Data-Driven Decision-Making and the Scenario Approach

M.C. Campi, A. Carè*

Dipartimento di Ingegneria dell'Informazione - Università di Brescia, via Branze 38, 25123 Brescia, Italia.

S. Garatti

Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italia.

Abstract

In the eye of many control scientists, the theory of the scenario approach is a tool for determining the sample size in certain randomized control-design methods, where an uncertain variable is replaced by a random sample of scenarios. This point of view is rooted in the history of the scenario approach and stands on a long track record of successful applications. However, in the last two decades, the theory of the scenario approach has gone beyond its original motivations and applications, and has unveiled some fundamental relationships between the complexity of a design and its generalization capabilities. The new knowledge brought by the theory provides a solid ground for a framework where data can be exploited in a flexible and wise manner, throughout a large variety of engineering activities. By this paper we aim at providing an access point to a set of state-of-the-art results in the theory of the scenario approach that can be valuable to target important challenges in modern control-design and decision-making at large. In the first part of the paper, we introduce a set-up for decision-making where the role of prior knowledge and user preferences can, and should, be distinguished from the role of data. Then, we show that the theory of the scenario approach offers a platform for conjugating heuristic approaches, which

*Corresponding author

in complex contexts are unavoidably based on incomplete and possibly imprecise information, with a solid theory for certifying the validity of the output of the decision process.

Keywords: data-driven methods, scenario approach, control, generalization.

---

1. Introduction

1.1. The era of data

Opportunities and challenges intertwine in the era of data. On the one hand, pervasive networks of smart sensors collect, process and store measurements in extremely fast, reliable and cheap ways. On the other hand, control scientists are haunted by the question of how one can take advantage of the increasing availability of data, either to improve existing control systems or to bring order to newly born systems of systems.

The weaving of opportunities with old and new challenges is apparent in many fields; we mention below but a few, in full awareness that many others could have deserved to be included in the list.

- **Automotive systems control.** The availability and smart processing of real-time traffic measurements has made it possible to deploy automatic and effective coordinated ramp metering strategies for freeways, Papageorgiou & Papamichail (2014); Seo et al. (2017). Complex systems of systems arise not only from the conscious striving to go beyond ordinary solutions, at the frontier of automotive technology (e.g., in the control of autonomous vehicles and of the systems arising from their interconnection Guanetti et al. (2018)), but also as technology-driven side-effects: for example, it has become urgent to deal with the disrupting effects of having many (ordinary) human drivers in (ordinary) cars, all relying simultaneously on the same (ordinary) mapping and route planning software, Macfarlane (2019).

- **Power generation and dispatchment.** At the dawn of the Second Industrial Revolution, connecting steam turbines to dynamos made large

scale power generation possible. Since then, energy has been generated and rationally distributed by leveraging well-understood physical laws, but the increasing penetration of renewable energy poses new and significant challenges to the planning and operation of modern power grids. In particular, power generation now depends on uncertain, hardly predictable phenomena, ranging from weather conditions to individual and social human behaviors. The underpinning for modern decisions and control schemes in this field has necessarily to include, besides physics, economic and social sciences, as well as conceptual tools to integrate historical and real-time data in the decision process, Li et al. (2020).

- **Real-time control of biological systems.** Feedback, as a mechanism to ensure homoeostasis, has long been recognized as a central feature of life, Cannon (1939); Hoagland & Dodson (1995), and control scientists have been increasingly at work to design feedback loops to re-establish homoeostasis in medical patients. A remarkable example is the development of an artificial pancreas for keeping blood glucose into a healthy range. The reader is referred to the recent review Quiroz (2019) for an account of how the control algorithms have evolved from the first closed-loop algorithms, which computed insulin infusion rates with poor information about the process to be controlled, towards personalized and data-based algorithms. Personalization has been made by a combination of a better understanding of glucose metabolism with the great availability of historical and real-time data that are offered by sensors now available in the market.

- **Medical computer-aided diagnosis.** A wise combination of domain knowledge, which remains fundamental to direct design efforts, and of data availability has fostered major progresses in medical computer-aided diagnosis. We just recall here the case of the automated classification of images of skin lesions as benign or malignant, which might soon become a widely available tool at the disposal of smartphone customer users, see

Esteva et al. (2017).

- **Machine-Learning-in-the-loop technologies.** Traditionally, an automatic classifier returns a result to a human being who is then responsible to make an informed decision (in the skin lesions example, this human being is a medical doctor or directly the patient). More and more often, however, classification algorithms, or other machine learning algorithms that are trained on data in a black-box manner, can be found in automated control loops, where they play the role of soft sensors, similarly to traditional state estimators. Examples range again from automotive systems (where classifiers are used *inter alia* for obstacle detection and avoidance, Gruyer et al. (2017); Devi et al. (2020)) to medical applications such as the artificial pancreas, Cappon et al. (2019). The more black-box learning algorithms enter safety-critical decision loops, the more urgent the need for a solid, scientific understanding of their limits and potentials.

Overall, it is clear that the increasing availability of computational power and of distributed large-scale optimization techniques enables the deployment of innovative data-driven decisions and control schemes, but it is also clear that this extraordinary potential cannot be fully expressed until it gets backed by a solid theoretical understanding.

1.2. Indirect vs. direct methods

In traditional model-based control, the "driving power" of data is often employed by engineers in the modeling of the reality to be controlled. In a typical workflow, first principles dictate the model class into which a reasonable representative of the reality should be sought, and experimental data are used to find the best parameter values to be plugged into the model equations. At the end of this modeling effort, the engineer has obtained a model that describes with a sufficient accuracy the way in which reality is expected to behave in response to signals that are under control. At the final stage of the workflow,

the engineer is called to make decisions, set-up policies and control mechanisms so as to optimize the objective values.

The increasing intricacy of many modern problems challenges this traditional design workflow: in fact, the input-output behavior of many modern systems is hardly captured by the class of models that can be found in the traditional system modeling toolboxes. Even in those lucky cases where, *in the eyes* of the data-analyst, available models look like acceptable descriptors of the reality, it is often the case that, *in the hands* of the control engineer, they lead to unsatisfactory control performances. The reason is that the metric according to which a model is said to capture "well" the hidden nature of the complex system is usually not tailored to the needs of the control engineer, whose goal is optimizing specific objective functions. This explains why the recent history of feedback control has witnessed a surge of the so-called "direct methods" (see e.g. Åström & Hägglund (1995); Safonov & Tsao (1995); Hjalmarsson et al. (1998); Guardabassi & Savaresi (2000); Gerencsér et al. (2002); Hjalmarsson (2002); Campi et al. (2002); Van Heusden et al. (2011); Bazanella et al. (2011); Moore (2012); Hou & Wang (2013); Formentin et al. (2014); Hori et al. (2016); Sutter et al. (2017); Karimi & Kammer (2017); Apkarian & Noll (2018); Novara & Milanese (2019); Formentin et al. (2019); Chiluka et al. (2021)), where one uses the available information that is carried by experiments directly to design a controller that minimizes the cost function of interest; this is opposed to the traditional indirect methods where significant effort is spent in preliminarily approximating reality by means of a model.

The development of "direct methods" for decisions and control is part of a broader scientific trend, which is largely technology-driven and is not limited to the control community (the reader is referred to the box "Direct, goal-oriented approaches: a cultural, technology-driven trend" for more discussion on this point).

Direct, goal-oriented approaches: a cultural, technology-driven trend.

According to a highly influential paper, Breiman et al. (2001), two cultures abide in statistical sciences. The classic one aims at using data for modeling the data generation mechanism: if successful, this culture generates powerful tools not only to control but also to describe reality (control engineering practice reflects this culture when, e.g., the transfer function of a simple second-order linear system is estimated from noisy data); on the other hand, as the complexity of the data generation mechanism increases, modeling becomes more and more an ambitious task, prone to detrimental oversimplifications. Hence, the other culture aims at using data to inform problem-oriented procedures, and to issue certificates on the statistical effectiveness of such procedures: this task can be accomplished under much milder and realistic assumptions even in the presence of very complex data generating mechanisms.

In the context of statistical learning, Vapnik, Vapnik (2013) (Section 1.9), formulated the following principle to be applied in the presence of restricted information: "When solving a given problem, try to avoid solving a more general problem as an intermediate step". In the wake of this advice, many recent advances in machine learning (in the fields of supervised classification and regression, Krizhevsky et al. (2012); clustering, Ghasedi Dizaji et al. (2017); reinforcement learning, Silver et al. (2016); etc.) share an agnostic approach with respect to the "true" or "best" description of the system at hand and focus on directly optimizing a cost function that maps the possible options that are available to the decision maker into a value that quantifies the level of satisfaction with the selected option.

Generally speaking, direct approaches become more urgent as technology enables one to address problems of increasing complexity, which is the trend that applied science is nowadays experiencing at an increasingly fast pace. The reader is also referred to Norvig (2017) for a thought-provoking discussion on matters related to the topic of this box.

1.3. This paper

In this paper, we present a framework for direct data-driven decision-making that can benefit from a series of technical results that have sprout from the theory of the scenario approach in the last two decades.

In the next Section 2, the scenario approach is introduced at a rather informal level as a general methodology to govern, in a well-grounded manner, the interplay between prior knowledge and data in decision-making. Some space (the whole Section 2.4) is devoted to introducing the probabilistic point of view, which allows one to assess the quality of a decision not only with respect to the data collected before the decision is made, but also with respect to the infinitely larger set of the unseen cases (those that can occur at the time the decision is applied). The section ends with a non-technical preview (on a simple example in Section 2.5) of the kind of statistical evaluations that are possible thanks to the theory of the scenario approach. This prepares the ground for the following, more technical, Section 3.

Section 3 provides an easily accessible, but technical, gallery of results: moving from simple decision schemes that are based on convex worst-case optimization, the reader is gradually introduced to the state-of-the-art of the theory, which encompasses non-convex optimization and general decision-making schemes.

2. A set-up for direct data-driven decision-making: the scenario approach

2.1. When is a decision good?

Many decision and control problems can be abstracted as the problem of choosing an object x from within a decision set $\mathcal{X}$. For example, in control-design, x is a vector of parameters representing a controller in a given class; in financial portfolio optimization, x is a vector where each element is the amount of money to be invested in a given asset; in classification, x encodes, according

to some predefined rule, a function that maps any possible relevant object into a label belonging to $\{0, 1\}$, etc.

We will denote by x^* the specific object that represents the final output of the decision process. The quality of x^* can be judged in relation to intrinsic and extrinsic criteria as described in what follows.

1) INTRINSIC QUALITY

We call "intrinsic" a quality of x that depends only on x itself, in the light of cemented knowledge and preferences that are available to the decision-maker. Here are some examples.

Example (Filter Design)**.** *If x is a digital filter, a good x may be expected to have a cut-off frequency in a certain range, to be physically realizable, possibly simple and cheap to implement, etc.*

Example (Prediction interval)**.** *If x is an interval used to predict an unknown variable, its width should be small for the prediction to be informative.*

Example (Home temperature control system)**.** *If x is a temperature controller, it should be designed so that, in nominal conditions, the control operates fast enough, overshooting and oscillations are limited, energy consumption is minimized, etc.*

2) EXTRINSIC QUALITY

The "extrinsic" quality refers to the performance of x in relation to various operating *situations* that may occur when x is applied. A bit more formally, we can think of the occurrence of a *situation* as an assignment of values to a vector of variables that we denote with the symbol δ, and, hence, extrinsic quality refers to the couple (x, δ): for every δ, x attains a performance as measured by a suitable indicator and the extrinsic quality refers to the variability of performances achieved by x as δ takes value in its range of variability.

Example (Filter Design)**.** *If the aim of the filter x is to work in a mobile device as an audio channel equalizer and the frequency response of the channel is δ,*

then the extrinsic quality of x refers to the performance of x in relation to a variety of channels δ and "high extrinsic quality" may refer to the ability of x to perform well over a large portion of δ's that may be encountered in the lifespan of the device.

Example (Prediction interval)**.** *Suppose that x is used to predict how effective a medical therapy is. Since the effectiveness depends on the patient δ to whom the therapy is administered, the extrinsic quality may refer to how large the portion of potential patients for which x provides a correct prediction is.*

Example (Home temperature control system)**.** *The performance of a building temperature controller x can be affected by the weather conditions and the minute actions of people in the building (which are partly unpredictable and will certainly differ from hour to hour and from day to day). Here, δ can be identified with a vector that includes quantities such as the external temperature and other weather conditions, the average amount of people in the building in a given time horizon, the number and the size of windows that happen to be open, etc., and the extrinsic quality may refer to the capability of the controller to keep the temperature within admissible limits for various δ.*

2.2. The limits of knowledge

In traditional decision processes, the intrinsic quality and the extrinsic quality of a candidate solution are often judged at the decision-making stage on the ground of available models.

Example (Traditional control design)**.** *Let x be a controller to be applied to a linear plant whose poles (expressed in a vector δ) are somewhat uncertain. If the variability of δ is known, imposing suitable phase and gain margins may ensure that the design will work well for all the relevant values of δ.*

On the other hand, when we move from simple to complex application domains, we often experience that δ refers to articulated and elusive portions of the real world for which it is extremely difficult to obtain a satisfactory and complete model and, hence, an *a priori* assessment of the extrinsic quality becomes

impractical. Under these circumstances, one may advocate the use of first-hand data, i.e., observed instances of δ, along a *direct decision-making approach*. This is the condition in which the scenario approach finds its natural application, as explained in the next section (see also the box "When should we consider using the scenario approach?" for a quick summary of the main ideas).

2.3. The principles of the scenario approach

We shall denote the empirical instances of δ that are available at the decision-making stage by $\delta^{(1)}, \ldots, \delta^{(N)}$, and refer to them as "*scenarios*". When the variability of δ is difficult to describe by means of a model, the fact that δ impacts on the extrinsic performance suggests a different way of proceeding and, at an informal level, the scenario approach prescribes to choose the candidate solution x that

(i) works well for the scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$

and, subject to (i),

(ii) optimizes the intrinsic quality.

In this way, the intrinsic quality is pursued directly while the scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$ are used as an *empirical substitute* of the infinite amount of situations δ that could occur in a future use of the decision to heuristically secure the extrinsic quality.

In applications, (i) and (ii) must be formulated quantitatively and choosing suitable indicators is highly problem-dependent. Moreover, while the application may suggest the meaning of the expression "x works well for δ", still the decision-maker retains the right of deciding whether to enforce that the solution works well for all the N scenarios $\delta^{(i)}$ or rather to neglect some of them. There is also much flexibility in constructing the indicator of the intrinsic quality and the domain in which the solution x is sought: they are typically based on prior knowledge and background preferences, but, as we shall see in Section 2.5, they can also be influenced by informal reasoning, second-hand information, guesses

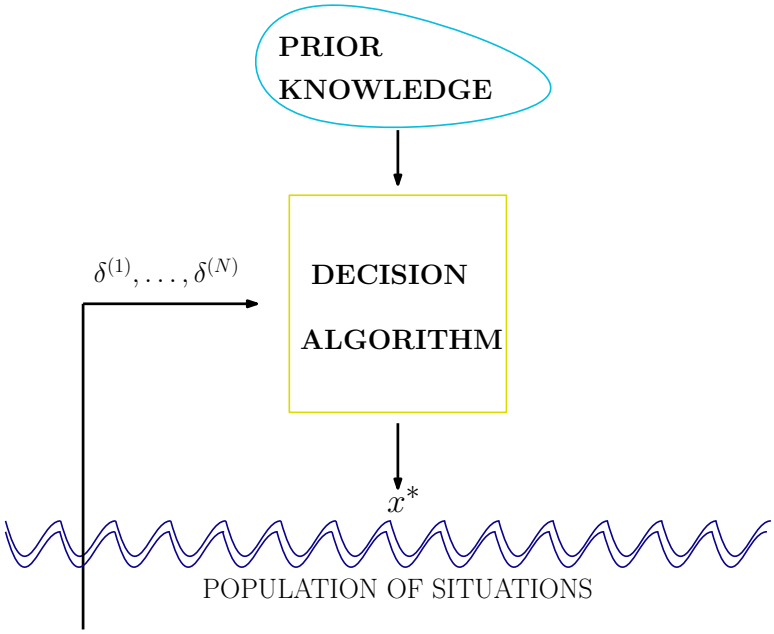

Figure 1: A decision is required to work well for a large amount of situations δ, characterized by high uncertainty and complexity. In the scenario approach, the decision x^* is obtained by an algorithm that incorporates prior knowledge to direct the decision, while instances $\delta^{(1)}, \ldots, \delta^{(N)}$ are used as a substitute of the multitude of unobserved situations.

and even conjectures about the variability of δ. The workflow for the scenario approach is illustrated in Figure 1.

Importantly, making a wise decision based on observations requires tools to gain confidence in that the decision is good both intrinsically and extrinsically. While the intrinsic quality is directly measured (and optimized), replacing the large set of the unseen situations δ's with a sample of scenarios is heuristic in nature and, when applied naively, may suffer from excessive empiricism: the final decision x^* performs well with respect to the instances $\delta^{(1)}, \ldots, \delta^{(N)}$ but it may fall short for other instances of δ that eventuates *after* the decision x^* has been implemented. The aim of the theory of the scenario approach is precisely that of taking control on this issue. This is important because only a trustful evaluation of the actual extrinsic quality along with the evidence about the intrinsic quality gives the decision-maker a handle to judge the solution. Based on a fair judgment, the decision-maker can:

(i) decide whether to "buy" the solution or to discard it;

(ii) decide how to subsequently use the solution, if the latter is part of a bigger decision process;

(iii) if the solution is not satisfactory, go back to the original choices in the problem formulation and re-calibrate them to construct solutions which are better aligned with the desires (for example, the definition of the intrinsic criterion can be modified, the size of the sample $\delta^{(1)}, \ldots, \delta^{(N)}$ can be re-tuned, etc.);

(iv) re-design the decision set $\mathcal{X}$ itself from which x^* is selected.

The main mathematical tool the scenario approach is based on is probability theory and the concepts introduced at a high-level in this section will be better formalized in the light of probability theory in the next section. Before proceeding, however, we feel advisable to highlight here two cornerstones of the theory as it will emerge in the remainder of this article.

1. While one way of quantifying the extrinsic quality is by validation, this requires using many data points for testing rather than designing. One key message of the scenario approach is that data can be used to simultaneously make a design while also evaluating its extrinsic quality. This is made possible by an exploitation of the information contained in the data beyond what traditional approaches can do.

2. In modern control and decision problems that deal with complex systems, besides scenarios one wants to exploit prior knowledge that comes from various sources, often including some that, while not completely trustworthy, can still be of help to obtain a satisfactory solution. If prior knowledge turns out to be lacking or even defective, its use may downgrade the quality of the solution. Importantly, in the scenario approach the correctness of the evaluation of the extrinsic quality (which is not directly observable) remains intact independently of the correctness of the prior. This fact

(which may came to the reader's surprise) is established in subsequent sections and is named "separation principle". The separation principle implies that the decision-maker can always correctly judge the impact of the prior information on the solution in terms of intrinsic (which is directly measurable) and extrinsic quality and decide to discard a prior if it is deemed unreliable. In this respect, it can be said that the scenario approach meets the pressing need in modern data-driven decision-making for a sound integration of domain knowledge and priors having uneven levels of trustworthiness with first-hand information given by data.

When should we consider using the scenario approach?

The scenario approach employs the data to directly target a design problem, without going through an intermediate step aiming at finding a description of the mechanism that generates δ. The scenario approach is of interest when:

(I) the value taken by δ can significantly impact on the performance;

(II) δ is made up of many variables, possibly interconnected, which take value according to mechanisms that are difficult, too expensive or even impossible to describe satisfactorily.

Because of (I), neglecting the variability of δ is not an option as this would lead to poor decisions. At the same time, (II) makes it impractical to find a description of the mechanism underlying the generation of δ. Hence, a good decision can be pursued heuristically driven by a direct use of the data. In this context, the scenario approach provides a theory to quantify the extrinsic quality of a decision and drives the user towards a wise selection.

Examples of applications where (I) and (II) apply are ubiquitous and can e.g. be found in:

- the stock market values; Pagnoncelli et al. (2012); Calafiore (2013); Ramponi & Campi (2018);

- the electrocardiogram, or a set of features extracted from an electrocardiogram, of a patient in cardiac arrest, Carè et al. (2018).
- the values of the power generation and load in a power system with intermittent and distributed power sources, Modarresi et al. (2019); Geng & Xie (2019);
- a demand profile for selling products, Carè et al. (2014);
- input-output data collected from a plant in different operating conditions, e.g., a car with different tires and road conditions, Rallo et al. (2016);
- inflows and outflows in a system of rivers or other weather-dependent phenomena, Nasir et al. (2018);
- the transfer function of a communication channel in mobile communication, Carè et al. (2015);
- instances of disturbances, Campi et al. (2009b), or actuation errors, Carè et al. (2019), in a largely unpredictable environment.

2.4. Mathematical foundations: a probabilistic framework

Probability is, first and foremost, a measuring tool that allows us to make statements such as "x^* performs well for a *large portion* of situations δ". In fact, if we accept that various instances of δ present themselves according to a probability distribution $\mathbb{P}$, i.e., that

δ is distributed according to $\mathbb{P}$,

then it is *possible* (at least in principle) to *quantify* the portion of the δ's for which x^* does not perform satisfactorily.

In what follows, the probability that the solution x^* does not perform satisfactorily is denoted with $V(x^*)$, which we call the *risk* of the solution x^*.

The risk $V(x^*)$ is a number between 0 and 1 and is an indicator of the extrinsic quality of x^*. A good enough extrinsic quality can then be formalized by means of a condition of the kind $V(x^*) \leq \bar{\epsilon}$, where $\bar{\epsilon}$ is a domain-dependent user-chosen threshold.

It must be remarked that, from the decision-maker point of view, there is a large gap between *accepting that δ is distributed according to some probability* $\mathbb{P}$ and *assuming that such* $\mathbb{P}$ *is known*. In fact, when δ is the outcome of a complex

generation mechanism, the availability of a satisfactory description of $\mathbb{P}$ is often precluded. The theory of the scenario approach recognizes this gap and is built upon the premise that

> $\mathbb{P}$ exists but is not known to the decision-maker.

$\mathbb{P}$ manifests itself through:

> observations $\delta^{(1)}, \ldots, \delta^{(N)}$, which are modeled as independent draws from $\mathbb{P}$ (according to a standard terminology, observations are independent and identically distributed, "i.i.d.").

While this i.i.d. assumption is limiting (relaxing this assumption is at present an open and thrilling research endeavor), it is worth remarking that many applications can be cast within, or drawn back to, this i.i.d. set-up. For example, stock prices at equispaced time intervals are definitely not independent; however, logarithmic return increments are independent according to the Black-Scholes model, Black & Scholes (1973) (the reader is referred to the box "How to get i.i.d. scenarios in practice" for more general strategies to recast a problem into an i.i.d. framework).

Draws $\delta^{(1)}, \ldots, \delta^{(N)}$ are first-hand knowledge on the problem and the scenario approach provides a well-principled framework to estimate $V(x^*)$ from the data. The estimate remains correct even when any partial or insecure knowledge on $\mathbb{P}$ that has been used at the time the decision problem was formulated turns out to be incorrect (the reader should trace this requirement back to point 2 at the end of the previous Section 2.3). Moreover, the theory is grounded on finite-sample results that are rigorously valid for any sample size (the reader is referred to the box "Asymptotic results in the era of data: the tantalizing horizon" for a discussion on the value of finite-sample results).

How to get i.i.d. scenarios in practice.

In many problems, scenarios are naturally i.i.d.; this is the case for example in all problems where data are draws from a population, with myriad

applications in machine learning, prediction and classification. Otherwise, various techniques can be used to draw the problem back to the i.i.d. set-up. Here, two families of strategies of wide applicability to recast non-independent data into an independent sample are briefly touched upon.

- **Prediction-based strategies:** Sometimes, at least a rough predictor for a forthcoming observation is available (e.g., the Weather Bureau provides us with weather predictions). Then, any observation at time t can be decomposed into a prediction part, based on the best of our knowledge until time t, and a prediction error. Often, prediction errors at different time instants are only lightly correlated and the sequence of the prediction errors can, at least in first approximation, be treated as an independent sequence.

- **Segmentation:** The states visited by a Markov Chain do not form an independent sequence, but an observed trajectory can be segmented into independent episodes by exploiting the visit of a recurrent state, restart events after entering an absorbing state, etc., see, e.g., Vidyasagar (2014).

Instead, even when the assumption that data are identically distributed is not overall realistic, still it is often an acceptable approximation over relatively short time windows, or after suitable domain-dependent preprocessing such as the removal of seasonal trends.

Asymptotic results in the era of data: the tantalizing horizon.

Traditionally, statistics has been dominated by asymptotic results. The usage of these results in practice always introduces approximations and is acceptable only when the number of data points is large compared to the dimension of the solution that is being tuned on the data set. The present era where data are ubiquitous and largely available may seem to have lessened the need for finite-sample results and have favored the usage

of asymptotics. However, this evaluation turns out to be incorrect. The reason lies in the fact that in our present times the greater availability of data fares hand in hand with the increasing scale of the problems. As problems become larger-scale, they require solutions of higher dimension to be satisfactorily resolved so that, as new data become available, the decision maker is tantalized to resort to more articulated and complex solutions sets. This results in a sort of receding horizon that jeopardizes the use of asymptotic statistics.

2.5. The operation of the scenario approach illustrated on a simple example

This section illustrates, and complements, various aspects touched upon in previous sections of this article. It is meant to provide the reader with a more transparent understanding of the operation of the method before delving into the more technically-oriented presentation of Section 3.

Suppose that the severity of a disease can be quantified by a real number y, and yet an accurate assessment of y requires a medical test that is too invasive to be applied on a vast scale. On the other hand, a simpler inspection delivers a number u that carries information on y and

> our aim is to construct an interval predictor that associates to a given value of u a range of values for y.

To do so, we follow a workflow in line with Figure 1.

Leveraging prior knowledge

We start by collecting the opinions of some experts. Most of them express the educated guess that y should increase linearly with u, while a few of them, with a reputation of being contrarians, are more doubtful and say that they could even expect a negative correlation between u and y for values of u that are above average. None of the experts expects that the dispersion of the values of y changes significantly with u. Moving from this latter observation,

we set out to construct an interval predictor that maps u into intervals $I(u) = [\varphi(u) - \frac{h}{2}, \varphi(u) + \frac{h}{2}]$ of fixed width h, where both the function $f(\cdot)$ and the width parameter h have to be suitably designed based on further assessments of the problem.

Next we decide to trust the first group of scientists and take $\varphi(u) = a + b \cdot u$, a linear function of u with a and b tunable parameters.

Owing to these choices, we are only left to select three parameters, a, b and h, which form the decision variable $x = (a, b, h)$.

The quality criteria

The indicator of the intrinsic quality of an interval predictor $I(u)$ is the width h: the smaller h the more accurate the prediction. While pursuing this intrinsic quality, we also have to keep control on the extrinsic quality, represented by the reliability of the predictor: we must ensure that the portion of patients for which $y \in I(u)$ is large enough. More formally, letting $V(x) = \mathbb{P}\{y \notin I(u)\}$, where $\mathbb{P}$ is the probability according to which patients (corresponding to pairs (u, y)) distribute, we would like that $V(x^*) \leq \bar{\epsilon}$ for a suitably small $\bar{\epsilon}$.

The scenario approach: let the data speak

Suppose that 100 patients are independently drawn from $\mathbb{P}$ and tested (with both the invasive and the simpler test), and the corresponding 100 scenarios $\delta^{(i)} = (u^{(i)}, y^{(i)}), i = 1, \ldots, 100$, are at our disposal. Based on these scenarios, we choose (a^*, b^*, h^*) by the following rule.

RULE: the parameters a^*, b^* and h^* are those that yield the interval predictor with minimum width h subject to the condition that $y^{(i)} \in I(u^{(i)}), i = 1, \ldots, 100$.

Figure 2 gives the result generated by the RULE using the data that are

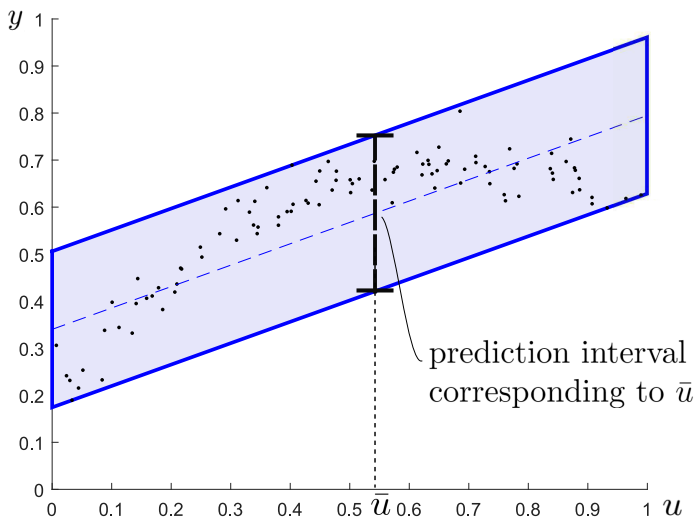

Figure 2: A sample of 100 scenarios: each scenario is a data point in the (u, y) space. The function $\varphi(u) = a^* + b^* \cdot u$ is represented by the dashed line. For each value of u the predicted interval is $I(u) = [\varphi(u) - \frac{h^*}{2}, \varphi(u) + \frac{h^*}{2}]$.

shown in the same picture.

Assessing the quality of the solution

The intrinsic quality can be assessed from the value of h^* as soon as the predictor is constructed. In the outcome shown in Figure 2, h^* turned out to be only moderately satisfactory (and indeed in this toy example it is even visually clear that the interval predictor includes large portions of empty space).

While h^* is an observable quantity, $V(x^*)$ cannot be directly computed because it depends on probability $\mathbb{P}$, which is not known.

An evaluation of $V(x^*)$ can however be performed by using the scenario theory and, to understand the type of results the scenario theory offers, in the next points we first analyze what has been observed in a campaign of simulated examples where $\mathbb{P}$ was artificially manufactured, and hence it was known.

The distribution of $V(x^*)$

Simulation campaign #1. Our first simulation campaign consists of repeated constructions of the predictor according to the RULE, with different sets of data.

- In $M = 1000$ repetitions, we generated a data set made up of $N = 100$ observations, $(\delta^{(1)}, \ldots, \delta^{(100)})$ according to the same distribution that was used to generate the data points in Figure 2;
- for each one of the 1000 data sets, we constructed a predictor by computing $x^* = (a^*, b^*, h^*)$ according to the RULE;
- for each x^*, we computed $V(x^*)$ (note that we can compute the exact value of $V(x^*)$ because we are running an artificial example where we know the distribution according to which the pairs (u, y) are generated; this would be impossible if data were real data generated from an unknown distribution).

The histogram of the 1000 values of $V(x^*)$ is given in Figure 3. The true distribution of $V(x^*)$ computed analytically is represented by the dashed line in the same figure (the histogram tends to this distribution as $M \to \infty$).

Let us consider the quality threshold $\bar{\epsilon} = 0.11$. In our 1000 simulations, the condition $V(x^*) < \bar{\epsilon}$ was always satisfied. By an analytical computation, we found that $\mathbb{P}^{100}\{(\delta^{(1)}, \ldots, \delta^{(100)}) : V(x^*) > \bar{\epsilon}\} = 7.73 \cdot 10^{-4}$ (note that the distribution of $(\delta^{(1)}, \ldots, \delta^{(100)})$ is $\mathbb{P}^{100}$ because scenarios are i.i.d. draws). Since $7.73 \cdot 10^{-4}$ is a small number, it is not surprising that this event did not happen in our 1000 experiments.

The knowledge of the value $7.73 \cdot 10^{-4}$ can be used to make statements like the following one:

> if we run an experiment and build a predictor based on 100 scenarios, the resulting predictor will have a risk smaller than 11% with confidence $1 - 7.73 \cdot 10^{-4}$.

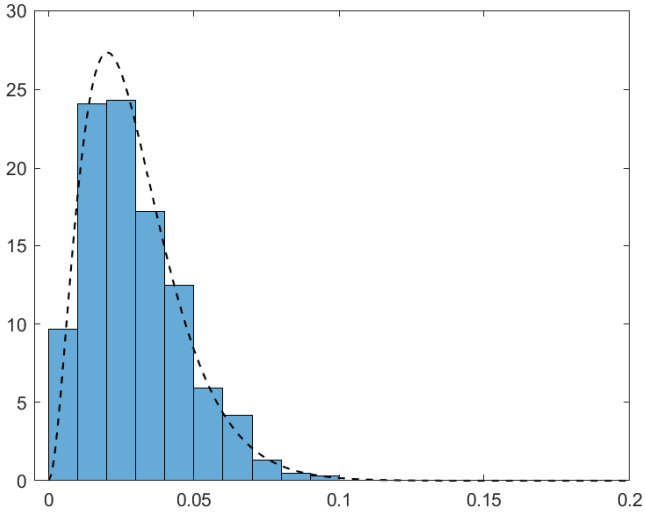

Figure 3: Histogram of $M = 1000$ values of $V(x^*)$. The distribution of $V(x^*)$, to which the histogram tends when $M \to \infty$, is represented by the dashed line.

A natural question now is how this story changes for a different data generation mechanism, that is, for a different $\mathbb{P}$.

Simulation campaign #2. We made a second simulation campaign as before but, this time, we replaced the probability distribution of (u, y) with a new one: u is uniform over $[0, 1]$ and the distribution of y given u is uniform over $[0.9u, 0.9u + 0.1]$. Unlike the distribution in campaign #1, this distribution fits very well our linear *a priori* belief leading to a small value of h^* (see Figure 4). The histogram of $V(x^*)$ for this second simulation campaign is shown in Figure 5, together with the exact distribution of $V(x^*)$.

Surprisingly, the true distribution is the same as that for campaign #1. This is a symptom of a general fact: any distribution of (u, y) with a density leads exactly to the same distribution of $V(x^*)$. More on the invariant distribution of $V(x^*)$ will be provided in the next technical Section 3, however, we anticipate that it is a Beta distribution with expected value equal to $\frac{3}{N+1}$. Interestingly, the number 3 at the numerator coincides with the number of optimization vari-

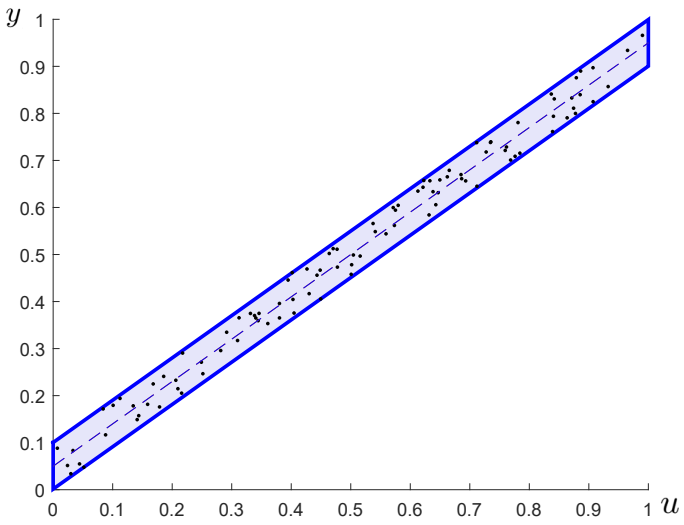

Figure 4: An interval predictor obtained in the second simulation campaign.

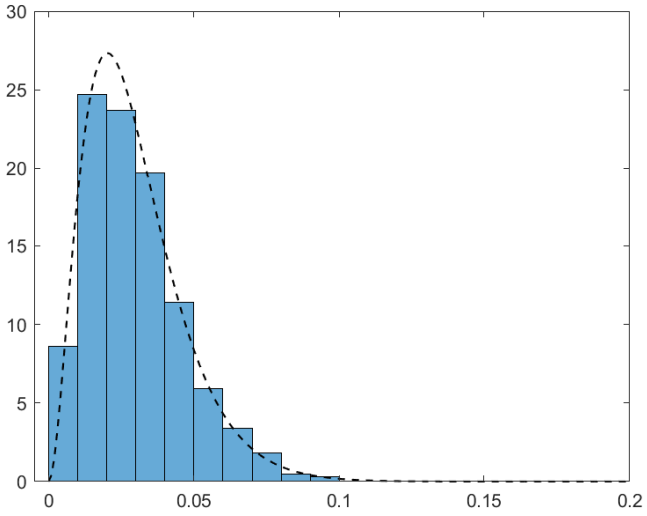

Figure 5: Histogram of $M = 1000$ values of $V(x^*)$ for another distribution of (u, y). The true distribution of $V(x^*)$, to which the histogram tends when $M \to \infty$, is given by the dashed line and it is exactly the same as that in Figure 3.

ables (a, b, h). This is not a fortuitous circumstance: should we use for example a predictor that has 4 optimization variables (a, b, c, h), $V(x^*)$ would distribute like a Beta with expected value equal to $\frac{4}{N+1}$. In this case, the probability of the event $V(x^*) > 0.11$ would increase to $3.4 \cdot 10^{-3}$.

The separation principle

The above simulation campaigns have revealed a fundamental, and unexpected, property of the RULE: the distribution of $V(x^*)$ remains the same irrespective of the actual mechanism by which data are generated. As a consequence, we can trust our judgment on $V(x^*)$ despite the possible incorrectness of the priors which were used in the formulation of our optimization problem. This is a manifestation of the so-called "separation principle".

Going back to our example, in the result in Figure 2 using the prior that u and y are linearly correlated led to a poor result in terms of the width of the prediction interval (which might suggest that our trust in the majority of the experts was misplaced). Nonetheless, the result that $V(x^*) \leq 0.11$ holds with high confidence remains intact. Next, we may want to give a chance to the minority opinion and move to consider a quadratic function $\varphi(u) = a+b\cdot u+c\cdot u^2$ so as to incorporate a possible negative correlation for high values of u. By using again the RULE, extended to the additional parameter c, we found the much thinner interval predictor in Figure 6. In this latter case, the confidence in the result that $V(x^*) \leq 0.11$ is $1 - 3.4 \cdot 10^{-3}$ (just slightly lower than before as a consequence of having used one more optimization variable).

From the simple example to more general problems

Many real life problems differ from the above simple example in some important respects:

- situations are typically described by large dimensional objects and not

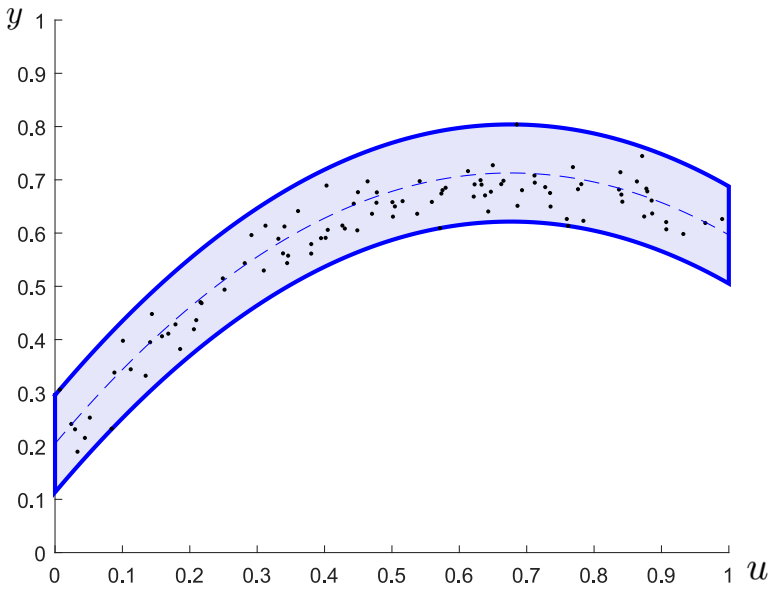

Figure 6: An interval predictor obtained by fitting a function $\varphi(u) = a + b \cdot u + c \cdot u^2$.

just (u, y) pairs;

- the impact of prior knowledge cannot always be described as clearly as in this simple example where there is a single decision variable, h, that accounts for the intrinsic quality of the solution, while prior beliefs affect the shape of the predictor through the other decision variables $(a, b, \ldots)$;
- the distribution of $V(x^*)$ is not an invariant as the distribution of data changes.

These differences do not prevent the theory of the scenario approach to be applicable. The last point is of particular interest and we anticipate that the reason why the distribution of $V(x^*)$ was invariant in the example of this section was that the *complexity* (as defined in the next technical section) of the solution was the same irrespective of the sample of scenarios at hand. Importantly, when this circumstance does not turn out to be correct, universal statements certifying $V(x^*)$ independently of the data generation mechanism are still possible. Moreover, the scenario theory is much more general than what the previous

example has shown and includes a rich framework for the evaluation of the risk based on the complexity of the solution when this is not constant, as well as schemes allowing for the removal of some data points with the aim to strike a suitable balance between reliability (extrinsic quality) and accuracy (intrinsic quality).

3. The Theory of the Scenario Approach

To better illustrate the nature of the results, we start from a special but notable set-up (which encompasses as a special case the simple example of Section 2.5) and we will then move gradually towards more general frameworks.

3.1. Convex worst-case optimization

Let $x \in \mathbb{R}^d$ be the decision vector. To make precise the somehow informal description of Section 2.3, let us introduce two real functions $c(x)$ and $f(x, \delta)$ as follows:

- $c(x)$ is a cost function (to be minimized), which is used as a quantifier of the intrinsic quality of x;
- $f(x, \delta)$ models the "regret" for employing x when δ occurs. Condition $f(x, \delta) \leq 0$ indicates a satisfactory performance and, correspondingly, $V(x) = \mathbb{P}\{\delta : f(x, \delta) > 0\}$ is the indicator of the extrinsic quality of x.

In this section, we assume that $\mathcal{X}$ is a convex set; $c(x)$ is convex; and $f(x, \delta)$ is convex in x for all δ (while the dependence of f on δ is arbitrary).[1]

The decision process aims at finding the decision with optimal intrinsic quality (i.e., minimum cost $c(x)$) among all the candidate decisions that perform "well" for all the observed scenarios. In formal terms, this leads to the following

[1] As we shall see in Section 3.3, recent developments of the scenario approach remove these convexity assumptions.

mathematical program:

$$\begin{aligned} &\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} c(x) \\ &\text{subject to: } f(x, \delta^{(i)}) \leq 0, \ i = 1, \ldots, N. \end{aligned} \tag{1}$$

As is clear, one primary goal behind using (1) is that of safeguarding against the worst, which motivates enforcing the constraints $f(x, \delta^{(i)}) \leq 0$ for *all* the scenarios. Although suitable to various contexts, this conservative standpoint may result in a poor value of $c(x^*)$, in which case one may want to resort to more flexible alternatives as discussed in Section 3.4.

In the remainder of this section we will assume that, for any N, the feasibility domain of the scenario program (1) has an interior point and that the solution to (1) is unique.[2].

Before proceeding with the description of the theoretical results, it is worth remarking that the example of Section 2.5 fits into this framework of convex worst-case optimisation if (a, b, h) is identified with x, and (u, y) with δ. In fact, building a predictor by applying the RULE in Section 2.5 amounts to solving (1) with the cost function $c(x) = h$ and the regret function $f(x, \delta) = |a + b \cdot u - y| - h$; note also that, in that context, the risk $V(x^*)$ (x^* is the solution to problem (1)) is the probability with which a new observation (u, y) falls outside the prediction model of width h^*.[3]

The scenario theory for convex worst-case optimization

[2]The uniqueness assumption is introduced to simplify the presentation; if the solution is not unique, a suitable tie-break rule can be introduced, Campi & Garatti (2008); other variations on the theme are possible, see e.g. Calafiore (2010b); Nasir et al. (2016) for contributions about unfeasible problems

[3]Interval Predictor Models as in Figures 2, 4 and 6 are obtained from convex worst-case optimisation problems that have a special structure for which specific theoretical results are available, see in particular Carè et al. (2015) and Carè et al. (2014). For other contributions on Interval Predictor Models, the reader is also referred to Campi et al. (2009a); Calafiore (2010a); Patelli et al. (2013); Crespo et al. (2016); Lacerda et al. (2018); Garatti et al. (2019); Wang et al. (2020).

In Campi & Garatti (2008), it is proven that the cumulative distribution function of $V(x^*)$ is (first-order) stochastically dominated by a Beta distribution with parameters $(d, N-d+1)$. In mathematical terms,

$$\mathbb{P}^N\{V(x^*) \leq v\} \geq 1 - \sum_{i=0}^{d-1} \binom{N}{i} v^i (1-v)^{N-i}. \tag{2}$$

Importantly, this result is *universal*, in the sense that it is valid irrespective of $\mathbb{P}$. An upper-bound $\bar{\epsilon}$ for $V(x^*)$ can then be obtained from (2) as follows. Let β be a small number, say $\beta = 10^{-5}$, and let $\bar{\epsilon}$ be the value that solves the equation

$$\sum_{i=0}^{d-1} \binom{N}{i} v^i (1-v)^{N-i} = \beta,^4 \tag{3}$$

then $V(x^*) \leq \bar{\epsilon}$ holds with high probability $1-\beta$. Since the Beta distribution is unimodal and thin tailed, $\bar{\epsilon}$ gets close to its mean (which is $\frac{d}{N+1}$) for a relatively small number of observations. Thus, roughly, $\bar{\epsilon}$ depends on the ratio between d and N.

In applications, $\bar{\epsilon}$ can be used as a relevant information to compare decisions coming from various decision spaces. For instance, in the example of Section 2.5 solving (3) for $\beta = \frac{1}{1000}$ gives $\bar{\epsilon} = 0.108$ when $d = 3$ (linear predictor) and $\bar{\epsilon} = 0.125$ when $d = 4$ (predictor centered around a quadratic function) and these values complement the information coming from the intrinsic quality. The reader is referred to the box "Dealing with multiple comparisons by the union bound" for a more detailed discussion on the guarantees that can be attached to a selection made out of competing alternatives.

As already mentioned, in the example of Section 2.5 the distribution of $V(x^*)$

[4] This equation can be easily solved by bisection. The MATLAB function betainv(1-β,d,$N-d+1$) solves it and returns directly the value of $\bar{\epsilon}$. An approximate formula that can be useful for a first pencil-and-paper estimation of $\bar{\epsilon}$ is

$$\bar{\epsilon} \leq \frac{1}{N}\left(d - 1 + \ln\frac{1}{\beta} + \sqrt{2(d-1)\ln\frac{1}{\beta}}\right),$$

see Alamo et al. (2015).

is exactly a Beta distribution with parameters $(d, N-d+1)$, irrespective of the data generation mechanism, provided $\mathbb{P}$ has density. Thus, in that case, the inequality "$\leq$" in (2) holds in fact with "=", i.e.,

$$\mathbb{P}^N\{V(x^*) \leq v\} = 1 - \sum_{i=0}^{d-1} \binom{N}{i} v^i (1-v)^{N-i} \tag{4}$$

(the density of this distribution when $N = 100$ and $d = 3$ is the dashed curve in Figures 3 and 5). The existence of problems for which the inequality in (2) is an equality (examples can be found for any d and N) shows that (2) cannot be improved unless it is specialized to subclasses of problems.

The notion of complexity

In the proof of (2) developed in Campi & Garatti (2008), a key role is played by the notion of *complexity*.

> The *complexity* of x^* is an integer $s^* \in \{0, 1, \ldots, N\}$ such that x^* can be obtained by solving a problem similar to (1) that only contains a subsample of scenarios from $\delta^{(1)}, \ldots, \delta^{(N)}$ whose cardinality is s^* while no subsamples of scenarios with cardinality lower than s^* exist that give the same solution x^*.[5]

For an example, we can go back to Figure 2 and note that the same solution (a^*, b^*, h^*) would have been obtained with just 3 points, those that lie on the boundary of the prediction model. Instead, any subsample of scenarios that does not include these 3 points leads to a different solution, so that $s^* = 3$ in this case.

The following is a key fact in convex worst-case scenario theory.

Key Fact. *For any problem in the form of* (1)*, it holds that* $s^* \leq d$.

[5] It is worth noticing that, in the convex set-up, the scenarios that suffice to reconstruct x^* are always a subset of the active constraints, which makes the evaluation of s^* computationally easy.

Looking at the proof of (2) in Campi & Garatti (2008), one sees that (2) deeply relies on the above Key Fact. Also, the fact that the predictor in Figure 2 turns out to have a complexity s^* that is equal to d (= 3) is not by chance: in Garatti et al. (2019), one can find a proof that this problem belongs to the class of *fully-supported problems*, for which $s^* = d$ happens with probability 1.[6] In Campi & Garatti (2008), it is proven that relation (4) holds for all fully-supported problems and that all other problems are dominated by the fully-supported class in the sense that (2) holds.

Dealing with multiple comparisons by the union bound.

In the prediction problem of Section 2.5, we know that $V(x^*) > \bar{\epsilon}$ happens with probability $\beta = \frac{1}{1000}$, where $\bar{\epsilon} = 0.108$ for $d = 3$ and $\bar{\epsilon} = 0.125$ for $d = 4$. If the selection between $d = 3$ and $d = 4$ is made *a posteriori* (after seeing the prediction interval), we might be advised by an "evil oracle" that indicates a "bad" choice (for which $V(x^*) > \bar{\epsilon}$) whenever one exists. However, no matter how evil the oracle is, the probability that a data set verifies the condition $V(x^*) > \bar{\epsilon}$ for one of the two choices cannot be larger than $\frac{1}{1000} + \frac{1}{1000} = \frac{2}{1000}$. Therefore, the certificate "$V(x^*) \leq \bar{\epsilon}$" is valid with probability at least $1 - 2 \cdot 10^{-3}$. This argument can be extended to the case where one chooses a solution from M possibilities leading to the conclusion that $\mathbb{P}^N\{V(x^*) \leq \bar{\epsilon}\} \geq 1 - M\beta$ (where, similarly to the above example with $d = 3$ or $d = 4$, the value of $\bar{\epsilon}$ depends on the choice made). Since small values of β (such as 10^{-7} or 10^{-8}) can be enforced

[6] A problem is fully-supported if $s^* = d$ with probability 1 and it is *non-degenerate*; a problem is non-degenerate if, for any N, there is with probability 1 a unique choice of indexes $i_1, i_2, \ldots, i_k$ (with $i_1 < i_2 < \ldots < i_k$) from $1, 2, \ldots, N$ such that: (a) problem (1) where only the constraints $f(x, \delta^{(i_j)}) \leq 0$, $j = 1, \ldots, k$, are enforced gives the same solution x^* as with all constraints; (b) if $k > 0$, discarding further indexes from $i_1, i_2, \ldots, i_k$ changes the solution (irreducibility of the set of indexes). (Note that for $k = s^*$ one certainly finds one such set of indexes, the definition of non-degeneracy requires that this set is unique within all subsets of indexes whose cardinality k is equal to s^* or larger.) Within convex optimization, degeneracy requires an anomalous accumulation of active constraints to happen and assuming non-degeneracy is therefore reasonable in many applications. Note for the reader: the definition of fully-supportedness given in the scenario literature is sometimes formulated differently from, but is equivalent to, that given here.

with reasonable sample sizes, the value of $M\beta$ can be easily kept small even when large set of choices are tested out.

3.2. The wait-and-judge perspective

The results in Section 3.1 stand on the observation that the complexity s^* is upper bounded by d (see "Key Fact") and culminate in result (2), which is tight for fully-supported problems, that is, (4) holds. On the other hand, (2) is only an upper bound for non fully-supported problems and the reader is referred to Figure 7 for the distribution of $V(x^*)$ in two non fully-supported examples.

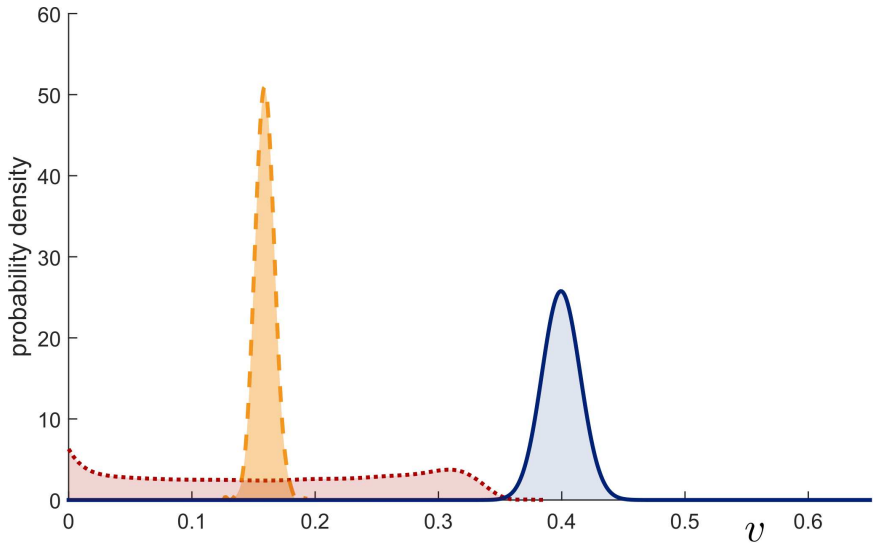

Figure 7: [This figure is taken from Garatti & Campi (2019)] The probability density function of the Beta distribution, as given by the right-hand side of (2) with $N = 1000$ and $d = 400$, is represented in blue continuous line. For a fully-supported problem, $V(x^*)$ distributes as the Beta distribution. The dotted and dashed lines show the distribution of $V(x^*)$ for two examples described in Garatti & Campi (2019): while these densities are dominated by the Beta distribution in the sense of (2), using the Beta distribution to estimate $V(x^*)$ produces conservative evaluations in these two cases.

The conservatism inherent in the Beta result normally worsens in large scale problems with a high dimensional optimization domain, which is a situation that is encountered with increasing frequency in modern applications.[7]

[7]In some cases, one can use a Beta in dimension $\tilde{d} < d$ (where d is the actual dimension of

In what follows, we present results for convex worst-case optimization that establish a connection between $V(x^*)$ and s^* (rather than a connection between $V(x^*)$ and the upper bound d on s^*). The use of these results is that one waits until the solution is determined and $V(x^*)$ is then judged from the complexity s^* that has been measured at the solution (see also the box "Compression and risk: a solid marriage").

Compression and risk: a solid marriage.

The complexity s^* is small when the sample $\delta^{(1)}, \ldots, \delta^{(N)}$ can be slimmed down to a small subsample that is sufficient to reconstruct the solution x^*. Therefore, the complexity is related to the *compressibility* of the sample $\delta^{(1)}, \ldots, \delta^{(N)}$. The idea that the compressibility of the information carried by the data is related to generalization properties is not new in machine learning, see e.g. Rissanen (1978); Littlestone & Warmuth (1986); Rissanen (1986); Ming & Vitányi (1990); Barron et al. (1998); Graepel et al. (2005); Moran & Yehudayoff (2016); Hanneke & Kontorovich (2019). What is new in scenario optimization is the vast applicability of this concept beyond a machine learning context and that compressibility leads to extraordinarily powerful results as in (2), (4). Recent research efforts have been geared towards extending these results to general schemes beyond convex optimization, see e.g. Campi (2010); Margellos et al. (2015); Esfahani et al. (2015); Grammatico et al. (2016); Carè et al. (2017); Ramponi & Campi (2018); Carè et al. (2018); Campi et al. (2018); Margellos et al. (2018); Paccagnan & Campi (2019). Part of these results are outlined in Sections 3.3, 3.4 and 3.5.

We start by illustrating a simulation example; later, we present the general theory. This section contains results from Campi & Garatti (2018); Garatti & Campi

the problem), in which case $\tilde{d}$ is called the "effective dimension". Regularization mechanisms have been used to achieve this result in Campi & Carè (2013), while Schildbach et al. (2013); Zhang et al. (2015) present studies in specific contexts in which $\tilde{d}$ is derived from using the concept of "support rank".

(2019, 2021), to which the reader is referred for the proofs and more details.

Example: orthant that includes random points

A population of points p in a 400-dimensional Euclidean space $\mathbb{R}^{400}$ is distributed according to a probability $\mathbb{P}$. We want to choose an $x \in \mathbb{R}^{400}$ such that $\sum_{j=1}^{400} x_j$ (subscript j denotes component) is minimized (intrinsic criterion) while relation $p_j - x_j \leq 0$, $j = 1, \ldots, 400$, holds with high probability (in other words, the negative orthant with vertex in x contains most of the probabilistic mass of $\mathbb{P}$ – extrinsic criterion).

Worst-case scenario solution

We collected $N = 1000$ points $p^{(1)}, \ldots, p^{(1000)}$ (these are the scenarios), and solved the scenario program:

$$\begin{aligned} &\min_{x \in \mathbb{R}^{400}} \sum_{j=1}^{400} x_j \\ &\text{subject to: } p_j^{(i)} - x_j \leq 0, \;\; j = 1, \ldots, 400 \;\; i = 1, \ldots, 1000, \end{aligned} \tag{5}$$

which is a convex scenario program in the form of (1) with $d = 400$ and $N = 1000$, $c(x) = \sum_{j=1}^{400} x_j$ and $f(x, \delta) = \max_{j=1,\ldots,400}(p_j - x_j)$ (where, clearly, $\delta = p$).

Results of two simulation campaigns

Two simulation campaigns were performed for two different probability distributions $\mathbb{P}_A$ and $\mathbb{P}_B$. For each simulation campaign, we repeated 100 000 times the sampling of $N = 1000$ scenarios, and computed the corresponding x^* and s^*. Every time, we also computed $V(x^*)$ by exploiting the privilege (due to the fact that we are in a simulated set-up) of knowing the real distribution of the

points.

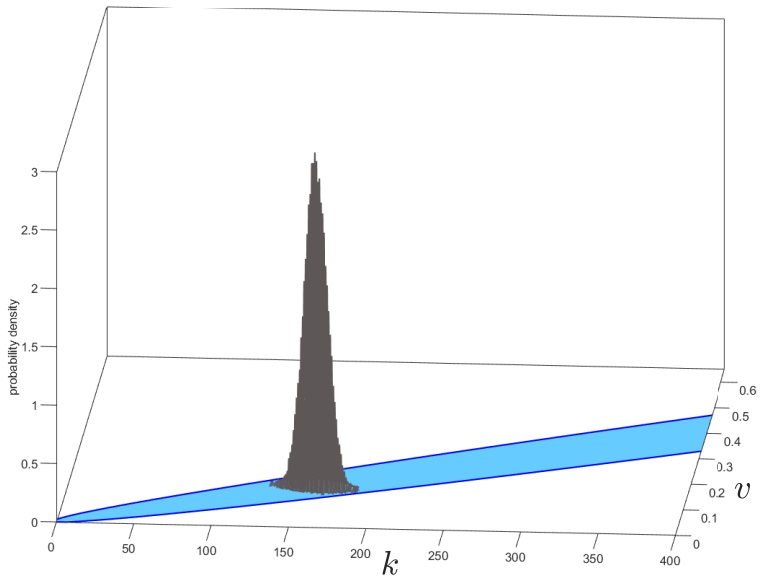

Figure 8: Empirical probability distribution of $(s^*, V(x^*))$ when points are generated by $\mathbb{P}_A$.

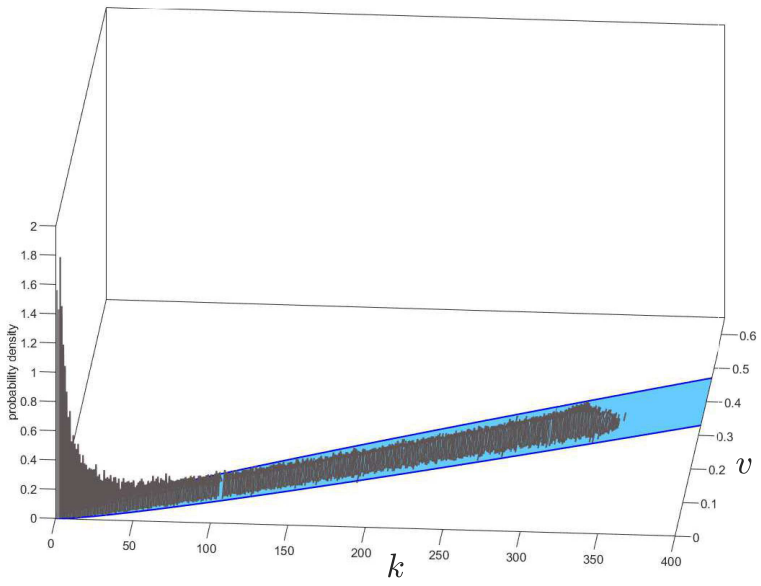

Figure 9: Empirical probability distribution of $(s^*, V(x^*))$ when points are generated by $\mathbb{P}_B$.

Figure 8 shows the empirical bivariate distribution of the values (k, v) taken by $(s^*, V(x^*))$ over the 100 000 trials in the case of probability $\mathbb{P}_A$. The reader can notice that a blue slanted region is also represented in the (k, v) plane. This region is precisely introduced in the theoretical developments presented below, and, for the time being, we just notice that the distribution of $(s^*, V(x^*))$ appears to be supported on this slanted region. A similar picture for $\mathbb{P}_B$ is given in Figure 9. In this second case, the slanted region is exactly as in the first case and it happens again that the support of the distribution of $(s^*, V(x^*))$ belongs to it.

We next make overt that the marginal distributions of $V(x^*)$ under $\mathbb{P}_A$ and $\mathbb{P}_B$ are those represented in Figure 7 as dashed yellow and dotted red lines respectively. As already discussed in Section 3.2, the Beta distribution of Figure 7 only sets an upper limit to these marginals. The present simulations in Figures 8 and 9 suggest that more information can be gained from the lens of a bivariate point of view where one variable, the risk, is estimated from the other, the com-

plexity (which is a measurable quantity). We anticipate that this result is true in high generality and that the precision in the evaluation of $V(x^*)$ that can be achieved thanks to this new lens is comparable to the precision that comes in fully-supported problems from using relation (4), indeed a remarkable finding.

A general result

As is clear from our simulations, the distribution of the pair $(s^*, V(x^*))$ can take various forms. However, in the two examples that we have just seen this distribution was confined in the slanted region which, for easy reference, is again displayed in Figure 10.

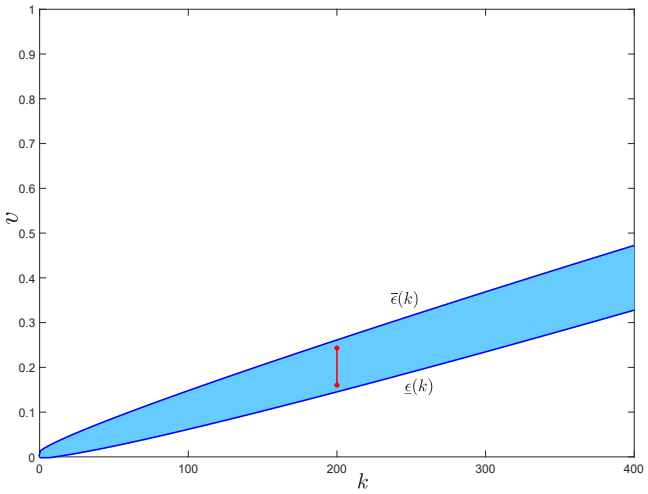

Figure 10: The slanted blue area is the region to which at least $1 - \beta = 99.9\%$ of the probabilistic mass of $(s^*, V(x^*))$ belongs for any non-degenerate convex scenario problem (as defined in Footnote 6) in the form of (1) when $d = 400$ and $N = 1000$. Moreover, for any problem in the form of (1) (degenerate or non-degenerate) at least $1 - \beta = 99.9\%$ of the probabilistic mass of $(s^*, V(x^*))$ lies below the upper boundary of the blue area. For the sake of comparison, in the figure a red interval is represented which is a 99.9% confidence interval of the risk for a fully-supported problem in dimension 200 (the interval has been obtained by using the Beta distribution).

The rule by which the boundaries of the slanted region are constructed is as follows.

Rule to compute the boundaries of the slanted region

Assume $N > d$. Given a confidence parameter $\beta \in (0,1)$, for any $k = 0, \ldots, d$ consider the polynomial equation in the t variable

$$\binom{N}{k} t^{N-k} - \frac{\beta}{2N} \sum_{i=k}^{N-1} \binom{i}{k} t^{i-k} - \frac{\beta}{6N} \sum_{i=N+1}^{4N} \binom{i}{k} t^{i-k} = 0. \qquad (6)$$

For any $k = 0, 1, \ldots, d$, equation (6) has exactly two solutions in $[0, +\infty)$, which we denote with $\underline{t}(k)$ and $\bar{t}(k)$ ($\underline{t}(k) \leq \bar{t}(k)$). Define $\underline{\epsilon}(k) := \max\{0, 1 - \bar{t}(k)\}$ and $\bar{\epsilon}(k) := 1 - \underline{t}(k)$, $k = 0, \ldots, d$. Function $\underline{\epsilon}(k)$ is the lower boundary of the slanted region and $\bar{\epsilon}(k)$ its upper boundary.

A MATLAB code for computing the values of $\underline{\epsilon}(k)$ and $\bar{\epsilon}(k)$ according to this rule is available in the Appendix A of Garatti & Campi (2019).

We now have the following result (it is assumed that the solution x^* exists and is unique, possibly after the use of a suitable tie-break rule).

Wait-and-judge result for convex scenario optimization. *For any non-degenerate problem (as defined in Footnote 6) in the form of* (1)*, it holds that*

$$\mathbb{P}^N\{\underline{\epsilon}(s^*) \leq V(x^*) \leq \bar{\epsilon}(s^*)\} \geq 1 - \beta; \qquad (7)$$

moreover, for any problem in the form of (1)*, it holds that*

$$\mathbb{P}^N\{V(x^*) \leq \bar{\epsilon}(s^*)\} \geq 1 - \beta. \qquad (8)$$

Hence, referring to Figure 10, in non-degenerate problems the distribution of $(s^*, V(x^*))$ is confined to the blue slanted region (but a small portion whose probability is no more than β) whereas in degenerate problems the distribution of $(s^*, V(x^*))$ can expand below the lower boundary of the slanted region, while the upper boundary, that sets a limit to $V(x^*)$, is always valid (this latter result is proven in the recent paper Garatti & Campi (2021)).

For the use of this result, the crucial fact to remark is that the quantity on the horizontal axis (complexity) is measurable, while the vertical axis corresponds to the value of the risk $V(x^*)$ that, in real life, is hidden to the decision-maker and can only be estimated. Thus, from Figure 10, one obtains a rule to bound $V(x^*)$ based on the observable s^* and the result in (7) provides sample-dependent bounds of the kind $\underline{\epsilon}(s^*) \leq V(x^*) \leq \bar{\epsilon}(s^*)$ (or just $V(x^*) \leq \bar{\epsilon}(s^*)$, in degenerate cases thanks to (8)) that are valid with high probability $1 - \beta$. The interval $[\underline{\epsilon}(k), \bar{\epsilon}(k)]$ for a given k is comparable to the one that can be generated when working with a fully-supported problem in dimension $d = k$. One of these intervals with $d = 200$, is represented in Figure 10. The quantitative similarity between $[\underline{\epsilon}(k), \bar{\epsilon}(k)]$ and the intervals in the fully-supported case is a quite remarkable fact and reveals the value of the information conveyed by the complexity. From results in Campi & Garatti (2018); Garatti & Campi (2019), one also sees that β impacts on $\underline{\epsilon}(k)$ and $\bar{\epsilon}(k)$ logarithmically so that taking very small values of β enlarges only marginally the interval $[\underline{\epsilon}(k), \bar{\epsilon}(k)]$; moreover, for any k,

$$\bar{\epsilon}(k) - \underline{\epsilon}(k) \to 0$$

as $N \to \infty$, i.e., the interval gets more and more informative as the number of data increases, see Campi & Garatti (2020).

3.3. Non-convex worst-case optimization

When the assumption of convexity is dropped, the bound $s^* \leq d$ loses validity. For example, let $x = (x_1, x_2) \in [-1, 1]^2$, $c(x) = x_2$, $\delta \in [-1, 1]$ and $f(x, \delta) = -|x_1 - \delta| - x_2$. An instance of the corresponding scenario program is pictured in Figure 11, and the reader can easily check that any subsample of the 6 observed scenarios yields a solution x^* different from the one obtained with all the scenarios, so that $s^* = 6 > 2 = d$ in this case. Moreover, the same example reveals that a constraint need not be active to be necessary to reconstruct the solution in the non-convex case. This circumstance suggests that degeneracy

becomes quite a common circumstance (the reader may want to think of the case in which one more constraint is added in Figure 11 that "shields" the global minimum that opens up after removing one of the 6 constraints in that figure and observe that this generates a degenerate situation).

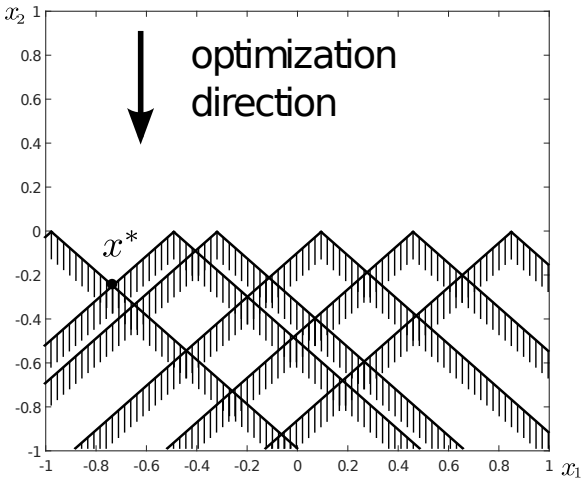

Figure 11: An instance of a non-convex scenario program where $d = 2$ but $s^* = 6$.

Remarkably, the results in the recent paper Garatti & Campi (2021) do not require that $s^* \leq d$ and are applicable even in the degenerate case, leading to the following result (x^* is assumed to exist and to be unique, possibly after the use of a suitable tie-break rule).

Wait-and-judge result for non-convex scenario optimization. *Use equation* (6) *to compute* $\bar{\epsilon}(k)$ *for* $k = 0, 1, \ldots, N$ *(note that the range of* k *is here extended till* N*).*[8] *Then, for any problem in the form of* (1)*, where* $c(x)$ *and* $f(x, \delta)$ *need not be convex in* x *and* $\mathcal{X}$ *is any set, it holds that*

$$\mathbb{P}^N\{V(x^*) \leq \bar{\epsilon}(s^*)\} \geq 1 - \beta. \tag{9}$$

[8] The construction for $k = 0, 1, \ldots, N-1$ is exactly as indicated in the "Rule to compute the boundaries of the slanted region"; for $k = N$, however, equation (6) has only one solution $\bar{t}(N)$, and one defines $\underline{t}(N) = 0$, so that $\bar{\epsilon}(N) = 1$.

The curve $\bar{\epsilon}(k)$ for $N = 1000$ and $\beta = 0.1\%$ is displayed in Figure 12. Note that, for k in the range $[0, 400]$, $\bar{\epsilon}(k)$ is exactly as in Figure 10.

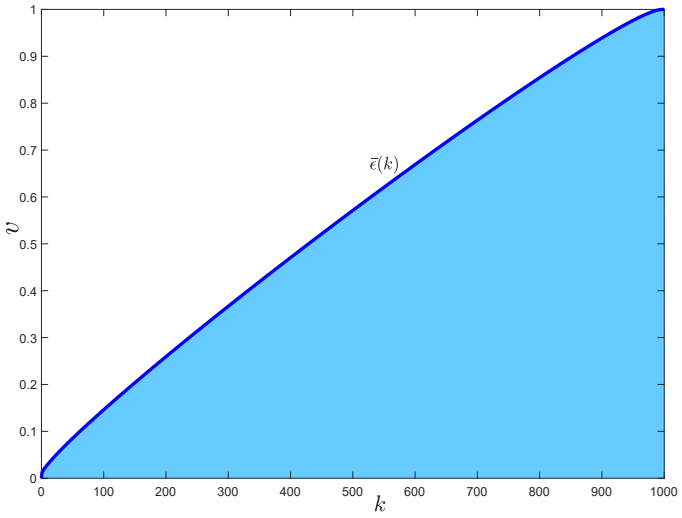

Figure 12: The blue area is the region to which at least $1-\beta = 99.9\%$ of the probabilistic mass of $(s^*, V(x^*))$ belongs for any non-convex scenario problem in the form of (1) when $d = 400$ and $N = 1000$.

3.4. Tuning of the extrinsic quality

In (1), all scenario constraints are rigidly enforced, so expressing an attitude to safeguard against the worst. This, however, can lead to conservative designs: the presence of just one ill scenario can forbid a whole set of candidate solutions and confine the choice to solutions with poor intrinsic quality $c(x)$ (see, e.g., Shapiro et al. (2009); Ramponi (2018); Assif et al. (2020)). On the other hand, in many applications, $c(x)$ and $V(x)$ are seen as conflicting objectives and in this section we focus on a class of decision schemes that allow the decision-maker to compromise between intrinsic and extrinsic quality.

According to the interpretation of $f(x, \delta)$ provided in Section 3.1, a violation of the condition $f(x, \delta^{(i)}) \leq 0$ yields a regret. The first scheme relaxes the rigid enforcement of constraints by minimizing a weighted combination of the cost $c(x)$ and the sum of the (positive) regrets $\sum_{i=1}^{N} \left[f(x, \delta^{(i)})\right]^+$ ($[\cdot]^+$ is positive part, that is, it returns its argument when it is positive and zero otherwise).

This yields the following optimization program where the weight α assigned to the regrets is a tunable trade-off parameter:[9]

$$\begin{aligned} \min_{\substack{x\in\mathcal{X}\subseteq\mathbb{R}^d,\\ \xi_i\geq 0, i=1,\ldots,N}} \quad & c(x)+\alpha\sum_{i=1}^{N}\xi_i \\ \text{subject to: } & f(x,\delta^{(i)})\leq\xi_i \;\; i=1,\ldots,N. \end{aligned} \tag{10}$$

Let $(x^*_\alpha, \xi^*_{i,\alpha})$ be the solution to (10) (assumed to exist and to be unique, possibly after the use of a suitable tie-break rule). In this context, a suitable generalization of the concept of complexity allows one to estimate the extrinsic quality of x^*_α (i.e., quantity $V(x^*_\alpha)=\mathbb{P}\{\delta : f(x^*_\alpha,\delta)>0\}$).[10] The *generalized complexity* s^*_α is the cardinality of a subsample S of scenarios from $\delta^{(1)},\ldots,\delta^{(N)}$ defined as follows: S contains all the violated scenarios (i.e., the scenarios for which $f(x^*_\alpha,\delta^{(i)})>0$)[11] plus a minimum subsample of the remaining scenarios such that the x-component of the solution to (10) with only S in place remains equal to x^*_α. Then, exactly the same result as in "Wait-and-judge result for non-convex optimization" holds (without requiring any convexity assumption) with the only warning that x^* and s^* must be replaced in the present context by x^*_α and s^*_α.[12] The value of α (which ranges from $\alpha=0$, corresponding to minimizing $c(x)$ with no concern for the observed scenarios, to $\alpha=\infty$, corresponding to the worst-case approach) influences s^*_α (and, thereby, the bounds on $V(x^*_\alpha)$) and $c(x^*_\alpha)$ and represents a tuning knob in the hands of the decision-maker. By plotting $c(x^*_{\alpha_1}), c(x^*_{\alpha_2}),\ldots$ against the bounds $\bar{\epsilon}(s^*_{\alpha_1}), \bar{\epsilon}(s^*_{\alpha_2}),\ldots$ for $V(x^*_{\alpha_1}), V(x^*_{\alpha_2}),\ldots$,

[9] The set of all the optimal decisions x^*_α obtained from (10) as a function of α is known as the Pareto frontier of the multi-objective problem of minimizing $c(x)$ and $\sum_{i=1}^{N}\left[f(x,\delta^{(i)})\right]^+$.

[10] The need for a generalization of the concept of complexity can be recognized by noting that problem (10) is not directly in the form of (1), in particular the number of optimization variables x, ξ_i increases with N.

[11] The subsample S is with repetitions, i.e., if two scenarios $\delta^{(i)}$ and $\delta^{(j)}$ for which $f(x^*_\alpha,\delta^{(i)})>0$ and $f(x^*_\alpha,\delta^{(j)})>0$ turn out to be coincident ($\delta^{(i)}=\delta^{(j)}$), then they both appear in S.

[12] Also in this context a concept of non-degeneracy can be applied leading to the stronger result that $V(x^*_\alpha)$ is both upper and lower bounded, see Garatti & Campi (2019).

one obtains the so-called *cost-risk plot* (see Garatti & Campi (2019) for an example) by which the user can perform a selection of a suitable value for α.

Program (10) is not the only scheme to tune the intrinsic vs. the extrinsic quality. Alternatively, one can discard some of the constraints from the worst-case program (1) and the reader is referred to the papers Campi & Garatti (2011); Garatti & Campi (2013); Picallo & Dörfler (2019); Romao et al. (2020) for studies in this direction.

3.5. A general theory for decision-making

The scenario theory carries over to an abstract decision-making framework that encompasses all previous set-ups and many others as special cases. Here, we briefly summarize the results in Garatti & Campi (2019, 2021), to which the reader is referred for details.

For any $N = 0, 1, 2, \ldots$ let M_N be a map from any set of scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$ to a decision $z^* \in \mathcal{Z}$, where $\mathcal{Z}$ is a generic decision set. The maps $M_0, M_1, \ldots$ describe the rule according to which decisions are made based on observations in an integrated data-driven set-up as in Figure 1. In order to evaluate whether the extrinsic criterion is met, there should be a rule to decide whether a decision z performs "well" when a situation δ occurs. This is expressed in mathematical terms by saying that, to any situation δ, there is associated a set $\mathcal{Z}_\delta \subseteq \mathcal{Z}$ which models the set of the decisions that perform well for δ. The risk of a decision $z \in \mathcal{Z}$ is then defined as $V(z) := \mathbb{P}\{\delta : z \notin \mathcal{Z}_\delta\}$.

No limiting assumptions on the domain of δ and on the map between δ and $\mathcal{Z}_\delta$ are necessary. The freedom of the decision-maker in choosing the maps $M_0, M_1, \ldots$ is also vast as long as the following three properties are satisfied:

- **Permutation invariance**:
  for every N, every $\delta^{(1)}, \ldots, \delta^{(N)}$ and every permutation $(i_1, \ldots, i_N)$ of $(1, \ldots, N)$ it holds that

$$M_N(\delta^{(1)}, \ldots, \delta^{(N)}) = M_N(\delta^{(i_1)}, \ldots, \delta^{(i_N)}).$$

- **Stability in the case of confirmation**:
  for every integers N and n, if $\delta^{(1)}, \ldots, \delta^{(N)}, \delta^{(N+1)}, \ldots, \delta^{(N+n)}$ are such that
  $$\forall i \in \{1, \ldots, n\} : \ M_N(\delta^{(1)}, \ldots, \delta^{(N)}) \in \mathcal{Z}_{\delta^{(N+i)}},$$
  then
  $$M_N(\delta^{(1)}, \ldots, \delta^{(N)}) = M_{N+n}(\delta^{(1)}, \ldots, \delta^{(N+n)}).$$

- **Responsiveness to contradiction**:
  for every integers N and n, if $\delta^{(1)}, \ldots, \delta^{(N)}, \delta^{(N+1)}, \ldots, \delta^{(N+n)}$ are such that
  $$\exists i \in \{1, \ldots, n\} : \ M_N(\delta^{(1)}, \ldots, \delta^{(N)}) \notin \mathcal{Z}_{\delta^{(N+i)}},$$
  then
  $$M_{N+n}(\delta^{(1)}, \ldots, \delta^{(N+n)}) \neq M_N(\delta^{(1)}, \ldots, \delta^{(N)}).$$

Note that the properties of M_N as written above *do not* imply that $z^* \in \mathcal{Z}_{\delta^{(i)}}$, $i = 1, \ldots, N$.

Complexity and degeneracy

Given the scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$, the *complexity* s^* of the decision $z^* = M_N(\delta^{(1)}, \ldots, \delta^{(N)})$ is the cardinality of a minimum subsample S (with repetitions) of scenarios from $\delta^{(1)}, \ldots, \delta^{(N)}$ such that $M_{s^*}(S) = z^*$. The decision-scheme is said to be *non-degenerate* if, for every N, there is with probability 1 a unique choice of indexes $i_1, i_2, \ldots, i_k$ (with $i_1 < i_2 < \ldots < i_k$) from $1, 2, \ldots, N$ such that $M_k(\delta^{(i_1)}, \ldots, \delta^{(i_k)}) = M_N(\delta^{(1)}, \ldots, \delta^{(N)})$ while, for $k > 0$, discarding further indexes from $i_1, i_2, \ldots, i_k$ changes the solution.

Guarantees

Similarly to "Wait-and-judge result for non-convex optimization" in Section

3.3, using (6) one computes $\bar{\epsilon}(k)$ for $k = 0, 1, \ldots, N$. Then, equation (9) applies where x^* is replaced in the present context by z^*, while s^* and $V(z^*)$ have to be interpreted according to the definitions of this section. Moreover, under non-degeneracy, one can also compute $\underline{\epsilon}(k)$, $k = 0, 1, \ldots, N$, by extending till $k = N$ the rule for $\underline{\epsilon}(k)$, $k = 0, 1, \ldots, d$, in "Wait-and-judge result for convex optimization" in Section 3.2 and equation (7) holds where again x^* has to be replaced by z^*, while s^* and $V(z^*)$ have to be interpreted according to the definitions of this section.

A bit of history

The "responsiveness to contradiction" property was a key property in important generalization results in the history of statistical learning, Littlestone & Warmuth (1986). This property, alone, is sufficient to derive various results within the scenario theory that are applicable to a wide range of decision schemes, Campi et al. (2018). On the other hand, the "stability in the case of confirmation" property plays a special role in obtaining tight results (and, under non-degeneracy, small ranges $[\underline{\epsilon}(k), \bar{\epsilon}(k)]$) and, when satisfied, confers considerable added value to the scenario approach. Being naturally satisfied in optimization problems, this property can be recognized as one of the main "secrets" behind the success of the scenario approach in many applications.

References

Alamo, T., Tempo, R., Luque, A., & Ramirez, D. R. (2015). Randomized methods for design of uncertain systems: Sample complexity and sequential algorithms. *Automatica*, *52*, 160–172.

Apkarian, P., & Noll, D. (2018). Structured H_∞-control of infinite-dimensional systems. *International Journal of Robust and Nonlinear Control*, *28*, 3212–3238.

Assif, M., Chatterjee, D., & Banavar, R. (2020). Scenario approach for min-max optimization with emphasis on the nonconvex case: Positive results and caveats. *SIAM Journal on Optimization*, *30*, 1119–1143.

Åström, K. J., & Hägglund, T. (1995). *PID controllers: theory, design, and tuning* volume 2. Instrument society of America, Research Triangle Park, NC.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, *44*, 2743–2760.

Bazanella, A. S., Campestrini, L., & Eckhard, D. (2011). *Data-driven Controller Design: the H_2 Approach*. Springer Science & Business Media.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*, 637–654.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*, 199–231.

Calafiore, G. C. (2010a). Learning noisy functions via interval models. *Systems & Control Letters*, *59*, 404 – 413.

Calafiore, G. C. (2010b). Random convex programs. *SIAM Journal on Optimization*, *20*, 3427–3464.

Calafiore, G. C. (2013). Direct data-driven portfolio optimization with guaranteed shortfall probability. *Automatica*, *49*, 370–380.

Campi, M. C. (2010). Classification with guaranteed probability of error. *Machine Learning*, *80*, 63–84.

Campi, M. C., Calafiore, G., & Garatti, S. (2009a). Interval predictor models: identification and reliability. *Automatica*, *45*, 382–392.

Campi, M. C., & Carè, A. (2013). Random convex programs with L_1-regularization: sparsity and generalization. *SIAM Journal on Control and Optimization*, *51*, 3532–3557.

Campi, M. C., & Garatti, S. (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, *19*, 1211–1230.

Campi, M. C., & Garatti, S. (2011). A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, *148*, 257–280.

Campi, M. C., & Garatti, S. (2018). Wait-and-judge scenario optimization. *Mathematical Programming*, *167*, 155–189.

Campi, M. C., & Garatti, S. (2020). Scenario optimization with relaxation: a new tool for design and application to machine learning problems. https://arxiv.org/abs/2004.05839.

Campi, M. C., Garatti, S., & Prandini, M. (2009b). The scenario approach for systems and control design. *Annual Reviews in Control*, *33*, 149 – 157.

Campi, M. C., Garatti, S., & Ramponi, F. A. (2018). A general scenario theory for nonconvex optimization and decision making. *IEEE Transactions on Automatic Control*, *63*, 4067–4078.

Campi, M. C., Lecchini, A., & Savaresi, S. M. (2002). Virtual Reference Feedback Tuning: a direct method for the design of feedback controllers. *Automatica*, *38*, 1337–1346.

Cannon, W. B. (1939). *The Wisdom of the Body*. Norton & Co.

Cappon, G., Facchinetti, A., Sparacino, G., Georgiou, P., & Herrero, P. (2019). Classification of postprandial glycemic status with application to insulin dosing in type 1 diabetes – an in silico proof-of-concept. *Sensors*, *19*, 3168.

Carè, A., Campi, M. C., & Garatti, S. (2017). A coverage theory for least squares. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 1367–1389.

Carè, A., Garatti, S., & Campi, M. C. (2014). FAST – Fast Algorithm for the Scenario Technique. *Operations Research*, *62*, 662–671.

Carè, A., Garatti, S., & Campi, M. C. (2015). Scenario min-max optimization and the risk of empirical costs. *SIAM Journal on Optimization*, *25*, 2061–2080.

Carè, A., Garatti, S., & Campi, M. C. (2019). The wait-and-judge scenario approach applied to antenna array design. *Computational Management Science*, *16*, 481–499.

Carè, A., Ramponi, F. A., & Marco, M. C. (2018). A new classification algorithm with guaranteed sensitivity and specificity for medical applications. *IEEE Control Systems Letters*, *2*, 393–398.

Chiluka, S. K., Ambati, S. R., Seepana, M. M., & Babu Gara, U. B. (2021). A novel robust Virtual Reference Feedback Tuning approach for minimum and non-minimum phase systems. *ISA Transactions*, *in press*. doi:`https://doi.org/10.1016/j.isatra.2021.01.018`.

Crespo, L. G., Kenny, S. P., & Giesy, D. P. (2016). Interval predictor models with a linear parameter dependency. *Journal of Verification, Validation and Uncertainty Quantification*, *1*, 021007.

Devi, S., Malarvezhi, P., Dayana, R., & Vadivukkarasi, K. (2020). A comprehensive survey on autonomous driving cars: A perspective view. *Wireless Personal Communication*, *114*, 2121–2133.

Esfahani, P. M., Sutter, T., & Lygeros, J. (2015). Performance bounds for the scenario approach and an extension to a class of non-convex programs. *IEEE Transactions on Automatic Control*, *60*, 46–58.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115–118.

Formentin, S., Campi, M. C., Carè, A., & Savaresi, S. M. (2019). Deterministic continuous-time Virtual Reference Feedback Tuning (VRFT) with application to PID design. *Systems & Control Letters*, *127*, 25–34.

Formentin, S., Heusden, K., & Karimi, A. (2014). A comparison of model-based and data-driven controller tuning. *International Journal of Adaptive Control and Signal Processing*, *28*, 882–897.

Garatti, S., & Campi, M. C. (2013). Modulating robustness in control design: principles and algorithms. *IEEE Control Systems Magazine*, *33*, 36–51.

Garatti, S., & Campi, M. C. (2019). Risk and complexity in scenario optimization. *Mathematical Programming*, *in press*. doi:`https://doi.org/10.1007/s10107-019-01446-4`.

Garatti, S., & Campi, M. C. (2021). The risk of making decisions from data through the lens of the scenario approach. In *Proceedings of the 19th IFAC Symposium on System Identification*. Padua, Italy.

Garatti, S., Campi, M. C., & Carè, A. (2019). On a class of interval predictor models with universal reliability. *Automatica*, *110*, 108542.

Geng, X., & Xie, L. (2019). Data-driven decision making in power systems with probabilistic guarantees: Theory and applications of chance-constrained optimization. *Annual Reviews in Control*, *47*, 341–363.

Gerencsér, L., Vágó, Z., & Hjalmarsson, H. (2002). Randomization methods in optimization and adaptive control. In B. Pasik-Duncan (Ed.), *Stochastic Theory and Control* (pp. 137–153). Springer Berlin Heidelberg.

Ghasedi Dizaji, K., Herandi, A., Deng, C., Cai, W., & Huang, H. (2017). Deep clustering via joint convolutional autoencoder embedding and relative entropy

minimization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 5736–5745).

Graepel, T., Herbrich, R., & Shawe-Taylor, J. (2005). PAC-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, *59*, 55–76.

Grammatico, S., Zhang, X., Margellos, K., Goulart, P. J., & Lygeros, J. (2016). A scenario approach for non-convex control design. *IEEE Transactions on Automatic Control*, *61*, 334–345.

Gruyer, D., Magnier, V., Hamdi, K., Claussmann, L., Orfila, O., & Rakotonirainy, A. (2017). Perception, information processing and modeling: Critical stages for autonomous driving applications. *Annual Reviews in Control*, *44*, 323 – 341.

Guanetti, J., Kim, Y., & Borrelli, F. (2018). Control of connected and automated vehicles: State of the art and future challenges. *Annual Reviews in Control*, *45*, 18–40.

Guardabassi, G. O., & Savaresi, S. M. (2000). Virtual reference direct design method: an off-line approach to data-based control system design. *IEEE Transactions on Automatic Control*, *45*, 954–959.

Hanneke, S., & Kontorovich, A. (2019). A sharp lower bound for agnostic learning with sample compression schemes. In *30th International Conference on Algorithmic Learning Theory* (pp. 489–505).

Hjalmarsson, H. (2002). Iterative Feedback Tuning — an overview. *International Journal of Adaptive Control and Signal Processing*, *16*, 373–395.

Hjalmarsson, H., Gevers, M., Gunnarsson, S., & Lequin, O. (1998). Iterative Feedback Tuning: theory and applications. *IEEE Control Systems Magazine*, *18*, 26–41.

Hoagland, M. B., & Dodson, B. (1995). *The Way Life Works*. Times Books.

Hori, T., Yubai, K., Yashiro, D., & Komada, S. (2016). Data-driven controller tuning for sensitivity minimization. In *2016 International Conference on Advanced Mechatronic Systems (ICAMechS)* (pp. 132–137).

Hou, Z.-S., & Wang, Z. (2013). From model-based control to data-driven control: Survey, classification and perspective. *Information Sciences*, *235*, 3–35.

Karimi, A., & Kammer, C. (2017). A data-driven approach to robust control of multivariable systems by convex optimization. *Automatica*, *85*, 227–233.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Lacerda, M. J., Martins, S. A. M., & Nepomuceno, E. G. (2018). Structure selection based on interval predictor model for recovering static non-linearities from chaotic data. *IET Control Theory & Applications*, *12*, 1889–1894(5).

Li, S., Lian, J., Conejo, A. J., & Zhang, W. (2020). Transactive energy systems: The market-based coordination of distributed energy resources. *IEEE Control Systems Magazine*, *40*, 26–52.

Littlestone, N., & Warmuth, M. (1986). Relating data compression and learnability. Technical report, University of California Santa Cruz.

Macfarlane, J. (2019). When apps rule the road: The proliferation of navigation apps is causing traffic chaos. it's time to restore order. *IEEE Spectrum*, *56*, 22–27.

Margellos, K., Falsone, A., Garatti, S., & Prandini, M. (2018). Distributed constrained optimization and consensus in uncertain networks via proximal minimization. *IEEE Transactions on Automatic Control*, *63*, 1372–1387.

Margellos, K., Prandini, M., & Lygeros, J. (2015). On the connection between compression learning and scenario based single-stage and cascading optimization problems. *IEEE Transactions on Automatic Control*, *60*, 2716–2721.

Ming, L., & Vitányi, P. M. B. (1990). Kolmogorov complexity and its applications. In *Algorithms and Complexity* (pp. 187–254). Elsevier.

Modarresi, M. S., Xie, L., Campi, M. C., Garatti, S., Carè, A., Thatte, A. A., & Kumar, P. R. (2019). Scenario-based economic dispatch with tunable risk levels in high-renewable power systems. *IEEE Transactions on Power Systems*, *34*, 5103–5114.

Moore, K. L. (2012). *Iterative Learning Control for Deterministic Systems*. Springer Science & Business Media.

Moran, S., & Yehudayoff, A. (2016). Sample compression schemes for VC classes. *Journal of the ACM*, *63*.

Nasir, H. A., Carè, A., & Weyer, E. (2018). A scenario-based stochastic MPC approach for problems with normal and rare operations with an application to rivers. *IEEE Transactions on Control Systems Technology*, *27*, 1397–1410.

Nasir, H. A., Garatti, S., & Weyer, E. (2016). Scenario based stochastic MPC schemes for rivers with feasibility assurance. In *2016 European Control Conference (ECC)* (pp. 1928–1933).

Norvig, P. (2017). On Chomsky and the two cultures of statistical learning. In W. Pietsch, J. Wernecke, & M. Ott (Eds.), *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data* (pp. 61–83). Wiesbaden: Springer Fachmedien Wiesbaden.

Novara, C., & Milanese, M. (2019). Control of MIMO nonlinear systems: A data-driven model inversion approach. *Automatica*, *101*, 417–430.

Paccagnan, D., & Campi, M. C. (2019). The scenario approach meets uncertain game theory and variational inequalities. In *2019 IEEE 58th Conference on Decision and Control (CDC)* (pp. 6124–6129).

Pagnoncelli, B. K., Reich, D., & Campi, M. C. (2012). Risk-return trade-off with the scenario approach in practice: A case study in portfolio selection. *Journal of Optimization Theory and Applications*, *155*, 707–722.

Papageorgiou, M., & Papamichail, I. (2014). Coordinated ramp metering for freeways. In T. Samad, & A. M. Annaswamy (Eds.), *The Impact of Control Technology, 2nd Edition*. IEEE Control Systems Society.

Patelli, E., Broggi, M., Tolo, S., & Sadeghi, J. (2013). Cossan software: A multi-disciplinary and collaborative software for uncertainty quantification. In *Proceedings of the 2nd ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering, UNCECOMP 2017, Rhodes Island, Greece*.

Picallo, M., & Dörfler, F. (2019). Sieving out unnecessary constraints in scenario optimization with an application to power systems. In *2019 IEEE 58th Conference on Decision and Control (CDC)* (pp. 6100–6105).

Quiroz, G. (2019). The evolution of control algorithms in artificial pancreas: A historical perspective. *Annual Reviews in Control*, *48*, 222 – 232.

Rallo, G., Formentin, S., Garatti, S., & Savaresi, S. M. (2016). Vehicle stability control via VRFT with probabilistic robustness guarantees. In *2016 IEEE 55th Conference on Decision and Control (CDC)* (pp. 7165–7170).

Ramponi, F. A. (2018). Consistency of the scenario approach. *SIAM Journal on Optimization*, *28*, 135–162.

Ramponi, F. A., & Campi, M. C. (2018). Expected shortfall: Heuristics and certificates. *European Journal of Operational Research*, *267*, 1003–1013.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 465–471.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, (pp. 1080–1100).

Romao, L., Margellos, K., & Papachristodoulou, A. (2020). Tight generalization guarantees for the sampling and discarding approach to scenario optimization.

In *2020 59th IEEE Conference on Decision and Control (CDC)* (pp. 2228–2233).

Safonov, M. G., & Tsao, T.-C. (1995). The unfalsified control concept: A direct path from experiment to controller. In *Feedback Control, Nonlinear Systems, and Complexity* (pp. 196–214). Springer.

Schildbach, G., Fagiano, L., & Morari, M. (2013). Randomized solutions to convex programs with multiple chance constraints. *SIAM Journal on Optimization*, *23*, 2479–2501.

Seo, T., Bayen, A. M., Kusakabe, T., & Asakura, Y. (2017). Traffic state estimation on highway: A comprehensive survey. *Annual Reviews in Control*, *43*, 128 – 151.

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2009). *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia, USA: MPS-SIAM.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–503.

Sutter, T., Kamoutsi, A., Esfahani, P. M., & Lygeros, J. (2017). Data-driven approximate dynamic programming: A linear programming approach. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)* (pp. 5174–5179).

Van Heusden, K., Karimi, A., & Bonvin, D. (2011). Data-driven model reference control with asymptotically guaranteed stability. *International Journal of Adaptive Control and Signal Processing*, *25*, 331–351.

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Vidyasagar, M. (2014). *Hidden Markov Processes: Theory and Applications to Biology*. Princeton University Press.

Wang, C., Shang, C., Yang, F., Huang, D., & Yu, B. (2020). Robust interval prediction model identification with a posteriori reliability guarantee. In *Proceedings of the 21th IFAC World Congress*. Berlin, Germany.

Zhang, X., Grammatico, S., Schildbach, G., Goulart, P. J., & Lygeros, J. (2015). On the sample size of random convex programs with structured dependence on the uncertainty. *Automatica*, *60*, 182–188.