, respectively. We let $\mathbb{R}^{n\times m}$ denote the space of $n \times m$ real matrices. For vectors $v \in \mathbb{R}^n$, we write $v \geqslant 0$ (resp. $v > 0$) for elementwise nonnegativity (resp. positivity). The sets of such vectors are denoted $\mathbb{R}^n_{\geqslant 0} := \{v \in \mathbb{R}^n | v \geqslant 0\}$ and $\mathbb{R}^n_{>0} := \{v \in \mathbb{R}^n | v > 0\}$. On the other hand, for $P \in \mathbb{R}^{n\times n}$, $P \geqslant 0$ (resp. $P > 0$) denotes that P is symmetric positive semi-definite (resp. definite). We denote the smallest singular value of $M \in \mathbb{R}^{n\times m}$ by $\sigma_-(M)$. For a set $\mathcal{S} \subseteq \mathbb{R}^n$ we denote the convex hull by $\operatorname{conv}(\mathcal{S})$ and the interior by $\operatorname{int}(\mathcal{S})$.

data informativity framework, e.g. [17], [18]. For instance, given $\delta \in \mathbb{R}$ and $\mathcal{S} \subseteq \mathbb{R}^n$, one could say that the data (Y, Φ) is *informative for δ-suboptimization on $\mathcal{S}$* if there exists $z \in \mathcal{S}$ such that $\phi^\gamma(z) \leqslant \delta$ for all $\gamma \in \Gamma$.

III. One-shot cautious suboptimization

This section addresses Problems 1.(a) and 1.(b). Consider

$$\phi^+(z;\Gamma) := \sup_{\gamma \in \Gamma} \phi^\gamma(z), \qquad \phi^-(z;\Gamma) := \inf_{\gamma \in \Gamma} \phi^\gamma(z), \quad (4)$$

which correspond to the elementwise worst-case realization of the unknown parameter $\hat{\gamma}$. Note that if Γ is compact, cf. Remark II.1, the supremum and infimum are both attained over Γ. Hence they can be replaced by maximum and minimum, respectively, which implies that both $\phi^+(z;\Gamma)$ and $\phi^-(z;\Gamma)$ are finite-valued functions.

Resolving Problem 1.(a) is equivalent to determining function values of $\phi^+(\cdot;\Gamma)$. Similarly, we can reformulate Problem 1.(b) as finding

$$\min_{z \in \mathcal{S}} \max_{\gamma \in \Gamma} \phi^\gamma(z) = \min_{z \in \mathcal{S}} \phi^+(z;\Gamma). \quad (5)$$

This problem takes the form of a minimax or bilevel optimization problem. In this section we first investigate the inner problem of finding values of $\phi^+(z;\Gamma)$, i.e., solving Problem 1.(a). After a detour regarding *uncertainty*, we check this function for convexity. Then, by explicitly finding gradients, we can efficiently resolve Problem 1.(b).

A. Verification of suboptimality

As a first step towards resolving the cautious suboptimization problem we investigate *verification* of suboptimality. That is, *given* $z \in \mathbb{R}^n$, test whether the unknown function is such that $\hat{\phi}(z) \leqslant \delta$. Recall that this problem can be resolved if we can explicitly find function values of the functions in (4). The following result provides closed-form expressions for these functions on the basis of measurements.

Theorem III.1 (Closed-form expressions for bounds)**.** *Assume that the measurements (Y, Φ) are such that $\Gamma = \mathcal{Z}(N)$ with $N_{22} < 0$. Then*

$$\phi^\pm(z;\Gamma) = -N_{12}N_{22}^{-1}b(z) \pm \sqrt{(N|N_{22})b(z)^\top(-N_{22}^{-1})b(z)}.$$

Note that the first term in the closed forms expressions of the functions $\phi^+(\cdot;\Gamma)$ and $\phi^-(\cdot;\Gamma)$ is the least squares estimate, cf. Remark II.2. The result in Theorem III.1 then shows that the difference between the values of true unknown function and the least squares estimate can be quantified in terms of the basis functions and the data, as expressed in N.

As a consequence of Theorem III.1, we have the following result expressing the gradient of the bounding functions.

Corollary III.2 (Gradients in terms of data)**.** *Assume that the measurements (Y, Φ) are such that $\Gamma = \mathcal{Z}(N)$ with $N_{22} < 0$. Let the basis functions ϕ_i be differentiable and such that $b(z) \neq 0$ for all $z \in \mathcal{S}$. Then,*

$$\nabla\phi^\pm(z;\Gamma) = [\nabla\phi_1(z) \cdots \nabla\phi_k(z)] \cdot \\ \left(-N_{22}^{-1}N_{21} \pm \sqrt{N|N_{22}}\frac{(-N_{22}^{-1})b(z)}{\sqrt{b(z)^\top(-N_{22}^{-1})b(z)}}\right).$$

B. Uncertainty of function values

The discussion in Section III-A allows us to find bounds for the unknown function $\hat{\phi}$, but does not consider how much these bounds deviate from its true value. By definition, we have $\phi^-(z;\Gamma) \leqslant \hat{\phi}(z) \leqslant \phi^+(z;\Gamma)$. Therefore, we define the *uncertainty at z* by

$$U(z;\Gamma) := \phi^+(z;\Gamma) - \phi^-(z;\Gamma)$$

to quantify how well we know the function value of $\hat{\phi}(z)$ at $z \in \mathbb{R}^n$. If the uncertainty at z is close to 0, then the function value of $\hat{\phi}(z)$ is quantifiably close to $\phi^{\text{lse}}(z;\Gamma) = -N_{12}N_{22}^{-1}b(z)$. To balance the demands of a low upper bound on the value of the true function with an associated low uncertainty, it is reasonable to consider the following generalization of the cautious suboptimization problem (5): for $\lambda \geqslant 0$, consider

$$\min_{z \in \mathcal{S}} \phi^+(z;\Gamma) + \lambda U(z;\Gamma). \quad (6)$$

Under the conditions of Theorem III.1, we obtain the following closed form for the objective function.

Lemma III.3 (Explicit forms and (6))**.** *Let the measurements (Y, Φ) be such that $\Gamma = \mathcal{Z}(N)$ with $N_{22} < 0$. Then $U(z;\Gamma) = 2\sqrt{(N|N_{22})b(z)^\top(-N_{22}^{-1})b(z)}$. Moreover, if*

$$N_\lambda := \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} + \begin{bmatrix} 4\lambda(1+\lambda)(N|N_{22}) & 0 \\ 0 & 0 \end{bmatrix},$$

and $\Gamma_\lambda := \mathcal{Z}(N_\lambda)$, then $\phi^+(z;\Gamma) + \lambda U(z;\Gamma) = \phi^+(z;\Gamma_\lambda)$.

Lemma III.3 means that, even though problem (6) is more general than the cautious suboptimization problem (5), both problems can be resolved in the same fashion.

C. Convexity and suboptimization

To provide efficient solutions to (5), we investigate when $\phi^+(\cdot;\Gamma)$ is convex. Towards this, we first investigate conditions under which we can guarantee that the true function $\hat{\phi}$ is convex. Since nonnegative combinations of convex functions are convex, the following result identifies conditions that ensure the set of parameters consistent with the measurements are nonnegative.

Lemma III.4 (Test for nonnegativity of parameters)**.** *Given measurements (Y, Φ), let $\Gamma = \mathcal{Z}(N)$, where N is as in* (3)*. Then, $\Gamma \subseteq \mathbb{R}^k_{\geqslant 0}$ if and only if Φ has full row rank and one of the following conditions hold*

1) *The matrix $N \leqslant 0$ and $-N_{22}^{-1}N_{21} \in \mathbb{R}^k_{\geqslant 0}$, or*
2) *The matrix N has one positive eigenvalue, $-N_{22}^{-1}N_{21} \in \mathbb{R}^k_{>0}$ and, for all $i = 1, \ldots, k$,*

$$N|N_{22} + \frac{(e_i^\top N_{22}^{-1}N_{21})^2}{e_i^\top N_{22}^{-1}e_i} \leqslant 0.$$

We can use this result to identify conditions that ensure the convexity of the unknown function and its upper bound.

Corollary III.5 (Convexity of the true function and upper bound)**.** *Suppose that the basis functions ϕ_i are convex and $\Gamma \subseteq \mathbb{R}^k_{\geqslant 0}$.*

- *Then, ϕ^γ is convex for all $\gamma \in \Gamma$ and $\phi^+(\cdot;\Gamma)$ is a finite-valued convex function;*
- *If, in addition, the functions ϕ_i are strictly convex and $0 \notin \Gamma$, then ϕ^γ is strictly convex for all $\gamma \in \Gamma$ and $\phi^+(\cdot;\Gamma)$ is strictly convex.*

Lemma III.4 and Corollary III.5 taken together mean that, if the basis functions are convex, we can test for convexity on the basis of data.

Recall that we are interested in the optimization problem (5), and therefore not necessarily in properties of the true function $\hat{\phi}$, but of its upper bound $\phi^+(\cdot;\Gamma)$. This motivates our ensuing discussion to provide conditions that ensure convexity of the upper bound instead. Note that, under the assumptions of Theorem III.1, we have

$$\phi^+(z;\Gamma) = \phi^{\mathrm{lse}}(z;\Gamma) + \tfrac{1}{2}U(z;\Gamma).$$

Thus, if (i) $\phi^{\mathrm{lse}}(\cdot;\Gamma) = -N_{12}N_{22}^{-1}b(\cdot)$ is convex and (ii) $U(\cdot;\Gamma)$ is convex, then so is $\phi^+(\cdot;\Gamma)$. Moreover, if in addition either is strictly convex, then so is $\phi^+(\cdot;\Gamma)$. Condition (i) could be checked directly if all basis functions ϕ_i are twice continuously differentiable by computing the Hessian of ϕ^{lse}. Here, we present the following simple criterion derived from composition rules, see e.g. [23, Example 3.14], to test for condition (ii).

Corollary III.6 (Convexity of the uncertainty)**.** *Assume that the measurements (Y,Φ) are such that $\Gamma = \mathcal{Z}(N)$ with $N_{22} < 0$. Then $U(\cdot;\Gamma)$ is convex if each basis function ϕ_i is convex and $-N_{22}^{-1}b(z) \geqslant 0$ for all $z \in \mathbb{R}^n$.*

Equipped with the results of this section, one can solve the cautious suboptimization problems (5) and (6) efficiently. Under the assumptions of Theorem III.1, we can write closed-form expressions for $\phi^+(\cdot;\Gamma)$. This, in turn, allows us to test for (strict) convexity using e.g., Corollaries III.5 or III.6. If so, we can apply (projected) gradient descent, using Corollary III.2, to resolve cautious suboptimization.

IV. Online cautious optimization

In this section, we develop an online optimization procedure on the basis of local measurements of the true function to refine the optimality gap. Specifically, we devise a procedure where we first collect data *near* a candidate optimizer, we update a convex upper bound of $\hat{\phi}$ on the basis of the measurements, and lastly we recompute the candidate optimizer on the basis of the updated upper bound.

To formalize this, we require some notation. Let $\mathcal{F} = \{f_i\}_{i=1}^T \subseteq \mathbb{R}^n$ be a finite set. For a given $z \in \mathbb{R}^n$ we measure the function at all points in $z + \mathcal{F}$. For this, define

$$\Phi^{\mathcal{F}}(z) := \begin{bmatrix} \phi_1(z+f_1) & \dots & \phi_1(z+f_T) \\ \vdots & & \vdots \\ \phi_k(z+f_1) & \dots & \phi_k(z+f_T) \end{bmatrix}.$$

Given an initial point z_0, consider measurements at step k,

$$Y_k = \hat{\gamma}^\top \Phi^{\mathcal{F}}(z_k) + W_k, \text{ with } W_k^\top \in \mathcal{Z}(\Pi), \qquad (7)$$

where Y_k and W_k are as in (1). Define the set Γ_k of parameters which are compatible with the k^{th} set of measurements,

$$\Gamma_k := \mathcal{Z}(N_k), \qquad (8)$$

where $N_k := \begin{bmatrix} 1 & Y_k \\ 0 & -\Phi^{\mathcal{F}}(z_k)\end{bmatrix} \Pi \begin{bmatrix} 1 & Y_k \\ 0 & -\Phi^{\mathcal{F}}(z_k)\end{bmatrix}^\top.$

The online optimization procedure then incrementally incorporates these measurements to refine the computation of the candidate optimizer. The following result investigates the properties of the resulting *online gradient descent*.

Theorem IV.1 (Online gradient descent)**.** *Let $\mathcal{F}$ be a finite set such that $0 \in \operatorname{int}(\operatorname{conv}\mathcal{F})$ and such that $\Phi^{\mathcal{F}}(z)$ has full row rank for all z. Define $\mathcal{S}(z) := z + \operatorname{conv}\mathcal{F}$. Consider an initial point $z_0 \in \mathbb{R}^n$ and suppose that the measurements $(Y_0, \Phi^{\mathcal{F}}(z_0))$ are such that ϕ^γ is strictly convex for all $\gamma \in \Gamma_0$. For $k \geqslant 1$, repeat the following two steps iteratively: first update the candidate optimizer*

$$z_k := \operatorname*{arg\,min}_{z \in \mathcal{S}(z_{k-1})} \ \phi^+(z;\Gamma_0 \cap \ldots \cap \Gamma_{k-1}) \qquad (9)$$

and secondly measure the function $\hat{\phi}$ as in (7) *and define Γ_k as in* (8)*. Then the following hold:*

1) *For any $k \geqslant 1$, the problem* (9) *is strictly convex;*
2) *For each $k \geqslant 1$, the algorithm provides an upper bound*

$$\min_{z \in \mathbb{R}^n} \hat{\phi}(z) \leqslant \phi^+(z_k;\Gamma_0 \cap \ldots \cap \Gamma_{k-1}); \qquad (10)$$

3) *The upper bounds are monotonically nonincreasing*

$$\phi^+(z_{k+1};\Gamma_0 \cap \ldots \cap \Gamma_k) \leqslant \phi^+(z_k;\Gamma_0 \cap \ldots \cap \Gamma_{k-1}) \quad (11)$$

4) *If $z_k \neq z_{k+1}$, then* (11) *holds with a strict inequality.*

The algorithm described in Theorem IV.1 provides a sequence of upper bounds to true function values, cf. (10), on the basis of *local* measurements. This means that after *any* number of iterations, we obtain a 'worst-case' estimate of the function value $\hat{\phi}(z_k)$ and, as such, for the minimum of $\hat{\phi}$. Given that this sequence of upper bounds is nonincreasing, cf. (11), and bounded below by the true minimum of $\hat{\phi}$, we can conclude that the algorithm converges. However, without further assumptions, one cannot guarantee convergence of the upper bounds to the minimal value of $\hat{\phi}$, or respectively of z_k to the global minimum of $\hat{\phi}$ (the simulations of Section V below show an example of this precisely).

The following result shows that if the uncertainty is sufficiently small near the optimizer, then the optimizer of the upper bound is close to the global optimizer of the true, unknown function.

Lemma IV.2 (Stopping criterion)**.** *Let Γ be compact and $\mathcal{S} \subseteq \mathbb{R}^n$ closed. Define*

$$\bar{z} := \operatorname*{arg\,min}_{z \in \mathcal{S}} \phi^+(z;\Gamma), \quad \hat{z} := \operatorname*{arg\,min}_{z \in \mathcal{S}} \hat{\phi}(z).$$

Then $\phi^+(\bar{z};\Gamma) \geqslant \hat{\phi}(\hat{z}) \geqslant \phi^+(\bar{z};\Gamma) - \max_{z \in \mathcal{S}} U(z;\Gamma)$.

From Lemma IV.2, we see that if the uncertainty on $\mathcal{S}$ is equal to zero, then the local minima of ϕ^+ and $\hat{\phi}$ coincide. In addition, if $\hat{\phi}$ is strictly convex, we have that any local minimum in the *interior* of $\mathcal{S}$ is equal to its global minimum.

When repeatedly collecting measurements, it seems reasonable to assume that the uncertainty would decrease. However, without making further assumptions, this is not necessarily the case. In particular, a situation might arise where repeated measurements corresponding to a worst-case, noise signal give rise to convergence to a fixed bound with nonzero uncertainty. To address this problem, we consider a scenario where the noise samples are not only bounded but distributed uniformly over the set $\mathcal{Z}(\Pi)$. To make this formal, suppose that Π is such that $\mathcal{Z}(\Pi)$ is bounded. Consider the measure μ and probability distribution over $\mathbb{R}^T$,

$$p(W) := \begin{cases} \frac{1}{\mu(\mathcal{Z}(\Pi))} & W^\top \in \mathcal{Z}(\Pi), \\ 0 & \text{otherwise.} \end{cases}$$

As a notational shorthand, we write $W^\top \sim \operatorname{Uni}(\mathcal{Z}(\Pi))$. The following result shows that, under uniformly distributed noise samples, the uncertainty does indeed decrease.

Theorem IV.3 (Uncertainty under repeated measurements)**.** *Under the assumptions of Theorem IV.1, suppose in addition that for $k \geqslant 1$, the measurements in* (7) *are such that $W_k^\top \sim \operatorname{Uni}(\mathcal{Z}(\Pi))$ and $\sigma_-(\Phi^{\mathcal{F}}(z)) \geqslant a$ for all z. Then for any $z \in \mathbb{R}^n$, the expected value of the uncertainty monotonically converges to 0, that is,*

$$U(z; \Gamma_0 \cap \ldots \cap \Gamma_{k-1}) \geqslant U(z; \Gamma_0 \cap \ldots \cap \Gamma_k),$$

and $\lim_{k\to\infty} \mathbb{E}(U(z; \Gamma_0 \cap \ldots \cap \Gamma_k)) = 0$.

As a consequence of Lemma IV.2 and Theorem IV.3, one can conclude that the expected difference between the optimal value $\min_{z\in\mathcal{S}} \hat{\phi}(z)$ of the unknown function and the optimal value $\min_{z\in\mathcal{S}} \phi^+(z; \Gamma_0 \cap \cdots \cap \Gamma_k)$ provided by online gradient descent both converge to zero.

V. Simulation examples

We illustrate here our results in a simple example. Let the unknown function $\hat{\phi} : \mathbb{R}^2 \to \mathbb{R}$ be given by $\hat{\phi}(z) = 1 + z^\top z$. Let $z = \begin{pmatrix} z_1 & z_2 \end{pmatrix}^\top$ and consider the basis functions

$$\phi_1(z) = 1, \quad \phi_2(z) = z_1, \quad \phi_3(z) = z_2, \quad \phi_4(z) = z^\top z.$$

The value of the true parameter $\hat{\gamma}$ is $\begin{pmatrix} 1 & 0 & 0 & 1 \end{pmatrix}^\top$. We sample the function at points in $z + \mathcal{F}$, where $\mathcal{F} = \{(0,0), (1,0), (0,1), (-1,-1)\}$ (i.e., we measure at the point itself and three points around it). For this choice, $\Phi^{\mathcal{F}}(z)$ has full row rank for all $z \in \mathbb{R}^2$. The measurements are corrupted by a large amount of noise: we assume a noise model of the form $W_k W_k^\top \leqslant 30$ for all $k \geqslant 0$ and that the noise is uniformly distributed in this set. This means that

$$\Pi = \begin{bmatrix} 30 & 0 \\ 0 & -I_4 \end{bmatrix} \text{ and } W_k^\top \sim \operatorname{Uni}(\mathcal{Z}(\Pi)).$$

For z_0, we collect uniform random noisy measurements $(Y_0, \Phi^{\mathcal{F}}(z_0))$, leading to a set of consistent parameters Γ_0. We take $z_0 = \begin{pmatrix} 3 & 3 \end{pmatrix}^\top$ and verify that Γ_0 is such that ϕ^γ is strictly convex for all $\gamma \in \Gamma_0$. On the basis of this, define

$$z_1 := \operatorname*{arg\,min}_{z\in\mathcal{S}(z_0)} \phi^+(z; \Gamma_0).$$

Moreover, we can determine the least-squares estimate of the parameter $\hat{\gamma}$ and the function $\hat{\phi}$. The latter is given by

$$\phi^{\text{lse}}(z; \Gamma_0) = 63.99\phi_1(z) - 23.27\phi_2(z) - 23.27\phi_3(z) + 5.1\phi_4(z).$$

Now we can evaluate the true function, least-squares estimate, and the upper bound, finding the following values:

	z_0	z_1	0
$\hat{\phi}(\cdot)$	19	13.48	1
$\phi^{\text{lse}}(\cdot; \Gamma_0)$	16.26	11.44	63.99
$\phi^+(\cdot; \Gamma_0)$	21.74	16.74	136.45

We can make a few observations. First, the least-squares estimate is quite close at the measured point z_0, yet far at the true minimum of $\hat{\phi}$ at the origin. Further, the upper bound indeed decreases monotonically. Lastly, while the upper bound majorizes $\hat{\phi}$, the least-squares estimate does not.

We run 100 iterations of the online gradient descent algorithm in Theorem IV.1 and plot the results in Figure 1. In spite of the relatively noisy data, we see rapid convergence of the parameters γ and the values of the estimate z_k. Moreover, the upper bound is close to the true value at the measured point. After just 10 steps, we see that the true function value at the estimate is already significantly lower than at the start. To show that this happens regardless of the choice of initial conditions, Figure 2 shows the trajectories of z_k resulting from 8 different initial conditions and the corresponding values of the upper bounds.

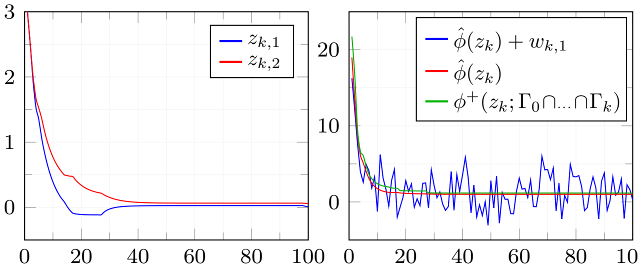

Fig. 1: The simulation results for the initial condition $z_0 = \begin{pmatrix} 3 & 3 \end{pmatrix}^\top$. In the first plot, the resulting trajectory of the estimate z_k, shown elementwise. The second plot shows the measurements corresponding to $z_k + f_1 = z_k$, that is, to $\hat{\phi}(z_k) + w_{k,1}$, where $w_{k,1}$ is the first element of W_k. These measurements are compared to the actual value of the function $\hat{\phi}$ and the current upper bound.

Lastly, we illustrate that worst-case, adversarial noise can lead to convergence to a suboptimal bound. For this, in the last set of simulations and for the same scenario, instead of generating noise randomly, we apply the same noise sample from $\mathcal{Z}(\Pi)$ at each step after $k = 1$. These results can be seen in Figure 3. In particular, note that z_k does not converge to the optimizer at the origin origin, but to the point $\begin{pmatrix} 0.1519 & 0.1519 \end{pmatrix}^\top$. Moreover, since the noise is constant, it can be seen that the upper bound does not converge to the optimal value of the true function.

VI. Conclusions

We have investigated suboptimization for unknown functions on the basis of measurements with bounded noise. Employing ideas from the informativity framework for data-driven control, the notions of set-valued regression and the cautious suboptimization problem were introduced. In short, the data gives rise to a set of possible parameters, and in order to draw conclusions regarding the true function we require bounds for all possible realizations of this parameter. Resolving this problem was shown to be equivalent

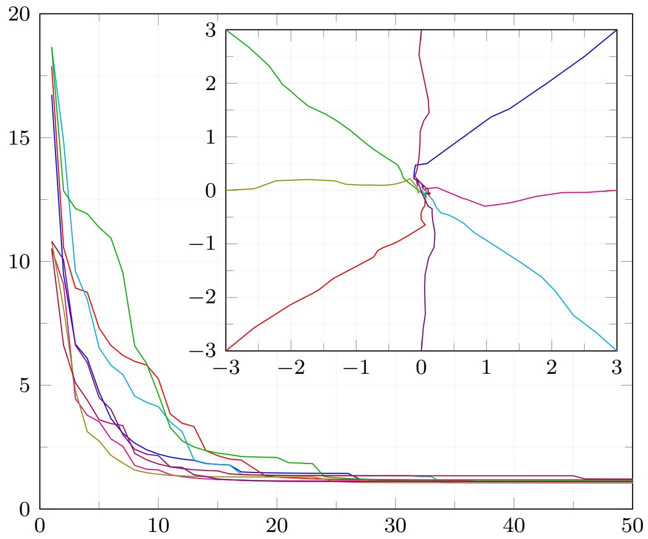

Fig. 2: The simulation results for eight different initial conditions. In the large plot we see the upper bounds corresponding to the different trajectories, which indeed decrease monotonically towards the true minimum $\hat{\phi}(0) = 1$. The inset image shows each of the corresponding trajectories of the estimate z_k, revealing that each tends to the true minimizer.

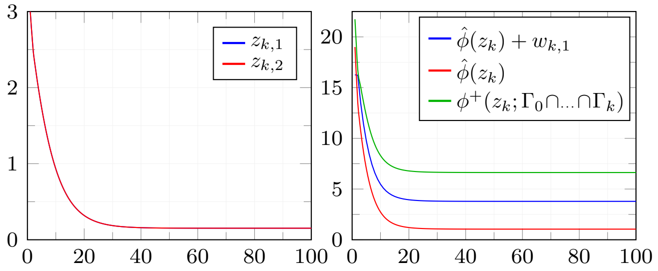

Fig. 3: The simulation results for the initial condition $z_0 = \begin{pmatrix} 3 & 3 \end{pmatrix}^\top$ with nonrandom noise. In particular, the noise samples are equal and proportional to the vector of ones. The figures correspond to those in Figure 1. Note that the elements of the vector z are equal for all time.

to minimization of the worst-case realization. For this, we provided explicit forms and convexity results, allowing efficient solutions. In an online setting, we investigated the iteration of cautious suboptimization and local collection of new measurements. This procedure gives rise to nonincreasing guaranteed upper bounds for the optimal value of the unknown function. Moreover, in the case that the noise is randomly generated, this procedure is proven to converge to the true optimal value.

A number of avenues for future work present themselves. As illustrated by set membership estimation (see e.g. [2]), the Lipschitz constant is a powerful tool for deriving local bounds on the basis of measurements. Indeed, Lipschitz constants of the parameterized functions can be derived from those of the basis functions, allowing for the determination of locally suboptimal values outside the scope of convex functions. Another extension would be to consider non-scalar functions, with an aim at analysis of nonlinear systems within a parameterized class. We also would like to characterize the sample efficiency and computational complexity of the proposed techniques. Regarding online methods, there are many possible extensions, including relaxations of the optimization problem to avoid intersections of a large amount of convex sets. This work investigates uniformly distributed noise, and investigation of different distributions and less conservative convergence results is a topic of interest.

References

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, Nov. 2005.

[2] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic systems with set membership uncertainty: An overview," *Automatica*, vol. 27, no. 6, pp. 997–1009, 1991.

[3] M. Milanese and C. Novara, "Set membership estimation of nonlinear regressions," in *IFAC World Congress*, vol. 35, no. 1, Barcelona, Spain, 2002, pp. 7–12.

[4] J. Calliess, S. J. Roberts, C. E. Rasmussen, and J. Maciejowski, "Lazily adapted constant kinky inference for nonparametric regression and model-reference adaptive control," *Automatica*, vol. 122, p. 109216, 2020.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer Series in Statistics. New York: Springer, 2013.

[6] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[7] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[8] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of Fluid Mechanics*, vol. 656, pp. 5–28, 2010.

[9] ——, "Dynamic mode decomposition and its variants," *Annual Review of Fluid Mechanics*, vol. 54, no. 1, pp. 225–254, 2022.

[10] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*, ser. Other Titles in Applied Mathematics. Philadelphia, PA: SIAM, 2016, vol. 149.

[11] M. R. Jovanović, P. J. Schmid, and J. W. Nichols, "Sparsity-promoting dynamic mode decomposition," *Physics of Fluids*, vol. 26, no. 2, p. 024103, 2014.

[12] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski, *Robust Optimization*, ser. Applied Mathematics Series. Princeton, NJ: Princeton University Press, 2009.

[13] D. Bertsimas, D. B. Brown, and C. Caramanis, "Theory and applications of robust optimization," *SIAM Review*, vol. 53, no. 3, pp. 464–501, 2011.

[14] M. Krstić and H.-H. Wang, "Stability of extremum seeking feedback for general nonlinear dynamic systems," *Automatica*, vol. 36, no. 4, pp. 595–601, 2000.

[15] K. B. Ariyur and M. Krstić, *Real-Time Optimization by Extremum-Seeking Control*. New York: Wiley, 2003.

[16] A. Teel and D. Popovic, "Solving smooth and nonsmooth multivariable extremum seeking problems by the methods of nonlinear programming," in *American Control Conference*, Arlington, VA, 2001, pp. 2394–2399.

[17] H. J. van Waarde, J. Eising, H. L. Trentelman, and M. K. Camlibel, "Data informativity: a new perspective on data-driven analysis and control," *IEEE Transactions on Automatic Control*, vol. 65, no. 11, pp. 4753–4768, 2020.

[18] H. J. van Waarde, M. K. Camlibel, J. Eising, and H. L. Trentelman, "Quadratic matrix inequalities with applications to data-based control," *arXiv preprint arXiv:2203.12959*, 2022.

[19] J. C. Willems, P. Rapisarda, I. Markovsky, and B. L. M. De Moor, "A note on persistency of excitation," *Systems & Control Letters*, vol. 54, no. 4, pp. 325–329, 2005.

[20] I. Markovsky, "Data-driven simulation of generalized bilinear systems via linear time-invariant embedding," *IEEE Transactions on Automatic Control*, vol. 68, no. 2, pp. 1101–1106, 2022.

[21] M. Guo, C. D. Persis, and P. Tesi, "Data-driven stabilization of nonlinear polynomial systems with noisy data," *IEEE Transactions on Automatic Control*, vol. 67, no. 8, pp. 4210–4217, 2021.

[22] C. De Persis, M. Rotulo, and P. Tesi, "Learning controllers from data via approximate nonlinearity cancellation," *IEEE Transactions on Automatic Control*, 2023, to appear.

[23] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2009.