UNIVERSITÀ DEGLI STUDI DI BRESCIA

Dipartimento di Ingegneria dell'Informazione

XXV CICLO DI DOTTORATO DI RICERCA IN
INGEGNERIA INFORMATICA ED AUTOMATICA

SSD: ING-INF/04

Data-Based Optimization for Applications to Decision-Making, Identification and Control

A Study of Coverage Properties

Relatore:
Prof. Marco C. Campi
Correlatore:
Ing. Simone Garatti
Coordinatore del dottorato:
Prof. Alfonso Gerevini

Dottorando:
Algo Carè

A.A. 2011/2012

To Faërie

Overview

In this work we provide new theoretical results fit for use by decision-makers called to cope with uncertainty. We will focus on single-objective decision problems where a cost function beset by uncertainty has to be minimized. In these contexts, which are common in control systems engineering, operations research, finance, etc., a widespread heuristic is that of making a decision based on a set of collected data called *scenarios*. We will focus on two important approaches to data-based decision-making: the *average approach* with quadratic cost function (least-squares decision rule) and the *worst-case approach* with convex cost function (min-max decision rule).
Once the optimal decision has been computed according to the selected data-based approach, we are interested in the probability that a new situation, i.e. a new uncertainty instance, carries a cost no higher than some empirically meaningful cost thresholds. The probability that a cost threshold is not exceeded is called *coverage*. By describing the coverage properties of meaningful cost thresholds, we gain quantitative information about the reliability of our decision. Some recent theoretical developments have shown that, under the hypothesis that the scenarios are drawn independently according to a fixed probability distribution, coverage properties can - in situations of great interest - be effectively studied in a distribution-free context, that is, without any knowledge of the probability distribution according to which data are generated. In this work, we will follow this same line. We will determine meaningful cost thresholds (that are, in statistical terms, meaningful statistics of the data) apt to characterize least-squares and min-max decisions, and provide the decision-maker with analytical tools to evaluate the distribution of the costs, without knowing the probability distribution of the data.

Description by chapters, and notes on original contributions

Chapter 1 - Decision-making in the presence of uncertainty
The mathematical framework is introduced, a common background for the concepts used throughout this work is provided. Motivations for our studies are given, and the results in the following chapters are surveyed. Some works related - by affinity or by opposition - to our approach are briefly discussed at the end of this chapter.

Chapter 2 - The coverage probabilities of the least squares residuals
This chapter deals with data-based least-squares decisions. An algorithm to compute characterizing statistics, having interesting distribution-free coverage properties, is provided. Results presented in this chapter are still unpublished.

Chapter 3 - On the reliability of data-based min-max decisions
This chapter deals with data-based min-max decisions with convex cost functions. The most important known result in this context is given by the theory of the scenario approach to constrained convex optimization (which is recalled for completeness' sake in the Appendix A), stating that the empirical worst-case cost has distribution-free coverage for a whole class of problems. This result is here extended to all the others empirical costs: the joint probability distribution of the coverages of all the empirical costs turns out to be an ordered Dirichlet distribution, independently of the distribution of the data. The material in this chapter has been partially published by the author of this thesis together with Simone Garatti and Marco C.Campi, [1].

Chapter 4 - Data-based min-max decisions with reduced sample complexity
In this chapter, we propose an algorithm that allows the decision-maker to characterize, by means of a statistic having guaranteed high coverage, a min-max decision even when the number of observed scenarios is smaller than that required by the standard approach. The idea presented in this chapter has been published, together with Simone Garatti and Marco C.Campi, in [2].

Appendix A - The scenario approach to constrained convex optimization problems
We summarize for easy reference the most important known results in the theory of the scenario approach for constrained convex optimization.

Appendix B - Generalized FAST algorithm
A more general version of the algorithm of Chapter 4 is presented.

Compendio

In questo lavoro si forniscono risultati utili al decisore chiamato ad affrontare situazioni di incertezza. Ci concentreremo su problemi decisionali a singolo obiettivo, nei quali si richiede di minimizzare una funzione di costo affetta da incertezza. In tali contesti, che accomunano decisori negli ambiti dell'ingegneria del controllo, della ricerca operativa, della finanza, etc., un procedimento euristico diffuso suggerisce di prendere una decisione basandosi su una raccolta di dati, detti *scenari*. Prenderemo in considerazione due filosofie comunemente adottate nel prendere decisioni sulla base degli scenari: l'approccio ai "minimi quadrati", che prescrive di scegliere la decisione che minimizza il costo medio rispetto agli scenari, e l'approccio del "caso peggiore", che prescrive di minimizzare il costo più alto tra quelli dei diversi scenari. Una volta calcolata la decisione secondo uno dei due approcci considerati, siamo interessati alla probabilità che una nuova situazione porti con sé un costo non più alto di certe soglie empiricamente significative. La probabilità che una soglia di costo non sia superata è detta *copertura*. Attraverso la descrizione delle proprietà di copertura di opportune soglie di costo, si acquista dunque un'informazione quantitativa circa l'affidabilità della decisione adottata. Alcuni recenti sviluppi teorici dimostrano che, sotto l'ipotesi che gli scenari siano osservati indipendentemente e in accordo con una stessa distribuzione di probabilità, importanti proprietà della copertura possono, in contesti di grande interesse, essere studiate prescindendo dalla conoscenza della distribuzione di probabilità dei dati. Questo lavoro si colloca in tale prospettiva. Individueremo delle soglie di costo significative (le quali altro non sono, in termini statistici, che statistiche significative dei dati), fornendo al decisore uno strumento per valutare la distribuzione dei costi in corrispondenza della decisione presa, a prescindere dalla distribuzione dei dati.

Acknowledgments

I would like to express my special appreciation and thanks to my advisor, Prof. Marco Campi, for his guidance, trust, patience, and the uncountably many enlightening conversations and ideas, shared with me on a plane of equality, in spite of my slow-paced intellect and pretty irresistible stubbornness. For definitely similar reasons my sincere thanks go to Simone Garatti - definitely, a special thankful mention deserves his tremendous skill in turning coffee into counterexamples.
I am tempted not to mention any friend or colleague explicitly: I have a remarkable list of them, which this margin is too small to contain[1]*. I just cannot omit Eliana, and her polymorphic support. Finally, I thank my family and ancestors, for life is worth living.*

Algo

[1] But heart is not.

Contents

Chapter 1

Decision-making in the presence of uncertainty

In this chapter we introduce the main concepts used throughout this work, and provide a common framework for the results proved in the remaining chapters. In the following Section 1.1 the problem of making decisions in the presence of uncertainty is formalized. This is done step-by-step, with the purpose of providing motivations for the theoretical set-up used. In Section 1.3 the formal position of our problem is summarized and specialized to the cases covered in this work. A brief review of existing results related to our research concludes the chapter.

1.1 Theoretical set-up

The problem of making a decision can be abstracted as the problem of choosing a value x from within a decision set $\mathcal{X}$. We are interested in decision problems where the decision-maker wants to choose x so as to minimize a *cost*. The dependence of the cost on the decision x can, in principle, be modeled through a real-valued function $\ell(x)$, so that the best solution is but the solution to the following minimization problem:

$$\min_{x \in \mathcal{X}} \ell(x).$$

However, such a formalization turns out to be naive in many situations, due to that the cost incurred by the decision-maker is commonly beset by uncertainty. Uncertainty can enter the problem at various levels, but there are mainly two sources of uncertainty that are worth mentioning:

- unpredictability of phenomena;
- modeling errors.

Indeed, circumstances that are not directly under the decision-maker control and are not completely predictable may influence the cost of a decision.

Example 1 (unpredictability of phenomena)**.** *River banks should be built up to reduce the costs incurred in case of floods. The higher the banks, the higher is the building cost. For a given water level, we can compute the banks height required to prevent severe floods. But water levels changes in the course of time depending on weather conditions, which are variable and unpredictable.*

⋆

Also, it is rare to have a perfect model of the reality underlying the decision problem: *modeling errors* normally occur when physical systems are involved.

Example 2 (modeling errors)**.** x *represents the tunable parameters of an electric controller and* $\ell(x)$ *the maximum output voltage overshoot. Since some physical aspects of the real system may elude the model underlying the cost function (e.g. a resistance differs from its nominal value by* 5%*; there may be small unmodeled nonlinearities, etc.), the cost incurred is normally affected by uncertainty.*

⋆

In real life, both sources of uncertainty combine[1] and we need a way to face uncertainty in a general manner. We model the effect of the uncertainty by introducing in the cost function an uncertainty variable δ taking values in the uncertainty set Δ. Thus, the cost function is redefined as a bivariate function $\ell(x, \delta)$, where the presence of δ shows that to a fixed decision x a variety of possible costs is associated, depending on the value assumed by δ.

Example 3. *In the case of Example 1,* Δ *is the set of possible water levels. In the case of Example 2,* Δ *may represent the space of all the possible models of the system, more concretely, an interval of the possible values for an uncertain resistance.*

⋆

The uncertainty δ can be faced according to different approaches, as will be discussed in Section 1.4. In this work, we focus on the *scenario approach*, introduced in the following section.

1.1.1 Scenario Approach

The *scenario approach* is an intuitive heuristic used at large in optimization problems affected by uncertainty. It prescribes to face uncertain decision problems based on a finite number N of instances of the uncertain parameter δ. These instances of δ, $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, are called *scenarios*, and sometimes we will denote them more compactly with D^N. In order to produce the final decision x^*,

[1] Though clear in words, the line of demarcation between modeling errors and unpredictable phenomena sometimes seems to blur. Indeed, effects of modeling errors are often modeled themselves as exogenous "noises".

the cost functions associated with the scenarios are altogether taken into account. Concretely, the decision x^* is obtained as follows:

$$x^* := \arg\min_{x \in \mathcal{X}} \mathcal{L}\left(\ell(x, \delta^{(1)}), \ell(x, \delta^{(2)}), \ldots, \ell(x, \delta^{(N)})\right),$$

where $\mathcal{L}(\cdot)$ is a suitable cost-aggregating function, summarizing the behavior of the N scenario-dependent cost functions. As principal examples, $\mathcal{L}$ can be the *averaging* function $\frac{1}{N}\sum_{i=1}^{N} \ell(x, \delta^{(i)})$, discussed in Section 1.3.1, or the *worst-case* function $\max_{i=1,\ldots,N} \ell(x, \delta^{(i)})$, discussed in Section 1.3.2. Throughout, the decision x^* will be also called the *scenario solution*.

Although, in principle, the N scenarios required may be constructed by the decision-maker according to some ad-hoc criterion, we are interested in the case where the N scenarios are collected observations, i.e. *data*. In this case, we can interpret the decision x^* as a decision based on past experience, and the main concern with x^* is that of assessing its reliability with respect to the whole set Δ of possible uncertainty instances. Note that, while the number of past observations is finite and equal to N, usually Δ *is an infinite set*. For example, a possible indicator of the reliability of x^* with respect to the unseen δ's is

$$c^* := \max_{i=1,\ldots,N} \ell(x^*, \delta^{(i)}),$$

i.e. the maximum cost associated to the decision x^* among the seen scenarios. The so defined c^*, however, is just an empirical quantity depending on $\delta^{(1)}, \ldots, \delta^{(N)}$, and it is clear that it is meaningful only if an assessment of the risk that a new uncertainty instance carries a cost higher than c^* is provided. Such a risk, clearly, depends on "how large" the set

$$\{\delta \in \Delta : \ \ell(x^*, \delta) \leq c^*\}$$

is, which we call *coverage set* associated to c^*. Our c^* is a significant but particular case of a data-dependent cost threshold that can be used to characterize a data-based decision x^*. Below, we formally define the concept of *coverage set* for general data-dependent cost thresholds.

Definition 1 (coverage set)**.** *Let* $\mathbf{c}$ *be a real function defined over* Δ^N*. For every data* $\mathsf{D}^N \in \Delta^N$*, the* coverage set *of* $\mathbf{c}(\mathsf{D}^N)$ *is defined as*

$$\{\delta \in \Delta : \ \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\},$$

where x^* *is the decision made based on* D^N*, according to some fixed decision-making algorithm (to be more explicit, we could have written* x^* *as* $x^*(\mathsf{D}^N)$*, but such a dependence is left implicit throughout).* $\star$

In conclusion, to characterize a solution x^* based on a cost threshold $\mathbf{c}(\mathsf{D}^N)$, we need a way to measure the coverage set of $\mathbf{c}(\mathsf{D}^N)$.

Our next step is to introduce a probabilistic framework that allows us, at the same time,

1. to interpret a data element $\mathsf{D}^N \in \Delta^N$ as the result of real observations,
2. to measure the coverage set of a data dependent threshold $\mathbf{c}(\mathsf{D}^N)$.

We will see that this can be done in a distribution-free context, that is, without assuming the knowledge of the distribution according to which data are observed.

1.1.2 Probabilistic framework

We will assume that the set Δ is endowed with a σ-algebra and a probability measure $\mathbb{P}_\Delta$, and that scenarios $\delta^{(1)}, \dots, \delta^{(N)}$ are independently chosen by Reality according to $\mathbb{P}_\Delta$, that is, D^N can be thought of as a sample from Δ^N according to the product measure $\mathbb{P}_\Delta^N$. The reader interested in a broader discussion of this assumption and its practical meaning is referred to Section 1.2. For basic concepts of probability, we refer the reader e.g. to [3]. Finally, we will not discuss explicitly measurability issues in this work: we limit ourselves to assuming that all sets considered are measurable.
Usually, the probability $\mathbb{P}_\Delta$ underlying the data-generation mechanism is unknown to the decision-maker. Consequently, we assume that $\mathbb{P}_\Delta$ *exists*, but that it remains *hidden* to the decision-maker. In view of these positions, it seems natural to consider the unknown quantity

$$\mathbb{P}_\Delta\{\delta \in \Delta : \ \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\}$$

as a suitable measure of the coverage set of $\mathbf{c}(\mathsf{D}^N)$. We denote this quantity as the "coverage probability of $\mathbf{c}(\mathsf{D}^N)$", or just as the "coverage of $\mathbf{c}(\mathsf{D}^N)$". Since D^N is random, the function $\mathbf{c}$ of the data D^N is, according to the statistical nomenclature, a *statistic*.

Definition 2 (coverage)**.** *Given a statistic* $\mathbf{c}$ *of the data* D^N*, the coverage of* $\mathbf{c}(\mathsf{D}^N)$ *is*

$$\mathbb{P}_\Delta\{\delta \in \Delta : \ \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\},$$

and it is denoted by $\mathcal{C}(\mathbf{c}(\mathsf{D}^N))$. ⋆

Clearly, given a statistic $\mathbf{c}$, its coverage $\mathcal{C}(\mathbf{c}(\mathsf{D}^N))$ is a random variable taking values in $[0, 1]$ with a distribution that, in general, depends on the specific problem at hand, in particular on the specific probability measure $\mathbb{P}_\Delta$. Nonetheless, the object of this work is to show that there are many cases of interest where statistics having an intuitive empirical interpretation, like for example $\mathbf{c}(\mathsf{D}^N) = c^*$, are such that much is known about their coverages, though nothing is known about $\mathbb{P}_\Delta$. In other words, we will focus on coverage properties that are *independent of* $\mathbb{P}_\Delta$.

1.1.3 Distribution-free coverage properties

We here define classes of statistics whose coverages have meaningful properties that hold true for every possible $\mathbb{P}_\Delta$, i.e. in a distribution-free manner.

A first, important quantity characterizing the distribution of $\mathcal{C}(\mathbf{c}(\mathsf{D}^N)$ is the mean coverage $\mathbb{E}_{\Delta^N}[\mathcal{C}(\mathbf{c}(\mathsf{D}^N))]$, i.e. the expected value of $\mathcal{C}(\mathbf{c}(\mathsf{D}^N)$ computed over all possible D^N. We introduce the following definition.

Definition 3 (distribution-free α-mean coverage statistic)**.** *Let $\alpha \in (0,1)$. For a fixed number of scenarios N, a statistic* $\mathbf{c}$ *has a distribution-free α-mean coverage if, for every probability measure $\mathbb{P}_\Delta$, it holds that*

$$\mathbb{E}_{\Delta^N}[\mathcal{C}(\mathbf{c}(\mathsf{D}^N))] \geq \alpha. \tag{1.1}$$

$\star$

Thus, on average, an α-mean coverage statistic is expected to "cover" at least a proportion α of the possible costs, no matter what $\mathbb{P}_\Delta$ is. We are particularly interested in (empirically meaningful) statistics that can be characterized *tightly* as distribution-free α-mean statistics, i.e. whose mean coverages are exactly equal to α for some probability measure $\mathbb{P}_\Delta$. Also, it is reasonable to look for statistics whose mean coverages are equal to α for *large classes* of probability measures of practical significance, and this is what we aim to do in the chapters that follow.[2]

If we know that $\mathbf{c}$ is a distribution-free α-mean coverage statistic, we can immediately answer to the following question:

> "What is the total probability of observing N scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$, consequently obtaining x^* and $\mathbf{c}(\mathsf{D}^N)$ based on them, and that a new observation δ carries a cost $\ell(x^*, \delta)$ not higher than $\mathbf{c}(\mathsf{D}^N)$?"

In fact, we have that

$$\begin{aligned}
\mathbb{E}_{\Delta^N}\left[\mathcal{C}(\mathbf{c}(\mathsf{D}^N))\right] &= \mathbb{E}_{\Delta^N}\left[\mathbb{P}_\Delta\{\delta \in \Delta :\ \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\}\right] \\
&= [\text{denoting with } \mathbb{1}\{\cdot\} \text{ the indicator function}] \\
&= \mathbb{E}_{\Delta^N}\left[\mathbb{E}_\Delta\left[\mathbb{1}\{\delta \in \Delta :\ \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\}\right]\right] \\
&= \mathbb{E}_{\Delta^{N+1}}\left[\mathbb{1}\{(\mathsf{D}^N, \delta) \in \Delta^N \times \Delta :\ \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\}\right] \\
&= \mathbb{P}_\Delta^{N+1}\{(\mathsf{D}^N, \delta) \in \Delta^N \times \Delta :\ \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\}, \qquad (1.2)
\end{aligned}$$

which is the sought probability. Therefore

$$\mathbb{P}_\Delta^{N+1}\{(\mathsf{D}^N, \delta) \in \Delta^N \times \Delta : \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\} \geq \alpha,$$

[2] Clearly, the requirement that $\mathbb{E}_{\Delta^N}[\mathcal{C}(\mathbf{c}(\mathsf{D}^N))]$ be exactly equal to α for *every* $\mathbb{P}_\Delta$ is too strong and cannot be achieved. In fact, for a given $\alpha \in (0,1)$, the class of statistics satisfying the property "$\mathbb{E}_{\Delta^N}[\mathcal{C}(\mathbf{c}(\mathsf{D}^N))] = \alpha$ for every $\mathbb{P}_\Delta$" is empty. To see this, take $\mathbb{P}_\Delta$ as a probability measure concentrated on a unique scenario $\bar{\delta}$. In this case, any statistic $\mathbf{c}$ takes a constant value, say $\bar{c}$, and has a deterministic coverage equal to 1, if $\bar{c} \geq f(x^*, \bar{\delta})$, or to 0, if $\bar{c} < f(x^*, \bar{\delta})$.

and the answer to the above question is (independently of $\mathbb{P}_\Delta$): "certainly no less than α."

In spite of the usefulness of distribution-free results about mean coverages, more refined distribution-free characterizations of $\mathcal{C}(\mathbf{c}(\mathsf{D}^N))$ can be obtained. We give the following definition.

Definition 4 (distribution-free (ϵ, β)-coverage statistic)**.** *Let $\epsilon, \beta \in (0, 1)$. For a fixed number of scenarios N, a statistic $\mathbf{c}$ has a distribution-free (ϵ, β)-coverage if relation*

$$\mathbb{P}_\Delta^N \left\{ \mathsf{D}^N \in \Delta^N : \mathcal{C}(\mathbf{c}(\mathsf{D}^N)) \geq 1 - \epsilon \right\} \geq 1 - \beta$$

holds for all probability measures $\mathbb{P}_\Delta$. $\star$

So, if $\mathbf{c}$ is a distribution-free (ϵ, β)-coverage statistic, we can claim that the coverage of $\mathbf{c}(\mathsf{D}^N)$ is at least $1 - \epsilon$ with *confidence* $1 - \beta$. Of major interest is the case where ϵ is "small", say 0.01, while β is "*very* small", say 10^{-7}, so that, no matter what $\mathbb{P}_\Delta$ is, the coverage of $\mathbf{c}(\mathsf{D}^N)$ is at least $1 - \epsilon$ "with a reasonable degree of certainty". For fixed ϵ and β, the so-defined statistics are tightly characterized if there exists some probability measure $\mathbb{P}_\Delta$ such that $\mathbb{P}_\Delta^N \left\{ \mathsf{D}^N \in \Delta^N : \mathcal{C}(\mathbf{c}(\mathsf{D}^N)) \geq 1 - \epsilon \right\} = 1 - \beta$. We will show that, for many important decision-making problems, it is possible to find tight statistics such that the condition $\mathbb{P}_\Delta^N \left\{ \mathsf{D}^N \in \Delta^N : \mathcal{C}(\mathbf{c}(\mathsf{D}^N)) \geq 1 - \epsilon \right\} = 1 - \beta$ holds true for large classes of probability measures that are of practical significance. Moreover, the whole probability distribution of $\mathcal{C}(\mathbf{c}(\mathsf{D}^N))$ may turn out to be the same for large classes of probability measures. We give the following definition.

Definition 5 (distribution-free coverage statistic)**.** *If the probability distribution of $\mathcal{C}(\mathbf{c}(\mathsf{D}^N))$ is the same for a whole class of probability measures $\mathbb{P}_\Delta$, we will say that the statistic $\mathbf{c}$ is a* distribution-free coverage statistic *for the class of problems characterized by those probability measures.* $\star$

Before proceeding, some terminological remarks are worthwhile.

Remark 1 (coverage)**.** *The term "coverage of the statistic $\mathbf{c}(\mathsf{D}^N)$" is quite intuitive and allows us to emphasize our interest in making statements, based on the finite set of observations D^N, that "covers" the unseen instances of δ. Nonetheless, it is a direct application of a term borrowed from statistical literature, since the coverage set of $\mathbf{c}(\mathsf{D}^N)$ in Definition 1 can be interpreted as a "tolerance region" in the space Δ, having indeed "coverage probability" equal to $\mathcal{C}(\mathbf{c}(\mathsf{D}^N))$, see e.g. [4, 5].* $\star$

Remark 2 (risk)**.** *In some contexts, the focus is on the complementary of the coverage set, that is, on the "bad" set of those δ's exceeding the cost threshold $\mathbf{c}(\mathsf{D}^N)$. The probability of this set is called* risk*. With this terminology, a distribution free (ϵ, β)-coverage statistic has a risk less than or equal to ϵ with confidence $1 - \beta$. We will prefer to deal with the* risk *instead of with the* coverage *in the min-max context, i.e. from Chapter 3 onward.* $\star$

Admittedly, the notation used up to now to indicate events is quite pedantic, since the sample space over which the probability measure is defined can usually be understood without ambiguity. Hence, throughout we will e.g. write more compactly

$$\mathbb{P}_{\Delta}^{N+1}\{\ell(x^*,\delta) \leq \mathbf{c}(\mathsf{D}^N)\}$$

in place of

$$\mathbb{P}_{\Delta}^{N+1}\{(\mathsf{D}^N,\delta) \in \Delta^N \times \Delta : \ell(x^*,\delta) \leq \mathbf{c}(\mathsf{D}^N)\},$$

and similarly in similar cases.

We are now ready to overview the results achieved in this work. This is done in Section 1.3. The following Section 1.2 goes into the interpretations and motivations of the probability framework introduced in this chapter, and can be skipped without loss of continuity.

1.2 Interpretation of the probabilistic framework

We recall the two main sources of uncertainty entering a decision problem:

- unpredictability of phenomena;
- modeling errors.

The first kind of uncertainty is commonly involved when the collected scenarios $\delta^{(1)},\dots,\delta^{(N)}$ are the results of field experiments performed in various environmental conditions, or are retrieved from historical series, see e.g. [6, 7]. Although the link between probability and physical world is subject to philosophical controversies[3], most engineers seem willing to admit that, in many situations of interest, the data that Reality provides us can be conveniently thought of as the outcomes of a random variable δ, sampled according to some arcane, but in some sense *existing*, probability distribution $\mathbb{P}_{\Delta}$. Arguing in favor of the existence of a certain $\mathbb{P}_{\Delta}$ underlying the generation of $\delta^{(1)},\dots,\delta^{(N)}$ commonly requires using application-domain knowledge and arguments. Once the existence of $\mathbb{P}_{\Delta}$ is accepted, we think that measuring the coverage sets introduced above based on $\mathbb{P}_{\Delta}$ is the most natural option. Moreover, since we do *not* assume that the decision-maker knows the arcane $\mathbb{P}_{\Delta}$, the only assumption requiring further justification is the assumption that $\delta^{(1)},\dots,\delta^{(N)}$ are independent. Still, the decision-maker can argue in favor of this assumption based on a-priori knowledge about reality and on his data-acquisition procedure. For example, when microscopic phenomena with very rapid dynamics (e.g. thermal agitation) are at the origin of noises, noises usually turn out to be independent processes when sampled on a macroscopic time scale: this argument has been used to justify the commonly accepted noise model in systems and control

[3] For a recent debate in the systems and control community, see [8] and the discussion paper [9], in particular the contribution of Jan C. Willems.

engineering, see e.g. the seminal paper [10]. In finance, according to the classical Black-Scholes model, logarithmic return increments at equispaced time intervals are independent, see [11]. Remarkably, independence can be induced by a suitable data-acquisition procedure, as, for example, in opinion polls, where the decision-maker can *randomly select* people to be interviewed. Generalization to contexts where scenarios are not independent or are not generated according the same $\mathbb{P}_\Delta$ is an open and stimulating research area, beyond the scope of the present work.
As for uncertainty due to modeling errors, we offer here a similar interpretation as that above. The space Δ can be thought of in the abstract as the set of all the models that are candidate to represent the physical reality determining the cost function, as in Example 2. More concretely, δ can be a vector of uncertain parameters, whose correct values can be estimated through identifications procedures, see e.g. [12]. The scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$ are then collected as the outputs of N independent identification procedures. Indeed, the output of an identification procedure is subject to some variability, since it depends on the kind of experiment performed, on environmental conditions, and so forth and so on. If we accept a probabilistic description for this lack of determinism, $\mathbb{P}_\Delta$ can be naturally defined as the (unknown) probability according to which the identification outputs are generated.

The interpretation of the probability framework advocated above is not the sole possible. Probability $\mathbb{P}_\Delta$ may be introduced by the user without any reference to a supposed state of the world, but just as a technical tool aimed at quantifying the relative importance of the possible occurrences of δ. According to this point of view, scenarios are artificially generated from $\mathbb{P}_\Delta$ in an independent way, so that more important values of δ are more likely to be considered in making the decision. A threshold $\mathbf{c}(\mathsf{D}^N)$ with guaranteed large coverage is thus associated with the obtained decision x^*, so that we can conclude that the most important instances of δ are likely to carry a cost no higher than $\mathbf{c}(\mathsf{D}^N)$. Indeed, even though our $\mathbb{P}_\Delta$-independent quantifications of the coverages are of particular interest when $\mathbb{P}_\Delta$ is unknown, they can still be useful when the probability is known, e.g. for computational reasons.

1.3 Introduction to the problems studied in this work

We recall for easy reference the main symbols introduced above.

$$\begin{aligned}
\mathcal{X} &: \text{decision set;}\\
x \in \mathcal{X} &: \text{decision variable;}\\
\Delta &: \text{uncertainty set;}\\
\delta \in \Delta &: \text{uncertainty parameter;}\\
(\Delta, \mathcal{D}, \mathbb{P}_\Delta) &: \text{probability space, where}\\
\mathcal{D} &: \sigma\text{-algebra over } \Delta,
\end{aligned}$$

$\mathbb{P}_\Delta$: probability measure over $(\Delta, \mathcal{D})$;
$\mathsf{D}^N \in \Delta^N$: a sample of N scenarios, independent and identically distributed according to $\mathbb{P}_\Delta$, i.e., D^N is short notation for $\delta^{(1)}, \ldots, \delta^{(N)}$;
$\ell : \mathcal{X} \times \Delta \to \mathbb{R}$: cost function - the cost depends on x and the uncertain δ;
$\mathcal{L} : \left(\mathbb{R}^{\mathcal{X}}\right)^N \to \mathbb{R}^{\mathcal{X}}$: cost-aggregating function - it aggregates $\ell(x, \delta^{(1)}), \ldots, \ell(x, \delta^{(N)})$.

We concentrate on the scenario solution, defined as

$$x^* := \arg\min_{x \in \mathcal{X}} \mathcal{L}\left(\ell(x, \delta^{(1)}), \ldots, \ell(x, \delta^{(N)})\right),$$

and study its coverage properties, in the light of Definitions 2, 4, 5 and 3, for two instantiations of the cost-aggregating function that are ubiquitous in applications.

1.3.1 Average setting

When an *average* cost-aggregating function is used, the decision x^* is chosen as the one that performs best *on average* for the N scenarios, that is

$$\begin{aligned} x^* &= \arg\min_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^{N} \ell(x, \delta^{(i)}) \\ &= \arg\min_{x \in \mathcal{X}} \sum_{i=1}^{N} \ell(x, \delta^{(i)}). \end{aligned} \tag{1.3}$$

Although many variations on the theme are possible and stimulating, in the next Chapter 2 we will study the case where the cost function $\ell(x, \delta)$ is a (convex) quadratic function in x for each δ. Moreover we will assume $\mathcal{X} = \mathbb{R}^d$. In optimization terminology, the problem (1.3) with $\ell(x^*, \delta)$ quadratic in $x \in \mathbb{R}^d$ is an *unconstrained quadratic optimization* problem, more commonly called a (classical) *least squares* problem. We will discuss the coverage properties of statistics defined to be as similar as possible to the empirical costs $\ell(x^*, \delta^{(1)}), \ldots, \ell(x^*, \delta^{(N)})$. In particular, we will focus on distribution-free $\frac{i}{N+1}$-mean coverage statistics, where $i = 1, \ldots, N$. It will be shown that empirical costs *do not have* the desired coverage properties, while $\frac{i}{N+1}$-mean coverage statistics can be obtained by considering suitable approximations of $\ell(x^*, \delta^{(1)}), \ldots, \ell(x^*, \delta^{(N)})$.

The focus on the particular case of least squares problems is justified because, since Gauss and Legendre proposed to solve regression problems through the least squares method at about the beginning of the XIX century, see e.g. [21], the least squares method has been used in countless applications. See Table 1.1 for just a few examples.

Table 1.1. A few examples of least squares problems.

	Interpretation of δ	Interpretation of x	Interpretation of $\ell(x, \delta)$	References
Linear regression theory	Data point	Coefficients weighting regressor functions	Regression error	Chapter 3 of [13], see also Section 2.1 of this work
Facility location	Position and weight of the demand points	Location for a new facility	Cost of the state of the world	[14, 15, 16, 17, 18], see also Sections 2.1 and 2.3 of this work
LQ regulator	Noises, uncertain model matrices	Control action	Quadratic performance index	Chapter 2 Section 4 of [19], see also Section 2.1 of this work
Receding-horizon estimation	Uncertain model matrices	Estimated state	Deviation of estimated outputs from measured outputs	[20]

1.3.2 Worst-case setting

When the *worst-case* cost-aggregating function is used, the decision x^* is the one minimizing the worst-case cost among those carried by the seen scenarios, i.e.

$$x^* = \arg\min_{x \in \mathcal{X}} \max_{i=1,\dots,N} \ell(x, \delta^{(i)}). \tag{1.4}$$

Clearly, the worst-case function leads to more cautionary results with respect to the average one. We will show that, in this context, stronger results than those of Chapter 2 can be obtained, under more general assumptions. Assumptions are relaxed by allowing for $\ell(x, \delta)$ to be a general convex function in x for each δ, and for x to be chosen in a constrained way, that is, $x \in \mathcal{X} \subseteq \mathbb{R}^d$, with convex and closed $\mathcal{X}$.

An immediate application of an already known result, borrowed by the theory recalled in Appendix A, establishes that, if $\ell(x, \delta)$ is convex, the statistic given by the highest empirical cost associated to x^*, that is

$$c^* = \max_{i=1,\dots,N} \ell(x^*, \delta^{(i)}),$$

is a distribution-free coverage statistic for a whole subset of problems. In general, it is a distribution-free $\frac{N-d}{N+1}$-mean coverage statistic, and can be tightly characterized as an (ϵ, β)-coverage statistic for very interesting values of (ϵ, β). In fact, for fixed d and ϵ, the confidence parameter β can be heavily reduced (i.e. the confidence can be heavily increased) by a small increase of N. The decision-maker can use this result about the coverage of c^* by associating with the decision x^* a performance c^* guaranteed at a certain level of probability ϵ. In Chapter 3, this result is extended to all the others empirical costs $\ell(x^*, \delta^{(1)}), \dots, \ell(x^*, \delta^{(N)})$. In particular, for the same set of problems for which c^* is a distribution-free coverage statistic, we have that the *joint probability distribution* of the coverages of all the empirical costs $\ell(x^*, \delta^{(1)}), \dots, \ell(x^*, \delta^{(N)})$ can be computed exactly and does *not* depend on $\mathbb{P}_\Delta$. Moreover, by denoting with $c_1^*, \dots, c_N^*$ all the empirical costs from the largest to the smallest, we have that $c_{d+1}^*, \dots, c_N^*$ are *in general*, that is under very mild assumptions, distribution-free coverage statistics, and the classic result, which characterizes c^* in full generality as a distribution-free (ϵ, β)-coverage statistic, turns out to be an immediate consequence of the trivial fact that $c^* \geq c_{d+1}^*$. In short, we

provide an extension and a reinterpretation of the classic result about the coverage of c^*, thus providing the decision-maker with an instrument to characterize in a distribution-free manner the coverages of all the costs associated with x^*.
A possible issue with this approach is that the number of scenarios N required to guarantee that the coverage of c^* is no smaller than $1-\epsilon$ with a good confidence depends on the problem dimension d as $\frac{d}{\epsilon}$: if d is large, it may become difficult to guarantee a high coverage. In Chapter 4, a way to face this problem is offered, by introducing a slightly modified decision-making algorithm and a statistic - a modified version of c^* - whose coverage is far less sensitive to an increasing of d than its counterpart c^*. This allows the decision-maker to characterize a worst-case decision by means of a high coverage statistic, using a relatively small number of scenarios N.

1.4 Review of the literature

Our work has a two-layered nature.

> **Layer of the decision-making approaches**: a decision in the presence of uncertainty has to be made, and we propose to make the decision according to an intuitive and common sense data-based algorithm (in particular, by minimizing the value of the *average* or of the *worst-case* aggregating-cost function);
>
> **Layer of the probabilistic guarantees**: our theoretical analysis, based on distribution-free coverage properties of empirically significant statistics, allows to characterize the decision x^*, which is made based on a finite data sample, with respect to the infinitely many *unseen* situations.

First, we consider some classic approaches to decision-making in the presence of uncertainty, and then we discuss studies about the probabilistic characterization of a solution obtained based on a finite data sample. At the end of this section we will give some more specific references related to the two settings (average setting with quadratic cost function and worst-case with convex cost function) considered in this work.

Decision-making approaches

Standard approaches to face uncertainty in optimization problems can be grouped into three classes:

- Robust optimization;
- Stochastic optimization;

- Chance-constrained optimization.

We aim not at providing a complete survey of them, but just at pointing out how the main ideas of these approaches reflect in our framework.

Robust optimization
In robust optimization, studied in [22, 23, 24, 25, 26, 27, 28], the main idea is that the decision-maker wants to be "robust" with respect to all the possible uncertainty instances $\delta \in \Delta$, so that the decision should be made by choosing

$$x^* := \arg\min_{x \in \mathcal{X}} \left(\max_{\delta \in \Delta} \ell(x, \delta) \right). \tag{1.5}$$

If the set Δ grasps the real range of uncertainty, the decision is certainly robust with respect to the worst case. However, taking into account all the (believed to be) possible realizations of δ can turn out to be too conservative an approach, sometimes nearly paralyzing the decision-making process. More sophisticated schemes has been proposed to increase the degrees of freedom of the decision-maker in trading robustness and conservatism, also with the support of probabilistic models for the uncertain parameters. For example, [27] assumes that uncertain parameters take values on intervals according to symmetric distributions. In all cases, the decision-maker is called to model suitably the uncertainty set Δ, and this is a delicate task prone to arbitrariness (e.g. in [6, 28], methods to build reasonable uncertainty sets according to a data-based criterion are suggested). The idea of the data-based worst-case approach, according to which x^* is chosen as in (1.4), is to simply replace the uncertainty set Δ in (1.5) with the data themselves, that is with $\{\delta^{(1)}, \ldots, \delta^{(N)}\}$. An important point is also that to solve (1.5) is in general computationally difficult, see e.g. [23, 24]: this was one of the motivations for the theoretical study of (1.4), see [29, 30] and references therein.

Stochastic and chance-constrained optimization
In stochastic optimization and chance-constrained optimization, the probabilistic framework is adopted, and the probability $\mathbb{P}_\Delta$ is assumed to be *known*. Stochastic optimization has been introduced in [31]. The basic idea, using our notation, is to choose

$$x^* := \arg\min_{x \in \mathcal{X}} \mathbb{E}_\Delta \left[\ell(x, \delta) \right]. \tag{1.6}$$

Clearly, the expected value of the cost has to be computed in order to be minimized and in general this involves the difficult computation of a multi-variable integral. This is why the expected value is sometimes replaced by an empirical mean over N scenarios, thus recovering a problem like (1.3) with average cost-aggregating function. In other words, the stochastic problem (1.6) can be solved through a Monte-Carlo method, where each Monte-Carlo sample $\delta^{(i)}$ can be interpreted as a scenario, see e.g. [32, 33]. However, a word of caution about terminology is necessary, because the term "scenario approach" in stochastic programming has usually

a slightly different meaning than that used in our work. In stochastic programming, the term "scenarios" is usually employed to indicate instances $\delta^{(1)}, \dots, \delta^{(N)}$ that are selected by the decision-maker, who assigns to each of them a probability, not necessarily equal to $\frac{1}{N}$. The probability is indeed assigned according to some (subjective or objective) criterion, and the decision x^* is commonly made by averaging over the cost functions *weighted* by the corresponding probability. However, if the scenarios are randomly sampled and the uniform sample distribution is assigned to them, the "scenario approach" in stochastic programming boils down to solving our average problem (1.3).
More in general, in stochastic optimization any other operator depending on $\mathbb{P}_\Delta$ could be placed instead of the expected value. A special case of stochastic optimization is chance-constrained optimization, see e.g. [34, 35, 36, 37, 38, 39]. The idea is to choose x^* by optimizing over the set Δ minus an ϵ-probability set of uncertainty instances of δ carrying the highest costs. Formally, for a fixed $\epsilon \in (0, 1)$, the chance-constrained problem writes as:

$$\begin{aligned} &\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d, \gamma \in \mathbb{R}} \gamma \\ &\text{subject to: } \mathbb{P}_\Delta\{\ell(x, \delta) \leq \gamma\} \geq 1 - \epsilon, \end{aligned} \tag{1.7}$$

whose solution $(x^*_{\text{cc}}, \gamma^*_{\text{cc}})$ is the pair consisting of the optimal chance-constrained decision x^*_{cc} and the corresponding cost γ^*_{cc} that can be exceeded with probability no greater than ϵ. Chance-constrained optimization is notoriously hard to solve in general, even though there are notable exceptions where the solution can actually be computed, see [36, 37, 40]. In Chapter 3, we show how to compute N so that, for a fixed ϵ and very small β, the worst-case cost c^* associated to the worst-case decision x^*, computed according to (1.4), is a distribution-free (ϵ, β)-coverage statistic. This allows us to interpret the scenario approach as a method to find a pair (x^*, c^*) being a *feasible* solution for the chance-constrained problem (1.7) with very high confidence $1 - \beta$, independently of $\mathbb{P}_\Delta$.

Probabilistic guarantees for sample-based solutions

In the following, we focus on main contributions aiming at a theoretical analysis of a decision x^* computed based on a finite data sample. In particular, we mention three kinds of theoretical studies:

- studies of asymptotic properties,
- studies in statistical learning theory,
- studies about a posteriori evaluations,

and show briefly in what they differ from ours.

Asymptotic properties and statistical learning theory
The study of asymptotic properties is the study of the properties of the decision x^* when it is made based on a number of data N that goes to infinity, while our work aims at providing results for finite, and possibly small N. In some cases, the decision of the problem with "an infinite number of scenarios" is considered as the ideal decision. For example, consider the solution $\bar{x}$ to the following stochastic optimization problem

$$\min_{x \in \mathcal{X}} \mathbb{E}_{\Delta}[\ell(x, \delta)].$$

Sometimes, because $\mathbb{P}_{\Delta}$ is unknown or for computational reasons, one tries to approximate the ideal $\bar{x}$ by means of the minimizer of the empirical mean, that is, one decides for x^*, computed as in (1.3), instead of for the ideal $\bar{x}$ (unattained or unattainable). In this case, convergence results guaranteeing that $x^* \to \bar{x}$ are in order, see e.g. Chapter 5 in [39]. Also, the decision-maker may be interested in knowing what the difference is between $\frac{1}{N}\sum_{i=1,\dots,N} \ell(x^*, \delta^{(i)})$, i.e. the mean of the empirical costs corresponding to the finite set of data D^N, and the true expected cost $\mathbb{E}_{\Delta}[\ell(x^*, \delta)]$, or even the true expected cost associated to the ideal decision $\bar{x}$, i.e. $\mathbb{E}_{\Delta}[\ell(\bar{x}, \delta)]$. Statistical-learning theory faces this and other similar problems, and studies the conditions under which it is possible to generalize from N finite data to a quantity depending on all the infinite possibile data, see the fundamental book [41], and [42, 43]. The basic assumptions of statistical-learning theory are the same as ours:

- $\mathbb{P}_{\Delta}$ exists but it is unknown;
- the samples $\delta^{(1)}, \dots, \delta^{(N)}$ are independent and identically distributed.

For example, for a given sample-based decision-making algorithm producing x^*, for a fixed η and a very small β one can find the suitable N such that

$$\mathbb{P}_{\Delta}^{N}\left\{\left|\mathbb{E}_{\Delta}[\ell(x^*, \delta)] - \frac{1}{N}\sum_{i=1}^{N} \ell(x^*, \delta^{(i)})\right| \leq \eta\right\} \geq 1 - \beta, \tag{1.8}$$

thus guaranteeing (with very high confidence $1 - \beta$) that the empirical mean and the real expected value differ at most by a small amount η. The magnitude (or even the finiteness) of N depends on boundedness properties of the cost function $\ell(x, \delta)$ and on its possible "variability" with respect to the random δ. A formula like (1.8), which represents a typical result that can be obtained through statistical learning theory, is out of the scope of our work, because it focuses on an expected value, that is an average quantity depending on all the unseen δ's, while we focus on the property of *a single* unseen uncertainty instance δ of being or not being covered by a given cost threshold $\mathbf{c}(\mathsf{D}^N)$. Statistical learning theory, however, *can* be used to study coverage properties, too. In fact, given a data-based decision x^* and a

cost threshold $\mathbf{c}(\mathsf{D}^N)$, we may define the indicator function of the coverage set of $\mathbf{c}(\mathsf{D}^N)$ as:

$$\mathbb{1}_{\mathcal{C}}(\delta) := \mathbb{1}\{\delta \in \Delta : \ \ell(x^*, \delta) \leq \mathbf{c}(\mathsf{D}^N)\}, \tag{1.9}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function of the set $\{\cdot\}$. Mathematically, the coverage of $\mathbf{c}(\mathsf{D}^N)$ is anything but the expected value of the binary function $\mathbb{1}_{\mathcal{C}}(\delta)$ with respect to the uncertain δ, that is $\mathcal{C}(\mathbf{c}(\mathsf{D}^N)) = \mathbb{E}_{\Delta}[\mathbb{1}_{\mathcal{C}}(\delta)]$. The difference between the empirical version of $\mathcal{C}(\mathbf{c}(\mathsf{D}^N))$, that is $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\mathcal{C}}(\delta^{(i)})$, and its true value, that is $\mathbb{E}_{\Delta}[\mathbb{1}_{\mathcal{C}}(\delta)]$, can indeed be studied by resorting to the statistical learning theory. However, in the contexts studied in our work, statistical learning theory is more conservative, and requires a large number of scenarios N (e.g. when $\ell(x, \delta)$ is a convex function the sought finite N may not even exist). Indeed, the usual approach to this kind of problems is based on uniform convergence results aiming at guaranteeing the convergence of the empirical to the true value not only for $\mathbb{1}_{\mathcal{C}}(\delta)$, which is the indicator function defined for *the* decision x^* and *the* cost $\mathbf{c}(\mathsf{D}^N)$ of interest, but also and *at the same time* for *all the other* possible decision-cost pairs (or, at least, for a large subset of them, as in [44]). We will formally show that results in statistical learning approach cannot improve the results presented in Chapters 3 and 4, by showing that our results are tight, i.e. not improvable at all. As for results in the average setting of Chapter 2, here we limit ourself to the observation that they do not depend directly on the size d of x, which enters instead the bounds that, to our knowledge, can be obtained according to the statistical learning theory. Nonetheless, situations that we have not considered in the present study (e.g. non-convex cost functions $\ell(x, \delta)$, unusual selection of the cost-aggregating function, etc.) can at least in principle be faced with the support of the statistical learning theory. For recent results on this topic, see e.g. [44]. Studies on data-dependent penalties may allow to reduce the gap between conservative uniform convergence results and approaches that guarantees one particular decision: see e.g. [45] and references therein.

A-posteriori assessments

Assume that x^* and $\mathbf{c}(\mathsf{D}^N)$ have been computed according to some rule, and that we have at our disposal an arbitrarily high number of additional scenarios (this is a rare situation if scenarios come from real experiments). Assume also that computing $\ell(x^*, \delta)$ for many δ's is easy (this is not always the case, see the example considered in [46]). Then, a posteriori assessments can be easily made. A Monte-Carlo evaluation can be run on M new independent uncertainty instances, e.g. $\delta^{(N+1)}, \ldots, \delta^{(N+M)}$, and the coverage estimated by the quantity $\frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\mathcal{C}}(\delta^{(N+i)})$, with $\mathbb{1}_{\mathcal{C}}(\delta)$ defined as in (1.9). In fact, an application of the classic Hoeffding's inequality (see e.g. [47] for results on concentrations inequalities) yields

$$\mathbb{P}_{\Delta}^{M}\left\{\left|\mathcal{C}(\mathbf{c}(\mathsf{D}^N)) - \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}_{\mathcal{C}}(\delta^{(N+i)})\right| \leq \eta\right\} \geq 1 - 2e^{-\frac{2\eta^2}{M}}.$$

In the following, we give some more specific references about the two distinct settings (average setting with quadratic cost function and worst-case with convex cost function) considered in this work.

1.4.1 Quadratic cost function

We mention here a pair of works about decision-making with uncertain quadratic cost function. Results for the robust problem (1.5) with quadratic $\ell(x, \delta)$ are presented in [48, 49], under some conditions about the structure of the uncertainty. In [48] it is also shown that the robust problem is in general NP-complete. In [50], the guaranteed residual-at-risk minimization of the quadratic $\ell(x, \delta)$ is introduced: similarly to chance-constrained optimization, one obtains the best pair $(x^*_{\mathrm{rr}}, \gamma^*_{\mathrm{rr}})$ such that $\mathbb{P}_\Delta\{\ell(x^*_{\mathrm{rr}}, \delta) \leq \gamma^*_{\mathrm{rr}}\} \geq 1 - \epsilon$, but, differently from the chance-constrained case, the condition $\mathbb{P}_\Delta\{\ell(x^*_{\mathrm{rr}}, \delta) \leq \gamma^*_{\mathrm{rr}}\} \geq 1 - \epsilon$ has to hold not only for *one* known $\mathbb{P}_\Delta$, but for a *set* of distributions $\mathbb{P}_\Delta$, i.e. those satisfying the condition that the uncertain parameter δ has a certain known expected value and a certain known variance. Moreover, the uncertainty is there assumed to be constrained according to a precise structure.

1.4.2 Average setting with quadratic cost function

About our work
Results presented in Chapter 2 have been inspired by investigations about order statistics and tolerance regions, see e.g. [4, 51, 5, 52], and conformal predictors, [53]. Indeed, the result presented in Chapter 2 can be thought of as a generalization of a classic result, there recalled, about order statistics to a context where the distribution of the cost is influenced by an optimization procedure. Also, as noted in Remark 1, coverage sets can be interpreted as tolerance regions. Nonetheless, such an interpretation fails to grasp the specificity of our approach: we are not interested in predicting future realizations of δ, but rather in the event that a cost threshold is not exceeded. The event of a cost not being exceeded is studied theoretically *as if it were* a tolerance region suitably tailored in the space Δ.

Some references to related results for x^* chosen as in (1.3) with quadratic cost function follows.

Known results in a very restricted context
In a very restricted context, when $x \in \mathbb{R}$, $\Delta = \mathbb{R}$ and $\ell(x, \delta) = (x-\delta)^2$, the classical theory of tolerance regions can be easily applied to find statistics similar to those obtained in Chapter 2 that have distribution-free $\frac{k}{N+1}$-mean coverages, too. For example, the tolerance regions defined as $T(\mathsf{D}^N) := [\min_{i=1,\dots,N} \delta^{(i)}, \max_{i=1,\dots,N} \delta^{(i)}]$ is known to have mean coverage no less than $\frac{N-1}{N+1}$. It is not difficult to see that the set $T(\mathsf{D}^N)$ is a subset of the coverage set of the empirical worst-case cost $c^* = \max_{i=1,\dots,N} \ell(x^*, \delta^{(i)})$. Hence, this proves that, in this restricted context, c^*

has distribution-free $\frac{N-1}{N+1}$-mean coverage, too. A result presented in [54] could also be applied in this same restricted scalar context to obtain statistics with distribution-free mean coverage: these statistics are by construction a scaled version of the estimated variance of the samples $\delta^{(1)}, \ldots, \delta^{(N)}$. For a comparison between prediction intervals obtained in [54] and ordinary tolerance regions see [55]. We limit ourselves to observing that the distribution-free $\frac{N-1}{N+1}$-mean coverage statistic obtained in Chapter 2, as well as c^*, remains bounded at the increasing of N whenever δ has bounded support, while the $\frac{N-1}{N+1}$-mean coverage statistic obtainable by using the results in [54] necessarily goes to infinity.

A related result in the general context
The general contexts with $\ell(x, \delta) = \|A(\delta)x - b(\delta)\|^2$, where $A \in \mathbb{R}^{n \times d}$ is an $n \times d$ matrix and $b \in \mathbb{R}^n$ a column vector, is studied in [20], where x^* is computed according to (1.3) and known results in statistical learning theory are applied, under some a-priori conditions, e.g. boundedness of the uncertainty set and of $\ell(x, \delta)$, in order to study how near $\mathbb{E}_\Delta[\ell(x^*, \delta)]$ is to $\min_x \mathbb{E}_\Delta[\ell(x, \delta)]$.

1.4.3 Worst-case setting with convex cost function

The results from Chapter 3 onward are in the vein of the so-called *theory of the scenario approach* for convex optimization, [29, 30, 56, 57], which, under the assumption that $\ell(x, \delta)$ is convex in x, provides the sharpest possible characterization of the coverage set of the worst-case cost statistic. The main results and other references to the scenario approach for general constrained convex problems are recalled in Appendix A.

Chapter 2

The coverage probabilities of the least squares residuals

In this chapter, we study a data-based *average* approach known as the least squares method, and show how the least squares solution can be characterized through suitably constructed distribution-free mean coverage statistics. In the following Section 2.1 the data-based least squares problem is formally stated, with examples, and our result is introduced. In Section 2.2 the main theorem is provided followed by a discussion, while proofs are postponed to Section 2.4. A numerical example is given in Section 2.3. Some pointers to possible future developments are briefly discussed in Section 2.5.

2.1 Introduction and problem position

We consider an uncertain optimization problems where a decision, modeled as the selection of a variable $x \in \mathbb{R}^d$, has to be made so as to minimize a (convex) quadratic cost function $\ell(x, \delta)$ that also depends on the uncertain random element δ. Whenever the uncertain cost function $\ell(x, \delta)$ is a non-negative function[1], we can, according to a standard notation, identify the uncertainty instance δ with a pair (A, b), where $A \in \mathbb{R}^{n \times d}$, i.e A is an $n \times d$ matrix, and $b \in \mathbb{R}^n$, i.e. b is a column vector, and rewrite $\ell(x, \delta)$ as a *squared residual*:

$$\ell(x, \delta) = \|Ax - b\|^2,$$

that is, as the squared Euclidean norm of the difference between Ax and b. Hence, given a certain, deterministic $\delta = (A, b)$, the best decision would be the minimizer of the *squared residual* of (A, b).
On the other hand, in the presence of uncertainty, the decision is made by consid-

[1]The assumption that $\ell(x, \delta) \geq 0$ will be maintained throughout, but it is immaterial for the validity of the theoretical results presented in this chapter.

ering N scenarios, i.e. N uncertainty instances

$$\mathsf{D}^N = \delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)} = (A_1, b_1), (A_2, b_2), \ldots, (A_N, b_N),$$

independently generated according to a probability $\mathbb{P}_\Delta$ over the uncertainty set $\Delta = \mathbb{R}^{n\times d} \times \mathbb{R}^n$. The data-based method of *least squares* prescribes to minimize the *average of the squared residuals* associated with the N scenarios D^N, so that

$$\begin{aligned} x^* &= \arg\min_x \frac{1}{N} \sum_{i=1}^{N} \|A_i x - b_i\|^2 \\ &= \arg\min_x \sum_{i=1}^{N} \|A_i x - b_i\|^2. \end{aligned} \tag{2.1}$$

The minimizer of (2.1) is[2] the scenario solution x^*.
A standard application of this approach is in linear regression.

Example 4 (linear regression)**.** *Let θ and y be two random variables. We want to regress y against a polynomial of order $d-1$ in θ. During a campaign of data acquisition, N independent observations $(\theta^{(1)}, y^{(1)}), \ldots, (\theta^{(N)}, y^{(N)})$ are collected. By letting*

$$A_i = \left(1, \theta^{(i)}, \left(\theta^{(i)}\right)^2, \ldots, \left(\theta^{(i)}\right)^{d-1}\right) \in \mathbb{R}^{1\times d}, \text{ for } i = 1, \ldots, N, \text{ and}$$
$$b_i = y^{(i)},$$

we can find the coefficients of the best fitting polynomial by solving

$$\min_x \sum_{i=1}^{N} \|A_i x - b_i\|^2.$$

So, writing the minimizer x^* explicitly as a vector $x^* = (\alpha_0, \alpha_1, \ldots, \alpha_{d-1})$, we have that

$$P(\theta) = \alpha_0 + \theta\alpha_1 + \theta^2\alpha_2 + \ldots + \theta^{d-1}\alpha_{d-1}$$

is the sought polynomial modeling[3] *the relationship between θ and y.*

$\star$

The following example is a well-known facility location problem and will be further developed in Section 2.3.

[2] If the minimizer is not unique, an arbitrary solution determined through a tie-break rule is taken and denoted by x^* throughout the paper.

[3] Indeed, the regression problem is a so-called *model fitting problem* where, strictly speaking, the scenario solution x^* represents a model more than a "decision", and the "cost" function $\ell(x, \delta)$ represents how badly a model x fits a data point δ.

Example 5 (Weber problem with squared Euclidean norm)**.** *There are m clients, located at points $p_1, \ldots, p_m$ in the space $\mathbb{R}^2$, to be served by a facility whose location $x \in \mathbb{R}^2$ has to be decided. If there is no uncertainty, the best x is chosen by minimizing*

$$\ell(x) = \sum_{i=1}^{m} \omega_i \|x - p_i\|^2,$$

where ω_i is a positive weight reflecting the relative importance of serving the client located at p_i. However, both the clients' locations and their relative importance can alter during the course of time, and are subject to uncertainty. Past locations and the associated weights are then obtained from historical data, so that N scenarios

$$\begin{matrix} p_1^{(i)}, \ldots, p_m^{(i)} \\ \omega_1^{(i)}, \ldots, \omega_m^{(i)} \end{matrix}, \quad \text{for } i = 1, \ldots, N,$$

can be used to compute the scenario solution:

$$\begin{aligned} x^* &= \arg\min_x \sum_{i=1}^{N} \left(\sum_{j=1}^{m} \omega_j^{(i)} \|x - p_j^{(i)}\|^2 \right) \\ &= \arg\min_x \sum_{i=1}^{N} \|A_i x - b_i\|^2, \end{aligned}$$

where we have posed

$$A_i^T = \begin{bmatrix} \sqrt{\omega_1^{(i)}} & 0 & \sqrt{\omega_2^{(i)}} & 0 & \ldots & \sqrt{\omega_n^{(i)}} & 0 \\ 0 & \sqrt{\omega_1^{(i)}} & 0 & \sqrt{\omega_2^{(i)}} & \ldots & 0 & \sqrt{\omega_n^{(i)}} \end{bmatrix},$$

$$b_i^T = \begin{bmatrix} \sqrt{\omega_1^{(i)}} p_1^{T(i)} & \sqrt{\omega_2^{(i)}} p_2^{T(i)} & \ldots & \sqrt{\omega_n^{(i)}} p_n^{T(i)} \end{bmatrix},$$

(A_i^T and b_i^T indicate the transposes of A_i and b_i). Note that we only require that scenarios *are independent one from the others, while e.g. the m clients' locations can be correlated as well as weights are allowed to be location-dependent in a very complicated way.* ⋆

Finally, we mention a well-known problem in systems and control theory that involves the least squares method.

Example 6 (finite-horizon linear quadratic regulator)**.** *Consider the following linear system*

$$\begin{aligned} &z_{t+1} = Fz_t + Bx_t + w_t, \text{ where} \\ &\quad F \in \mathbb{R}^{m \times m}, B \in \mathbb{R}^{m \times n}, \\ &\quad z_t \in \mathbb{R}^m, w_t \in \mathbb{R}^m \; t = 0, 1, \ldots, T_f. \end{aligned}$$

We look for a control action $x = (x_0^T, x_1^T, \ldots, x_{T_f-1}^T)^T$, $x_t \in \mathbb{R}^n$, *for* $t = 0, 1, \ldots, T_f - 1$, *that keeps* z_t *close to* 0 *with a moderate control effort. Quantitatively, the best control action is defined as the one minimizing the following cost function*

$$\ell(x) = \sum_{t=0}^{T_f-1} \left(z_t^T Q z_t + x_t^T R x_t \right) + z_{T_f}^T Q_{T_f} z_{T_f}, \tag{2.2}$$

where $Q_t \in \mathbb{R}^{m\times m}$, $t = 0, \ldots, T_f$, *are fixed positive definite matrices penalizing deviations of the state variables* z_t *from* 0 *at each time* t, *and* $R_t \in \mathbb{R}^{n\times n}$, $t = 0, \ldots, T_f - 1$, *are fixed positive definite matrices penalizing the control effort. Denote with* $I_{m\times n}$ *the* $m \times n$ *rectangular identity matrix and by* I_m *the squared* $m \times m$ *identity matrix, and, similarly, with* $0_{m\times n} \in \mathbb{R}^{m\times n}$ *and* $0_m \in \mathbb{R}^{m\times m}$ *the matrices of zeros. In view of the system equations we have that*

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_{T_f} \end{bmatrix} = \begin{bmatrix} B & 0_{m\times n} & \cdots & 0_{m\times n} \\ FB & B & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0_{m\times n} \\ F^{T_f-1}B & F^{T_f-2}B & \cdots & B \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{T_f-1} \end{bmatrix}$$
$$+ \left(\begin{bmatrix} F \\ F^2 \\ \vdots \\ F^{T_f} \end{bmatrix} z_0 + \begin{bmatrix} I_m & 0_m & \cdots & 0_m \\ I_m & I_m & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0_m \\ I_m & I_m & \cdots & I_m \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \cdots \\ w_{T_f} \end{bmatrix} \right),$$

which, by denoting with ξ *the term in parentheses, with* $\mathcal{G}$ *the matrix multiplying* x, *and by defining* $z := (z_1^T, \ldots, z_{T_f}^T)^T$, *writes*

$$z = \mathcal{G}x + \xi.$$

Letting $\mathcal{Q}$ *and* $\mathcal{R}$ *be the following* $mT_f \times mT_f$ *and* $nT_f \times nT_f$ *matrices*

$$\mathcal{Q} := \begin{bmatrix} Q_1 & 0_m & \cdots & 0_m \\ 0_m & Q_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_m \\ 0_m & \cdots & 0_m & Q_{T_f} \end{bmatrix}^{\frac{1}{2}} \quad \mathcal{R} := \begin{bmatrix} R_0 & 0_n & \cdots & 0_n \\ 0_n & R_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_n \\ 0_n & \cdots & 0_n & R_{T_f-1} \end{bmatrix}^{\frac{1}{2}},$$

the cost function $\ell(x)$ *can be written as*

$$\begin{aligned} \ell(x) =& z_0^T Q_0 z_0 + \|\mathcal{Q}(\mathcal{G}x + \xi)\|^2 + \|\mathcal{R}x\|^2 \\ =& \left\| \begin{bmatrix} 0_{1\times nT_f} \\ \mathcal{Q}\mathcal{G} \\ \mathcal{R} \end{bmatrix} x - \begin{bmatrix} -Q_0^{\frac{1}{2}} z_0 \\ -\mathcal{Q}\xi \\ 0_{nT_f\times 1} \end{bmatrix} \right\|^2, \end{aligned}$$

which is clearly in the form $\|Ax-b\|^2$*, with* $A \in \mathbb{R}^{(m+n+1)T_f \times nT_f}$*,* $b \in \mathbb{R}^{(m+n+1)T_f}$ *and* $x \in \mathbb{R}^{nT_f}$*. If the initial state* z_0 *or the matrices* F, B *or the vector* $(w_0^T, \ldots, w_{T_f}^T)$ *are affected by uncertainty, the control action (decision)* x^* *can be chosen as the one that behaves best on average with respect to the observed realizations of* z_0, F, B *and* $(w_0^T, \ldots, w_{T_f}^T)$*. This is done by minimizing* $\sum_{i=1}^N \|A_i x - b_i\|^2$*, where* $(A_1, b_1), \ldots, (A_N, b_N)$ *are built as above for every observed uncertainty instance.* $\star$

Some references about the examples above can be found in Table 1.1 in Chapter 1.

Now, let us denote with $\mathbf{q}_i$ the squared residual of (A_i, b_i) evaluated at x^*, i.e.

$$\mathbf{q}_i := \|A_i x^* - b_i\|^2, \quad i = 1, \ldots, N$$

and consider a new instance of (A, b) sampled from $\mathbb{P}_\Delta$ independently of D^N. The *squared residual* of (A, b) evaluated at x^* is

$$\mathbf{q} := \|Ax^* - b\|^2.$$

$\mathbf{q}_1, \ldots, \mathbf{q}_N$ are statistics of $(A_1, b_1), \ldots, (A_N, b_N)$. Overall, $\mathbf{q}_1, \ldots, \mathbf{q}_N, \mathbf{q}$ are (univariate) random variables that depend on $(A_1, b_1), \ldots, (A_N, b_N), (A, b)$. In particular, each of them depends on all the data D^N through the decision x^*. In analogy with a classic result about order statistics, we would ask if the probability that $\mathbf{q}$ exceeds $\mathbf{q}_i$ can be studied independently of $\mathbb{P}_\Delta$. First, let us recall this classic result. Given a univariate sample $r_1, r_2, \ldots, r_N$, we denote with $r_{(1)}, r_{(2)}, \ldots, r_{(N)}$ the order statistics of the r_i's, that is, $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(N)}$. A similar notation is in force throughout for all univariate samples. The following theorem holds true.

Theorem 1. *Let* $r_1, r_2, \ldots, r_N$ *be a random sample from a distribution* F *on* $\mathbb{R}$*. For a new* r *sampled from* F *independently of* $r_1, r_2 \ldots, r_N$ *it holds that*

$$\mathbb{P}_F^{N+1}\{r \leq r_{(i)}\} \geq \frac{i}{N+1}, \quad i = 1, \ldots, N, \tag{2.3}$$

where $\mathbb{P}_F^{N+1}\{r \leq r_{(i)}\}$ *is the total probability of seeing* $r_1, \ldots, r_N$ *and* r *such that* $r \leq r_{(i)}$*.* $\star$

Equality in (2.3) holds whenever F is continuous, see e.g. [52], Chapter 3.

However, this result does not apply to our problem. Indeed, our decision x^* is chosen to minimize the average of the squared residuals, so that, in general, the squared residuals cannot but be biased toward small values when evaluated at that same x^*, and $\mathbb{P}_\Delta^{N+1}\{\mathbf{q} \leq \mathbf{q}_{(i)}\}$ is normally *less* than $\frac{i}{N+1}$. The following is a simple example illustrating this fact.

Example 7. *Consider having two data,* $\mathsf{D}^N = (A_1, b_1), (A_2, b_2)$*, i.e.* $N = 2$*. We assume that, with probability* 1*,* $A_1 = A_2 = 1$ *and* $b_1 \neq b_2$*. Based on* D^N*, the least squares solution* x^* *and the squared residuals* $\mathbf{q}_1, \mathbf{q}_2$ *are computed. We will evaluate the probability that a new instance* $(1, b)$ *is such that* $\mathbf{q} \leq \mathbf{q}_{(2)}$ *and*

show that it is strictly less than $\frac{2}{3}$*. First, notice that conditionally to any set of three instances, let's say* $S = \{(1, b'), (1, b''), (1, b''')\}$*, the probability of each permutation of the elements in* S *is the same, that is, the role of the* new *instance* $(1, b)$ *is played by each element of* S *with probability* $\frac{1}{3}$*. As a consequence, for any set of three instances, the three situations represented in Fig. 2.1 are equally likely and, since* $\mathbf{q} \leq \mathbf{q}_{(2)}$ *holds in one out of the three cases, integrating over all possible set of three instances* Δ^3 *yields* $\mathbb{P}^3_\Delta\{\mathbf{q} \leq \mathbf{q}_{(2)}\} = \frac{1}{3}$*.* ⋆

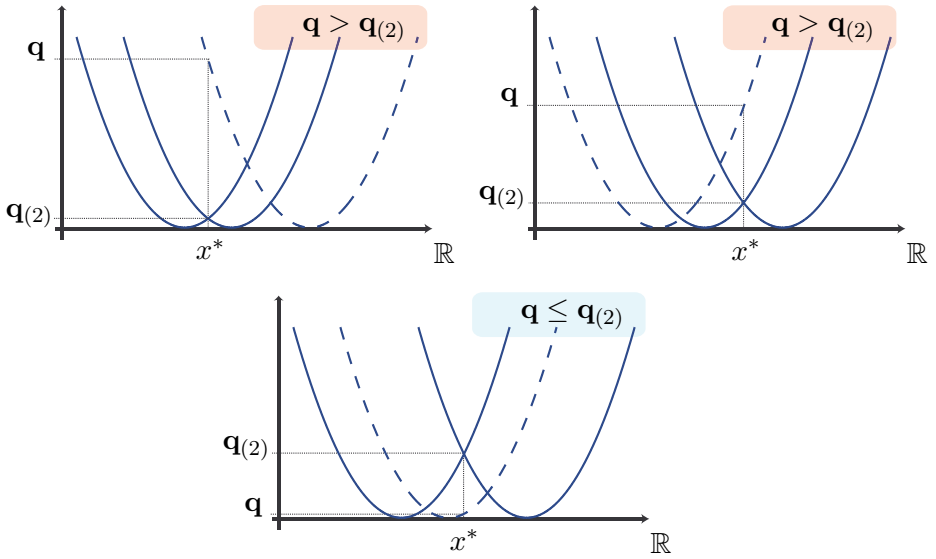


Figure 2.1. Given three uncertainty instances, the figure shows the possible relations between the statistic $\mathbf{q}_{(2)}$ of two of them (which play as the data $\mathsf{D}^N = (1, b_1), (1, b_2)$) and the squared residual $\mathbf{q}$ of the remaining instance (which plays as the new instance $(1, b)$). Squared residuals, as functions of x, are parabolae: the dashed parabola is $(x - b)^2$, associated with the new instance $(1, b)$, while the other two correspond to the data D^N.

In this chapter we provide statistics $\bar{\mathbf{q}}_{(i)}$, $i = 1, \ldots, N$, such that

$$\mathbb{P}^{N+1}_\Delta\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}\} \geq \frac{i}{N+1},$$

for every possible $\mathbb{P}_\Delta$. These statistics are obtained by adding a *margin* to the $\mathbf{q}_{(i)}$'s, according to a rule that does not depend on $\mathbb{P}_\Delta$. We will see that the margin is small in many situations, so that a good characterization of x^* through a finite and even small number of scenarios N is possible.

As already remarked in Chapter 1, the new instance (A, b) can be interpreted as the datum (A_{N+1}, b_{N+1}) observed immediately after the decision has been made based on $\mathsf{D}^N = (A_1, b_1), \ldots, (A_N, b_N)$. For example, in the regression problem considered in Example 4, the new instance corresponds to the next observed data point $(\theta^{(N+1)}, y^{(N+1)})$. In that context, our result guarantees being at least $\frac{i}{N+1}$ the probability of the event that

the data points D^N are observed, the coefficients x^* of the polynomial $P(\theta)$ and the statistic $\bar{\mathbf{q}}_{(i)}$ are computed depending on D^N only, and the next data point $(\theta^{(N+1)}, y^{(N+1)})$ is such that the squared difference between $P(\theta^{(N+1)})$ and $y^{(N+1)}$ is greater than $\bar{\mathbf{q}}_{(i)}$.

According to the Definition 3 in Section 1.1.3, the statistic $\bar{\mathbf{q}}_{(i)}$ obtained in this work is a *distribution-free* $\frac{i}{N+1}$*-mean coverage statistic*. Indeed, the coverage of $\bar{\mathbf{q}}_{(i)}$ is a function of the data D^N, defined as $\mathcal{C}(\bar{\mathbf{q}}_{(i)}) = \mathbb{P}_\Delta\{(A,b) \in \Delta : \mathbf{q} \leq \bar{\mathbf{q}}_{(i)}\}$ - see Definition 2 in Chapter 1. In view of (1.2), we have that $\mathbb{E}_{\Delta^N}[\mathcal{C}(\bar{\mathbf{q}}_{(i)})] = \mathbb{P}_\Delta^{N+1}\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}\}$, and therefore the value $\frac{i}{N+1}$ lower bounds the mean coverage of $\bar{\mathbf{q}}_{(i)}$.

Distributions-free results are of great importance since prior assumptions about $\mathbb{P}_\Delta$ are usually unrealistic. However, one may expect some conservatism, due to the pretension of guaranteeing the mean coverage against all possible $\mathbb{P}_\Delta$. Intuitively, we limit the conservatism by the fact of using statistics mimicking those of Theorem 1. This point will be discussed in more detail later on in Section 2.2.3, after the main theorem is stated.

2.2 Main result

First of all, we recall some frequently used notations.

2.2.1 Frequently used matrix notations

1. I denotes the identity matrix.

2. For a matrix M:

   M^T = transpose matrix of M;

   $M^\dagger$ = Moore Penrose generalized inverse of M;

   $\|M\|$ = spectral norm = $\sup_{\|x\|=1} \|Mx\|$, where the norm in the right-hand side is the Euclidean norm;

   $\lambda_{\max}(M)$ = maximum eigenvalue of M (M square matrix).

3. For a symmetric matrix M, $M \succ 0$ ($M \succeq 0$) means M positive definite (positive semi-definite). $P \succ Q$ ($P \succeq Q$) means $P - Q$ positive definite (positive semi-definite).

For further information on matrix concepts see e.g. [58, 59].

2.2.2 Main theorem

To ease the interpretation of the results here exposed, $\|A_i x - b_i\|^2$ will be conveniently written as: $\|A_i x - b_i\|^2 = (x - v_i)^T K_i (x - v_i) + h_i$, with $K_i = A_i^T A_i$, $v_i = A_i^\dagger b_i$, $h_i = \|A_i v_i - b_i\|^2$. Observe that, in general, we have $K_i \succeq 0$ and not $K_i \succ 0$. For example, in the regression problem of Example 4, K_i is always a rank 1 matrix, so that $K_i \not\succ 0$ for every $d > 1$.
Now, let us define the following N statistics of the data D^N, for $i = 1, \dots, N$,

$$\bar{\mathbf{q}}_i := \begin{cases} (x^* - v_i)^T \bar{K}_i (x^* - v_i) + h_i \\ \text{with } \bar{K}_i := K_i + 6K_i \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^N K_\ell \right)^{-1} K_i & \text{if } K_i \prec \frac{1}{6} \sum_{\substack{\ell=1 \\ \ell \neq i}}^N K_\ell \\ +\infty & \text{otherwise.} \end{cases} \tag{2.4}$$

Theorem 2. *For every probability measure $\mathbb{P}_\Delta$, with the notation above it holds that*

$$\mathbb{P}_\Delta^{N+1} \{ \mathbf{q} \leq \bar{\mathbf{q}}_{(i)} \} \geq \frac{i}{N+1}, \quad i = 1, \dots, N. \tag{2.5}$$

$\star$

The proof is given in Section 2.4 where a slightly stronger (but more cumbersome) result than that of Theorem 2 is also proved. Statistics $\bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_N$, as well as their ordered versions $\bar{\mathbf{q}}_{(1)}, \dots, \bar{\mathbf{q}}_{(N)}$, have a straight geometric interpretation. The squared residual $\mathbf{q}_i$ is the value of the paraboloid $(x - v_i)^T K_i (x - v_i) + h_i$ at $x = x^*$. According to the Theorem 2, the corresponding $\bar{\mathbf{q}}_i$ is obtained by evaluating at $x = x^*$ a *steepened* version of the paraboloid, obtained by replacing the matrix K_i with $\bar{K}_i$, see Fig. 2.2. The modified $\bar{K}_i$ is given by the original K_i *plus* a term whose magnitude depends on the comparison between K_i and $\sum_{\substack{\ell=1 \\ \ell \neq i}}^N K_\ell$, that is, between the steepness of the i-th paraboloid and the steepness of all the others as a whole. Intuitively, if K_i is "small" with respect to $\sum_{\substack{\ell=1 \\ \ell \neq i}}^N K_\ell$, then $\bar{K}_i \approx K_i$, so that $\bar{\mathbf{q}}_i \approx \mathbf{q}_i$ (i.e. the margin is small), otherwise, $\bar{\mathbf{q}}_i$ may become large, or even infinite if $K_i \not\prec \frac{1}{6} \sum_{\substack{\ell=1 \\ \ell \neq i}}^N K_\ell$. The so-obtained $\bar{\mathbf{q}}_1, \dots, \bar{\mathbf{q}}_N$ are finally ordered and $\bar{\mathbf{q}}_{(i)}$ is a distribution-free $\frac{i}{N+1}$-mean coverage statistic. Some remarks are in order.

Remark 3 (characterization of the margin)**.** *Under very mild assumptions, at the increasing of N the sum $\sum_{\substack{\ell=1 \\ \ell \neq i}}^N K_\ell$ becomes larger and larger with respect to a fixed K_i, so that the term $K_i \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^N K_\ell \right)^{-1} K_i$ in the definition of $\bar{K}_i$ tends to zero and $\bar{K}_i \to K_i$, for every i, yielding $\bar{\mathbf{q}}_{(i)} - \mathbf{q}_{(i)} \to 0$. Hence, our result mimics the classic result of Theorem 1. In Section 2.4.3 we prove formally the convergence of the margin between $\bar{\mathbf{q}}_{(i)}$ and $\mathbf{q}_{(i)}$ to zero under the hypothesis that the distributions of the K_i's and the v_i's have exponential tails. The rate of convergence of the margin to zero is problem dependent, as illustrated by the comparison between the two examples below: in Example 8, the margin goes to zero as $1/N$, while in Example 9, the margin is exactly zero for any $N \geq 8$.* $\star$

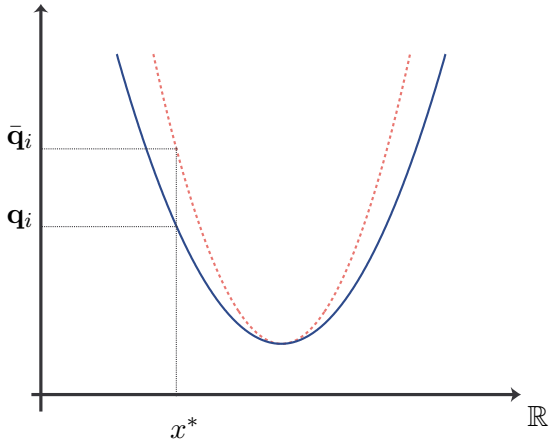


Figure 2.2. The parabola $(x - v_i)^T K_i (x - v_i) + h_i$ associated with the i-th instance (continuous line) is compared with its steepened version $(x - v_i)^T \bar{K}_i (x - v_i) + h_i$ (dashed line). At $x = x^*$, their values are, respectively, the squared residual $\mathbf{q}_i$ and $\bar{\mathbf{q}}_i$ as defined in (2.4).

In the following two examples, the statistics guaranteed by Theorem 1 can be very easily computed by hand, and they are representative of many possible situations where $K_i \approx I$, $i = 1, \dots, N$.

Example 8 (parabolae with coplanar vertexes and identity K_i)**.** *Assume that $A_i = I$, $i = 1, \dots, N$. Thus, $K_i = I$, $v_i = b_i$, $h_i = 0$. See Fig. 2.3(a) for a visualization of the associated cost functions $\|A_i x - b_i\|^2$.*
Observe that $K_i \prec \frac{1}{6}\sum_{\ell \neq i} K_\ell \iff N \geq 8$, hence, according to Theorem 2, we have:

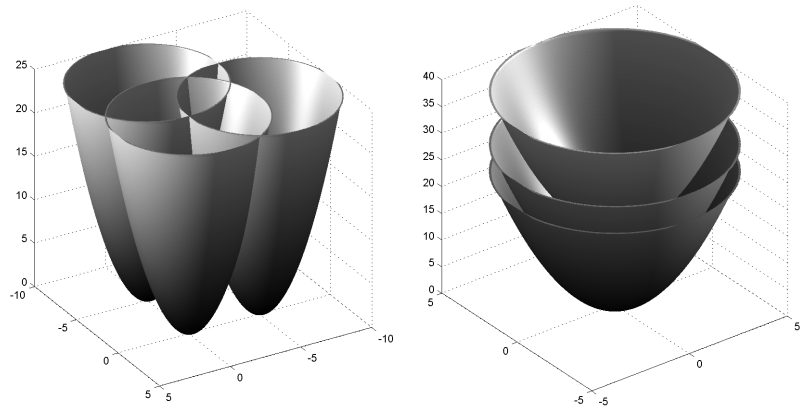
$$\bar{K}_i = \frac{N+5}{N-1} I,$$
$$\bar{\mathbf{q}}_{(i)} = \frac{N+5}{N-1} \mathbf{q}_{(i)},$$

whenever $N \geq 8$. Clearly, the margin $\bar{\mathbf{q}}_{(i)} - \mathbf{q}_{(i)}$ goes to zero as $1/N$, e.g. the margin is less than the 10% of $\mathbf{q}_{(i)}$ with $N = 62$, less than 1% with $N = 602$, etc. $\star$

Example 9 (stack of parabolae)**.** *Assume that the scenarios D^N are such that, for $i = 1, \dots, N$, we have*

$$A_i = \begin{bmatrix} I_{d\times d} \\ 0_{1\times d} \end{bmatrix} \text{ and } \quad b_i = \begin{bmatrix} 0_{d\times 1} \\ u_i \end{bmatrix},$$

where the subscripts denote the matrix dimensions (e.g. $0_{1\times d}$ is a row vector of zeros) and $u_1, \dots, u_N$ are scalar values. Thus, $K_i = I_{d\times d}$, $v_i = 0$, $h_i = u_i^2$. See



(a) (b)

Figure 2.3. In (a), three instances of the cost functions $\|A_i x - b_i\|^2$, with $A_i = I$ as in Example 8, are shown. In (b), cost functions like those in Example 9 are shown.

Fig. 2.3(b) for a visualization of the corresponding cost functions. Observe that $\frac{1}{6}\sum_{\substack{\ell=1\\ \ell\neq i}}^{N} K_\ell \succ K_i \iff N \geq 8$*, from which, according to Theorem 2, we have:*

$$\bar{K}_i = \frac{N+5}{N-1} I_{d\times d},$$
$$\bar{\mathbf{q}}_{(i)} = \mathbf{q}_{(i)},$$

whenever $N \geq 8$*. Thus, for* $N \geq 8$*, it holds that*

$$\mathbb{P}_{\Delta}^{N+1}\left\{\mathbf{q} \leq \mathbf{q}_{(i)}\right\} \geq \frac{i}{N+1},$$

i.e., there is no margin *and, in this situation, the result of Theorem 1 is recovered.*

$\star$

Remark 4 (the role of dimension d)**.** *By its definition* (2.4)*,* $\bar{\mathbf{q}}_i$ *has a finite value if* $\sum_{\substack{\ell=1\\ \ell\neq i}}^{N} K_\ell$ *is "sufficiently large" with respect to* K_i*. This is a technical fact with an intuitive interpretation. Consider the regression problem of Example 4. We have already observed that* K_i *has always rank 1: hence, the paraboloid associated with the* i*-th scenario is flat with respect to* $d-1$ *orthogonal directions, and it does not influence the solution* x^* *with respect to these directions. Thus, in this case, a necessary condition for the matrix inequality* $K_i \prec \frac{1}{6}\sum_{\substack{\ell=1\\ \ell\neq i}}^{N} K_\ell$ *to be true, and for* $\bar{\mathbf{q}}_i$ *to be significant, is that we have more than* d *observations, so that* $\sum_{\substack{\ell=1\\ \ell\neq i}}^{N} K_\ell$ *may span all the directions. However, this is* not *a general fact. Indeed, every time that* K_i *is nonsingular, each scenario brings information on every direction at the same*

time: in Examples 8 and 9 above, we have seen that for every $N \geq 8$ the condition $K_i \prec \frac{1}{6}\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell$ holds true independently of d. Hence, in general, the minimum N such that the considered statistics are significant does not depend directly on d, *but rather it depends on the problem structure and on the amount of information brought by each scenario.*

$\star$

2.2.3 Distribution-free results and conservatism

The statistics $\bar{\mathbf{q}}_{(1)}, \ldots, \bar{\mathbf{q}}_{(N)}$, defined in (2.4), can be computed without using any knowledge about $\mathbb{P}_\Delta$. Hence, in the light of the distribution-free result of Theorem 2, a decision-maker that is looking for a statistic whose mean coverage is guaranteed against all possible probability measures $\mathbb{P}_\Delta$ can always rely on $\bar{\mathbf{q}}_{(i)}$. However, an expected drawback of distribution-free results is conservatism, that is to say, given a problem and *a fixed* $\mathbb{P}_\Delta$, there may exist a statistic $\mathbf{c}$ "better" than $\bar{\mathbf{q}}_{(i)}$. Formally, $\mathbf{c}$ is better than $\bar{\mathbf{q}}_{(i)}$ if $\mathbf{c}$ satisfies the conditions

$$\mathbb{P}_\Delta^{N+1}\{\mathbf{q} \leq \mathbf{c}(\mathsf{D}^N)\} \geq \frac{i}{N+1} \tag{2.6}$$

and

$$\mathbf{c}(D^N) \leq \bar{\mathbf{q}}_{(i)} \text{ holds for every data } \mathsf{D}^N \tag{2.7}$$

and

$$\mathbb{P}_\Delta^{N+1}\{\mathbf{c}(\mathsf{D}^N) < \bar{\mathbf{q}}_{(i)}\} > 0. \tag{2.8}$$

Informally, we say that $\bar{\mathbf{q}}_{(i)}$ is *significantly* conservative if we actually have

$$\mathbb{P}_\Delta^{N+1}\{\mathbf{c}(\mathsf{D}^N) \ll \bar{\mathbf{q}}_{(i)}\} \gg 0,$$

rather than simply (2.8) - the symbol $\ll$ ($\gg$) stands for "significantly less (more) than". Otherwise, we can consider the conservatism to be practically negligible. As a starting point in order to study to what extent $\bar{\mathbf{q}}_{(i)}$ may be conservative, let us first consider the statistic $\mathbf{q}_{(i)}$. The following Theorem 3, proved in Section 2.4.2, holds under broad assumptions about the distribution of the data. In particular, we assume that the squared residuals do not accumulate at the same value, and this, in view of Theorem 3, entails that the mean coverage of $\mathbf{q}_{(i)}$ is always *no greater* than $\frac{i}{N+1}$, independently of $\mathbb{P}_\Delta$.

Theorem 3. *For any $\mathbb{P}_\Delta$ such that*

$$\mathbb{P}_\Delta^{N+1}\{\mathbf{q} \neq \mathbf{q}_\ell \text{ and } \mathbf{q}_\ell \neq \mathbf{q}_{\ell'} \text{ for every } \ell, \ell' \in \{1, \ldots, N\},\ \ell \neq \ell'\} = 1,$$

it holds that

$$\mathbb{P}_\Delta^{N+1}\{\mathbf{q} \leq \mathbf{q}_{(i)}\} \leq \frac{i}{N+1}, \quad i = 1, \ldots, N.$$

$\star$

The fact that the mean coverage of $\mathbf{q}_{(i)}$ *cannot* be greater than $\frac{i}{N+1}$ entails that, for any *specific* $\mathbb{P}_\Delta$, a statistic $\mathbf{c}$ that satisfies $\mathbf{c}(\mathsf{D}^N) \leq \mathbf{q}_{(i)}$ for all D^N cannot satisfy the condition $\mathbb{P}_\Delta^{N+1}\{\mathbf{q} \leq \mathbf{c}(\mathsf{D}^N)\} \geq \frac{i}{N+1}$ and, at the same time, the condition $\mathbb{P}_\Delta^{N+1}\{\mathbf{c}(\mathsf{D}^N) < \mathbf{q}_{(i)}\} > 0$. Hence, if $\bar{\mathbf{q}}_{(i)}$ is always close to $\mathbf{q}_{(i)}$, $\bar{\mathbf{q}}_{(i)}$ cannot be significantly conservative. On the other hand, note that the presence of a (possibly large) margin $\bar{\mathbf{q}}_{(i)} - \mathbf{q}_{(i)} > 0$ *is not* by itself a symptom of conservatism. Indeed, the intuition, corroborated by cases like Example 7, tells us that, even when a statistic $\mathbf{c}$ is constructed based on the knowledge of the specific $\mathbb{P}_\Delta$, a strictly positive margin $\mathbf{c}(\mathsf{D}^N) - \mathbf{q}_{(i)}$ is usually *necessary* to guarantee $\mathbb{P}_\Delta^{N+1}\{\mathbf{q} \leq \mathbf{c}(\mathsf{D}^N)\} \geq \frac{i}{N+1}$. The conclusion is that the only source of conservatism for $\bar{\mathbf{q}}_{(i)}$ can be a *larger-than-necessary* margin with respect to $\mathbf{q}_{(i)}$. However, we have shown that in many cases the margin $\bar{\mathbf{q}}_{(i)} - \mathbf{q}_{(i)}$ is small and tends to zero at the increasing of N, see Remark 3. Thus, the result of Theorem 2, though distribution-free, in many relevant cases is *not* significantly conservative.

2.3 Numerical example

The problem here considered is an instance of the Weber problem presented in Example 5. It is inspired by the problem example presented in [60].

2.3.1 An application to facility location

With reference to Example 5 above, we have 8 demand points (clients) whose weights and locations are uncertain.

We face the problem based on $N = 15$ scenarios, independently collected during a data acquisition campaign. In Fig. 2.4 the locations of the 8 clients in each of the 15 scenarios are showed. The corresponding weights are listed in Table 2.1. The least squares solution turns out to be $x^* = (0.6208, 0.5967)$. The values of the statistics $\bar{\mathbf{q}}_{(i)}$, $i = 1, \ldots, N$, having distribution-free $\frac{i}{N+1}$-mean coverages, are then computed. Fig. 2.5 compares the values of $\bar{\mathbf{q}}_{(1)}, \ldots, \bar{\mathbf{q}}_{(N)}$ with those of the ordered empirical least squares residuals $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(N)}$. The margins $\bar{\mathbf{q}}_{(i)} - \mathbf{q}_{(i)}$ turn out to be small.

2.3.2 Monte-Carlo tests

Usually, in real applications the distribution of the uncertainty, $\mathbb{P}_\Delta$, is unknown. However, the 15 scenarios used above have been randomly generated by simulation according to a known distribution, for illustration purpose. The nominal values of the uncertain locations and weights are reported in Table 2.2. We know that locations have been generated independently from eight Gaussian symmetric distributions centered in the nominal values, with standard deviation 0.11. Weights in each scenario have been generated according to a multivariate Gaussian distribution (truncated to positive values) such that each weight has a standard deviation

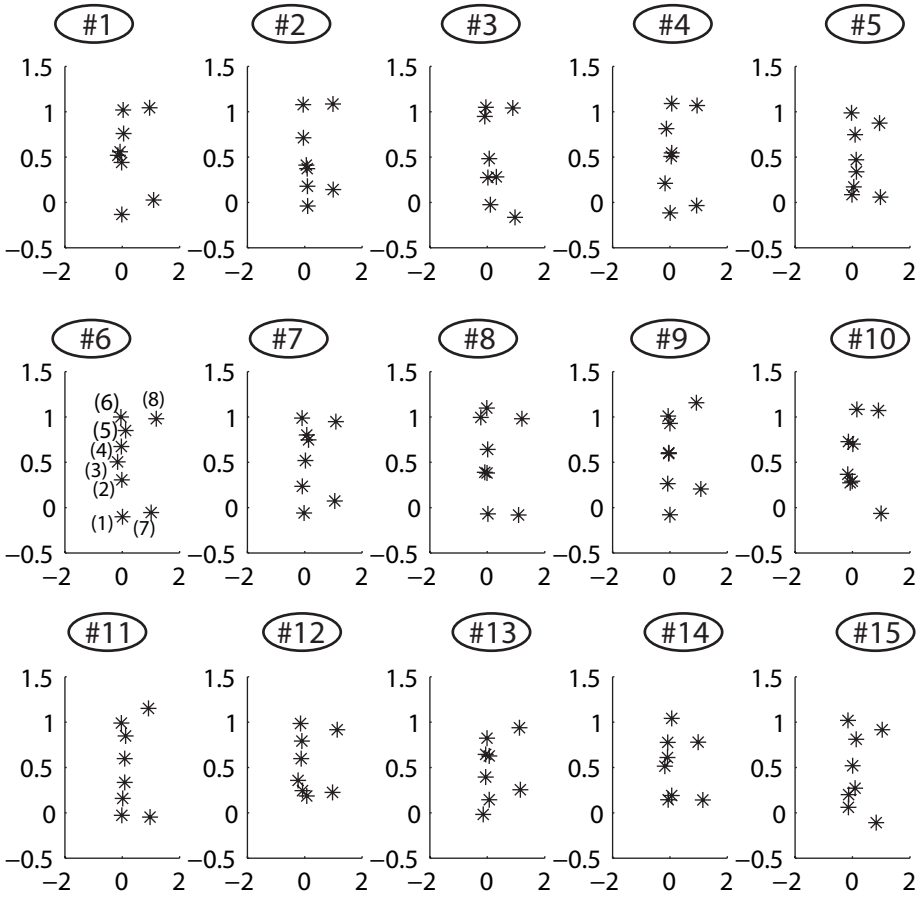


Figure 2.4. The clients' locations in $\mathbb{R}^2$ are shown for each of the 15 observed scenarios. In the representation of the scenario $\#6$, the identities of the 8 clients are indicated explicitly. Elsewhere they are omitted for ease of reading.

equal to half its nominal value, while the correlation coefficient between any two weights is $\rho = 0.1$.

Since Theorem 2 holds true for every possible $\mathbb{P}_\Delta$, all the results and considerations in Section 2.3.1 hold even in the absence of information about the real distribution. However, since we know the underlying data-generating mechanism, we can study the coverage properties of the cost thresholds $\bar{\mathbf{q}}_{(1)}, \ldots, \bar{\mathbf{q}}_{(N)}$ in the light of this knowledge. For example, a Monte-Carlo test based on $M = 2 \cdot 10^6$ trials allows us to estimate with an accuracy of 0.002 the mean coverages of $\bar{\mathbf{q}}_{(1)}, \ldots, \bar{\mathbf{q}}_{(N)}$ and of $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(N)}$ (with a confidence greater than $1 - 10^{-5}$). We obtain that the mean coverage of ant statistic $\mathbf{q}_{(i)}$ *exceeds* $\frac{i}{N+1}$, see Fig. 2.7, so that a margin *is necessary* to guarantee a mean coverage of at least $\frac{i}{N+1}$.

Moreover, we can estimate the whole coverage distribution of a given statistic $\bar{\mathbf{q}}_{(i)}$. For $M = 1000$ times, we generate $N = 15$ scenarios and compute the cov-

client:	1	2	3	4	5	6	7	8
scen. #1:	14.57	1.58	1.72	1.15	2.10	8.08	15.66	35.24
#2:	10.99	0.56	0.80	1.63	0.29	5.32	17.17	29.39
#3:	15.73	2.42	1.87	0.85	1.17	15.85	11.62	26.37
#4:	10.04	0.98	1.25	1.21	0.79	10.12	9.06	27.67
#5:	4.47	0.70	1.10	1.56	1.57	3.09	13.67	24.59
#6:	5.03	0.72	0.92	0.72	1.90	6.00	12.16	25.30
#7:	10.92	1.25	0.59	0.14	0.88	9.60	9.74	26.61
#8:	7.35	0.71	1.56	0.88	1.30	12.94	23.61	37.20
#9:	6.77	1.17	1.79	1.06	0.37	5.86	14.66	6.45
#10:	7.35	1.07	0.97	1.51	1.36	11.46	9.48	13.09
#11:	6.73	1.10	1.26	0.66	0.03	7.35	6.97	27.58
#12:	6.99	0.55	1.29	0.69	1.01	9.60	6.82	22.25
#13:	11.66	1.46	0.96	1.42	0.86	4.82	14.14	6.03
#14:	13.48	1.08	1.32	1.39	0.23	8.56	6.27	34.01
#15:	8.97	0.69	1.42	1.67	1.34	8.39	11.00	1.44

Table 2.1. The table reports the weights associated with each of the 8 clients, for each of the 15 observed scenarios.

client:	1	2	3	4	5	6	7	8
location:	(0.0,0.0)	(0.0,0.2)	(0.0,0.4)	(0.0,0.6)	(0.0,0.8)	(0.0,1.0)	(1.0,0.0)	(1.0,1.0)
weight:	10	1	1	1	1	10	10	20

Table 2.2. The table shows the nominal values of the locations and weights of the 8 clients. These numerical values are taken from [60] (however, note that in [60] only the weights are stochastic, while the locations are considered to be deterministic).

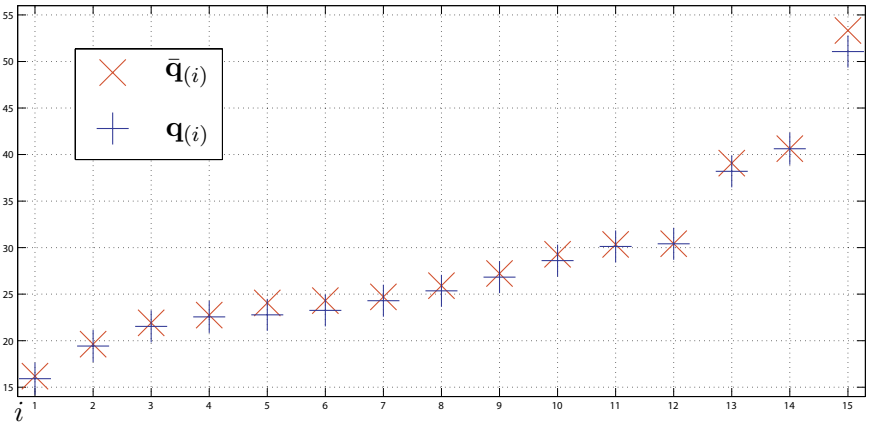


Figure 2.5. The figure shows a comparison between the values of $\bar{\mathbf{q}}_{(i)}$ and $\mathbf{q}_{(i)}$, for each $i = 1, \dots, N$ (i is in abscissa). We know that each $\bar{\mathbf{q}}_{(i)}$ has mean coverage guaranteed to be no less than $\frac{i}{N+1}$. For example, $\bar{\mathbf{q}}_{(14)}$ has distribution-free $\frac{14}{15}$-mean coverage. With the present data, $\bar{\mathbf{q}}_{(14)}$ is practically indistinguishable from $\mathbf{q}_{(14)} \approx 40$.

erage of $\bar{\mathbf{q}}_{(14)}$, i.e. of the $\frac{7}{8}$-mean coverage statistic. The histogram obtained from the $M = 1000$ trials is shown in Fig. 2.6(a). We perform the same Monte-Carlo test with $N = 31$ and build the histogram of the coverage of the $\frac{7}{8}$-mean coverage statistic, which, in this case, is $\bar{\mathbf{q}}_{(28)}$, see Fig. 2.6(b). Finally, we compute the histogram with $N = 63$ and the corresponding $\frac{7}{8}$-mean coverage statistic, which is $\bar{\mathbf{q}}_{(56)}$, see Fig. 2.6(c). We notice that the dispersion of the coverage distribution decreases sensibly for increasing N.

2.4 Proofs

2.4.1 Proof of Theorem 2

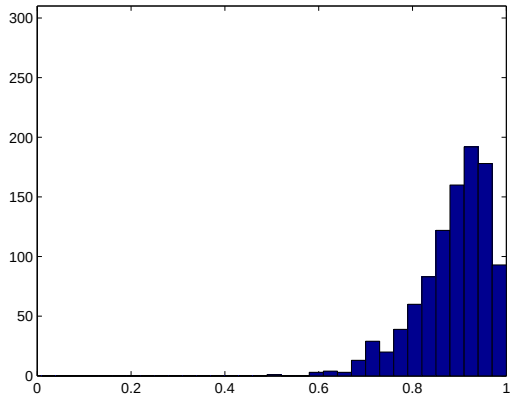
We prove below a Theorem 4 which is slightly stronger than Theorem 2, and show that Theorem 2 follows from Theorem 4. Throughout, we use the notation

$$\sum K_\ell \text{ for } \sum_{\ell=1}^{N} K_\ell$$

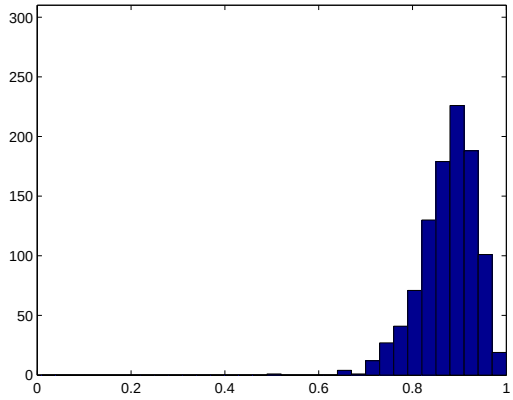
and

$$\sum_{\ell \neq i} K_\ell \text{ for } \sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell.$$

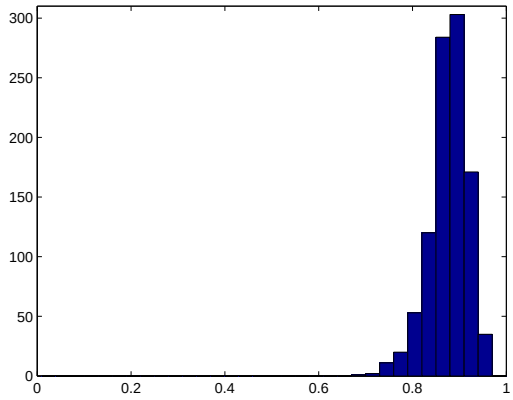
We start with a Lemma.



(a)



(b)



(c)

Figure 2.6. Histograms of the coverage of $\bar{\mathbf{q}}_{(14)}$ when $N = 15$, (a); of $\bar{\mathbf{q}}_{(28)}$ when $N = 31$, (b); of $\bar{\mathbf{q}}_{(56)}$ when $N = 63$, (c). In all the three cases, the statistics considered have distribution-free $\frac{7}{8}$-mean coverage.

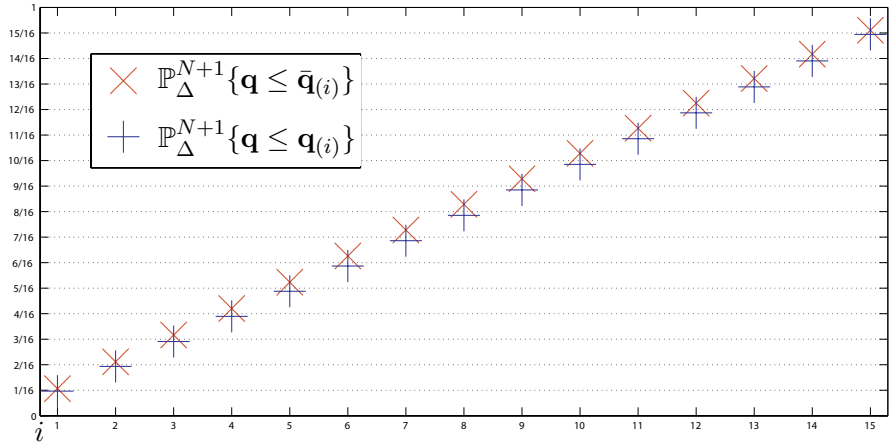


Figure 2.7. The mean coverages of $\bar{\mathbf{q}}_{(i)}$ and $\mathbf{q}_{(i)}$, $i = 1, \ldots, 15$ are here compared.

Lemma 1. *Assume that* $\sum_{\ell \neq i} K_\ell \succ 0$. *For any* $\gamma \geq 0$, *the following equivalences hold:*

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \prec \gamma I \iff K_i \prec \gamma \sum_{\ell \neq i} K_\ell, \tag{2.9}$$

and

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \preceq \gamma I \iff K_i \preceq \gamma \sum_{\ell \neq i} K_\ell. \tag{2.10}$$

$\star$

Proof. For $\gamma = 0$ the result is trivial. Suppose $\gamma > 0$. We prove (2.9); (2.10) can be proved similarly. Suppose first that $K_i \succ 0$. Multiplying the two sides of the inequality $K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \prec \gamma I$ on the left and on the right by $K_i^{-\frac{1}{2}}$, the equivalent inequality $\left(\sum_{\ell \neq i} K_\ell \right)^{-1} \prec \gamma K_i^{-1}$ follows. For positive definite matrices A and B, $A \prec B$ is equivalent to $B^{-1} \prec A^{-1}$ (see e.g. [58], Section 7.7), so that the last inequality can be reversed to $K_i \prec \gamma \sum_{\ell \neq i} K_\ell$, which is the inequality on the right-hand side of (2.9).

Suppose now that $K_i \succeq 0$. From $K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \prec \gamma I$ it follows that $(K_i + \epsilon I)^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} (K_i + \epsilon I)^{\frac{1}{2}} \prec \gamma I$ for some $\epsilon > 0$ small enough. Since $K_i + \epsilon I \succ 0$, from the first part of the proof we obtain $K_i + \epsilon I \prec \gamma \sum_{\ell \neq i} K_\ell$, which implies $K_i \prec \gamma \sum_{\ell \neq i} K_\ell$. Conversely, start from $K_i \prec \gamma \sum_{\ell \neq i} K_\ell$. Then, $K_i \prec \gamma' \sum_{\ell \neq i} K_\ell$ for some $\gamma' < \gamma$ close enough to γ, and further $K_i + \epsilon I \prec \gamma' \sum_{\ell \neq i} K_\ell$ for any $\epsilon > 0$ small enough. Since $K_i + \epsilon I \succ 0$, from the first part of

the proof we obtain $(K_i+\epsilon I)^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}(K_i+\epsilon I)^{\frac{1}{2}} \prec \gamma' I$.[4] Letting $\epsilon\to 0$ gives ${K_i}^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}{K_i}^{\frac{1}{2}} \preceq \gamma' I \prec \gamma I$, that is the left-hand side of (2.9). □

Whenever $\sum_{\ell\neq i}K_\ell \succ 0$, let

$$\gamma_i := \lambda_{\max}\left({K_i}^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}{K_i}^{\frac{1}{2}}\right),$$
$$W_i := K_i + (4+2\gamma_i)K_i\left(\sum_{\ell\neq i}K_\ell\right)^{-1}K_i. \tag{2.11}$$

Define

$$\tilde{\mathbf{q}}_i := \begin{cases} (x^*-v_i)^T\tilde{K}_i(x^*-v_i)+h_i \\ \text{with } \tilde{K}_i := W_i + W_i(2\sum K_\ell - W_i)^{-1}W_i & \text{if } \sum_{\ell\neq i}K_\ell \succ 0 \text{ and } \gamma_i < \frac{1}{\sqrt{2}} \\ +\infty & \text{otherwise.} \end{cases} \tag{2.12}$$

Note that $\tilde{K}_i$ in (2.12) is well-defined, that is, the inverse in the definition of $\tilde{K}_i$ exists. To show this, remember that γ_i is the maximum eigenvalue of ${K_i}^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}{K_i}^{\frac{1}{2}}$, so that

$$K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}K_i^{\frac{1}{2}} \preceq \gamma_i I, \tag{2.13}$$

and hence

$$\begin{aligned} W_i &= K_i + (4+2\gamma_i)K_i^{\frac{1}{2}}\left(K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}K_i^{\frac{1}{2}}\right)K_i^{\frac{1}{2}} \\ &\preceq K_i + (4+2\gamma_i)\gamma_i K_i \\ &= (1+4\gamma_i+2\gamma_i^2)K_i. \end{aligned} \tag{2.14}$$

Applying Lemma 1 to (2.13) gives $K_i \preceq \gamma_i\sum_{\ell\neq i}K_\ell$, from which $K_i \preceq \frac{\gamma_i}{1+\gamma_i}\sum K_\ell$. Substituting in the previous formula yields

$$W_i \preceq (1+4\gamma_i+2\gamma_i^2)\frac{\gamma_i}{1+\gamma_i}\sum K_\ell \prec [\text{since } \gamma_i < \frac{1}{\sqrt{2}}] \prec 2\sum K_\ell, \tag{2.15}$$

and the matrix that is inverted in (2.12) is therefore positive definite.

[4] Note that $(K_i+\epsilon I)^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}(K_i+\epsilon I)^{\frac{1}{2}} \prec \gamma' I$ for a given $\epsilon>0$ is not sufficient to conclude that $K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}K_i^{\frac{1}{2}} \preceq \gamma' I$. Indeed, XAX, $A\succ 0$, $X\succeq 0$, is not monotonic in X in general.

Theorem 4. *For every probability measure $\mathbb{P}_\Delta$, with the notation above it holds that*

$$\mathbb{P}_\Delta^{N+1}\{\mathbf{q} \le \tilde{\mathbf{q}}_{(i)}\} \ge \frac{i}{N+1}, \quad i = 1, \dots, N.$$

$\star$

Before proving the theorem, we show that Theorem 2 follows from Theorem 4. To prove this, it is enough to show that $\tilde{\mathbf{q}}_i \le \bar{\mathbf{q}}_i$. When $\bar{\mathbf{q}}_i = +\infty$, this is trivially true, so we consider the case when $\bar{\mathbf{q}}_i$ is finite, which holds if $K_i \prec \frac{1}{6}\sum_{\ell\neq i} K_\ell$. In view of Lemma 1, condition $K_i \prec \frac{1}{6}\sum_{\ell\neq i} K_\ell$ implies that $\gamma_i < \frac{1}{6}$. We show that, for $\gamma_i < \frac{1}{6}$, $\tilde{K}_i \preceq \bar{K}_i$ from which $\tilde{\mathbf{q}}_i \le \bar{\mathbf{q}}_i$.
Due to that $\gamma_i < \frac{1}{6}$, (2.14) gives $W_i \preceq 2K_i$, so that

$$2\sum K_\ell - W_i \succeq 2\sum K_\ell - 2K_i = 2\sum_{\ell\neq i} K_\ell.$$

Thus,

$$\begin{aligned}
\tilde{K}_i &= W_i + W_i\Big(2\sum K_\ell - W_i\Big)^{-1} W_i \preceq W_i + W_i\left(2\sum_{\ell\neq i} K_\ell\right)^{-1} W_i \\
&= [\text{substitute (2.11) for } W_i \text{ and let } \Phi = {K_i}^{\frac{1}{2}}\left(\sum_{\ell\neq i} K_\ell\right)^{-1} {K_i}^{\frac{1}{2}}] \\
&= K_i + {K_i}^{\frac{1}{2}}\left(\frac{9+4\gamma_i}{2}\Phi + (4+2\gamma_i)\Phi^2 + 2(2+\gamma_i)^2\Phi^3\right){K_i}^{\frac{1}{2}} \\
&\preceq [\text{since } \Phi \preceq \gamma_i I] \\
&\preceq 1 + {K_i}^{\frac{1}{2}}\left(\frac{9+4\gamma_i}{2}\Phi + (4+2\gamma_i)\gamma_i\Phi + 2(2+\gamma_i)^2\gamma_i^2\Phi\right){K_i}^{\frac{1}{2}} \\
&= K_i + (4.5 + 6\gamma_i + 10\gamma_i^2 + 8\gamma_i^3 + 2\gamma_i^4)K_i\left(\sum_{\ell\neq i} K_\ell\right)^{-1} K_i \\
&\preceq [\text{since } 4.5 + 6\gamma_i + 10\gamma_i^2 + 8\gamma_i^3 + 2\gamma_i^4 < 6 \text{ for } \gamma_i < \frac{1}{6}] \\
&\preceq \bar{K}_i.
\end{aligned}$$

□

Proof of Theorem 4
To ease the notation, let

$$\begin{aligned}
\mathbf{Q}_i(x) &:= (x - v_i)^T K_i (x - v_i) + h_i = \|A_i x - b_i\|^2, \text{ and} \\
\mathbf{Q}(x) &:= (x - v)^T K (x - v) + h = \|Ax - b\|^2.
\end{aligned}$$

With these positions,

$$x^* = \arg\min_x \sum_{i=1}^{N} \mathbf{Q}_i(x),\ \mathbf{q}_i = \mathbf{Q}_i(x^*),\ \mathbf{q} = \mathbf{Q}(x^*).$$

It is also convenient to introduce the minimizer of the least squares cost augmented with $\mathbf{Q}(x)$, namely,

$$\hat{x} := \arg\min_x \left\{ \sum_{i=1}^{N} \mathbf{Q}_i(x) + \mathbf{Q}(x) \right\}.$$

Finally, denote

$$\hat{x}^{[i]} := \arg\min_x \left\{ \sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} \mathbf{Q}_\ell(x) + \mathbf{Q}(x) \right\},\ i = 1, \ldots, N.$$

The following random variables $\mathbf{m}$ and $\mathbf{m}_1, \ldots, \mathbf{m}_N$, allow us to establish a ranking among $\mathbf{Q}(x), \mathbf{Q}_1(x), \ldots, \mathbf{Q}_N(x)$. Define:

$$\mathbf{m} := \begin{cases} \mathbf{Q}(x^*) + [\mathbf{Q}(x^*) - \mathbf{Q}(\hat{x})] & \text{if } \sum K_\ell \succ 0 \\ \infty & \text{otherwise,} \end{cases} \tag{2.16}$$

$$\mathbf{m}_i := \begin{cases} \mathbf{Q}_i(\hat{x}^{[i]}) + \left[\mathbf{Q}_i(\hat{x}^{[i]}) - \mathbf{Q}_i(\hat{x})\right] & \text{if } \sum_{\ell \neq i} K_\ell + K \succ 0 \\ \infty & \text{otherwise,} \end{cases} \tag{2.17}$$

for $i = 1, \ldots, N$.

Lemma 2. *For every probability measure $\mathbb{P}_\Delta$, with the notation above it holds that*

$$\mathbb{P}_\Delta^{N+1}\{\mathbf{m} \leq \mathbf{m}_{(i)}\} \geq \frac{i}{N+1},\ i = 1, \ldots, N.$$

$\star$

Proof. The random variables $\mathbf{m}$ and $\mathbf{m}_i$, $i = 1, \ldots, N$, are constructed from $\mathbf{Q}(x)$ and $\mathbf{Q}_i(x)$, $i = 1, \ldots, N$, and each of them depends on all the $\mathbf{Q}(x)$ and $\mathbf{Q}_i(x)$, $i = 1, \ldots, N$, directly and through $\hat{x}$, $\hat{x}^{[i]}$, and x^*. To indicate this, more explicitly write $\mathbf{m} = \mathbf{M}\left(\mathbf{Q}(x), \mathbf{Q}_1(x), \ldots, \mathbf{Q}_N(x)\right)$ and $\mathbf{m}_i = \mathbf{M}_i\left(\mathbf{Q}(x), \mathbf{Q}_1(x), \ldots, \mathbf{Q}_N(x)\right)$, $i = 1, \ldots, N$. On the other hand, an inspection of the definitions (2.16) and (2.17) reveals that each of the $\mathbf{M}_i\left(\mathbf{Q}(x), \mathbf{Q}_1(x), \ldots, \mathbf{Q}_N(x)\right)$, $i = 1, \ldots, N$, is but the function $\mathbf{M}$ applied to a permutation of the $\mathbf{Q}(x), \mathbf{Q}_1(x), \ldots, \mathbf{Q}_N(x)$:

$$\mathbf{m}_i = \mathbf{M}(\pi_i(\mathbf{Q}(x), \mathbf{Q}_1(x), \ldots, \mathbf{Q}_N(x))),\ \text{for suitable permutations } \pi_i,$$

$i = 1, \ldots, N$. Owing to that $\mathbf{Q}(x), \mathbf{Q}_1(x), \ldots, \mathbf{Q}_N(x)$ are independent and identically distributed, it follows that

$$\mathbb{P}_\Delta^{N+1}\{\mathbf{m} \leq \mathbf{m}_{(i)}\} = \mathbb{P}_\Delta^{N+1}\{\mathbf{m}_\ell \leq \mathrm{ord}_{(i)}\left[\mathbf{m}, \mathbf{m}_1, \ldots, \mathbf{m}_{\ell-1}, \mathbf{m}_{\ell+1}, \ldots, \mathbf{m}_N\right]\},$$

where $\mathrm{ord}_{(\mathrm{i})}$ is the i-th order statistic of the listed elements. Hence,

$$
\begin{aligned}
&\mathbb{P}_{\Delta}^{N+1}\{\mathbf{m} \leq \mathbf{m}_{(i)}\} \\
&\quad = \frac{1}{N+1}\left(\mathbb{P}_{\Delta}^{N+1}\{\mathbf{m} \leq \mathbf{m}_{(i)}\}\right. \\
&\qquad \left. + \sum_{\ell=1}^{N} \mathbb{P}_{\Delta}^{N+1}\{\mathbf{m}_{\ell} \leq \mathrm{ord}_{(i)}\left[\mathbf{m}, \mathbf{m}_1, \ldots, \mathbf{m}_{\ell-1}, \mathbf{m}_{\ell+1}, \ldots, \mathbf{m}_N\right]\}\right) \\
&\quad = [\mathbb{1}\{\cdot\} = \text{indicator function}] \\
&\quad = \frac{1}{N+1}\left(\mathbb{E}_{\Delta^{N+1}}\left[\mathbb{1}\left\{\mathbf{m} \leq \mathbf{m}_{(i)}\right\}\right]\right. \\
&\qquad \left. + \sum_{\ell=1}^{N} \mathbb{E}_{\Delta^{N+1}}\left[\mathbb{1}\left\{\mathbf{m}_{\ell} \leq \mathrm{ord}_{(i)}\left[\mathbf{m}, \mathbf{m}_1, \ldots, \mathbf{m}_{\ell-1}, \mathbf{m}_{\ell+1}, \ldots, \mathbf{m}_N\right]\right\}\right]\right) \\
&\quad = \frac{1}{N+1}\mathbb{E}_{\Delta^{N+1}}\left[\mathbb{1}\left\{\mathbf{m} \leq \mathbf{m}_{(i)}\right\}\right. \\
&\qquad \left. + \sum_{\ell=1}^{N} \mathbb{1}\left\{\mathbf{m}_{\ell} \leq \mathrm{ord}_{(i)}\left[\mathbf{m}, \mathbf{m}_1, \ldots, \mathbf{m}_{\ell-1}, \mathbf{m}_{\ell+1}, \ldots, \mathbf{m}_N\right]\right\}\right] \\
&\quad \geq \frac{i}{N+1},
\end{aligned}
$$

where the last inequality holds because at least i among the $\mathbf{m}$ and $\mathbf{m}_{\ell}$, $\ell = 1, \ldots, N$, are in one of the first i positions (they can be more than i when some assume the same value). □

Now, for $i = 1, \ldots, N$, define

$$
\begin{aligned}
\nu_i := \quad & \sup_{K,v,h} \mathbf{Q}(x^*) \\
& \text{subject to: } \mathbf{m} \leq \mathbf{m}_{(i)}.
\end{aligned}
\tag{2.18}
$$

Note that in the definition of ν_i, sup is taken with respect to (K, v, h), so that ν_i is a function of $(K_1, v_1, h_1), \ldots, (K_N, v_N, h_N)$. If we prove that

$$
\tilde{\mathbf{q}}_{(i)} \geq \nu_i, \tag{2.19}
$$

then

$$
\begin{aligned}
\mathbb{P}_{\Delta}^{N+1}\{\mathbf{q} \leq \tilde{\mathbf{q}}_{(i)}\} &= \mathbb{P}_{\Delta}^{N+1}\{\mathbf{Q}(x^*) \leq \tilde{\mathbf{q}}_{(i)}\} \\
&\geq \mathbb{P}_{\Delta}^{N+1}\{\mathbf{Q}(x^*) \leq \nu_i\} \\
&\geq \mathbb{P}_{\Delta}^{N+1}\{\mathbf{m} \leq \mathbf{m}_{(i)}\} \\
&\geq \frac{i}{N+1},
\end{aligned}
$$

where the last inequality follows from Lemma 2, and hence Theorem 4 is proved. Thus, in what follows, we concentrate on proving (2.19).

Let, for $\ell = 1, \ldots, N$,

$$\begin{aligned} \mu_\ell := \quad & \sup_{K,v,h} \mathbf{Q}(x^*) \\ & \text{subject to: } \mathbf{m} \leq \mathbf{m}_\ell. \end{aligned} \tag{2.20}$$

We show that $\mu_{(i)} \geq \nu_i$, $i = 1 \ldots, N$.
Assume for simplicity that $\sup$ in (2.18) is actually a $\max$ (if not, the proof follows by a limiting argument), and let (K^*, v^*, h^*) be the maximizer. At $(K, v, h) = (K^*, v^*, h^*)$, we have $\mathbf{m} \leq \mathbf{m}_{(i)}$, which entails that (K^*, v^*, h^*) is feasible for at least $N - i + 1$ values of ℓ in (2.20). Hence, $\mu_\ell \geq \mathbf{Q}^*(x^*) = \nu_i$ for at least $N - i + 1$ values of ℓ. Thus, $\mu_{(i)} \geq \nu_i$. Since $\mu_{(i)} \geq \nu_i$, it is enough, in order to prove (2.19), to show that $\tilde{\mathbf{q}}_{(i)} \geq \mu_{(i)}$. The remainder of the proof amounts to showing that $\tilde{\mathbf{q}}_i \geq \mu_i$, $i = 1, \ldots, N$, which plainly entails $\tilde{\mathbf{q}}_{(i)} \geq \mu_{(i)} \geq \nu_i$.

If $\sum_{\ell \neq i} K_\ell \not\succ 0$ or $\gamma_i \geq \frac{1}{\sqrt{2}}$, then $\tilde{\mathbf{q}}_i = +\infty$ and $\tilde{\mathbf{q}}_i \geq \mu_i$ is trivially verified (see (2.12)). Hence, we work under the condition

$$\sum_{\ell \neq i} K_\ell \succ 0 \text{ and } \gamma_i < \frac{1}{\sqrt{2}}.$$

By substituting in (2.20) the expressions (2.16) and (2.17) for $\mathbf{m}$ and $\mathbf{m}_i$, we have

$$\begin{aligned} \mu_i = \quad & \sup_{K,v,h} \mathbf{Q}(x^*) \\ & \text{subject to: } \mathbf{Q}(x^*) \leq \mathbf{Q}(\hat{x}) - \mathbf{Q}(x^*) + 2\mathbf{Q}_i(\hat{x}^{[i]}) - \mathbf{Q}_i(\hat{x}), \end{aligned}$$

i.e. μ_i is computed as the supremum of $\mathbf{Q}(x^*)$ over the values of K, v, h such that $\mathbf{Q}(x^*)$ is less than or equal to the bounding function in the right-hand side of the inequality. This entails that

$$\mu_\ell \leq \sup_{K,v,h} \left\{ \mathbf{Q}(\hat{x}) - \mathbf{Q}(x^*) + 2\mathbf{Q}_i(\hat{x}^{[i]}) - \mathbf{Q}_i(\hat{x}) \right\}, \tag{2.21}$$

where the right-hand side is an unconstrained supremum problem, which can be more easily handled than (2.20). We need now to write explicitly the dependence of the right-hand side of (2.21) on the optimization variables K, v, h. Note that:

$$\begin{aligned} x^* &= \Big(\sum K_\ell \Big)^{-1} \sum K_\ell v_\ell, \\ \hat{x} &= \Big(\sum K_\ell + K \Big)^{-1} \Big(\sum K_\ell v_\ell + Kv \Big), \\ \hat{x}^{[i]} &= \left(\sum_{\ell \neq i} K_\ell + K \right)^{-1} \left(\sum_{\ell \neq i} K_\ell v_\ell + Kv \right), \end{aligned}$$

so that

$$\begin{aligned}
\mathbf{Q}(x^*) =& \left(\left(\sum K_\ell \right)^{-1} \sum K_\ell v_\ell - v \right)^T K \left(\left(\sum K_\ell \right)^{-1} \sum K_\ell v_\ell - v \right) + h, \\
\mathbf{Q}_i(\hat{x}^{[i]}) =& \left(\left(\sum_{\ell \neq i} K_\ell + K \right)^{-1} \left(\sum_{\ell \neq i} K_\ell v_\ell + Kv \right) - v_i \right)^T K_i \\
& \cdot \left(\left(\sum_{\ell \neq i} K_\ell + K \right)^{-1} \left(\sum_{\ell \neq i} K_\ell v_\ell + Kv \right) - v_i \right) + h_i, \\
\mathbf{Q}(\hat{x}) =& \left(\left(\sum K_\ell + K \right)^{-1} \left(\sum K_\ell v_\ell + Kv \right) - v \right)^T K \\
& \cdot \left(\left(\sum K_\ell + K \right)^{-1} \left(\sum K_\ell v_\ell + Kv \right) - v \right) + h, \\
\mathbf{Q}_i(\hat{x}) =& \left(\left(\sum K_\ell + K \right)^{-1} \left(\sum K_\ell v_\ell + Kv \right) - v_i \right)^T K_i \\
& \cdot \left(\left(\sum K_\ell + K \right)^{-1} \left(\sum K_\ell v_\ell + Kv \right) - v_i \right) + h_i.
\end{aligned}$$

By letting

$$w := w(K, v) := \left(\sum K_\ell + K \right)^{-1} \left(\sum K_\ell v_\ell + Kv \right) - v, \tag{2.22}$$

$$w_i := w_i(K, v) := \left(\sum K_\ell + K \right)^{-1} \left(\sum K_\ell v_\ell + Kv \right) - v_i, \tag{2.23}$$

and by noting that

$$\begin{aligned}
&\left(\sum K_\ell \right)^{-1} \sum K_\ell v_\ell - v \\
&\quad = \left(\sum K_\ell \right)^{-1} \left(\sum K_\ell v_\ell - \sum K_\ell v \right) \\
&\quad = \left(\sum K_\ell \right)^{-1} \left(\sum K_\ell + K \right) \left(\sum K_\ell + K \right)^{-1} \left[\sum K_\ell v_\ell + Kv - \left(\sum K_\ell + K \right) v \right] \\
&\quad = \left(I + \left(\sum K_\ell \right)^{-1} K \right) \left[\left(\sum K_\ell + K \right)^{-1} \left(\sum K_\ell v_\ell + Kv \right) - v \right] \\
&\quad = \left(I + \left(\sum K_\ell \right)^{-1} K \right) w,
\end{aligned}$$

and that

$$\left(\sum_{\ell\neq i}K_\ell + K\right)^{-1}\left(\sum_{\ell\neq i}K_\ell v_\ell + Kv\right) - v_i = \text{[same calculations as before]}$$
$$= \left(I + \left(\sum_{\ell\neq i}K_\ell + K\right)^{-1}K_i\right)w_i,$$

$\mathbf{Q}(x^*)$, $\mathbf{Q}_i(\hat{x}^{[i]})$, $\mathbf{Q}(\hat{x})$ and $\mathbf{Q}_i(\hat{x})$ can be rewritten as:

$$\mathbf{Q}(x^*) = w^T\left(I + \left(\sum K_\ell\right)^{-1}K\right)^T K\left(I + \left(\sum K_\ell\right)^{-1}K\right)w + h,$$
$$\mathbf{Q}_i(\hat{x}^{[i]}) = w_i^T\left(I + \left(\sum_{\ell\neq i}K_\ell + K\right)^{-1}K_i\right)^T K_i\left(I + \left(\sum_{\ell\neq i}K_\ell + K\right)^{-1}K_i\right)w_i + h_i,$$
$$\mathbf{Q}(\hat{x}) = w^T K w + h, \quad \mathbf{Q}_i(\hat{x}) = w_i^T K_i w_i + h_i.$$

Substituting in (2.21) and noting that by taking the difference between $\mathbf{Q}(\hat{x})$ and $\mathbf{Q}(x^*)$ the dependence on h is lost, we have that

$$\mu_i \leq \sup_{K,v}\left\{w^T\left(K - \left(I + \left(\sum K_\ell\right)^{-1}K\right)^T K\left(I + \left(\sum K_\ell\right)^{-1}K\right)\right)w\right.$$
$$+ w_i^T\left(2\left(I + \left(\sum_{\ell\neq i}K_\ell + K\right)^{-1}K_i\right)^T K_i\left(I + \left(\sum_{\ell\neq i}K_\ell + K\right)^{-1}K_i\right)\right.$$
$$\left.\left. - K_i\right)w_i + h_i\right\} \tag{2.24}$$

In taking the sup in (2.24), we need to recall that $w = w(K, v)$ and $w_i = w_i(K, v)$, see (2.22) and (2.23). On the other hand, (2.22) defines a bijection between the pairs (K, v) and the pairs (K, w), since

$$w = \left(\sum K_\ell + K\right)^{-1}\left(\sum K_\ell v_\ell + Kv\right) - v$$
$$v = \left(\sum K_\ell\right)^{-1}\sum K_\ell v_\ell - \left(I + \left(\sum K_\ell\right)^{-1}K\right)w.$$

Therefore, the supremum with respect to (K, v) in (2.24) can be replaced by the supremum with respect to (K, w) as long as w_i is written as a function of (K, w)

by substituting in (2.23) the expression for v. We have

$$\begin{aligned}
w_i &= \Big(\sum K_\ell + K\Big)^{-1}\Bigg(\sum K_\ell v_\ell + K \\
&\quad \cdot \Bigg[\Big(\sum K_\ell\Big)^{-1}\sum K_\ell v_\ell - \Big(I + \Big(\sum K_\ell\Big)^{-1}K\Big)w\Bigg]\Bigg) - v_i \\
&= \Big(\sum K_\ell + K\Big)^{-1}\Big(I + K\Big(\sum K_\ell\Big)^{-1}\Big)\sum K_\ell v_\ell \\
&\quad - \Big(\sum K_\ell + K\Big)^{-1}\Big(K + K\Big(\sum K_\ell\Big)^{-1}K\Big)w - v_i \\
&= \Big(\sum K_\ell + K\Big)^{-1}\Big(\sum K_\ell + K\Big)\Big(\sum K_\ell\Big)^{-1}\sum K_\ell v_\ell \\
&\quad - \Big(\sum K_\ell + K\Big)^{-1}\Big(\sum K_\ell + K\Big)\Big(\sum K_\ell\Big)^{-1}Kw - v_i \\
&= \Big(\sum K_\ell\Big)^{-1}\sum K_\ell v_\ell - v_i - \Big(\sum K_\ell\Big)^{-1}Kw \\
&= x^* - v_i - \Big(\sum K_\ell\Big)^{-1}Kw.
\end{aligned}$$

By letting

$$V(K) := 2\left(I + \left(\sum_{\ell\neq i} K_\ell + K\right)^{-1} K_i\right)^T K_i \left(I + \left(\sum_{\ell\neq i} K_\ell + K\right)^{-1} K_i\right) - K_i, \tag{2.25}$$

(2.24) can be rewritten as (recall that K and K_i are symmetric)

$$\begin{aligned}
\mu_i \leq &\sup_{K,w}\Bigg\{w^T\Bigg(K - \Bigg(K + 2K\Big(\sum K_\ell\Big)^{-1}K + K\Big(\sum K_\ell\Big)^{-1}K\Big(\sum K_\ell\Big)^{-1}K\Bigg)\Bigg)w \\
&\quad + \Big(x^* - v_i - \Big(\sum K_\ell\Big)^{-1}Kw\Big)^T V(K)\Big(x^* - v_i - \Big(\sum K_\ell\Big)^{-1}Kw\Big) + h_i\Bigg\} \\
= &\sup_{K,w}\Bigg\{w^T K\Big(\sum K_\ell\Big)^{-1}\Big(V(K) - 2\sum K_\ell - K\Big)\Big(\sum K_\ell\Big)^{-1}Kw \\
&\quad - 2(x^* - v_i)V(K)\Big(\sum K_\ell\Big)^{-1}Kw + (x^* - v_i)^T V(K)(x^* - v_i) + h_i\Bigg\}.
\end{aligned}$$

Finally, letting

$$A(K) := V(K) - 2\sum K_\ell - K, \tag{2.26}$$

$$B(K) := -2(x^* - v_i)^T V(K), \tag{2.27}$$

$$C(K) := (x^* - v_i)^T V(K)(x^* - v_i)^T + h_i, \tag{2.28}$$

we have that

$$\mu_i \leq \sup_{K,w} \left\{ w^T K \Big(\sum K_\ell \Big)^{-1} A(K) \Big(\sum K_\ell \Big)^{-1} K w + B(K) \Big(\sum K_\ell \Big)^{-1} K w + C(K) \right\}. \tag{2.29}$$

For every $K \succeq 0$, let

$$\mathcal{M}(K) := \left\{ y \in \mathbb{R}^d : y = \Big(\sum K_\ell \Big)^{-1} K w, w \in \mathbb{R}^d \right\},$$

that is, $\mathcal{M}(K)$ is the image set of w through $(\sum K_\ell)^{-1} K$. Clearly, for every fixed K,

$$\begin{aligned} &\sup_{w} \left\{ w^T K \Big(\sum K_\ell \Big)^{-1} A(K) \Big(\sum K_\ell \Big)^{-1} K w + B(K) \Big(\sum K_\ell \Big)^{-1} K w + C(K) \right\} \\ &\quad = \sup_{y \in \mathcal{M}(K)} \left\{ y^T A(K) y + B(K) y + C(K) \right\} \\ &\quad \leq \sup_{y} \left\{ y^T A(K) y + B(K) y + C(K) \right\}, \end{aligned}$$

where the last inequality is an equality when $K \succ 0$ since in this case $\mathcal{M}(K) = \mathbb{R}^d$. Hence, by continuity in $K \succeq 0$ of the $\sup$ argument, we have

$$\begin{aligned} &\sup_{K,w} \left\{ w^T K \Big(\sum K_\ell \Big)^{-1} A(K) \Big(\sum K_\ell \Big)^{-1} K w + B(K) \Big(\sum K_\ell \Big)^{-1} K w + C(K) \right\} \\ &\quad = \sup_{K \succ 0, w} \left\{ w^T K \Big(\sum K_\ell \Big)^{-1} A(K) \Big(\sum K_\ell \Big)^{-1} K w + B(K) \Big(\sum K_\ell \Big)^{-1} K w \right. \\ &\qquad\qquad \left. + C(K) \right\} \\ &\quad = \sup_{K \succ 0, y} \left\{ y^T A(K) y + B(K) y + C(K) \right\} \\ &\quad = \sup_{K,y} \left\{ y^T A(K) y + B(K) y + C(K) \right\}. \end{aligned} \tag{2.30}$$

For every fixed K, $y^T A(K) y + B(K) y + C(K)$ admits a maximizer, say $y_{\max}$, because $A(K) \prec 0$, as stated by the following Lemma 3.

Lemma 3. *If $\sum_{\ell \neq i} K_\ell \succ 0$ and $\gamma_i < \frac{1}{\sqrt{2}}$, then $A(K) \prec 0$, $\forall K \succeq 0$.*

$\star$

Proof. From (2.26) and (2.25) we have that

$$\begin{aligned}
A(K) &= 2\left(I+\left(\sum_{\ell\neq i}K_\ell+K\right)^{-1}K_i\right)^T K_i\left(I+\left(\sum_{\ell\neq i}K_\ell+K\right)^{-1}K_i\right)-K_i \\
&\quad -2\sum K_\ell - K \\
&= 2K_i^{\frac{1}{2}}\left(I+K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell+K\right)^{-1}K_i^{\frac{1}{2}}\right)^2 K_i^{\frac{1}{2}} - K_i - 2\sum K_\ell - K \\
&\preceq 2K_i^{\frac{1}{2}}\left(I+K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell+K\right)^{-1}K_i^{\frac{1}{2}}\right)^2 K_i^{\frac{1}{2}} - K_i - 2\sum K_\ell.
\end{aligned} \tag{2.31}$$

Observe that

$$\begin{aligned}
&I+K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell+K\right)^{-1}K_i^{\frac{1}{2}} \\
&\quad\preceq [\text{since } \sum_{\ell\neq i}K_\ell+K\succeq\sum_{\ell\neq i}K_\ell \Rightarrow \left(\sum_{\ell\neq i}K_\ell+K\right)^{-1}\preceq\left(\sum_{\ell\neq i}K_\ell\right)^{-1}] \\
&\quad\preceq I+K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}K_i^{\frac{1}{2}} \\
&\quad\prec [\text{by (2.13) and using the assumption that } \gamma_i<\frac{1}{\sqrt{2}}] \\
&\quad\prec \left(1+\frac{1}{\sqrt{2}}\right)I,
\end{aligned}$$

so that

$$\left(I+K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell+K\right)^{-1}K_i^{\frac{1}{2}}\right)^2 \prec \left(1+\frac{1}{\sqrt{2}}\right)^2 I.$$

Substituting in (2.31), we obtain

$$A(K)\preceq 2\left(1+\frac{1}{\sqrt{2}}\right)^2 K_i - K_i - 2\sum K_\ell = 2\Big((\sqrt{2}+1)K_i-\sum K_\ell\Big)\prec 0,$$

where the last inequality follows since $\gamma_i<\frac{1}{\sqrt{2}}$ implies $K_i^{\frac{1}{2}}\left(\sum_{\ell\neq i}K_\ell\right)^{-1}K_i^{\frac{1}{2}}\prec\frac{1}{\sqrt{2}}I$, from which, in view of Lemma 1, $K_i\prec\frac{1}{\sqrt{2}}\sum_{\ell\neq i}K_\ell$, entailing in turn that $(\sqrt{2}+1)K_i\prec\sum K_\ell$. □

Clearly,

$$y_{\max} = -\frac{1}{2}A(K)^{-1}B(K)^T,$$

yielding

$$\begin{aligned}
&\sup_y \left\{y^T A(K)y + B(K)y + C(K)\right\} \\
&\quad = \max_y \left\{y^T A(K)y + B(K)y + C(K)\right\} \\
&\quad = -\frac{1}{4}B(K)A(K)^{-1}B(K)^T + C(K) \\
&\quad = [\text{from (2.27)}] \\
&\quad = -(x^* - v_i)^T V(K)A(K)^{-1}V(K)(x^* - v_i) + C(K) \\
&\quad = [\text{from (2.28)}] \\
&\quad = (x^* - v_i)^T \big(V(K) - V(K)A(K)^{-1}V(K)\big)(x^* - v_i) + h_i. \qquad (2.32)
\end{aligned}$$

Finally, from (2.29) and (2.30) we conclude

$$\mu_i \le \sup_K (x^* - v_i)^T \big(V(K) - V(K)A(K)^{-1}V(K)\big)(x^* - v_i) + h_i. \qquad (2.33)$$

The final step amounts to showing that $\forall K \succeq 0$

$$(x^* - v_i)^T \big(V(K) - V(K)A(K)^{-1}V(K)\big)(x^* - v_i) + h_i \le \tilde{\mathbf{q}}_i,$$

thus concluding the proof. This is done in view of the following lemma.

Lemma 4. *Assume that* $\sum_{\ell \neq i} K_\ell \succ 0$ *and* $\gamma_i < \frac{1}{\sqrt{2}}$. *Let* W *be a symmetric matrix,* $W \succeq 0$, *such that*

1. $W \prec 2\sum K_\ell$,
2. $V(K) \preceq W, \forall K \succeq 0$.

It holds that

$$V(K) - V(K)A(K)^{-1}V(K) \preceq W - W\Big(W - 2\sum K_\ell\Big)^{-1}W, \quad \forall K \succeq 0.$$

$\star$

Proof. Suppose first $K_i \succ 0$. Then, $V(K) \succ 0$ (see (2.25)), and, from $V(K) \preceq W$, we get (see e.g. [58], Section 7.7)

$$V(K)^{-1} \succeq W^{-1},$$

from which it follows that

$$\begin{aligned}
&V(K)^{-1} - \Big(2\sum K_\ell + K\Big)^{-1} \\
&\quad \succeq W^{-1} - \Big(2\sum K_\ell + K\Big)^{-1} \\
&\quad \succeq [2\sum K_\ell + K \succeq 2\sum K_\ell \succ 0 \Rightarrow \Big(2\sum K_\ell + K\Big)^{-1} \preceq \Big(2\sum K_\ell\Big)^{-1}] \\
&\quad \succeq W^{-1} - \Big(2\sum K_\ell\Big)^{-1},
\end{aligned}$$

where the latter matrix is positive definite because $0 \prec W \prec 2\sum K_\ell$. This entails that

$$\Big(V(K)^{-1} - \Big(2\sum K_\ell + K\Big)^{-1}\Big)^{-1} \preceq \Big(W^{-1} - \Big(2\sum K_\ell\Big)^{-1}\Big)^{-1},$$

which, by applying the Matrix Inversion Lemma (see [61]), gives

$$V(K) - V(K)\Big(V(K) - 2\sum K_\ell - K\Big)^{-1}V(K) \preceq W - W\Big(W - 2\sum K_\ell\Big)^{-1}W,$$

which is the Lemma statement in view of (2.26).

When $K_i \succeq 0$, since $V(K) \preceq W \prec 2\sum K_\ell$, it holds that

$$0 \prec V(K) + \epsilon I \preceq W + \epsilon I \prec 2\sum K_\ell,$$

for any $\epsilon > 0$ small enough. Repeating the argument above with $V(K) + \epsilon I$ and $W + \epsilon I$ in place of $V(K)$ and W yields

$$\begin{aligned}
&V(K) + \epsilon I - (V(K) + I\epsilon)\Big(V(K) + \epsilon I - 2\sum K_\ell - K\Big)^{-1}(V(K) + \epsilon I) \\
&\qquad \preceq W + \epsilon I - (W + \epsilon I)\Big(W + \epsilon I - 2\sum K_\ell\Big)^{-1}(W + \epsilon I),
\end{aligned}$$

and the sought result is obtained letting $\epsilon \to 0$. □

Consider now

$$W_i = K_i + (4 + 2\gamma_i)K_i\left(\sum_{\ell \neq i} K_\ell\right)^{-1} K_i,$$

as defined in (2.11). By (2.15) it holds that $W_i \prec 2\sum K_\ell$. We now prove that

$$V(K) \preceq W_i, \ \forall K \succeq 0, \tag{2.34}$$

so that, by Lemma 4, it follows that

$$V(K) - V(K)A(K)^{-1}V(K) \preceq W_i - W_i\Big(W_i - 2\sum K_\ell\Big)^{-1}W_i, \ \forall K \succeq 0,$$

which, together with (2.33), yields

$$\mu_i \leq (x^* - v_i)^T \left(W_i - W_i \Big(W_i - 2 \sum K_\ell \Big)^{-1} W_i \right) (x^* - v_i) + h_i = \tilde{\mathbf{q}}_i. \quad (2.35)$$

To prove (2.34), rewrite $V(K)$ defined in (2.25) as

$$V(K) = K_i + 4K_i \left(\sum_{\ell \neq i} K_\ell + K \right)^{-1} K_i + 2K_i^{\frac{1}{2}} \left(K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell + K \right)^{-1} K_i^{\frac{1}{2}} \right)^2 K_i^{\frac{1}{2}}.$$

Since

$$K_i \left(\sum_{\ell \neq i} K_\ell + K \right)^{-1} K_i \preceq K_i \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i,$$

and since

$$\left(K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell + K \right)^{-1} K_i^{\frac{1}{2}} \right)^2 \preceq \gamma_i K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}},$$

because

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell + K \right)^{-1} K_i^{\frac{1}{2}} \preceq K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \preceq \text{[by (2.13)]} \preceq \gamma_i I,$$

it holds that

$$\begin{aligned} V(K) &\preceq K_i + 4K_i \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i + 2\gamma_i K_i \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i \\ &= W_i. \end{aligned}$$

□

2.4.2 Proof of Theorem 3

To ease the notation, let

$$\mathbf{Q}_i(x) := (x - v_i)^T K_i (x - v_i) + h_i = \|A_i x - b_i\|^2, \ i = 1, \ldots, N,$$

and, by symmetry reasons,

$$\mathbf{Q}_{N+1}(x) := (x - v)^T K (x - v) + h = \|Ax - b\|^2.$$

Moreover, define

$$\hat{x}^{[i]} := \arg\min_{x} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{N+1} \mathbf{Q}_\ell(x),\ i = 1, \dots, N+1. \tag{2.36}$$

Observe that, with this notation, $\hat{x}^{[N+1]} = x^*$. Finally, for every $k = 1, \dots, N+1$, let

$$\mathbf{q}^{[k]} := \mathbf{Q}_k(\hat{x}^{[k]})$$

and

$$\mathbf{q}_i^{[k]} := \begin{cases} \mathbf{Q}_i(\hat{x}^{[k]}) & \text{if } i \leq k-1 \\ \mathbf{Q}_{i+1}(\hat{x}^{[k]}) & \text{otherwise.} \end{cases}, \quad i = 1, \dots, N.$$

Note that $\mathbf{q}^{[N+1]} = \mathbf{q}$ and $\mathbf{q}_i^{[N+1]} = \mathbf{q}_i$. As usual, $\mathbf{q}_{(i)}^{[k]}$ denotes the i-th order statistic of $\mathbf{q}_1^{[k]}, \mathbf{q}_2^{[k]}, \dots, \mathbf{q}_N^{[k]}$.
Fix a value for i. We have that

$$\begin{aligned}
&\mathbb{P}_\Delta^{N+1}\{\mathbf{q} \leq \mathbf{q}_{(i)}\} \\
&\quad = \mathbb{P}_\Delta^{N+1}\{\mathbf{q}^{[N+1]} \leq \mathbf{q}_{(i)}^{[N+1]}\} \\
&\quad = [\text{by exchangeability of } (K_1, v_1, h_1), \dots, (K_N, v_N, h_N), (K, v, h)] \\
&\quad = \mathbb{P}_\Delta^{N+1}\{\mathbf{q}^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\}, \quad \forall k = 1, \dots, N+1 \\
&\quad = \frac{1}{N+1} \sum_{k=1}^{N+1} \mathbb{P}_\Delta^{N+1}\{\mathbf{q}^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\} \\
&\quad = [\mathbb{1}\{\cdot\} \text{ indicator function}] \\
&\quad = \frac{1}{N+1} \sum_{k=1}^{N+1} \mathbb{E}_{\Delta^{N+1}}\left[\mathbb{1}\left\{\mathbf{q}^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\right\}\right] \\
&\quad = \frac{1}{N+1} \mathbb{E}_{\Delta^{N+1}}\left[\sum_{k=1}^{N+1} \mathbb{1}\left\{\mathbf{q}^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\right\}\right].
\end{aligned} \tag{2.37}$$

It is a fact that

$$\sum_{k=1}^{N+1} \mathbb{1}\left\{\mathbf{q}^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\right\} \leq i \ \text{ almost surely,} \tag{2.38}$$

so that $\sum_{k=1}^{N+1} \mathbb{1}\left\{\mathbf{q}^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\right\} \leq i$. This latter, plugged in (2.37) gives the theorem statement. To conclude the proof, we now show that (2.38) holds. Fix $(K_1, v_1, h_1), \dots, (K_N, v_N, h_N), (K, v, h)$ such that

$$\mathbf{q}^{[k]} \neq \mathbf{q}_\ell^{[k]} \text{ and } \mathbf{q}_\ell^{[k]} \neq \mathbf{q}_{\ell'}^{[k]},\ \text{for every } \ell, \ell' \in \{1, \dots, N\},\ \ell \neq \ell', \tag{2.39}$$

for every $k \in \{1, \dots, N+1\}$. Note that in view of the proposition assumption, (2.39) holds true with probability 1 by exchangeability of $(K_1, v_1, h_1), \dots, (K_N, v_N, h_N), (K, v, h)$. We show that $\sum_{k=1}^{N+1} \mathbb{1}\left\{\mathbf{q}^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\right\} \leq i$ holds true for the fixed $(K_1, v_1, h_1), \dots, (K_N, v_N, h_N), (K, v, h)$, by exhibiting indexes k_1, $k_2, \dots, k_{N+1-i}$ such that

$$\mathbf{q}^{[k_j]} > \mathbf{q}_{(i)}^{[k_j]}, \quad j = 1, \dots, N+1-i.$$

Define

$$S_k := \sum_{\ell=1}^{N} \mathbf{q}_{\ell}^{[k]}, \quad k = 1, \dots, N+1,$$

and let the $k_1, \dots, k_{N+1-i}$ be the indexes such that

$$S_{k_j} = S_{(j)}, \quad j = 1, \dots, N+1-i.$$

For the purpose of contradiction, assume that there exists a k_j, $j = 1, \dots, N+1-i$, such that

$$\mathbf{q}^{[k_j]} \leq \mathbf{q}_{(i)}^{[k_j]}.$$

Equality can be excluded in view of (2.39), that is, it must hold

$$\mathbf{q}^{[k_j]} < \mathbf{q}_{(i)}^{[k_j]}.$$

Then, by definition of order statistics $\mathbf{q}_{(1)}^{[k_j]}, \dots, \mathbf{q}_{(N)}^{[k_j]}$, we have

$$\mathbf{q}^{[k_j]} < \mathbf{q}_{(i)}^{[k_j]} < \mathbf{q}_{(i+1)}^{[k_j]} < \cdots < \mathbf{q}_{(N)}^{[k_j]}. \tag{2.40}$$

Hence, for any $\tau \in \{i, i+1, \dots, N\}$,

$$\begin{aligned}
S_{k_j} &= S_{k_j} - \mathbf{q}_{(\tau)}^{[k_j]} + \mathbf{q}_{(\tau)}^{[k_j]} \\
&> \text{[by (2.40)]} \\
&> S_{k_j} - \mathbf{q}_{(\tau)}^{[k_j]} + \mathbf{q}^{[k_j]} \\
&= \text{[letting } \rho_\tau \in \{1, \dots, N+1\} \setminus \{k_j\} \text{ be the index such that } \mathbf{Q}_{\rho_\tau}(\hat{x}^{[k_j]}) = \mathbf{q}_{(\tau)}^{[k_j]}\text{]} \\
&= S_{k_j} - \mathbf{Q}_{\rho_\tau}(\hat{x}^{[k_j]}) + \mathbf{Q}_{k_j}(\hat{x}^{[k_j]}) \\
&= \text{[since } S_{k_j} = \sum_{\ell=1}^{N+1} \mathbf{Q}_{\ell}(\hat{x}^{[k_j]}) - \mathbf{Q}_{k_j}(\hat{x}^{[k_j]})\text{]} \\
&= \sum_{\ell=1}^{N+1} \mathbf{Q}_{\ell}(\hat{x}^{[k_j]}) - \mathbf{Q}_{\rho_\tau}(\hat{x}^{[k_j]}) \\
&\geq \min_{x} \left\{ \sum_{\ell=1}^{N+1} \mathbf{Q}_{\ell}(x) - \mathbf{Q}_{\rho_\tau}(x) \right\}
\end{aligned}$$

$$
\begin{aligned}
&= \text{[by (2.36)]} \\
&= \sum_{\ell=1}^{N+1} \mathbf{Q}_\ell(\hat{x}^{[\rho_\tau]}) - \mathbf{Q}_{\rho_\tau}(\hat{x}^{[\rho_\tau]}) \\
&= S_{\rho_\tau},
\end{aligned}
$$

that is, S_{k_j} is greater than $N+1-i$ values among $S_1, \ldots, S_N$. This contradicts the fact that $S_{k_j} = S_{(j)}$, with $j \leq N+1-i$.

□

2.4.3 Asymptotic result

Here we show that $\bar{\mathbf{q}}_{(i)} \xrightarrow[N\to\infty]{} \mathbf{q}_{(i)}$, under suitable assumptions.

Theorem 5 (asymptotic convergence)**.** *Assume that* (K_i, v_i, h_i), $i = 1, 2, \ldots, N$, *are random elements independently and identically distributed according to* $\mathbb{P}_\Delta$, *such that*

$$\mu_K := \mathbb{E}_\Delta[K_1] = \cdots = \mathbb{E}_\Delta[K_N] \succ 0, \tag{2.41}$$

$$\exists \alpha, \bar{\chi} > 0 \text{ such that } \forall \chi > \bar{\chi} \quad \mathbb{P}_\Delta\{\|K_i\| > \chi\} \leq e^{-\alpha\chi},\ i = 1, \ldots, N, \tag{2.42}$$

$$\exists \beta, \bar{\nu} > 0 \text{ such that } \forall \nu > \bar{\nu} \quad \mathbb{P}_\Delta\{\|v_i\| > \nu\} \leq e^{-\beta\nu},\ i = 1, \ldots, N. \tag{2.43}$$

It holds that $\bar{\mathbf{q}}_{(i)} \xrightarrow[N\to\infty]{} \mathbf{q}_{(i)}$ *almost surely, for* $i = 1, \ldots, N$.

⋆

Proof. Condition (2.42) guarantees that the strong law of large numbers (see e.g. Theorem 3, §3, Chapter IV in [3]) applies, so that

$$\frac{\sum_{\ell=1}^N K_\ell}{N} \xrightarrow[N\to\infty]{} \mu_K \text{ almost surely.} \tag{2.44}$$

Since $\mu_K \succ 0$, we have almost surely that $\sum_{\ell=1}^N K_\ell \succ 0$ as well as $\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell \succ 0$ for N large enough. Moreover, for any positive function $f(N)$ and for N large enough, we have also that

$$
\begin{aligned}
\mathbb{P}_\Delta^N \left\{ \frac{1}{f(N)} \max_{i=1,\ldots,N} \|K_i\| > \frac{\ln N^3}{\alpha f(N)} \right\} &= \mathbb{P}_\Delta^N \left\{ \max_{i=1,\ldots,N} \|K_i\| > \frac{\ln N^3}{\alpha} \right\} \\
&\leq N \mathbb{P}_\Delta \left\{ \|K_i\| > \frac{\ln N^3}{\alpha} \right\} \\
&\leq N e^{-\alpha \frac{\ln N^3}{\alpha}} \\
&= \frac{1}{N^2},
\end{aligned}
$$

and, since $\sum_{N=1}^{\infty} \frac{1}{N^2} < \infty$, in view of the Borel-Cantelli Lemma (see e.g. Corollary 2, §10, Chapter II in [3]), we can conclude that:

$$\text{if } \frac{\ln N^3}{\alpha f(N)} \to 0, \text{ then } \frac{1}{f(N)} \max_{i=1,\dots,N} \|K_i\| \to 0 \text{ almost surely.} \tag{2.45}$$

Similarly, using (2.43) in place of (2.42), it can be proved that

$$\text{if } \frac{\ln N^3}{\beta f(N)} \to 0, \text{ then } \frac{1}{f(N)} \max_{i=1,\dots,N} \|v_i\| \to 0 \text{ almost surely.} \tag{2.46}$$

Taking $f(N) = N$, by (2.44) and (2.45), it holds almost surely that

$$\frac{1}{N} K_i \prec \frac{1}{7} \frac{\sum_{\ell=1}^{N} K_\ell}{N}, \ i = 1, \dots, N,$$

for N large enough. Since

$$\frac{1}{N} K_i \prec \frac{1}{7} \frac{\sum_{\ell=1}^{N} K_\ell}{N} \iff K_i \prec \frac{1}{7} \sum_{\ell=1}^{N} K_\ell \iff K_i \prec \frac{1}{6} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell,$$

we have almost surely that $\bar{\mathbf{q}}_i = (x^* - v_i)^T \bar{K}_i (x^* - v_i) + h_i, \ i = 1, \dots, N$, for N large enough, see (2.4). Hence, for each $i = 1, \dots, N$, the following bound holds:

$$\begin{aligned}
|\bar{\mathbf{q}}_i - \mathbf{q}_i| &= |(x^* - v_i)^T \bar{K}_i (x^* - v_i) + h_i - \big((x^* - v_i)^T K_i (x^* - v_i) + h_i\big)| \\
&= |(x^* - v_i)^T (\bar{K}_i - K_i)(x^* - v_i)| \\
&\leq [\text{since } \bar{K}_i = K_i + 6K_i \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell \right)^{-1} K_i] \\
&\leq 6\|x^* - v_i\| \|K_i\| \left\| \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell \right)^{-1} \right\| \|K_i\| \|x^* - v_i\| \\
&= 6 \frac{\|x^* - v_i\|}{N^{\frac{1}{4}}} \frac{\|K_i\|}{N^{\frac{1}{4}}} \left\| \left(\frac{\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell}{N} \right)^{-1} \right\| \frac{\|K_i\|}{N^{\frac{1}{4}}} \frac{\|x^* - v_i\|}{N^{\frac{1}{4}}},
\end{aligned}$$

where the last term tends to zero almost surely in view of (2.44), (2.45) and (2.46), and because $x^* = \left(\sum_{\ell=1}^{N} K_\ell \right)^{-1} \left(\sum_{\ell=1}^{N} K_\ell v_\ell \right)$ converges almost surely. This proves that $\bar{\mathbf{q}}_i \xrightarrow[N \to \infty]{} \mathbf{q}_i$ almost surely and the theorem statement immediately follows. □

2.5 Perspectives for future works

In this chapter we have studied the coverage properties of statistics, close to the empirical costs $\ell(x^*, \delta^{(1)}), \ldots, \ell(x^*, \delta^{(N)})$, that can be used to characterize a least squares solution. We have limited ourselves to proving *mean* coverage properties, while the distributions of the coverages has been studied only a posteriori, through Monte-Carlo experiments. The theoretical characterization of the statistics presented in this chapter as distribution-free (ϵ, β)-coverage statistics is still object of research. For example, an immediate but weak result can be obtained by a direct application of the Markov's inequality (see e.g. [47]), yielding the following inequality

$$\mathbb{P}_{\Delta}^{N}\{\mathcal{C}(\bar{\mathbf{q}}_{(i)}) \geq 1 - \epsilon\} \geq 1 - \frac{1}{\epsilon}\left[\frac{N+1-i}{N+1}\right],$$

entailing e.g. that $\bar{\mathbf{q}}_{(N)}$ is a distribution-free (ϵ, β)-coverage statistic if $N \geq \frac{1}{\epsilon\beta} - 1$, i.e. N scales linearly with $\frac{1}{\epsilon\beta}$. However, the dependence on $\frac{1}{\beta}$ is a particularly noxious fact, since it makes high confidence statements very expensive in terms of number of scenarios. Hopefully, an in-depth studying of the higher moments of the coverage distribution may lead to better results. Another possible way in the quest for a better (ϵ, β)-characterization is more radical and consists in modifying (as little as possible) the statistics themselves according to some suitable scheme: technically, the key point for this purpose is the problem (2.18) at the core of the proof in Section 2.4.

On the other hand, in the following Chapter 3, we will show that a complete characterization of the coverages of the empirical costs $\ell(x^*, \delta^{(1)}), \ldots, \ell(x^*, \delta^{(N)})$ is possible when a worst-case approach is followed.

Chapter 3

On the reliability of data-based min-max decisions

In this and the following chapter, the data-based *worst-case* approach is studied for general convex cost functions. The following Section 3.1 is introductory. In Section 3.2 we offer some background knowledge and state the main results, followed by a discussion. Section 3.3 provides a numerical example, while Section 3.4 is devoted to the proofs. Section 3.5 suggests possible applications and developments of the results here offered and provides a bridge to the next chapter.

3.1 Introduction and problem position

We consider uncertain optimization problems where a decision, modeled as the selection of a variable x belonging to a convex and closed set $\mathcal{X} \subseteq \mathbb{R}^d$, has to be made so as to minimize a cost function $\ell(x, \delta)$, convex in x, that also depends on the uncertainty parameter δ. Precisely, $\ell(x, \delta)$ is real, convex and continuous in x, for each possible δ. The uncertain δ is a random element that takes values in a generic set Δ according to a probability measure $\mathbb{P}_\Delta$.
The decision x^* is made by considering N scenarios, i.e. N instances of δ, say $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, independently generated according to $\mathbb{P}_\Delta$, and minimizing the worst-case cost over these scenarios, that is, by solving:

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \max_{i=1,\ldots,N} \ell(x, \delta^{(i)}). \tag{3.1}$$

The scenario solution x^* can be computed by rewriting (3.1) in epigraphic form as

$$\begin{aligned} \text{EPI}_N : \quad & \min_{c \in \mathbb{R}, x \in \mathcal{X} \subseteq \mathbb{R}^d} c \\ & \text{subject to: } \ell(x, \delta^{(i)}) \leq c, \quad i = 1, \ldots, N, \end{aligned} \tag{3.2}$$

and then by resorting to standard numerical solvers, [62]. See Table 3.1 for some examples of min-max problems arising in various applicative contexts.

Table 3.1. A few examples of min-max problems.

	Interpretation of δ	Interpretation of x	Interpretation of $\ell(x,\delta)$	References
Linear regression theory	Data point	Coefficients of the regression functions	Regression error	[63, 64, 65]
Investment theory	Asset return	Proportion of the assets in a portfolio	Investment loss	[66, 67]
Control theory	Disturbance realization	Controller parameters	Output variance	[68, 30]

As already discussed in Chapter 1, a possible indicator of the quality of the decision x^* is $c^* := \max_{i=1,\dots,N} \ell(x^*, \delta^{(i)})$, i.e. the worst cost among those carried by the seen scenarios. c^*, however, is just an *empirical* quantity and an assessment of the risk that a new uncertainty instance δ carries a cost $\ell(x^*, \delta)$ greater than c^* is needed in order to gain information on the reliability of x^*. Quantitatively, this entails to study the coverage of c^*, or, equivalently, the variable $R := \mathbb{P}_\Delta\{\delta \in \Delta : \ell(x^*, \delta) > c^*\}$, which is called the *risk* associated with c^*. We prefer to focus on the risk of c^*, instead of on its coverage, which is clearly equal to $1 - R$ (see the Remark 2 on page 6), because this is more in line with previous literature, which the theory presented in this and the following chapter builds on. We recall that R is a random variable since it depends on x^* and c^*, which in turn depend on the random sample $\mathsf{D}^N = \delta^{(1)}, \dots, \delta^{(N)}$. A fundamental result in the theory of the scenario approach to convex problems establishes that, irrespective of $\mathbb{P}_\Delta$, the probability distribution function of R is always equal to or bounded by a Beta probability distribution with parameters $d + 1$ and $N - d$ (recall that d is the dimension of the decision variable). Thus, we have that the worst-case cost c^* is a distribution-free coverage statistic in many cases, while in general we have that it is a distribution-free (ϵ, β)-coverage statistic for any N satisfying

$$N \geq \frac{e}{e-1} \frac{1}{\epsilon} \left(d + \ln \frac{1}{\beta} \right). \tag{3.3}$$

Due to the logarithmic dependence of N on β, the statistic c^* is very useful in characterizing x^* with very high confidence, even for relatively small N (clearly, this is true on condition that d is not too large: in Chapter 4 we will deal with this question).

Despite the sharp theoretical result offered by the theory of the scenario approach above mentioned, it may be advisable to study other indicators besides c^*. In particular, we here consider the whole set of costs $\ell(x^*, \delta^{(1)}), \dots, \ell(x^*, \delta^{(N)})$ associated with the various scenarios $\delta^{(1)}, \delta^{(2)}, \dots, \delta^{(N)}$. In the following, these costs, *sorted from largest to smallest*, will be indicated by $c_1^*, c_2^*, \dots, c_N^*$, see Fig. 3.1. With this notation, it always holds that $c_1^* = c^*$.
As is clear from an intuitive point of view, $c_1^*, c_2^*, \dots, c_N^*$ all together provide a more sensible characterization of x^* than by using c^* only, since they provide empirical evidence on how $\ell(x^*, \delta)$ distributes with respect to the variability of δ. Assume for instance that the gap between the maximum cost c^* and the second greatest cost and, similarly, other gaps between costs are large. Then, one expects

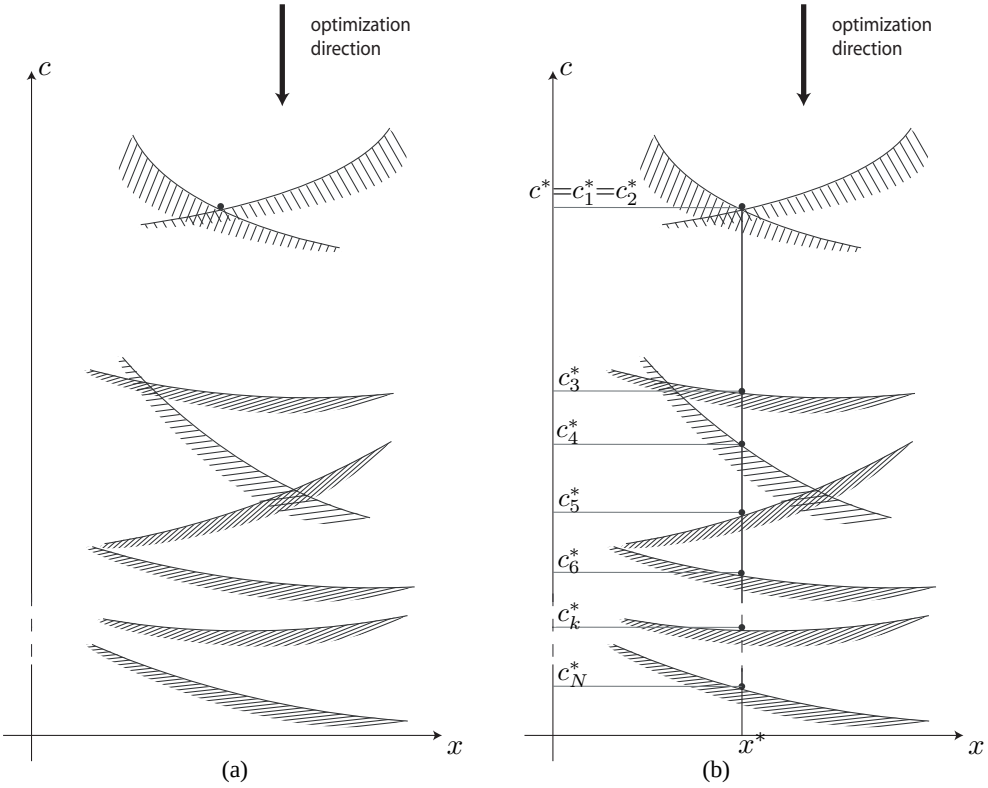


Figure 3.1. On the left, a pictorial representation of the optimization problem (3.2), where each scenario $\delta^{(i)}$ corresponds to a constraint of the form $\ell(x, \delta^{(i)}) \leq c$, here represented with a shaded area. On the right, the costs of x^* are put in evidence.

that a new δ carries a cost $\ell(x^*, \delta)$ significantly smaller than c^* with a high probability. On the contrary, when the values $\ell(x^*, \delta^{(i)})$ concentrate all around c^*, it is apparent that $\ell(x^*, \delta)$ will be almost always close to c^*. A similar idea is followed e.g. in [69], where empirical costs distribution are used for financial decision optimization. In order to put such kind of reasoning on a solid quantitative ground, the risk R_k associated with the costs c_k^*, i.e. the probability to observe an uncertainty instance δ carrying a cost higher than c_k^*, must be evaluated simultaneously for $k = 1, \ldots, N$. However, the existing result applies to the sole c^* and does not provide any characterization of the risks associated with other costs. We fill this gap by studying the *joint probability distribution* of all the risks $R_1, R_2, \ldots, R_N$. Our main achievement is that, no matter what the probability measure $\mathbb{P}_\Delta$ is, the joint probability distribution of $R_{d+1}, R_{d+2}, \ldots, R_N$ is equal to an *ordered Dirichlet distribution* whose parameters depend on the number of scenarios N and the number of decision variables d only. Based on this result, the distribution of the variables $R_1, \ldots, R_N$ can be tightly kept under control, and our conclusions can be employed to support decisions in many real cases even for small sizes of N.
To sum up, two kinds of quantities are central in the characterization of the reliability of x^*:

$$\begin{pmatrix} c_1^* \\ c_2^* \\ \vdots \\ c_N^* \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_N \end{pmatrix},$$

i.e. the vectors of the costs and of the associated risks. While the costs are known as soon as the optimal decision variable x^* is computed, the corresponding risks are hidden to the decision maker. Nevertheless, their joint probability distribution is known (as given by the theory here developed) so that the risks can be kept under control. In particular, since the ordered Dirichlet distribution is thin tailed, the risks can be bounded with high confidence by dropping the tails of the probability distribution. This way, a complete characterization of the reliability of x^* is obtained, and important information about the effective distribution of all the possible uncertain costs $\ell(x^*, \delta)$ is acquired.

In the next Section 3.2, the main result about the risk R of c^* is recalled more in depth, our achievements are formally stated and some relevant aspects are discussed.

3.2 Main results

We first give the formal definition of the costs $c_1^*, \ldots, c_N^*$ and of their risks.

Definition 6 (costs)**.** *We define the* costs *of the optimal decision variable x^* as $c_k^* := \max\left\{c \in \mathbb{R} : \; c \leq \ell(x^*, \delta^{(j)})$ for a choice of k indexes j among $\{1, \ldots, N\}\right\}$, for $k = 1, \ldots, N$.* $\star$

Clearly, $c^* = c_1^* \geq c_2^* \geq \cdots \geq c_N^*$.

Definition 7. *We denote with R_k the risk of the (empirical) cost c_k^* of the optimal decision x^*, formally*

$$R_k := \mathbb{P}_\Delta\{\delta \in \Delta : \ell(x^*, \delta) > c_k^*\}, \quad k = 1, \ldots, N.$$

$\star$

Clearly, c_k^* is a statistic of the data $\mathsf{D}^N = \delta^{(1)}, \ldots, \delta^{(N)}$, having coverage $\mathcal{C}(c_k^*) = 1 - R_k$, and each R_k is a random variable that depends on the random sample D^N through x^* and c_k^*.
Our results are all given under the following assumption.

Assumption 1 (existence and uniqueness)**.** *For every value of N and for every value of $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, the optimal solution to* EPI_N *in* (3.2) *exists and is unique.*

$\star$

This assumption can be relaxed, but we here prefer to maintain it to avoid technical complications.

We now show that the risk of c^* can be studied in the light of the theory of the scenario approach for general constrained convex problems. In particular, in the following, we will reformulate in the present min-max context the main result provided by that theory. For further details on the original result and others related, the reader is referred to Appendix A. First, we need to formulate in the present context the definition of *support scenario* and of *fully-supported* problem.

Definition 8 (support scenario)**.** *For given scenarios $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, the scenario $\delta^{(r)}$, $r \in \{1, \ldots, N\}$, is called a support scenario for the min-max problem* (3.1) *if its removal changes the solution of* EPI_N *in* (3.2)*.*

$\star$

Loosely speaking, support scenarios are those corresponding to the uppermost cost functions, preventing the solution from moving to any improving direction. The number of support scenarios can be bounded a-priori. Indeed, for every value of $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, the number of support scenarios for the min-max problem (3.1) is at most $d + 1$ (see Proposition 2 in the Appendix A). We say that the min-max problem is *fully-supported* if, for all $N \geq d + 1$, with probability one with respect to the possible $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, it has exactly $d + 1$ support scenarios.

Now, consider, for any given pair (x, c), $x \in \mathbb{R}^d$ and $c \in \mathbb{R}$, the function defined as follows

$$V(x, c) := \mathbb{P}_\Delta\{\delta \in \Delta : \ell(x, \delta) > c\}.$$

According to the scenario approach terminology, with reference to problem (3.2), $V(x, c)$ is the *violation probability* of the pair $z = (x, c)$. With this notation, the

risk R of the worst empirical cost c^* corresponding to the optimal decision x^* is given by

$$R = V(x^*, c^*),$$

i.e. R is the violation of the optimal solution $z^* = (x^*, c^*)$ to the problem (3.2). The main result recalled in the Appendix A (Theorem 12) deals with $V(x^*, c^*)$ and, in our context, boils down to the fact that, whenever the min-max problem (3.1) is fully-supported, the equality

$$\mathbb{P}_\Delta^N\{R \leq \epsilon\} = 1 - \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \tag{3.4}$$

holds true, that is, the probability distribution function of R is equal to a Beta with parameters $(d+1, N-d)$ independently of $\mathbb{P}_\Delta$ and of the specific problem considered. For non fully-supported problems, the result holds as a bound (Theorem 13):

$$\mathbb{P}_\Delta^N\{R \leq \epsilon\} \geq 1 - \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}. \tag{3.5}$$

In the rest of this chapter, we will show that, by a slight specialization of the fully-supportedness assumption, a broader result than (3.4) holds, and the whole *joint* probability distribution function of the risks associated with all the costs can be exactly computed without relying on the knowledge of $\mathbb{P}_\Delta$.

Assumption 2 (specialized fully-supportedness)**.** *Let consider the min-max problem* (3.1) *for all $N \geq d+1$. With probability one with respect to the possible data samples D^N, it holds that:*

i) it has exactly $d+1$ support scenarios;

ii) for every $\gamma \in \mathbb{R}$, $\mathbb{P}_\Delta\{\ell(x^*, \delta) = \gamma\} = 0$.

$\star$

Point $i)$ of Assumption 2 is the classic fully-supportedness assumption (see Assumption 3 in Appendix A) and, since, by definition, the support scenarios carry the same cost c^*, it implies that the first $d+1$ costs are equal, i.e. $c_1^* = c_2^* = \cdots = c_{d+1}^* = c^*$ (see Fig. 3.1). Since $c_1^* = c_2^* = \cdots = c_{d+1}^*$, the associated risks $R_1, R_2, \ldots, R_{d+1}$ are equal too. Point $ii)$ instead is a non-degeneracy condition (satisfied in many practical problems) asking that the possible values of the cost function at x^*, conditionally on the data D^N, do not accumulate over the same point. It is easy to show that, when $ii)$ is satisfied, the remaining costs $c_{d+1}^*, c_{d+2}^*, \ldots, c_N^*$ are all different from one another.

Before giving Theorem 6, we recall that the *ordered $(N-d)$-variate Dirichlet distribution* is the probability distribution having density function

$$p(\nu_{d+1}, \nu_{d+2}, \ldots, \nu_N) = \frac{N!}{d!} \nu_{d+1}^d \mathbb{1}\{0 \leq \nu_{d+1} \leq \nu_{d+2} \leq \cdots \leq \nu_N \leq 1\},$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, see e.g. [51], page 182. The cumulative distribution function of the ordered $(N-d)$-variate Dirichlet distribution will be denoted by $\mathrm{CDF}_d(\eta_{d+1},\dots,\eta_N)$, i.e.

$$\mathrm{CDF}_d(\eta_{d+1},\dots,\eta_N) = \frac{N!}{d!}\int_0^{\eta_{d+1}} \nu_{d+1}^d \int_0^{\eta_{d+2}} \cdots \int_0^{\eta_N} \mathbb{1}\{0 \le \nu_{d+1} \le \cdots \le \nu_N \le 1\} \mathrm{d}\nu_N \cdots \mathrm{d}\nu_{d+2}\mathrm{d}\nu_{d+1}. \tag{3.6}$$

See Section 3.2.3 for additional information about Dirichlet distributions.

Theorem 6. *Under Assumptions 1 and 2, the joint probability distribution function of $R_{d+1},\dots,R_N$ is as follows:*

$$\mathbb{P}_\Delta^N\{R_{d+1} \le \epsilon_{d+1}, R_{d+2} \le \epsilon_{d+2}, \dots, R_N \le \epsilon_N\} = \mathrm{CDF}_d(\epsilon_{d+1},\epsilon_{d+2},\dots,\epsilon_N), \tag{3.7}$$

so that

$$\begin{aligned}\mathbb{P}_\Delta^N\{&R_1 \le \epsilon_1, R_2 \le \epsilon_2, \dots, R_{d+1} \le \epsilon_{d+1}, R_{d+2} \le \epsilon_{d+2}, \dots, R_N \le \epsilon_N\} \\ &= \mathrm{CDF}_d(\underline{\epsilon}, \epsilon_{d+2}, \dots, \epsilon_N),\end{aligned}$$

where $\underline{\epsilon} := \min\{\epsilon_1, \epsilon_2 \dots, \epsilon_{d+1}\}$.

$\star$

Proof. See Section 3.4.

Theorem 6 states that for the class of problems satisfying Assumptions 1 and 2, the risks $R_1,\dots,R_N$ are pivotal quantities, since their joint probability distribution function is the same independently of the specific problem at hand and, in particular, independently of the probability measure $\mathbb{P}_\Delta$. It is well known, see e.g. [70], that the marginal distributions of an ordered Dirichlet distribution are Beta distributions. Hence, it can be inferred that the probability distribution function of R_k is a Beta with parameters $(k, N-k+1)$, $k = d+1,\dots,N$, that is,

$$\mathbb{P}_\Delta^N\{R_k \le \epsilon\} = 1 - \sum_{i=0}^{k-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}. \tag{3.8}$$

Specializing (3.8) for $k = d+1$ and recalling that $R_{d+1} = R$ because $c_{d+1}^* = c^*$, we have that

$$\mathbb{P}_\Delta^N\{R \le \epsilon\} = \mathbb{P}_\Delta^N\{R_{d+1} \le \epsilon\} = 1 - \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i},$$

i.e. the result (3.4) is recovered from Theorem 6.
The class of problems satisfying Assumptions 1 and 2 is neither empty nor "pathological". Notable examples, like the following one, arise in min-max linear regression:

$$\min_{(x_1,\dots,x_d)\in\mathbb{R}^d} \max_{i=1,\dots,N} \left| y_i - \left[x_1 + \theta^{(i)} x_2 + \left(\theta^{(i)}\right)^2 x_3 + \cdots + \left(\theta^{(i)}\right)^{d-1} x_d \right] \right|,$$

where $N \geq d+1$, and the points $(\theta^{(i)}, y^{(i)}) = \delta^{(i)}$, $i = 1, \ldots, N$, are sampled from $\Delta = \mathbb{R}^2$ according to a probability $\mathbb{P}_\Delta$ that admits a density, see [64]. As is clear, however, specialized fully-supported problems does not cover the whole realm of problems encountered in the practice of optimization.
Remarkably, Theorem 6 is just a corollary of a more general result that continues to hold even when Assumption $2.i$ is dropped and the sole non-degeneracy condition $2.ii$ is preserved.

Theorem 7. *Under Assumptions 1 and $2.ii$, the joint probability distribution function of $R_{d+1}, \ldots, R_N$ is as follows:*

$$\mathbb{P}_\Delta^N\{R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} = \mathrm{CDF}_d(\epsilon_{d+1}, \epsilon_{d+2}, \ldots, \epsilon_N). \tag{3.9}$$

$\star$

Proof. See Section 3.4.1.

Although equation (3.9) of Theorem 7 and equation (3.7) of Theorem 6 are formally the same, the conveyed information is different because, without Assumption $2.i$, it is no longer true that $c^* = c_1^* = c_2^* = \cdots = c_{d+1}^*$, and (3.9) does not determine the probability distribution of all the risks including the first d. In fact, under Assumptions 1 and $2.ii$ only, the distribution of $R_1, \ldots, R_d$ is intrinsically problem-dependent.
Because of (3.9), the marginal distribution of R_k, $k = d+1, \ldots, N$, is still a Beta as in (3.8). Under the assumptions of Theorem 7, we can only conclude that $c^* \geq c_{d+1}^*$, so that $R \leq R_{d+1}$, entailing that the probability distribution of R is dominated by that of R_{d+1}, that is

$$\mathbb{P}_\Delta^N\{R \leq \epsilon\} \geq \mathbb{P}_\Delta^N\{R_{d+1} \leq \epsilon\} = 1 - \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}.$$

Hence,

$$\mathbb{P}_\Delta^N\{R \leq \epsilon\} \geq 1 - \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$$

and the inequality (3.5) for general (i.e. not necessarily fully-supported) problems is recovered. Furthermore, our study has thus shown that the right-hand side of (3.5) is the *exact* probability distribution of the risk associated with a cost lower than c^*. Moreover, by observing that necessarily

$$R_1 \leq \cdots \leq R_d \leq R_{d+1},$$

because $c_1^* \geq c_2^* \geq \cdots \geq c_{d+1}^*$, we have that $R_{d+1} \leq \epsilon$ implies not only $R \leq \epsilon$, but also $R_i \leq \epsilon$, $\forall i \leq d$, and the joint probability distribution function of *all* the

risks $R_1, \ldots, R_{d+1}, \ldots, R_N$ including the first d can be bounded as follows:

$$\begin{aligned}
&\mathbb{P}_\Delta^N\{R_1 \leq \epsilon_1, \ldots, R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} \\
&\geq [\text{let } \underline{\epsilon} := \min\{\epsilon_1, \epsilon_2 \ldots, \epsilon_{d+1}\}] \\
&\geq \mathbb{P}_\Delta^N\{R_1 \leq \underline{\epsilon}, \ldots, R_{d+1} \leq \underline{\epsilon}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} \\
&= \mathbb{P}_\Delta^N\{R_{d+1} \leq \underline{\epsilon}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} \\
&= [\text{by using (3.9)}] \\
&= \mathrm{CDF}_d(\underline{\epsilon}, \epsilon_{d+2}, \ldots, \epsilon_N). \qquad (3.10)
\end{aligned}$$

This conclusion is formally stated in the following corollary.

Corollary 1. *Under Assumption 1 and 2.ii, the joint probability distribution function of the risks $R_1, \ldots, R_N$ is lower bounded by $\mathrm{CDF}_d(\underline{\epsilon}, \epsilon_{d+2}, \ldots, \epsilon_N)$, i.e.*

$$\begin{aligned}
&\mathbb{P}_\Delta^N\{R_1 \leq \epsilon_1, \ldots, R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} \\
&\qquad \geq \mathrm{CDF}_d(\underline{\epsilon}, \epsilon_{d+2}, \ldots, \epsilon_N), \qquad (3.11)
\end{aligned}$$

where $\underline{\epsilon} := \min\{\epsilon_1, \epsilon_2 \ldots, \epsilon_{d+1}\}$.

$\star$

Clearly, bound (3.11) is tight (i.e. cannot be improved without introducing additional assumptions) since it holds with equality for problems satisfying also Assumption 2.i.

3.2.1 Relaxing the non-degeneracy assumption

The non-degeneracy Assumption 2.ii is strictly required for the equalities in Theorems 6 and 7 to hold true. Indeed, if for example the probability measure $\mathbb{P}_\Delta$ is concentrated on a unique scenario $\bar{\delta}$, the costs $c_1^*, c_2^*, \ldots, c_N^*$ collapse to the same value $c^* = c_1^* = c_2^* = \cdots = c_N^*$ having zero risk and $\mathbb{P}_\Delta\{R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} = 1$. In this case, though (3.7) and (3.9) are violated, the distribution of the risks $R_{d+1}, R_{d+2}, \ldots, R_N$ is still trivially dominated by the ordered Dirichlet distribution. Actually, it is a general fact that if the non-degeneracy assumption is dropped, then the cumulative probability distribution function of the risks remains lower bounded by the ordered Dirichlet cumulative distribution function, as formally stated in the next theorem.

Theorem 8. *Under Assumption 1 only, the joint probability distribution function of $R_{d+1}, \ldots, R_N$ is lower bounded by $\mathrm{CDF}_d(\epsilon_{d+1}, \ldots, \epsilon_N)$, i.e.*

$$\mathbb{P}_\Delta^N\{R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, \ldots, R_N \leq \epsilon_N\} \geq \mathrm{CDF}_d(\epsilon_{d+1}, \epsilon_{d+2}, \ldots, \epsilon_N). \qquad (3.12)$$

$\star$

Proof. See Section 3.4.3.

Mimicking (3.10), the following corollary is easily obtained.

Corollary 2. *Under Assumption 1 only, the joint probability distribution function of $R_1, \ldots, R_N$ is lower bounded as follows*

$$\mathbb{P}_\Delta^N\{R_1 \leq \epsilon_1, \ldots, R_{d+1} \leq \epsilon_{d+1}, \ldots, R_N \leq \epsilon_N\} \geq \mathrm{CDF}_d(\underline{\epsilon}, \epsilon_{d+2}, \ldots, \epsilon_N),$$

where $\underline{\epsilon} := \min\{\epsilon_1, \epsilon_2 \ldots, \epsilon_{d+1}\}$.

$\star$

3.2.2 Practical use of the theoretical results

The theory developed above can be applied in various ways. The following two are especially useful in contexts where the uncertainty instances $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$ come as observations obtained from a data acquisition experiment.

Post-experiment analysis
The decision-maker has collected N scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$ and has solved the min-max problem (3.1) obtaining x^* and the corresponding costs c_k^*, $k = 1, \ldots, N$. He fixes a confidence parameter $\beta \in (0, 1)$ to a very small value, e.g. $\beta = 10^{-5}$ or $\beta = 10^{-7}$, and determines $\epsilon_{d+1}, \ldots, \epsilon_N$ such that $\mathrm{CDF}_d(\epsilon_{d+1}, \ldots, \epsilon_N)$ is bigger than or equal to $1 - \beta$. By appealing to Corollary 2, the decision-maker can claim with high confidence $1 - \beta$ that, simultaneously for $k = 1, \ldots, N$, the risk R_k of each cost c_k^* is no larger than the respective ϵ_k (taking $\epsilon_k = \epsilon_{d+1}$ when $k < d + 1$).

Experiment design
The decision-maker fixes a very small $\beta \in (0, 1)$, e.g. $\beta = 10^{-5}$ or $\beta = 10^{-7}$. Then he fixes the desired upper bounds on the risks of the first m costs, that is a vector of m increasing elements, $0 \leq \epsilon_1 \leq \epsilon_2 \leq \cdots \leq \epsilon_m \leq 1$. By letting $\epsilon_h = 1$ for $h > m$, he computes the minimum number N of scenarios guaranteeing that $\mathrm{CDF}_d(\epsilon_1, \epsilon_{d+2}, \ldots, \epsilon_N)$ is no less than $1 - \beta$. If N instances of δ are indeed observed and the min-max problem is solved, then, in the light of Corollary 2, the obtained x^* and the corresponding costs are such that $R_k \leq \epsilon_k$, $k = 1, \ldots, N$, simultaneously with high confidence $1 - \beta$.

In both cases, the decision-maker can link the solution x^* and the costs c_k^*'s obtained through the optimization procedure to the values ϵ_k's that limit the corresponding risks R_k's.
Now, let us consider the cumulative distribution function of the cost $\ell(x^*, \delta)$ incurred at the optimal solution x^*, defined as $F_\ell(c) := \mathbb{P}_\Delta\{\delta \in \Delta : \ell(x^*, \delta) \leq c\}$. Interestingly enough, the risks give us a lot of information about $F_\ell(c)$ because, by

the definition of risk, we have

$$\begin{aligned} R_k \leq \epsilon_k &\iff 1 - R_k \geq 1 - \epsilon_k \\ &\iff \mathbb{P}_\Delta\{\ell(x^*, \delta) \leq c_k^*\} = F_\ell(c_k^*) \geq 1 - \epsilon_k, \end{aligned}$$

so that, with confidence $1 - \beta$, we have also that

$$F_\ell(c_k^*) \geq 1 - \epsilon_k, \quad \text{for all } k = 1, \ldots, N. \tag{3.13}$$

Moreover, by observing that $F_\ell(c)$ is monotonic, (3.13) implies that

$$F_\ell(c) \geq \begin{cases} 1 - \epsilon_1 & \text{if } c \geq c_1^* \\ 1 - \epsilon_k & \text{if } c_k^* \leq c < c_{k-1}^*,\ k = 2, \ldots, N \\ 0 & \text{if } c < c_N^* \end{cases}$$

i.e., we have found a step function that, with confidence $1 - \beta$, lower bounds the cumulative distribution function of the cost $\ell(x^*, \delta)$. This provides strong knowledge on the performance of the decision variable x^* without any further sampling effort.

This result can be further refined in many situations, that is, when the assumptions of Theorem 7 are known to be satisfied. Theorem 7, indeed, provides the exact distribution of the risks $R_{d+1}, \ldots, R_N$, thus allowing the decision-maker to compute *two-sided* confidence intervals for $R_1, \ldots, R_N$, i.e. it is possible to compute $\bar{\epsilon}_1, \bar{\epsilon}_2, \ldots, \bar{\epsilon}_N$ (with $\bar{\epsilon}_1 = \bar{\epsilon}_2 = \ldots = \bar{\epsilon}_d = \bar{\epsilon}_{d+1}$) and $\underline{\epsilon}_1, \underline{\epsilon}_2, \ldots, \underline{\epsilon}_N$ (with $\underline{\epsilon}_1 = \underline{\epsilon}_2 = \ldots = \underline{\epsilon}_d = 0$) so that $R_k \in [\underline{\epsilon}_k, \bar{\epsilon}_k]$ simultaneously for $k = 1, \ldots, N$ with confidence $1 - \beta$. This is equivalent to building the "probability box"

$$F_\ell(c) \geq \begin{cases} 1 - \bar{\epsilon}_1 & \text{if } c \geq c_1^* \\ 1 - \bar{\epsilon}_k & \text{if } c_k^* \leq c < c_{k-1}^*,\ k = 2, \ldots, N \\ 0 & \text{if } c < c_N^* \end{cases}$$

and

$$F_\ell(c) \leq \begin{cases} 1 & \text{if } c > c_1^* \\ 1 - \underline{\epsilon}_k & \text{if } c_{k+1}^* < c \leq c_k^*,\ k = 1, \ldots, N-1 \\ 1 - \underline{\epsilon}_N & \text{if } c \leq c_N^* \end{cases} \tag{3.14}$$

enveloping the cumulative distribution function of $\ell(x^*, \delta)$ with confidence at least $1 - \beta$, see Fig. 3.2. See [71] for a discussion of "probability boxes" and their usefulness in risk-evaluation problems.

3.2.3 Some useful properties

In this section, we highlight some properties of the probability distribution function of the risks $R_{d+1}, \ldots, R_N$ as given by equation (3.9) of Theorem 7 that may be useful in practice.

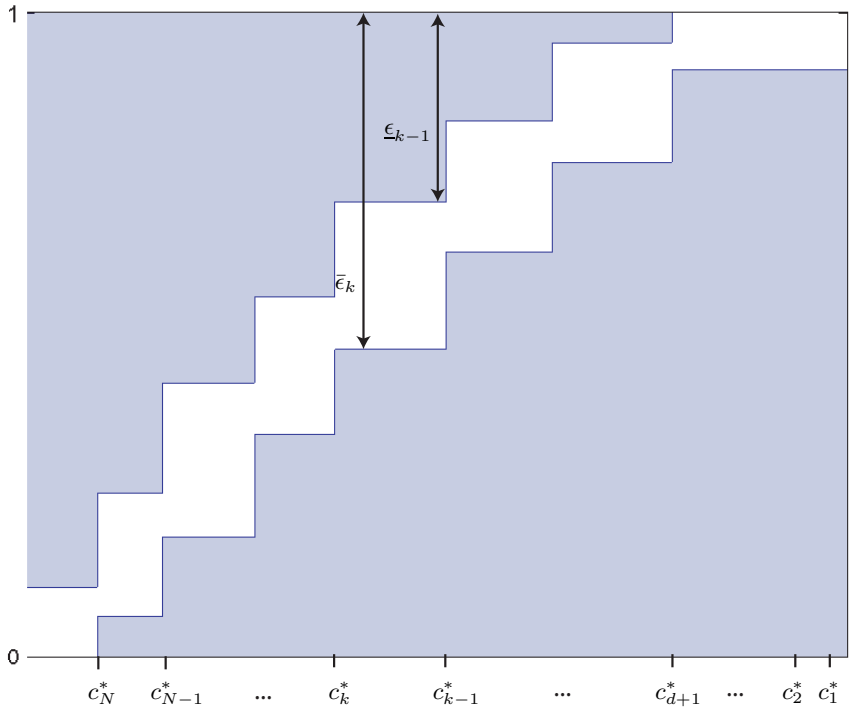


Figure 3.2. A "probability box" for the cumulative distribution function of the cost, $F_\ell(c) = \mathbb{P}_\Delta\{\ell(x^*, \delta) \leq c\}$. The graph of $F_\ell(c)$ is within the white area with confidence $1 - \beta$. The box is built based on the empirical costs c_k^* and the values $\bar{\epsilon}_k, \underline{\epsilon}_k$ that limit the corresponding risks $R_k = 1 - F_\ell(c_k^*)$. The probability box in this figure is a stylized representation, for a real instance see Fig. 3.6 in Section 3.3.

Comments on Dirichlet distributions

Equation (3.9) states that the random vector $R_{d+1}, R_{d+2}, \ldots, R_N$ is distributed according to the $(N-d)$*-variate ordered Dirichlet distribution function* $\mathrm{CDF}_d(\epsilon_{d+1}, \ldots, \epsilon_N)$. By applying the following transformation to the random variables R_k's

$$\begin{aligned} D_N &= 1 - R_N \\ D_{N-1} &= R_N - R_{N-1} \\ &\vdots \\ D_{d+1} &= R_{d+2} - R_{d+1} \end{aligned}$$

the vector $D_{d+1}, D_{d+2}, \ldots, D_N$ is obtained, which is distributed according to the so-called *Dirichlet distribution*, [51, 72]. Hence, the evaluation of an ordered Dirichlet distribution function can be converted to the problem of evaluating a Dirichlet distribution function. The reader is referred to [73, 74, 75, 76] and references therein for studies on computational issues about Dirichlet distributions.

Beta distributions as marginals

We have already observed that the marginal probability distribution function of R_k is a Beta with parameters $(k, N-k+1)$, for each $k = d+1, \dots, N$, see (3.8). Notably, the right-hand side of (3.8) can be easily evaluated by means of common tools, like the betainc function in MATLAB, [77], or pbeta in R, [78]. Such Beta distributions have known expected values, precisely:

$$\mathbb{E}_\Delta[R_k] = \frac{k}{N+1}, \quad k = d+1, \dots, N,$$

hence, c_k^* is a distribution free $\frac{N+1-k}{N+1}$-mean coverage statistic, satisfying (by the same reasoning as in (1.2))

$$\mathbb{P}_\Delta^{N+1}\{\ell(x^*, \delta) \le c_k^*\} = \frac{N+1-k}{N+1}, \quad k = d+1, \dots, N.$$

As is clear, a lower bound for the joint distribution function of $R_{d+1}, \dots, R_N$ is given by the sum of the marginals, i.e.

$$\begin{aligned} &\mathbb{P}_\Delta^N\{R_{d+1} \le \epsilon_{d+1}, \dots, R_N \le \epsilon_N\} \\ &\ge 1 - \sum_{k=d+1}^{N} \mathbb{P}_\Delta^N\{R_k > \epsilon_k\} \\ &= 1 - \sum_{k=d+1}^{N} \sum_{i=0}^{k-1} \binom{N}{i} \epsilon_k^i (1-\epsilon_k)^{N-i}, \end{aligned} \tag{3.15}$$

and (3.15) may indeed be an acceptable approximation of $\mathrm{CDF}_d(\epsilon_{d+1}, \dots, \epsilon_N)$ in some practical cases (see also Section 3.3).

Based on (3.15), we show in the following that for a given $\beta \in (0,1)$, if

$$N \ge \max_{k=d+1,\dots,N} N^{(k)}, \tag{3.16}$$

where

$$N^{(k)} := \left\lfloor \frac{2}{\epsilon_k}\left(k + \ln\frac{1}{\beta}\right) + \frac{4}{\epsilon_k}\ln\left(\frac{2}{\epsilon_k}\left(k + \ln\frac{1}{\beta}\right)\right)\right\rfloor + 1$$

($\lfloor\cdot\rfloor$ denotes integer part), then $\mathbb{P}_\Delta^N\{R_{d+1} \le \epsilon_{d+1}, \dots, R_N \le \epsilon_N\} \ge 1-\beta$, i.e. conditions $R_k \le \epsilon_k$, $k = d+1, \dots, N$, hold simultaneously with high confidence $1-\beta$. Although (3.16) may be loose, it reveals the logarithmic dependence of N on β by which it is possible to enforce *very* high confidence without affecting too much the sampling effort.

Proof. The fact that (3.16) entails $\mathbb{P}_\Delta^N\{R_{d+1} \le \epsilon_{d+1}, \dots, R_N \le \epsilon_N\} \ge 1-\beta$ is now proved by following almost verbatim the proof in Appendix B of [79], which is in a context different from our own but involves the same mathematical steps.

Note that, by (3.16) and the definition of $N^{(k)}$, for every $k = d+1, \ldots, N$ we have

$$\begin{aligned}
N \geq & \frac{2}{\epsilon_k}\left(k + \ln\frac{1}{\beta}\right) + \frac{4}{\epsilon_k}\ln\left(\frac{2}{\epsilon_k}\left(k + \ln\frac{1}{\beta}\right)\right) \\
= & [\text{letting } a_k = k + \ln\frac{1}{\beta}] \\
= & \frac{2}{\epsilon_k}a_k + \frac{2}{\epsilon_k}\cdot 2\cdot\ln\left(\frac{2a_k}{\epsilon_k}\right) \\
\geq & [\text{since } 2 \geq \frac{a_k}{a_k - 1}] \\
\geq & \frac{2}{\epsilon_k}a_k + \frac{2}{\epsilon_k}\cdot\frac{a_k}{a_k-1}\cdot\ln\left(\frac{2a_k}{\epsilon_k}\right) \\
= & \frac{2}{\epsilon_k}\frac{a_k}{a_k-1}\left(a_k - 1 + \ln\left(\frac{2a_k}{\epsilon_k}\right)\right) \\
= & \frac{1}{\frac{\epsilon_k}{2} - \frac{1}{\frac{2a_k}{\epsilon_k}}}\left(a_k - 1 + \ln\left(\frac{2a_k}{\epsilon_k}\right)\right),
\end{aligned}$$

so that

$$\begin{aligned}
\frac{\epsilon_k}{2}\cdot N \geq & \left(a_k - 1 + \ln\left(\frac{2a_k}{\epsilon_k}\right)\right) + \frac{1}{\frac{2a_k}{\epsilon_k}}\cdot N \\
\geq & [\text{since } 1 - \frac{2a_k}{\epsilon_k} \leq 0] \\
\geq & a_k - 1 + \ln\left(\frac{2a_k}{\epsilon_k}\right) + \frac{1}{\frac{2a_k}{\epsilon_k}}\cdot\left(N + 1 - \frac{2a_k}{\epsilon_k}\right) \\
\geq & [\text{since } \ln x + \frac{1}{x}(y-x) \geq \ln y \text{ (by the concavity of } \ln x)] \\
\geq & a_k - 1 + \ln(N+1) \\
= & k - 1 + \ln\frac{1}{\beta} + \ln(N+1).
\end{aligned}$$

Hence,

$$\frac{\epsilon_k N}{2} - (k-1) \geq \ln\frac{N+1}{\beta},$$

and, by observing that $\frac{(\epsilon_k N - (k-1))^2}{2\epsilon_k N} \geq \frac{\epsilon_k N}{2} - (k-1)$, we have

$$e^{-\frac{(\epsilon_k N - (k-1))^2}{2\epsilon_k N}} \leq \frac{\beta}{N+1}.$$

Since $N > \frac{k}{\epsilon_k}$, by the Chernoff's bound (see e.g. [43], Chapter 2, Section 3), it holds that

$$\sum_{i=0}^{k-1}\binom{N}{i}\epsilon_k^i(1-\epsilon_k)^{N-i} \leq e^{-\frac{(\epsilon_k N - (k-1))^2}{2\epsilon_k N}},$$

and, by recalling that

$$\mathbb{P}_{\Delta}^{N}\{R_k > \epsilon_k\} = \sum_{i=0}^{k-1} \binom{N}{i} \epsilon_k^i (1-\epsilon_k)^{N-i},$$

we conclude that

$$\mathbb{P}_{\Delta}^{N}\{R_k > \epsilon_k\} \leq \frac{\beta}{N+1} \leq \frac{\beta}{N-d}.$$

From (3.15) it follows that $\mathbb{P}_{\Delta}^{N}\{R_{d+1} \leq \epsilon_{d+1}, \ldots, R_N \leq \epsilon_N\} \geq 1-\beta$. □

Connection with order statistics

Consider the sampling of N random variables, uniformly and independently distributed in $[0,1]$, and sort them in order of magnitude,

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)},$$

$X_{(i)}$ being the $i-$th smallest value, i.e. the $i-$th order statistic. It is well known, [51, 52], that order statistics have joint ordered Dirichlet distribution with unitary parameters, that is $\mathbb{P}_{\Delta}\{X_{(1)} \leq \epsilon_1, X_{(2)} \leq \epsilon_2, \ldots, X_{(d+1)} \leq \epsilon_{d+1}, \ldots, X_{(N)} \leq \epsilon_N\}$ can be expressed as

$$N! \int_0^{\epsilon_1} \int_0^{\epsilon_2} \cdots \int_0^{\epsilon_N} \mathbb{1}\{0 \leq x_1 \leq \cdots \leq x_N \leq 1\} \mathrm{d}x_N \cdots \mathrm{d}x_2 \mathrm{d}x_1. \qquad (3.17)$$

If $\epsilon_1 = \epsilon_2 = \cdots = \epsilon_{d+1}$, then, by integrating with respect to the first $d+1$ components, (3.17) becomes

$$\frac{N!}{d!} \int_0^{\epsilon_{d+1}} x_{d+1}^d \int_0^{\epsilon_{d+2}} \cdots \int_0^{\epsilon_N} \mathbb{1}\{0 \leq x_{d+1} \leq \cdots \leq x_N \leq 1\} \mathrm{d}x_N \cdots \mathrm{d}x_{d+2} \mathrm{d}x_{d+1},$$

which is exactly $\mathrm{CDF}_d(\epsilon_{d+1}, \ldots, \epsilon_N)$. In short, the computation of $\mathrm{CDF}_d(\epsilon_{d+1}, \ldots, \epsilon_N)$ can be reduced to the well known problem of computing the joint cumulative distribution function of order statistics, see e.g. [73, 75]. The freely distributed package μtoss for R, [80, 81, 78], contains the function jointCDF.orderedUnif, which computes (3.17), though, because of numerical issues, it is reliable for $N \leq 100$ only.

Computability through Monte-Carlo methods

By virtue of the analogy with the distribution of order statistics, even Monte-Carlo methods can be employed to evaluate $\mathrm{CDF}_d(\epsilon_{d+1}, \ldots, \epsilon_N)$. Indeed, one can repeat a large number of times, say M times, the following steps (C is a counter initially set to 0):

- draw a sequence of N independent samples from a uniform distribution in $[0,1]$;
- sort the sequence, i.e. compute all the order statistics from $X_{(1)}$ (the smallest value) to $X_{(N)}$ (the largest);

- evaluate the condition $X_{(i)} \leq \epsilon_i$ for $i = d+1, \ldots, N$, and increment the counter C by 1 if it is satisfied for every value of the index i.

Then, $\hat{P} := \frac{C}{M}$ is an estimate of the sought probability $P := \mathrm{CDF}_d(\epsilon_{d+1}, \ldots, \epsilon_N)$. $\hat{P}$ and P are related by the Hoeffding's inequality (see [82, 83]), which guarantees that $P \geq \hat{P} - \gamma$ holds with confidence $1 - \eta$ (e.g. $\eta = 10^{-6}$) as long as the number of experiments is large enough (precisely, as long as $M \geq \frac{1}{2\gamma^2} \ln \frac{2}{\eta}$). This method becomes increasingly impractical as γ gets smaller, and more advanced randomized schemes must be considered if lowering γ under 10^{-4} is needed.

3.3 An application to audio equalization

In this section, we shall employ the main results of Section 3.2 in the characterization of the solution to an *equalizer design* problem, [84].

3.3.1 Problem formulation

In a digital communication system, [85, 86], a signal $u(t)$, $t = 0, \pm 1, \pm 2, \ldots$, is sent from a *transmitter* to a *receiver* through a communication *channel* C, see Fig. 3.3(a). In general, the signal at the receiver end, say $\tilde{u}(t)$, is different from the transmitted signal owing to the distortion introduced by the channel. This latter, indeed, acts approximately as a linear frequency filter and is completely characterized by its frequency response $C(\omega)$, which is a complex-valued function of $\omega \in [-\pi, \pi]$ linking the Fourier transform of $u(t)$, say $U(\omega)$, to the Fourier transform of $\tilde{u}(t)$, say $\tilde{U}(\omega)$, according to the equation $\tilde{U}(\omega) = C(\omega)U(\omega)$. If the distortion introduced by the channel is unacceptably high, a device E called *equalizer* can be added at the receiver end to improve the quality of the transmission, see Fig. 3.3(b).

The equalizer E is a frequency filter too, whose frequency response is denoted by $E(\omega)$. In particular, we consider a so-called d-tap FIR (Finite Impulse Response) equalizer:

$$E(\omega) = \sum_{k=0}^{d-1} x_k e^{-ik\omega}, \tag{3.18}$$

where i is the imaginary unit and $x_0, x_1, \ldots, x_{d-1}$ are real parameters through which the frequency response can be shaped. Here, we restrict to the case $d = 10$. Overall, the frequency response of the equalized channel in Fig. 3.3(b) linking $U(\omega)$ and $\tilde{U}(\omega)$ turns out to be the product $C(\omega)E(\omega)$, and the aim is to design the equalizer E by choosing the vector x of the parameters $x_0, x_1, \ldots, x_{d-1}$ so as to make $C(\omega)E(\omega)$ as similar as possible to a *desired frequency response*. Clearly, this can be cast as an optimization problem where the dissimilarity between $C(\omega)E(\omega)$ and the desired frequency response is measured by a suitable cost function to be minimized. In the line of [84], we regard $e^{-iD\omega}$ as the desired frequency response ($e^{-iD\omega}$ is the frequency response of a pure delay of D time

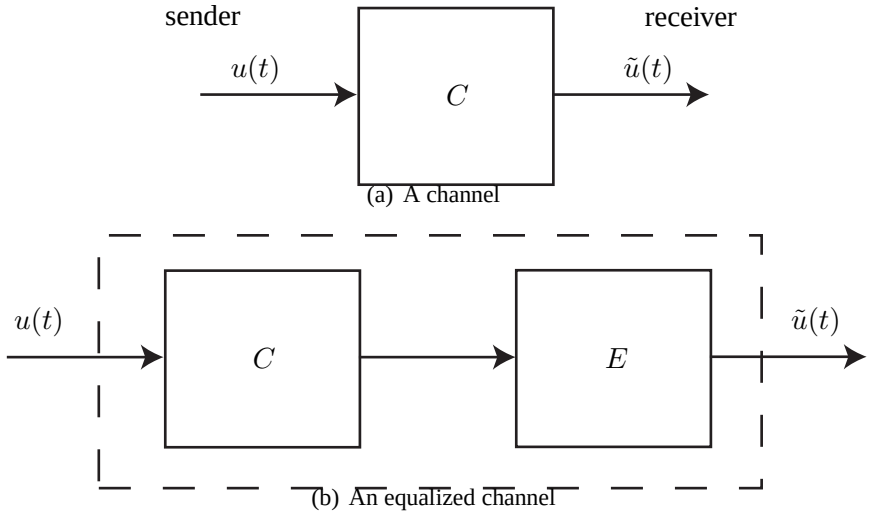


Figure 3.3. Channel equalization

steps), while, as cost function, we choose the sum of the *maximum* and the *average absolute deviation* between $C(\omega)E(\omega)$ and $e^{-iD\omega}$, formally

$$\mathrm{MAAD}(x) := \max_{k=-n,\ldots,0,\ldots,n} |C(\omega_k)E(\omega_k) - e^{-iD\omega}| + \lambda \frac{1}{2n+1} \sum_{k=-n}^{n} |C(\omega_k)E(\omega_k) - e^{-iD\omega}|, \tag{3.19}$$

where λ is a normalizing coefficient and $\omega_k = \frac{k}{n}\pi$, $k = 0, \pm 1, \ldots, \pm n$, is a gridding of $[-\pi, \pi]$. Throughout, n is set to 100, while $\lambda = 1$ and $D = 8$. In the MAAD cost function, the average absolute deviation takes care of the global behavior, over the whole range of frequencies, of the equalized channel, while the maximum absolute deviation explicitly penalizes the presence of resonant peaks, which are undesirable because they generate annoying whistling noise in audio communications.

The problem with (3.19) is that, in real-world applications, the frequency response of the channel is not exactly known because of imperfections in the estimation procedure used to retrieve $C(\omega)$ or to an intrinsic variability of the environment, as, for example, in mobile communication.
Hence, $C(\omega) = C(\omega, \delta)$ where δ is an uncertain parameter and the cost function should be more properly written as $\mathrm{MAAD}(x, \delta)$ so as to highlight the dependence on the uncertainty besetting the channel. We are thus facing an uncertain optimization problem and we resort to the scenario approach to deal with it.

3.3.2 Scenario Approach

Problem solution

The only requirement is the availability of N independent scenarios $C(\omega, \delta^{(1)})$, $C(\omega, \delta^{(2)}), \ldots, C(\omega, \delta^{(N)})$ of the uncertain frequency response to rely on. Notably, the scenario approach can be employed without a full knowledge of the probabilistic description of the uncertainty, and, in principle, the collected scenarios may be the results of field experiments performed in various environmental conditions (data-based optimization). If instead the probability distribution $\mathbb{P}_\Delta$ of δ is known, then the scenarios can be artificially generated. In this example, we suppose to be in this second case and that $C(\omega, \delta)$ is a second-order frequency response of the type:

$$C(\omega, \delta) = \frac{1}{e^{i2\omega} + \delta_1 e^{i\omega} + \delta_2},$$

where the uncertain parameter $\delta = (\delta_1, \delta_2)$ is uniformly distributed over $[-0.4, 0.4] \times [0.5, 0.8]$. $N = 3000$ scenarios are thus obtained through a random number generator.

According to the scenario approach, the optimal equalizer E^* is the one whose design parameter vector x^* solves the convex problem

$$\min_{x \in \mathbb{R}^{10}} \max_{j=1,\ldots,3000} \mathrm{MAAD}(x, \delta^{(j)}). \tag{3.20}$$

The solution in our simulation is $x^* = (7.08 \cdot 10^{-2}, 1.00 \cdot 10^{-3}, -6.64 \cdot 10^{-2}, 1.42 \cdot 10^{-3}, 4.71 \cdot 10^{-2}, 3.73 \cdot 10^{-4}, 8.37 \cdot 10^{-1}, 2 \cdot 10^{-3}, 5.09 \cdot 10^{-1}, -3.46 \cdot 10^{-4})$. The costs $c_1^*, \ldots, c_{3000}^*$ are then computed according to Definition 6, i.e. $c_k^* = \max\left\{c \in \mathbb{R} : c \leq \mathrm{MAAD}(x^*, \delta^{(j)}) \text{ for a choice of } k \text{ indexes } j \text{ among } \{1, \ldots, 3000\}\right\}$. This amounts to evaluating the costs $\mathrm{MAAD}(x^*, \delta^{(1)}), \ldots, \mathrm{MAAD}(x^*, \delta^{(3000)})$ and sorting their values in decreasing order. The reliability of the designed equalizer E^* is next evaluated in light of the results of this chapter.

Upper bounding the risks

We choose the vector of risk thresholds $\bar{\boldsymbol{\epsilon}} = (\bar{\epsilon}_{11}, \ldots, \bar{\epsilon}_{3000})$ according to the following rule: a parameter $\beta' \in [0, 1]$ is fixed and, for each $k = 11, 12, \ldots, 3000$, $\bar{\epsilon}_k \in [0, 1]$ is selected such that

$$\sum_{i=0}^{k-1} \binom{N}{i} \bar{\epsilon}_k^i (1 - \bar{\epsilon}_k)^{N-i} = \frac{\beta'}{2989}.$$

In words, the rule consists in choosing $\bar{\boldsymbol{\epsilon}}$ so that the marginal probability $\mathbb{P}_\Delta^N\{R_k > \bar{\epsilon}_k\}$ is equal to $\frac{\beta'}{2989}$ for all $k = 11, 12, \ldots, 3000$.
According to (3.10) and (3.15), and by posing $\bar{\epsilon}_1 = \bar{\epsilon}_2 = \cdots = \bar{\epsilon}_{11}$, the adopted

choice for $\bar{\epsilon}$ entails that

$$\begin{aligned}
&\mathbb{P}_{\Delta}^{N}\{R_1 \leq \bar{\epsilon}_1, \ldots, R_{11} \leq \bar{\epsilon}_{11}, \ldots, R_{3000} \leq \bar{\epsilon}_{3000}\} \\
&\qquad \geq \mathbb{P}_{\Delta}^{N}\{R_{11} \leq \bar{\epsilon}_{11}, \ldots, R_{3000} \leq \bar{\epsilon}_{3000}\} \\
&\qquad \geq 1 - \sum_{k=11}^{N} \mathbb{P}_{\Delta}^{3000}\{R_k > \bar{\epsilon}_k\} \\
&\qquad = 1 - \sum_{k=11}^{3000} \sum_{i=0}^{k-1} \binom{3000}{i} \bar{\epsilon}_k^i (1-\bar{\epsilon}_k)^{3000-i} \\
&\qquad = 1 - \sum_{k=11}^{3000} \frac{\beta'}{2989} \\
&\qquad = 1 - \beta', \qquad (3.21)
\end{aligned}$$

i.e. the risks R_k's are simultaneously less than the corresponding $\bar{\epsilon}_k$'s with confidence at least $1-\beta'$. For example, we have confidence 0, 0.9, 0.99 for $\beta' = 1, 10^{-1}, 10^{-2}$ respectively. A more refined evaluation of $\mathbb{P}_{\Delta}^{N}\{R_1 \leq \bar{\epsilon}_1, \ldots, R_{11} \leq \bar{\epsilon}_{11}, \ldots, R_{3000} \leq \bar{\epsilon}_{3000}\}$ is obtained through $\mathrm{CDF}_{10}(\bar{\epsilon}_{11}, \ldots, \bar{\epsilon}_{3000})$ as stated by Theorem 7. Indeed, by computing $\mathrm{CDF}_{10}(\bar{\epsilon}_{11}, \ldots, \bar{\epsilon}_{3000})$ with the Monte-Carlo algorithm in Section 3.2.3, it turns out that the conditions $R_1 \leq \bar{\epsilon}_1, \ldots, R_{3000} \leq \bar{\epsilon}_{3000}$ simultaneously hold with confidence equal to 0.98 (as opposed to 0) when $\beta' = 1$, confidence 0.997 (as opposed to 0.9) when $\beta' = 10^{-1}$, and confidence 0.9997 (as opposed to 0.99) when $\beta' = 10^{-2}$. Fig. 3.4 shows the values of $\bar{\epsilon}_k$ for $\beta' = 1$, 10^{-1}, and 10^{-2}. As it is apparent, the values of $\bar{\epsilon}_k$ are quite insensitive to the value of β' so that enforcing a high confidence only marginally impacts on the $\bar{\epsilon}_k$'s.

We select $\beta' = 10^{-2}$, so that confidence is 0.9997 and we can reasonably suppose the risks of the empirical costs $c_1^*, \ldots, c_N^*$ are simultaneously upper bounded by thresholds $\bar{\epsilon}_1, \bar{\epsilon}_2, \ldots, \bar{\epsilon}_N$.

Thus, by linking $c_1^*, c_2^*, \ldots, c_N^*$ to $\bar{\epsilon}_1, \bar{\epsilon}_2, \ldots, \bar{\epsilon}_N$, as in Fig. 3.5, we can e.g. claim that the risk that the equalizer E^* carries a cost greater than $c_{11}^* = 1.298$ is just at most 1.09%, i.e. cost 1.298 is guaranteed for about the 99% of the random instances of the channel frequency response $C(\omega, \delta)$, while, at the same time, cost $c_{99}^* = 1.252$ is guaranteed for the 95% of the channel frequency responses, cost $c_{229}^* = 1.230$ is guaranteed for the 90% of them, and so forth and so on.

Cost distribution

The evaluation of the reliability of E^* can be further refined according to the discussion in Section 3.2.2. In fact, in the same line as above, a lower bounding vector $\underline{\epsilon}_{11}, \ldots, \underline{\epsilon}_{3000}$ can be chosen such that the marginal probability $\mathbb{P}_{\Delta}^{N}\{R_k \leq \underline{\epsilon}_k\}$ is

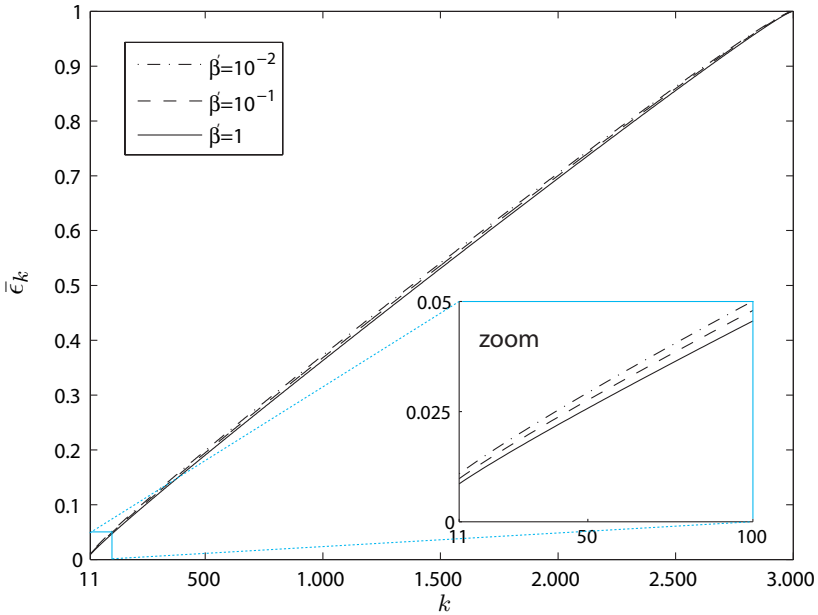


Figure 3.4. Values of $\bar{\epsilon}_k$, $k = 11, \ldots, 3000$, for $\beta' = 1$ (solid line), $\beta' = 10^{-1}$ (dashed line), and $\beta' = 10^{-2}$ (dash-dotted line).

equal to $\frac{\beta'}{2989}$ for all $k = 11, 12, \ldots, 3000$, i.e., by recalling (3.8), such that

$$1 - \sum_{i=0}^{k-1} \binom{N}{i} \underline{\epsilon}_k^i (1 - \underline{\epsilon}_k)^{N-i} = \frac{\beta'}{2989}.$$

By taking $\underline{\epsilon}_1 = \underline{\epsilon}_2 = \cdots = \underline{\epsilon}_{10} = 0$, we can compute

$$\begin{aligned}&\mathbb{P}_\Delta^N\{\underline{\epsilon}_1 \le R_1 \le \bar{\epsilon}_1, \underline{\epsilon}_2 \le R_2 \le \bar{\epsilon}_2, \ldots, \underline{\epsilon}_{3000} \le R_{3000} \le \bar{\epsilon}_{3000}\} \\ &\quad = \mathbb{P}_\Delta^N\{\underline{\epsilon}_{11} \le R_{11} \le \bar{\epsilon}_{11}, \underline{\epsilon}_{12} \le R_{12} \le \bar{\epsilon}_{12}, \ldots, \underline{\epsilon}_{3000} \le R_{3000} \le \bar{\epsilon}_{3000}\}\end{aligned}$$

based on $\mathrm{CDF}_d(\eta_{d+1}, \ldots, \eta_N)$ given by Theorem 7, and it turns out that $\mathbb{P}_\Delta^N\{\underline{\epsilon}_1 \le R_1 \le \bar{\epsilon}_1, \underline{\epsilon}_2 \le R_2 \le \bar{\epsilon}_2, \ldots, \underline{\epsilon}_{3000} \le R_{3000} \le \bar{\epsilon}_{3000}\}$ is 0.9993. In other words, we have that $\underline{\epsilon}_k \le R_k \le \bar{\epsilon}_k$ simultaneously hold for every $k = 1, \ldots, N$ with confidence 0.9993, a quite high value, which permits us to be reasonably sure that risks are indeed lower and upper bounded as indicated. Moreover, by recalling the relation between $F_\ell(c) = \mathbb{P}_\Delta\{\mathrm{MAAD}(x^*, \delta) \le c\}$ and the risks, according to equation (3.14), the probability box containing $F_\ell(c)$ with high confidence 0.9993 can be computed, see Figs. 3.6 and 3.7. This result is a sharp characterization of the probability distribution of the cost associated with the designed equalizer and provides full information on the reliability of E^*.

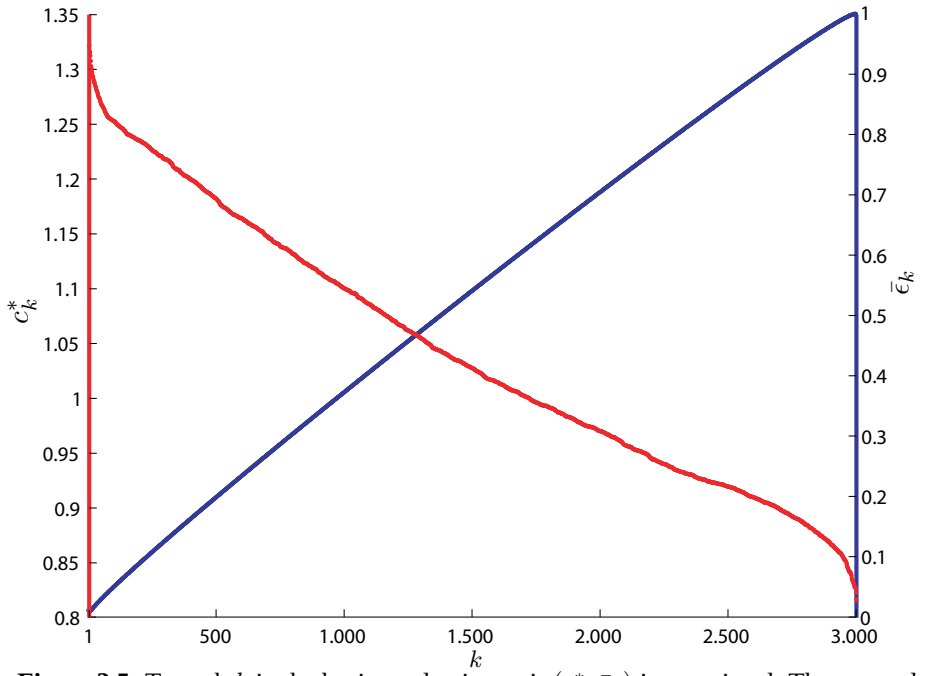


Figure 3.5. To each k in the horizontal axis a pair $(c_k^*, \bar{\epsilon}_k)$ is associated. The cost value c_k^* (red coloured) can be read on the left ordinate, while the risk threshold $\bar{\epsilon}_k$ (blue coloured) can be read on the right ordinate.

3.4 Proofs

We first prove in next Section 3.4.1 the fundamental Theorem 7 by computing

$$\mathbb{P}_\Delta^N\{R_{d+1} \leq \epsilon_{d+1}, R_{d+2} \leq \epsilon_{d+2}, R_{d+3} \leq \epsilon_{d+3}, \ldots, R_N \leq \epsilon_N\} \tag{3.22}$$

under Assumptions 1 and 2.ii. Based on this result, the proof of Theorem 8 is developed in Section 3.4.3 by releasing Assumption 2.ii. Theorem 6 immediately follows from Theorem 7 by noting that under the assumptions of Theorem 6 it holds that $R_1 = R_2 = \cdots = R_{d+1}$.

3.4.1 Proof of Theorem 7

For any fixed $(x, \underline{c}, \bar{c}) \in \mathbb{R}^{d+2}$, let $D(x, \underline{c}, \bar{c}) := \mathbb{P}_\Delta\{\delta \in \Delta : \underline{c} < \ell(x, \delta) \leq \bar{c}\}$ and, for any integer k such that $d + 1 \leq k \leq N$, let

$$D_k := D(x^*, c_{k+1}^*, c_k^*) \tag{3.23}$$

where c_{N+1}^* is defined to be equal to $-\infty$. Similarly to the R_k's, D_k's are random variables, since they depend on the sample $(\delta^{(1)}, \ldots, \delta^{(N)})$ through $x^*, c_{d+1}^*, \ldots, c_N^*$, and, indeed, D_k is the conditional probability with respect to x^*, c_{k+1}^*, c_k^* that a newly extracted uncertainty instance δ carries a cost between c_k^* and c_{k+1}^*. The

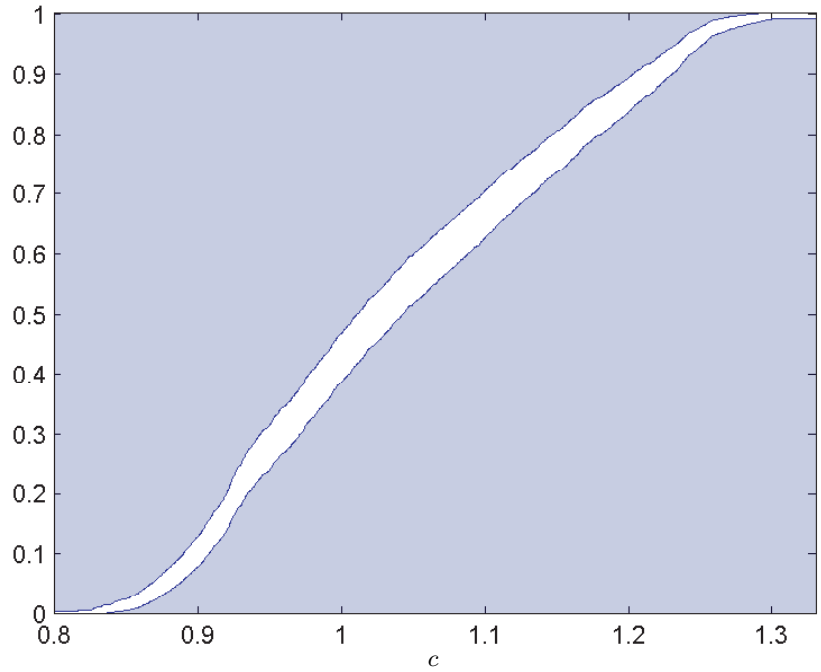


Figure 3.6. The graph of $F_\ell(c)$, the cumulative distribution function of the cost at $x = x^*$, lays in the white strip with confidence 0.9993. Thus, for each value of c on the abscissa, $F_\ell(c)$ belongs to an interval bounded from above and below. For a zoomed view of the probability box see Fig. 3.7.

variables D_k's and the R_k's are related by the following simple linear transformations

$$\begin{aligned}
R_{d+1} &= 1 - \sum_{i=d+1}^{N} D_k & D_{d+1} &= R_{d+2} - R_{d+1} \\
R_{d+2} &= 1 - \sum_{i=d+2}^{N} D_k & D_{d+2} &= R_{d+3} - R_{d+2} \\
&\vdots & &\vdots \\
R_{N-1} &= 1 - \sum_{i=N-1}^{N} D_k, & D_{N-1} &= R_N - R_{N-1} \\
R_N &= 1 - D_N, & D_N &= 1 - R_N.
\end{aligned} \tag{3.24}$$

Thanks to (3.24), the joint probability distribution function of the R_k's can be easily derived from the joint probability distribution function of the D_k's and vice versa. We, hence, proceed by computing the joint probability distribution function of the D_k's first. In order to do so, we consider $\mathbb{E}_{\Delta^N}[D_{d+1}^{k_{d+1}} \cdots D_N^{k_N}]$, the multivariate

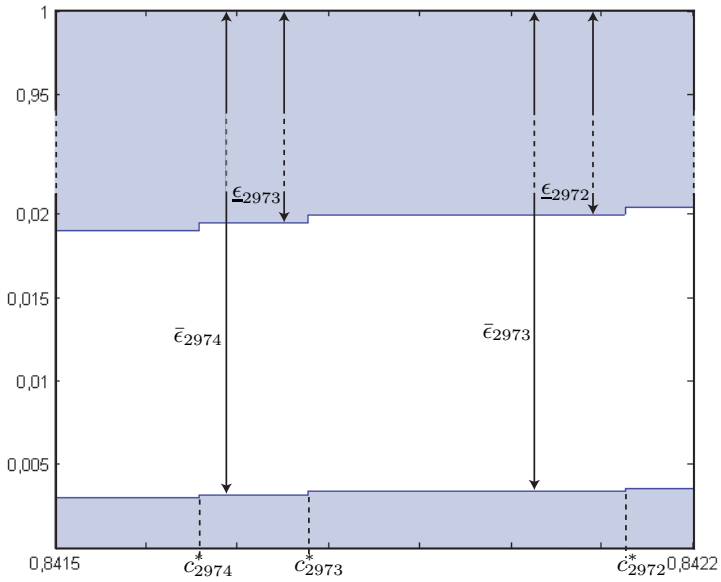


Figure 3.7. A zoomed detail view of the probability box in Fig. 3.6. The probability box for $F_\ell(c)$ in $[0.8415, 0.8422]$ is here represented and some empirical costs c^*_k, together with the lower ($\underline{\epsilon}_k$) and upper ($\bar{\epsilon}_k$) bounds to their risks, are put in evidence.

moment of $D_{d+1}, \ldots, D_N$, and evaluate it for each possible assignment of non-negative integers $k_{d+1}, \ldots, k_N$. The joint distribution function of $D_{d+1}, \ldots, D_N$ can then be deduced from the resulting moment problem.
To ease the notation, define: $M_d = N$, $M_{d+1} = N + k_{d+1}$, $M_{d+2} = N + k_{d+1} + k_{d+2}$, etc., until $M_N = N + \sum_{i=d+1}^{N} k_i$. By (3.23), the product $D_{d+1}^{k_{d+1}} D_{d+2}^{k_{d+2}} \cdots D_N^{k_N}$ gives the conditional probability with respect to x^*, $c^*_{d+1}, \ldots, c^*_N$, i.e. with respect to the data samples $(\delta^{(1)}, \ldots, \delta^{(N)})$, that $M_N - N$ new independently extracted uncertainty instances from Δ, say $\delta^{(N+1)}, \ldots, \delta^{(M_N)}$, are such that the first k_{d+1} (i.e. $\delta^{(N+1)}, \ldots, \delta^{(M_{d+1})}$) carry a cost between c^*_{d+1} and c^*_{d+2}, the next k_{d+2} (i.e. $\delta^{(M_{d+1}+1)}, \ldots, \delta^{(M_{d+2})}$) carry a cost between c^*_{d+2} and c^*_{d+3}, and so forth and so on till the last k_N carrying a cost below c^*_N (recall that $c^*_{N+1} = -\infty$). Therefore, the product $D_{d+1}^{k_{d+1}} D_{d+2}^{k_{d+2}} \ldots D_N^{k_N}$ can be expressed as

$$\prod_{i=d+1}^{N} D_i^{k_i} = \mathbb{P}_\Delta^{M_N - N}\{c^*_{i+1} < \ell(x^*, \delta^{(j)}) \le c^*_i,\ i = d+1, \ldots, N,\ j = M_{i-1}+1, \ldots, M_i\}, \tag{3.25}$$

where $\mathbb{P}_\Delta^{M_N - N} = \mathbb{P}_\Delta \times \cdots \times \mathbb{P}_\Delta$ denotes as usual the product probability measure of $\delta^{(N+1)}, \ldots, \delta^{M_N}$. Expressing probability as the integral of an indicator function and using the compact notation $\boldsymbol{\delta}_m^n$ to indicate $(\delta^{(m)}, \delta^{(m+1)}, \ldots, \delta^{(n)})$ and $\Delta_m^n =$

$\Delta \times \Delta \times \cdots \times \Delta$ to indicate the domain for $\boldsymbol{\delta}_m^n$, (3.25) can be rewritten as

$$\prod_{i=d+1}^{N} D_i^{k_i} = \int_{\Delta_{N+1}^{M_N}} \mathbb{1}\{c_{i+1}^* < \ell(x^*, \delta^{(j)}) \leq c_i^*,\ i = d+1, \ldots, N, \\ j = M_{i-1}+1, \ldots, M_i\} \mathbb{P}_{\Delta}^{M_N - N}\{\mathrm{d}\boldsymbol{\delta}_{N+1}^{M_N}\}.$$

As $(\delta^{(1)}, \ldots, \delta^{(N)})$ is let vary in Δ^N, $\prod_{i=d+1}^{N} D_i^{k_i}$ takes on various values and we are interested in computing its expected value, i.e.

$$\mathbb{E}_{\Delta^N}\left[\prod_{i=d+1}^{N} D_i^{k_i}\right] = \int_{\Delta_1^N} \prod_{i=d+1}^{N} D_i^{k_i} \mathbb{P}_{\Delta}^{N}\{\mathrm{d}\boldsymbol{\delta}_1^N\} \\ = \int_{\Delta_1^N} \int_{\Delta_{N+1}^{M_N}} \mathbb{1}\{c_{i+1}^* < \ell(x^*, \delta^{(j)}) \leq c_i^*,\ i = d+1, \ldots, N, \\ j = M_{i-1}+1, \ldots, M_i\} \mathbb{P}_{\Delta}^{M_N - N}\{\mathrm{d}\boldsymbol{\delta}_{N+1}^{M_N}\} \mathbb{P}_{\Delta}^{N}\{\mathrm{d}\boldsymbol{\delta}_1^N\},$$

which, by Tonelli's theorem, can be restated as

$$\int_{\Delta_1^{M_N}} \mathbb{1}\{c_{i+1}^* < \ell(x^*, \delta^{(j)}) \leq c_i^*,\ i = d+1, \ldots, N, \\ j = M_{i-1}+1, \ldots, M_i\} \mathbb{P}_{\Delta}^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\},$$

that is the moment $\mathbb{E}_{\Delta^N}[D_{d+1}^{k_{d+1}} \cdots D_N^{k_N}]$ is nothing but the *total* probability with respect to all variables $\delta^{(1)}, \ldots, \delta^{(N)}, \delta^{(N+1)}, \ldots, \delta^{(M_N)}$ that $\delta^{(N+1)}, \ldots, \delta^{(M_{d+1})}$ carry a cost between c_{d+1}^* and c_{d+2}^*, $\delta^{(M_{d+1}+1)}, \ldots, \delta^{(M_{d+2})}$ carry a cost between c_{d+2}^* and c_{d+3}^*, and so forth and so on. Now, let $\bar{S} = \{j_1, \ldots, j_N\}$ be a generic subset of N indexes taken from $\{1, \ldots, M_N\}$ and let $z_{|\bar{S}}^* = (x_{|\bar{S}}^*, c_{|\bar{S}}^*)$ be the optimal solution to problem

$$\mathrm{EPI}_{|\bar{S}}: \quad \min_{c \in \mathbb{R}, x \in \mathcal{X} \subseteq \mathbb{R}^d} c \\ \text{subject to: } \ell(x, \delta^{(i)}) \leq c, \quad i \in \bar{S}. \tag{3.26}$$

Moreover, for $k = 1, \ldots, N$, let $c_{k|\bar{S}}^* = \max\{c \in \mathbb{R} :\ c \leq \ell(x_{|\bar{S}}^*, \delta^{(i)})$ for a choice of k indexes i among $\bar{S}\}$, i.e. the $c_{k|\bar{S}}^*$ are the costs associated with $x_{|\bar{S}}^*$, and let $c_{N+1|\bar{S}}^* = -\infty$. Eventually, for each $i = d+1, \ldots, N$, let $S_i = \{j_1, \ldots, j_{k_i}\}$ be a subset of k_i indexes from $\{1, \ldots, M_N\} \setminus \bar{S}$ such that $S_m \cap S_n = \emptyset$ if $m \neq n$. Due to the i.i.d. (independent and identically distributed) nature of the uncertainty instances $(\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(M_N)})$, the total probability that the instances with indexes in S_{d+1} carry a cost between $c_{d+1|\bar{S}}^*$ and $c_{d+2|\bar{S}}^*$, those with indexes in S_{d+2} carry a cost between $c_{d+2|\bar{S}}^*$ and $c_{d+3|\bar{S}}^*$, and so forth and so on till those with indexes in

S_N carrying a cost between $c^*_{N|\bar{S}}$ and $c^*_{N+1|\bar{S}}$, does not depend in any way on the choice of $\bar{S}, S_{d+1}, \ldots, S_N$. Whence

$$\begin{aligned}
&\int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1} < \ell(x^*, \delta^{(j)}) \leq c^*_i,\ i = d+1, \ldots, N, \\
&\qquad j = M_{i-1}+1, \ldots, M_i\} \mathbb{P}_\Delta^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\} \\
&= \int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1|\bar{S}} < \ell(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i|\bar{S}},\ i = d+1, \ldots, N, \\
&\qquad\qquad j \in S_i\} \mathbb{P}_\Delta^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\}\ \ \forall (\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S},
\end{aligned}$$

where $\mathcal{S}$ is the set of all feasible choices of $\bar{S}, S_{d+1}, \ldots, S_N$ from $\{1, \ldots, M_N\}$. Indicating with $|\mathcal{S}|$ the cardinality of $\mathcal{S}$, we have

$$\begin{aligned}
&\mathbb{E}[D_{d+1}^{k_{d+1}} \cdots D_N^{k_N}] \\
&= \int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1} < \ell(x^*, \delta^{(j)}) \leq c^*_i,\ i = d+1, \ldots, N, \\
&\qquad j = M_{i-1}+1, \ldots, M_i\} \mathbb{P}_\Delta^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\} \\
&= \frac{1}{|\mathcal{S}|} \sum_{(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}} \int_{\Delta_1^{M_N}} \mathbb{1}\{c^*_{i+1|\bar{S}} < \ell(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i|\bar{S}}, \\
&\qquad\qquad i = d+1, \ldots, N,\ j \in S_i\} \mathbb{P}_\Delta^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\} \\
&= \frac{1}{|\mathcal{S}|} \int_{\Delta_1^{M_N}} \sum_{(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}} \mathbb{1}\{c^*_{i+1|\bar{S}} < \ell(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i|\bar{S}},\ i = d+1, \ldots, N, \\
&\qquad\qquad j \in S_i\} \mathbb{P}_\Delta^{M_N}\{\mathrm{d}\boldsymbol{\delta}_1^{M_N}\}.
\end{aligned}$$

For a fixed sample $\boldsymbol{\delta}_1^{M_N}$ the inner sum

$$\sum_{(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}} \mathbb{1}\{c^*_{i+1|\bar{S}} < \ell(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i|\bar{S}}, i = d+1, \ldots, N,\ j \in S_i\}$$

counts the number of partitions of uncertainty instances $\delta^{(1)}, \ldots, \delta^{(M_N)}$ into sets $\bar{S}, S_{d+1}, \ldots, S_N$, such that the costs associated with the instances in $S_{d+1}, \ldots, S_N$ fit into the costs computed based on the instances in $\bar{S}$ according to the condition

$$c^*_{i+1|\bar{S}} < \ell(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i|\bar{S}}, \quad i = d+1, \ldots, N, \quad j \in S_i. \tag{3.27}$$

It is a fact that such number is almost surely equal to 1 as formally stated in the next proposition, whose proof is postponed to next Section 3.4.2 in order to first draw the conclusion.

Proposition 1. *It holds with probability* 1 *that*

$$\sum_{(\bar{S}, S_{d+1}, \ldots, S_N) \in \mathcal{S}} \mathbb{1}\{c^*_{i+1|\bar{S}} < \ell(x^*_{|\bar{S}}, \delta^{(j)}) \leq c^*_{i|\bar{S}},\ i = d+1, \ldots, N,\ j \in S_i\} = 1.$$

$\star$

Thanks to Proposition 1, we have

$$\mathbb{E}_{\Delta^N}[D_{d+1}^{k_{d+1}} \cdots D_N^{k_N}] = \frac{1}{|\mathcal{S}|} = \frac{1}{\binom{M_N}{N, k_{d+1}, \ldots, k_N}}, \tag{3.28}$$

where the last equality follows from the evaluation of $|\mathcal{S}|$, through a multinomial coefficient, see e.g. [87]:

$$|\mathcal{S}| = \binom{M_N}{N, k_{d+1}, \ldots, k_N} = \prod_{i=0}^{N-d-1} \binom{M_{N-i}}{k_{N-i}} = \frac{M_N!}{N! k_{d+1}! \cdots k_N!}.$$

Note that (3.28) holds true for every value $k_{d+1}, \ldots, k_N$ so that (3.28) provides the infinite multivariate moments of $D_{d+1}, \ldots, D_N$. The probability distribution function of $D_{d+1}, \ldots, D_N$ then is uniquely determined, [88]. In particular, by integration one can check that the density of the Dirichlet distribution,

$$p_D(x_{d+1}, x_{d+2}, \ldots, x_N) = \frac{N!}{d!} \left(1 - \sum_{i=d+1}^{N} x_i\right)^d \mathbb{1}\left\{\sum_{i=d+1}^{N} x_i \leq 1, \quad 0 \leq x_i \leq 1\right\},$$

satisfies the moment problem posed by (3.28). By applying the transformation (3.24), we obtain the joint density p_R of $R_{d+1}, \ldots, R_N$:

$$\begin{aligned} &p_R(r_{d+1}, r_{d+2}, \ldots, r_N) \\ &\quad = p_D(r_{d+2} - r_{d+1}, r_{d+3} - r_{d+2}, \ldots, r_N - r_{N-1}, 1 - r_N) \\ &\quad = \frac{N!}{d!} r_{d+1}^d \mathbb{1}\{0 \leq r_{d+1} \leq r_{d+2} \leq \cdots \leq r_N \leq 1\}, \end{aligned} \tag{3.29}$$

and equation (3.9) follows by integrating (3.29).

3.4.2 Proof of Proposition 1

Consider the optimization problem with all the M_N uncertainty instances $\delta^{(1)}, \ldots, \delta^{(N)}$, $\delta^{(N+1)}, \ldots, \delta^{(M_N)}$ in place:

$$\begin{aligned} \text{EPI}_{M_N}: \quad &\min_{c \in \mathbb{R}, x \in \mathcal{X} \subseteq \mathbb{R}^d} c \\ &\text{subject to: } \ell(x, \delta^{(i)}) \leq c, \ i = 1, \ldots, M_N, \end{aligned} \tag{3.30}$$

and let $(\tilde{x}, \tilde{c})$ be the optimal solution. Moreover, let $\tilde{c}_k = \max\{c \in \mathbb{R} : \ell(\tilde{x}, \delta^{(i)}) \geq c$ for a choice of k indexes i among $\{1, \ldots, M_N\}\}$, $k = 1, \ldots, M_N$, be the costs associated with $\tilde{x}$. Plainly, $\tilde{c}_k \leq \tilde{c}_{k'}$ when $k > k'$. Assumption $2.ii$ implies that the following strict ordering holds true almost surely:

$$\tilde{c}_{d+1} > \tilde{c}_{d+2} > \cdots > \tilde{c}_{M_N}. \tag{3.31}$$

Proof of (3.31). For every fixed $\delta^{(1)}, \ldots, \delta^{(M_N)}$, at least one basis of $d+1$ instances can always be found such that the solution to the optimization problem when only those instances are considered is the same as the solution to (3.30) with all M_N instances in place, see e.g. [89]. Consider now the subset of the $(\delta^{(1)}, \ldots, \delta^{(M_N)}) \in \Delta_1^{M_N}$ violating condition (3.31) and whose first $d+1$ instances form a basis, so that $\tilde{x}$ is determined once the values of $\delta^{(1)}, \ldots, \delta^{(d+1)}$ are fixed. This subset has zero probability because, conditionally to $\delta^{(1)}, \ldots, \delta^{(d+1)}$, the probability that $\ell(\tilde{x}, \delta^{(k)}) = \ell(\tilde{x}, \delta^{(h)})$ for some $k \in \{d+2, \ldots, M_N\}$ and some $h \in \{1, \ldots, M_N\}$, $h \neq k$, is zero thanks to Assumption 2.ii, and, hence, the probability that two costs among $\tilde{c}_{d+1}, \ldots, \tilde{c}_N$ are equal is zero as well. The same reasoning permits one to conclude that all the subsets of $\delta^{(1)}, \ldots, \delta^{(M_N)}$ violating (3.31) and whose instances $\delta^{(i_1)}, \ldots, \delta^{(i_{d+1})}$ form a basis, for all possible choices of $i_1, \ldots, i_{d+1}$ in $\{1, \ldots, M_N\}$, have zero probability. The thesis follows by noting that the instances $\delta^{(1)}, \ldots, \delta^{(M_N)}$ violating (3.31) are obtained as the union of these subsets. □

In order for (3.27) to hold, observe that $\bar{S}$ must be such that $(\tilde{x}, \tilde{c})$, the optimal solution to (3.30), is equal to $(x^*_{|\bar{S}}, c^*_{|\bar{S}})$, the optimal solution computed with the uncertainty instances in $\bar{S}$ in place only. Indeed, if this were not the case, there would be an instance $\delta^{(j)}$ in one of the sets $S_{d+1}, \ldots, S_N$ violating $(x^*_{|\bar{S}}, c^*_{|\bar{S}})$, i.e. $\ell(x^*_{|\bar{S}}, \delta^{(j)}) > c^*_{|\bar{S}}$. By definition of $c^*_{d+1|\bar{S}}$, this would entail $\ell(x^*_{|\bar{S}}, \delta^{(j)}) > c^*_{d+1|\bar{S}}$, which does not fit (3.27).

Moreover, in order for (3.27) to hold, $\bar{S}$ must contain the set $\mathcal{H} = \{i \in \{1, \ldots, M_N\} : \ \ell(\tilde{x}, \delta^{(i)}) \geq \tilde{c}_{d+1}\}$, i.e. the set of all indexes of uncertainty instances corresponding to the uppermost costs in correspondence of $\tilde{x}$. Indeed, suppose that one or more instances in $\mathcal{H}$ do not belong to $\bar{S}$, and that $(x^*_{|\bar{S}}, c^*_{|\bar{S}}) = (\tilde{x}, \tilde{c})$. Since by (3.31) the cardinality of $\mathcal{H}$ is exactly $d+1$, then the costs $c^*_{d+1|\bar{S}}$ must take on a value strictly below $\tilde{c}_{d+1}$. Hence, for a $j \in \mathcal{H} \setminus \bar{S}$, we would have $\ell(x^*_{|\bar{S}}, \delta^{(j)}) = \ell(\tilde{x}, \delta^{(j)}) \geq \tilde{c}_{d+1} > c^*_{d+1|\bar{S}}$, again violating (3.27). If instead we impose $\mathcal{H} \subseteq \bar{S}$, we have that $(x^*_{|\bar{S}}, c^*_{|\bar{S}}) = (\tilde{x}, \tilde{c})$ because a simple inspection reveals that $\mathcal{H}$ contains all uncertainty instances corresponding to active constraints in (3.30). The other instances with indexes not belonging to $\mathcal{H}$ carry costs in correspondence of $\tilde{x}$ which are equal to costs $\tilde{c}_{d+2}, \ldots, \tilde{c}_{M_N}$ and which are strictly ordered by (3.31). By adding to S_{d+1} the k_{d+1} indexes of uncertainty instances having costs $\tilde{c}_{d+2}, \ldots, \tilde{c}_{d+2+k_{d+1}-1}$, then to $\bar{S}$ the next one having cost $\tilde{c}_{d+2+k_{d+1}}$, then to S_{d+2} the following k_{d+2}, then to $\bar{S}$ the next one having cost $\tilde{c}_{d+2+k_{d+1}+k_{d+2}}$, and so forth and so on till the last k_N are put into S_N, one obtains a partition $\bar{S}, S_{d+1}, \ldots, S_N$ satisfying (3.27). Due to the ordering of costs of instances not in $\mathcal{H}$, this partition is the sole possible one. □

3.4.3 Proof of Theorem 8

When Assumption 2 is completely dropped, the assessment of the probability distribution function of $R_{d+1}, \ldots, R_N$ can be obtained by mimicking the reasoning used in [56] to prove the general result recalled in Section 3.2, equation (3.5). The idea of [56] was to infinitesimally perturb the constraints of the scenario optimization problem ("heating") so as to go back to a setting where the needed assumption is verified and then to infer the sought result via a limiting process. Here, this reasoning can be decidedly simplified by perturbing constraints of (3.2) just along the direction of component $c \in \mathbb{R}$. The proof is now sketched, pointing out the differences from [56].

Heating

Consider $H = [-\rho, \rho]$, $\rho > 0$, and $\delta' = (\delta, h) \in \Delta'$, with $\Delta' = \Delta \times H$. We define, for each $x \in \mathcal{X}$ and $\delta' = (\delta, h)$, the function $\ell'(x, \delta') = \ell(x, \delta) + h$. Finally, indicating with $\mathbb{U}$ the uniform measure on H, the probability $\mathbb{P}'_\Delta = \mathbb{P}_\Delta \times \mathbb{U}$ is defined over Δ'. Clearly, ℓ' and $\mathbb{P}'_\Delta$ are such that Assumption $2.ii$ holds, since for any (x, c) we have $\mathbb{P}'_\Delta\{\ell'(x, \delta') = c\} = 0$. The problem obtained by extracting N constraints from Δ' is called the heated scenario problem and is as follows:

$$\text{H-EPI}_N : \quad \min_{c \in \mathbb{R}, x \in \mathcal{X} \subseteq \mathbb{R}^d} c \qquad \text{subject to: } c \geq \ell(x, \delta^{(i)}) + h^{(i)} \quad i = 1, \ldots, N. \tag{3.32}$$

The solution to (3.32) is indicated by (x'^*, c'^*). For this problem Theorem 7 is valid. Hence, letting c'^*_k, $k = 1, \ldots, N$, be the costs of the heated scenario problem and letting R'_k, $k = 1, \ldots, N$, be the corresponding risks (i.e. $R'_k = \{\delta' \in \Delta' : \ell'(x'^*, \delta') > c'^*_k\}$), the joint probability distribution function $\mathbb{P}'^N_\Delta\{R'_{d+1} \leq \epsilon_{d+1}, \ldots, R'_N \leq \epsilon_N\}$ can be exactly computed and is given by (3.9).

Convergence of the heated solution to the original solution

Fix a $\delta^{(1)}, \ldots, \delta^{(N)}$, and compute the solution of EPI_N, (x^*, c^*), as well as the costs $c^*_{d+1}, \ldots, c^*_N$. Let ρ_n be a sequence of heating parameters monotonically decreasing to zero. For every n, pick any N numbers $h_n^{(1)}, \ldots, h_n^{(N)}$ from the interval $H_n = [-\rho_n, \rho_n]$, and let (x'^*, c'^*) and $c'^*_{d+1}, \ldots, c'^*_N$ be the solution and the costs of problem (3.32), where $\delta'^{(1)} = (\delta^{(1)}, h_n^{(1)}), \ldots, \delta'^{(N)} = (\delta^{(N)}, h_n^{(N)})$. By mimicking [56], it is easy to show that the heated solution as well as the heated costs converge to the original solution and costs as the heating parameter ρ_n tends to zero: $\forall \delta^{(1)}, \ldots, \delta^{(N)} \in \Delta^N$,

$$\lim_{n \to \infty} \sup_{h_n^{(1)}, \ldots, h_n^{(N)} \in H_n} ||(x'^*, c'^*, c'^*_{d+1}, \ldots, c'^*_N) - (x^*, c^*, c^*_{d+1}, \ldots, c^*_N)|| = 0. \tag{3.33}$$

In particular, the convergence of the costs $c_k'^*$ to c_k^*, $k = d+1, \dots, N$, comes as a consequence of the continuity of $\ell(x, \delta)$ in x.

Derivation of (3.12)
Fix a data sample $\mathsf{D}^N = (\delta^{(1)}, \dots, \delta^{(N)})$ that is *bad*, i.e. such that the condition $R_j > \epsilon_j$ is true for at least one $j \in \{d+1, \dots, N\}$. As above, consider a sequence of heating parameters $\rho_n \downarrow 0$. In line with [56], it can be shown that, thanks to (3.33), there exists a big enough $\bar{n}$ such that, $\forall n > \bar{n}$, and for every choice of $h_n^{(1)}, \dots, h_n^{(N)}$, the heated data sample $((\delta^{(1)}, h_n^{(1)}), \dots, (\delta^{(N)}, h_n^{(N)}))$ is such that $R_j' > \epsilon_j$, i.e. it is bad in the heated setting. To conclude the theorem, note that

$$
\begin{aligned}
&(\mathbb{P}_\Delta \times \mathbb{U})^N\{\exists j : R_j' > \epsilon_j\} \\
&= \int_{\Delta^N} \int_{H_n^N} \mathbb{1}\{\exists j : R_j' > \epsilon_j\} \frac{\mathrm{d}\boldsymbol{h}_1^N}{(2\rho_n)^N} \, \mathbb{P}_\Delta^N\{\mathrm{d}\boldsymbol{\delta}_1^N\} \\
&\geq \int_{\Delta^N} \mathbb{1}\{\exists j : R_j > \epsilon_j\} \int_{H_n^N} \mathbb{1}\{\exists j : R_j' > \epsilon_j\} \frac{\mathrm{d}\boldsymbol{h}_1^N}{(2\rho_n)^N} \, \mathbb{P}_\Delta^N\{\mathrm{d}\boldsymbol{\delta}_1^N\}.
\end{aligned}
$$

The outer indicator function limits the integration domain to data samples in Δ^N that are bad. For every fixed data sample in this domain the inner integral is equal to 1 for a sufficiently large n. Thus, by the dominated convergence theorem, taking the limit for $n \to \infty$ it holds that

$$
\begin{aligned}
&(\mathbb{P}_\Delta \times \mathbb{U})^N\{\exists j : R_j' > \epsilon_j\} \\
&\quad \geq \int_{\Delta^N} \mathbb{1}\{\exists j : R_j > \epsilon_j\} \mathbb{P}_\Delta\{\mathrm{d}\boldsymbol{\delta}_1^N\} \\
&\quad = \mathbb{P}_\Delta^N\{\exists j : R_j > \epsilon_j\} = 1 - \mathbb{P}_\Delta^N\{\forall j \, R_j \leq \epsilon_j\}.
\end{aligned}
$$

Equation (3.12) follows since

$$
\begin{aligned}
&1 - \mathrm{CDF}_d(\epsilon_{d+1}, \dots, \epsilon_N) \\
&\quad = 1 - (\mathbb{P}_\Delta \times \mathbb{U})^N\{R_j' \leq \epsilon_j \, \forall j \in \{d+1, \dots, N\}\} \\
&\quad = (\mathbb{P}_\Delta \times \mathbb{U})^N\{\exists j : R_j' > \epsilon_j\}.
\end{aligned}
$$

□

3.5 Perspectives for future work and an open issue

The results presented in this chapter are full-fledged results for the kind of decision problems considered, and our analysis has shown that they are not improvable in the absence of further assumptions and before any data is observed. Moreover, in our specific context, they throw a new light on already known results. The theory presented is particularly useful in data-based optimization, where it can

be employed to characterize in-depth a worst-case scenario solution. We believe that a significant application may be in the comparative evaluation of manifold data-based decisions, where a theoretically sound evaluation of the respective cost distributions may be crucial.[1]

If some a-priori knowledge about the distribution of the uncertainty is at disposal, it can be integrated in the present framework. For example, bounds on the possible values attained by the cost function $\ell(x, \delta)$ can be integrated with the probability boxes obtained in Section 3.2.2 so as to allow the computation of useful confidence intervals for $\mathbb{E}_\Delta[\ell(x^*, \delta)]$.
Finally, the success in studying min-max convex decision problems suggests to investigate whether a larger class of problems can be studied with the same or similar tools, to achieve similarly strong results.

3.5.1 Shortage of samples

The number N of scenarios required to guarantee that c^* has a desired "high-coverage property" is usually called *sample complexity*. In large-scale problems with very large d, the sample complexity may become too high, especially if the scenarios come from real, expensive experiments. Indeed, we have seen that the risk of c^* is tightly bounded by the risk R_{d+1} of c^*_{d+1}, which is large if d is comparable with N, see (3.3). This problem can be tackled in different ways. An idea is to try to introduce a mathematical machinery similar to that used in Chapter 2. In fact, the theory in Chapter 2, though at present limited to mean coverages, leads to results that do not depend directly on the dimension d, as discussed in Remark 4. However, it is an open problem to what extent similar results can be obtained without restricting a-priori the class of the possible cost functions, i.e. without abandoning the general convex context studied in the present chapter. In the following Chapter 4, we will work under the same assumptions as in the present chapter and we will focus on the coverage of the worst-case cost. We will show how to compute a data-based decision similar to the worst-case decision x^* accompanied by a cost threshold similar to c^* but with the desired coverage properties even when a relatively small number of scenarios is at our disposal. The algorithm presented in the following chapter is an instance of a more general idea, which consists in suitably exploiting some structure that can be revealed by data. Indeed, as a matter of fact, very often reality is redundant: this means that even a few randomly observed scenarios may be sufficient to betray the structure of the whole unseen reality. If suitable mechanisms are introduced to reveal and exploit such a structure, cost thresholds with good coverage properties can be provided even if N is small.

[1] A viable approach is that of combining the characterization of the cost distribution here offered with results in the line of Theorem 14. Indeed, Theorem 14 is used in [79] to compare various possible decisions based on cost-risk pairs. It is viable to extend such an idea in the light of the theory here presented so as to characterize each possible decision based not only on a cost-risk pair but on the full distribution of the costs.

Chapter 4

Data-based min-max decisions with reduced sample complexity

In this chapter, we focus on the sample-complexity of data-based convex min-max problems. For a fixed (usually small) ϵ and a fixed (usually very small) β, the *sample-complexity* is the number N of scenarios needed in order for the coverage of the empirical worst-case cost c^* to be no smaller than $1-\epsilon$ with confidence $1-\beta$. The sample complexity of the data-based min-max optimization rapidly increases with the dimension d of the decision variable, and this may pose a hurdle to its applicability to medium and large scale problems. We here introduce FAST (Fast Algorithm for the Scenario Technique), a variant of the min-max decision-making algorithm with reduced sample complexity.

4.1 Introduction and problem position

We will work in the same framework of Chapter 3. We recall briefly the fundamental facts. Given a cost function $\ell(x, \delta)$, convex in the decision variable x for any value of the uncertainty variable δ, we have studied the properties of the decision x^*, solution to the min-max problem

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \max_{i=1,...,N} \ell(x, \delta^{(i)}), \tag{4.1}$$

where $\mathcal{X}$ is a convex and closed set, and $\delta^{(1)}, ..., \delta^{(N)}$ are instances of the uncertainty variable δ independently generated according to a probability measure $\mathbb{P}_\Delta$. We have seen that the worst-empirical cost c^*, i.e. the optimal value of (4.1), allows for a probabilistic characterization of x^* in a distribution-free manner. In particular, if N is suitably chosen, relation $\ell(x^*, \delta) \leq c^*$ holds with probability $1 - \epsilon$ with respect to δ (with very high confidence $1 - \beta$). That is, c^* is a cost guaranteed with probability $1 - \epsilon$ when decision x^* is made. It turns out that this "suitable N" is inversely proportional to ϵ and is proportional to d, the number of components in the optimization variable x, i.e. N scales as $\frac{1}{\epsilon} \cdot d$, see (3.3) on page

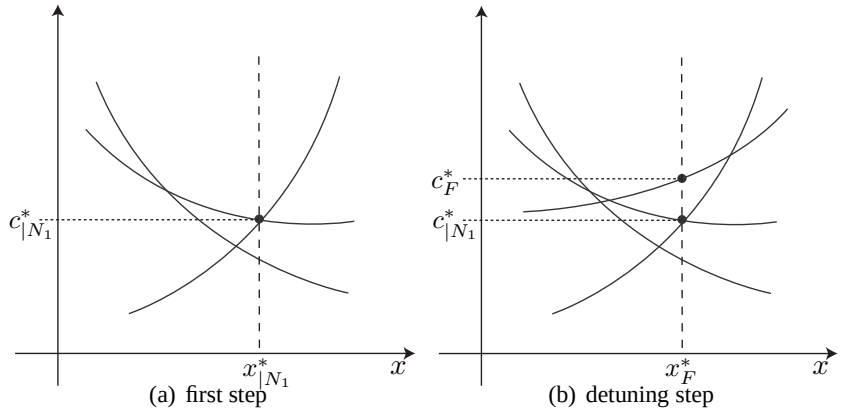
56. However, as noted also in [90, 91], this dependence on ϵ and d may result in too many scenarios for large scale problems with large d, thus posing a difficulty in practice. In fact, we may not have enough scenarios at our disposals. Moreover, even if we can sample an arbitrary number of scenarios, it can be hard in practice to solve the min-max problem with so many scenarios for computational reasons, since it involves solving a convex problem with so many constraints. In both cases, we would like to make a decision with an associated guaranteed cost based on a smaller amount of scenarios than that required by the "classical" decision-making algorithm.

In the present chapter a modified version of the worst-case decision-making algorithm, called FAST (Fast Algorithm for the Scenario Technique), is introduced in order to get around this difficulty. FAST associates to a min-max decision x_F^* a cost c_F^* still having coverage no less than $1-\epsilon$, i.e such that $\ell(x_F^*, \delta) \leq c_F^*$ holds with probability $1-\epsilon$, with a sample complexity N that exhibits a dependence on ϵ and d of the form $\frac{1}{\epsilon} + d$. This significantly reduces the sample complexity in large scale optimization problems.

4.1.1 The idea behind FAST

FAST operates in two steps. First, a moderate number N_1 of scenarios $\delta^{(i)}$ are considered and problem (4.1) with $N = N_1$ is solved so generating a solution $x^*_{|N_1}$ and an optimal value $c^*_{|N_1}$, refer to Figure 4.1(a). This first step is carried out at a low effort due to the moderate number N_1 of scenarios involved. On the other hand, $\ell(x^*_{|N_1}, \delta) \leq c^*_{|N_1}$ is not guaranteed with the desired probability level $1-\epsilon$ since N_1 is too low for this to happen. Then, a *detuning* step is started where N_2 additional scenarios are considered and the smallest value c_F^* such that $\ell(x^*_{|N_1}, \delta^{(i)}) \leq c_F^*$, $i = 1, \ldots, N_1 + N_2$, is computed, see Figure 4.1(b). The algorithm returns the solution $x_F^* = x^*_{|N_1}$ and the value c_F^*. The theory in Section 4.2 shows that $\ell(x_F^*, \delta) \leq c_F^*$ holds with the desired probability $1-\epsilon$. In this construction, N_1 and N_2 scale as d and $\frac{1}{\epsilon}$ respectively, leading to an overall number of scenarios $N = N_1 + N_2$ that is typically much smaller than that required by the classical worst-case approach. Moreover, choosing a small ϵ does not affect N_1 and only results in a large N_2 value, which corresponds to having many scenarios in the detuning step, a step that is a simple detuning procedure that can be solved efficiently even for large values of N_2.

The remainder of the chapter is organized as follows. In next Section 4.2, the FAST algorithm is presented in detail. In Section 4.2.1 theoretical results are presented, and a discussion about the practical use of FAST follows in Section 4.2.2. The proofs are in Section 4.3. In the Appendix B, an extension of FAST to the more general set-up presented in Appendix A is provided. A simulation example, in the general case, is also given in Appendix B.4.



(a) first step (b) detuning step

Figure 4.1. Illustration of FAST. Each line represents a function $\ell(x, \delta^{(i)})$.

4.2 The FAST algorithm

We maintain here for simplicity the assumption that any problem of the form

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \max_{i=1,\dots,m} \ell(x, \delta^{(i)})$$

has a unique solution for any m and any choice of $\delta^{(1)}, \dots, \delta^{(m)}$. Moreover, the following notation is in force to express shortly the fundamental Beta distribution function:

$$B_\epsilon^{N,d} := \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}. \tag{4.2}$$

The FAST algorithm follows

The FAST algorithm

- INPUT:
  - $\epsilon \in]0,1[$, risk parameter;
  - $\beta \in]0,1[$, confidence parameter;
  - N_1, an integer such that $N_1 \geq d+1$.

1. Compute the smallest integer N_2 such that

$$N_2 \geq \frac{\ln \beta - \ln B_\epsilon^{N_1,d}}{\ln(1-\epsilon)}, \tag{4.3}$$

   where $B_\epsilon^{N_1,d}$ is as in equation (4.2).

2. Sample N_1+N_2 independent constraints $\delta^{(1)}, \dots, \delta^{(N_1)}, \delta^{(N_1+1)}, \dots, \delta^{(N_1+N_2)}$, according to $\mathbb{P}_\Delta$.

3. Solve problem (4.1) with $N = N_1$; let $x^*_{|N_1}$ be the solution.

4. (Detuning step) Compute

$$c_F^* := \max_{i=1,\dots,N_1+N_2} \ell(x^*_{|N_1}, \delta^{(i)}).$$

- OUTPUT:

  · $(x_F^*, c_F^*) := (x^*_{|N_1}, c_F^*)$.

4.2.1 Theoretical results

Consider the risk of the cost c_F^* defined as

$$R_F := \mathbb{P}_\Delta\{\delta \in \Delta : \ell(x_F^*, \delta) > c_F^*\}.$$

Clearly, R_F is a random variable that depends on the samples $\delta^{(1)}, \dots, \delta^{(N_1+N_2)}$. The following theorem bounds the probability that $R_F > \epsilon$.

Theorem 9. *The following relation holds*

$$\mathbb{P}_\Delta^{N_1+N_2}\{R_F > \epsilon\} \leq (1-\epsilon)^{N_2} \cdot B_\epsilon^{N_1,d}. \tag{4.4}$$

$\star$

The proof of Theorem 9 is given in Section 4.3. It is a fact that the bound on the right-hand side of (4.4) is not improvable. Indeed, the following Theorem 10 holds for the class of problems satisfying the specialized fully-supportedness assumption. We recall that a min-max problem (3.1) satisfies the fully supported assumption (Assumption 2) if for all $N \geq d+1$, and with probability one with respect to the possible scenarios,

i) it has exactly $d+1$ support scenarios;

ii) for every $\gamma \in \mathbb{R}$, $\mathbb{P}_\Delta\{\ell(x^*, \delta) = \gamma\} = 0$.

For a discussion see page 60.

Theorem 10. *We have that relation*

$$\mathbb{P}_\Delta^{N_1+N_2}\{R_F > \epsilon\} = (1-\epsilon)^{N_2} \cdot B_\epsilon^{N_1,d} \tag{4.5}$$

holds under Assumption 2. $\star$

For a proof see Section 4.3.
Thus, Theorem 10 states that c_F^* is a distribution-free statistic, and, implicitly, that the bound in Theorem 9 is tight, i.e. it cannot be improved without further information on $\mathbb{P}_\Delta$ or the structure of the problem considered.

To prove the following main Theorem 11 from Theorems 9 and 10, let us consider the way N_2 is selected in point 1 of the FAST algorithm of Section 4.2. An easy computation shows that equation (4.3) is equivalent to

$$(1-\epsilon)^{N_2} \cdot B_\epsilon^{N_1,d} \leq \beta.$$

Thus, an application of Theorem 9 shows that

$$\mathbb{P}_\Delta^{N_1+N_2}\{R_F > \epsilon\} \leq \beta.$$

On the other hand, since N_2 is the smallest integer such that (4.3) holds, any $N_2' < N_2$ gives

$$(1-\epsilon)^{N_2'} \cdot B_\epsilon^{N_1,d} > \beta,$$

and, in light of Theorem 10, this implies that

$$\mathbb{P}_\Delta^{N_1+N_2'}\{R_F > \epsilon\} > \beta$$

when Assumption 2 is satisfied. We have proved the following theorem.

Theorem 11 (main theorem on the FAST algorithm)**.** *In the current set-up, it holds that*

$$\mathbb{P}_\Delta^{N_1+N_2}\{R_F > \epsilon\} \leq \beta. \tag{4.6}$$

Moreover, N_2 given in point 1 of the FAST algorithm cannot be improved in the sense that there are problems for which no N_2 smaller than that given in point 1 of the FAST algorithm makes (4.6) *true.* $\star$

4.2.2 Discussion

In the FAST algorithm, the user solves problem (4.1) with N_1 scenarios, and computes N_2 through (4.3). N_1 is decided by the user, while N_2 depends on N_1, ϵ, and β. In this section, guidelines on how to select N_1, and a handier formula for N_2, are provided. Moreover, the pros and cons with using FAST are also discussed.

Selection of N_1

Computational reasons suggest that N_1 should be chosen as small as possible, but other requirements also apply. If N_1 is too large, solving (4.1) for $x^*_{|N_1}$ becomes expensive so losing the advantages of using FAST. On the other hand, if N_1 is too small, $x^*_{|N_1}$ is poorly selected, and this in turn leads to a large cost value c_F^* after the detuning phase in FAST is carried out. As a rule of thumb out of empirical experience, we suggest to take $N_1 = 20d$. Notice that the theoretical result in Theorem 11 remains valid for any choice of N_1.

A handier formula for N_2

To a first approximation, in point 1 of the FAST algorithm, equation (4.3) can be substituted by the handier formula

$$N_2 \geq \frac{1}{\epsilon} \ln \frac{1}{\beta}. \tag{4.7}$$

In fact,

$$\frac{\ln \beta - \ln B_\epsilon^{N_1,d}}{\ln(1-\epsilon)} \leq \frac{\ln \beta}{\ln(1-\epsilon)} \leq \frac{1}{\epsilon} \ln \frac{1}{\beta},$$

showing that an N_2 satisfying (4.7) also satisfies (4.3). (4.7) is easier to apply than (4.3) since (4.3) also involves computing the term $B_\epsilon^{N_1,d}$.

Advantages with using FAST

Reduced sample size requirements

The FAST algorithm provides a cheaper way to find solutions to medium and large scale problems than the classical scenario approach. Indeed, one can choose $N_1 = Kd$, where K is a user-selected number normally set to 20, while, using (4.7), N_2 can be taken as the first integer bigger than or equal to $\frac{1}{\epsilon} \ln \frac{1}{\beta}$. Hence, a handy formula to estimate the overall number of scenarios needed with FAST is

$$Kd + \frac{1}{\epsilon} \ln \frac{1}{\beta}.$$

A comparison with the evaluation of the sample complexity in the classic approach (see (3.3) in Chapter 3), i.e.

$$N \geq \frac{e}{e-1} \frac{1}{\epsilon} \left(d + \ln \frac{1}{\beta} \right),$$

shows the key point that, with FAST, the critical multiplicative dependence on $\frac{1}{\epsilon} \cdot d$ is replaced by an additive dependence on $\frac{1}{\epsilon}$ and d.

Possibility to reduce ϵ to small values

The detuning step 4 of FAST is a simple maximization problem. Therefore, running step 4 with a large N_2 can be done at low computational effort so that ϵ can be reduced to values much smaller than with the classical scenario approach.

Suboptimality of FAST

Figure 4.2 represents the solution obtained using FAST. $c^*_{|N_1}$ is the cost value for the problem with N_1 scenarios, and c^*_F is the cost value after the introduction of N_2 extra scenarios in the detuning step. In white is the region above all cost functions $\ell(x, \delta^{(i)})$, $i = 1, \ldots, N_1 + N_2$. To achieve the same level of risk as in FAST, with the classical scenario approach additional scenarios have to be introduced, so reducing the white region in which (x^*, c^*), the cost-decision pair computed according to classical approach, will have to lie. From an inspection of Figure

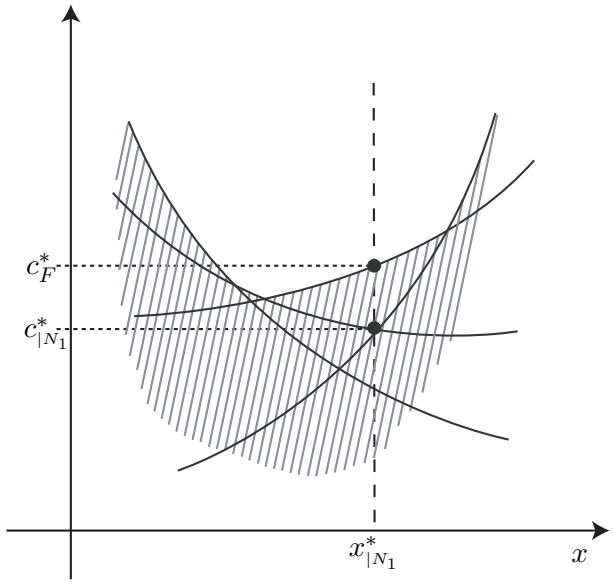


Figure 4.2. Comparison between FAST and the classical worst-case approach.

4.2 it appears that the classical approach may outperform FAST, that is, it may happen that $c^* < c_F^*$. If so, however, it certainly holds that $c_F^* - c^* < c_F^* - c_{|N_1}^*$. Consequently, the decision-maker has a simple way to evaluate the potential suboptimality of FAST by computing $c_F^* - c_{|N_1}^*$. Empirical evidence shows that c_F^* and c^* are often close to each other, and suboptimality is negligible.

4.3 Proofs

Theorems 9 and 10 are proved together in the following Section 4.3.1.

4.3.1 Proof of Theorems 9 and 10

Recall that Δ is the uncertainty domain where the random variable δ takes value, and define, for brevity, $\boldsymbol{\delta}_m^n := (\delta^{(m)}, \delta^{(m+1)}, ..., \delta^{(n)})$, so that $\boldsymbol{\delta}_m^n \in \Delta^{n-m+1}$.

We want to compute the probability of set

$$H := \left\{ \boldsymbol{\delta}_1^{N_1+N_2} : R_F > \epsilon \right\}.$$

Now, consider, for any given pair (x, c), $x \in \mathbb{R}^d$ and $c \in \mathbb{R}$, the violation probability function

$$V(x, c) := \mathbb{P}_\Delta\{\delta \in \Delta : \ell(x, \delta) > c\}.$$

With this notation, $R_F = V(x_F^*, c_F^*)$. For a given x_F^*, consider the set

$$L := \{c \in \mathbb{R} : V(x_F^*, c) > \epsilon\}.$$

L is a random set, depending on $\boldsymbol{\delta}_1^{N_1}$ through $x_F^* = x_{|N_1}^*$. Once $x_{|N_1}^*$ is fixed, $1 - V(x_{|N_1}^*, c)$, as a function of c, is clearly the cumulative distribution function of the random variable $\ell(x_{|N_1}^*, \delta)$. Hence, $V(x_{N_1}^*, c)$ is right-continuous and non-increasing in $\mathbb{R}$, entailing that L can be written as $L =]-\infty, \bar{c}[$. The following property provides a useful characterization of set H.

Property 1. *$\boldsymbol{\delta}_1^{N_1+N_2} \in H$ if and only if $V(x_{|N_1}^*, c_{|N_1}^*) > \epsilon$ and $\ell(x_{|N_1}^*, \delta^{(i)}) \in L$, $\forall i \in \{N_1+1, ..., N_1+N_2\}$.* $*$

Proof. At the detuning step 4, the FAST algorithm computes

$$c_F^* = \max_{i=1,\ldots,N_1+N_2} \ell(x_{|N_1}^*, \delta^{(i)}),$$

i.e. $c_F^* = \max\{c_{|N_1}^*, \max_{i=N_1+1,\ldots,N_1+N_2} \ell(x_{|N_1}^*, \delta^{(i)})\}$. If $V(x_{|N_1}^*, c_{|N_1}^*) > \epsilon$, we have $c_{|N_1}^* < \bar{c}$. If $\ell(x_{|N_1}^*, \delta^{(i)}) \in L$, $\forall i \in \{N_1+1, ..., N_1+N_2\}$, we have $\max_{i=N_1+1,\ldots,N_1+N_2} \ell(x_{|N_1}^*, \delta^{(i)}) < \bar{c}$. Thus, we have $c_F^* < \bar{c}$, i.e. $c_F^* \in L$, when both conditions hold true simultaneously, yielding $\boldsymbol{\delta}_1^{N_1+N_2} \in H$. Vice versa, if $V(x_{|N_1}^*, c_{|N_1}^*) \leq \epsilon$ we have $c_{|N_1}^* > \bar{c}$, so that $c_F^* \geq c_{|N_1}^* > \bar{c}$, i.e. $c_F^* \notin L$ and $\boldsymbol{\delta}_1^{N_1+N_2}$ is not in H; on the other hand, if $\ell(x_{|N_1}^*, \delta^{(\bar{i})}) \notin L$ for some $\bar{i} \in \{N_1+1, ..., N_1+N_2\}$ we have $\ell(x_{|N_1}^*, \delta^{(\bar{i})}) > \bar{c}$, so that $c_F^* \geq \ell(x_{|N_1}^*, \delta^{(\bar{i})}) > \bar{c}$, i.e. $c_F^* \notin L$ and $\boldsymbol{\delta}_1^{N_1+N_2}$ is not in H. □

Based on Property 1 we proceed now to evaluate the probability of H:

$$\begin{aligned}
&\mathbb{P}_\Delta^{N_1+N_2}\{H\} \\
&= [\mathbb{1}\{\cdot\} = \text{indicator function}] \\
&= \int_{\Delta^{N_1+N_2}} \mathbb{1}\{V(x_{|N_1}^*, c_{|N_1}^*) > \epsilon \text{ and } \ell(x_{|N_1}^*, \delta^{(i)}) \in L, \\
&\qquad\qquad \forall i \in \{N_1+1, \ldots, N_1+N_2\}\} \mathbb{P}_\Delta^{N_1+N_2}\{\mathrm{d}\boldsymbol{\delta}_1^{N_1+N_2}\} \\
&= \int_{\Delta^{N_1+N_2}} \mathbb{1}\{V(x_{|N_1}^*, c_{|N_1}^*) > \epsilon\} \cdot \mathbb{1}\{\ell(x_{|N_1}^*, \delta^{(i)}) \in L, \\
&\qquad\qquad \forall i \in \{N_1+1, \ldots, N_1+N_2\}\} \mathbb{P}_\Delta^{N_1}\{\mathrm{d}\boldsymbol{\delta}_1^{N_1}\} \mathbb{P}_\Delta^{N_2}\{\mathrm{d}\boldsymbol{\delta}_{N_1+1}^{N_1+N_2}\} \\
&= [\text{using Fubini's theorem}] \\
&= \int_{\Delta^{N_1}} \mathbb{1}\{V(x_{|N_1}^*, c_{|N_1}^*) > \epsilon\} \\
&\quad \left[\int_{\Delta^{N_2}} \mathbb{1}\{\ell(x_{|N_1}^*, \delta^{(i)}) \in L, \forall i \in \{N_1+1, \ldots, N_1+N_2\}\} \mathbb{P}_\Delta^{N_2}\{\mathrm{d}\boldsymbol{\delta}_{N_1+1}^{N_1+N_2}\}\right] \\
&\quad \mathbb{P}_\Delta^{N_1}\{\mathrm{d}\boldsymbol{\delta}_1^{N_1}\}. \qquad (4.8)
\end{aligned}$$

As we show below in this proof, the inner integral in the square brackets is upper-bounded by $(1-\epsilon)^{N_2}$ for any $\boldsymbol{\delta}_1^{N_1}$, and it is exactly equal to $(1-\epsilon)^{N_2}$ when Assumption 2 is satisfied. Therefore, $\mathbb{P}_\Delta^{N_1+N_2}\{H\}$ is upper-bounded as follows

$$\mathbb{P}_\Delta^{N_1+N_2}\{H\} \leq (1-\epsilon)^{N_2} \int_{\Delta^{N_1}} \mathbb{1}\{V(x^*_{|N_1}, c^*_{|N_1}) > \epsilon\}\mathbb{P}_\Delta^{N_1}\{\mathrm{d}\boldsymbol{\delta}_1^{N_1}\}. \tag{4.9}$$

The integral in (4.9) is $\mathbb{P}_\Delta^{N_1}\{V(x^*_{N_1}, c^*_{N_1}) > \epsilon\}$, that is, the complementary distribution function, evaluated at ϵ, of the risk of $c^*_{|N_1}$ when the min-max decision $x^*_{|N_1}$ is made based on N_1 scenarios. According to (3.4), this quantity is upper-bounded by $B_\epsilon^{N_1,d}$, while it is exactly equal to $B_\epsilon^{N_1,d}$ whenever Assumption 2 is satisfied, see (3.5). Thus, from (4.9) we conclude that

$$\mathbb{P}_\Delta^{N_1+N_2}\{H\} \leq (1-\epsilon)^{N_2} \cdot B_\epsilon^{N_1,d},$$

which is the statement of Theorem 9, while, if Assumption 2 is satisfied, we have equality, i.e.

$$\mathbb{P}_\Delta^{N_1+N_2}\{H\} = (1-\epsilon)^{N_2} \cdot B_\epsilon^{N_1,d},$$

and Theorem 10 is proved.

To complete the proof we have to evaluate the inner integral in (4.8).

In what follows, we take a fixed $\boldsymbol{\delta}_1^{N_1}$ - so that $x^*_{N_1}$ is fixed - and the result is proved by working conditionally with respect to $\boldsymbol{\delta}_1^{N_1}$.

By the independence of the samples,

$$\int_{\Delta^{N_2}} \mathbb{1}\{\ell(x^*_{|N_1}, \delta^{(i)}) \in L, \forall i \in \{N_1+1, \ldots, N_1+N_2\}\}\mathbb{P}_\Delta^{N_2}\{\mathrm{d}\boldsymbol{\delta}_{N_1+1}^{N_1+N_2}\} \tag{4.10}$$

$$= \left(\int_\Delta \mathbb{1}\{\ell(x^*_{|N_1}, \delta) \in L\}\mathbb{P}_\Delta\{\mathrm{d}\delta\}\right)^{N_2}$$

$$= \left(\mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) \in L\}\right)^{N_2}$$

$$= \left(\mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) < \bar{c}\}\right)^{N_2}. \tag{4.11}$$

By Assumption 2.*ii*, $\mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) = c\} = 0$ so that $V(x^*_{|N_1}, c) = \mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) > c\}$ is a continuous function of c. Since $\bar{c}$ is the extreme point of the set where $V(x^*_{|N_1}, c) > \epsilon$, by continuity it follows that $V(x^*_{|N_1}, \bar{c}) = \epsilon$. Hence, $\mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) < \bar{c}\} = \mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) \leq \bar{c}\} = 1 - \mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) > \bar{c}\} = 1 - V(x^*_{|N_1}, \bar{c}) = 1 - \epsilon$ and the right-hand side of (4.10) equals $(1-\epsilon)^{N_2}$. If Assumption 2 does not hold, we prove that $\mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) < \bar{c}\} \leq 1 - \epsilon$. To this end, define the sets $L_n := \left]-\infty, \bar{c} - \frac{1}{n}\right]$ for $n > 1$. Clearly, $L_n \subseteq L$, and $\mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) \in L_n\} = 1 - V\left(x^*_{|N_1}, \bar{c} - \frac{1}{n}\right) < 1 - \epsilon$. Applying the σ-additivity of $\mathbb{P}_\Delta$, we conclude that

$$\mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) \in L\} = \mathbb{P}_\Delta\{\ell(x^*_{|N_1}, \delta) \in \cup_{n=1}^\infty L_n\}$$

$$= \lim_{n\to\infty}\left[1 - V\left(x^*_{|N_1}, \bar{c} - \frac{1}{n}\right)\right] \leq 1 - \epsilon$$

and the right-hand of (4.10) is upper-bounded by $(1-\epsilon)^{N_2}$. □

4.4 Conclusion and perspectives for future work

In this chapter we have presented a decision-making algorithm that is a modification of the classical data-based min-max algorithm, working with reduced sample complexity. Indeed, the obtained cost threshold c_F^* for the FAST solution x^* has a risk that concentrates on smaller values than does the classical worst-case cost c^* for the classical worst-case decision x^*. Moreover, we have proved that c_F^* is a distribution-free coverage statistic for a whole (non-pathological) class of problems. The main idea behind FAST is that of exploiting redundancy in the data. In particular, in FAST this is done by solving two problems in cascade: at the first stage a classical min-max decision problem is solved and x_F^* obtained; at the second stage a simple one-dimensional optimization is performed, leading to the guaranteed cost c_F^*. From a theoretical point of view, the most interesting fact about FAST is the possibility of *exactly* computing the distribution of the risk of the two-step solution (x_F^*, c_F^*), thus suggesting that ideas similar to that behind FAST can be used to design decision-making algorithms with *tight* guarantees on risks.

A possible difficulty with FAST arises when the decision-maker incurs an unacceptable suboptimality. We recall that suboptimality can be easily detected during the decision process by evaluating how large the difference $c_F^* - c_{|N_1}^*$ is. If $c_F^* - c_{|N_1}^*$ is too high, the decision-maker may want to improve its decision x_F^*, e.g. by iterating the FAST algorithm with a larger N_1. FAST can then be used as the building block for an iterative procedure where a limited number of iterations can be performed, thus allowing the user to trade the sample complexity with the quality of the decision. Clearly, the confidence in the risk being no higher than a given ϵ should be accordingly computed (or, at least, bounded by a simple application of the union bound).

In Appendix B the FAST algorithm is presented for more general convex *and constrained* optimization problem, and an example is given in that context.

Conclusions and future developments

We have studied the theoretical properties of decisions made according to two different data-based approaches: a least-squares approach and a worst-case approach with convex cost function. In particular, we have shown that it is possible to evaluate in a distribution-free manner, tightly and before any data is observed, the coverage probabilities of meaningful cost thresholds, which are associated with a data-based decision according to suitably defined rules. In the least-squares context, we have provided an algorithm to compute cost thresholds with guaranteed mean coverage, and we have shown that these thresholds are close to the empirical costs. In the worst-case approach, we have extended to all the empirical costs a known result about the exact probability distribution of the coverage of the worst empirical cost. We have introduced a version of the worst-case approach that allows for reliable decision-making even when data are few.

All our results hold under the hypothesis that data are independent and identically distributed (i.i.d.). Since we do not assume that the probability according to which data are generated is known to the decision-maker, we believe that, in the situations considered, we have provided useful theoretical tools to put data-based decision-making on a solid theoretical ground. Also, the theory presented can be applied to wider contexts than decision-making. For instance, the author of this thesis is particularly interested in the *model selection problem*, which is now briefly outlined.

The model selection problem

In Chapter 2, Example 4, we have mentioned an application of our theory to a regression problem, that is, to a *model fitting* problem. In a model fitting problem, the decision variable $x \in \mathbb{R}^d$ represents a model, characterized by d parameters, of the data $\mathsf{D}^N = \delta^{(1)}, \ldots, \delta^{(N)}$ observed, and the cost function $\ell(x, \delta)$ measures how badly a data point δ fits the model x. The best model x^* can be chosen according to the average approach or the worst-case approach studied in this work. Our theory allows us to associate with the model x^* a certificate about the reliability of the model x^*, i.e. a value $\mathbf{c}(\mathsf{D}^N)$ such that, for a new data point δ, it holds that $\ell(x, \delta) \leq \mathbf{c}(\mathsf{D}^N)$ with a guaranteed probability α. Different models can be obtained by selecting the best model from various (say k) model classes of increasing

complexity, e.g. $\mathcal{M}^{(1)} \subseteq \mathcal{M}^{(2)} \subseteq \cdots \subseteq \mathcal{M}^{(k)}$, thus obtaining k best-fitting models , $x^{(1)^*}, \ldots, x^{(k)^*}$. To each model a cost can be associated in the light of our theory, i.e. $\mathbf{c}^{(1)}(\mathsf{D}^N), \ldots, \mathbf{c}^{(k)}(\mathsf{D}^N)$, so that each of these costs is guaranteed at the same level of probability α. The comparison among $\mathbf{c}^{(1)}(\mathsf{D}^N), \ldots, \mathbf{c}^{(k)}(\mathsf{D}^N)$ provides a first criterion to select a good model among various classes of models. For a basic introduction to the model selection problem, see e.g. [13], Chapter 7. $\star$

Our work can be continued along several directions. The theory presented in Chapter 2, dealing with mean coverages, waits to be completed with theoretically sound results about other properties of the coverages, e.g. variances. Algorithms alternative to FAST, presented in Chapter 4, can be studied to face the problem of the dependence on the problem dimension in the theory of Chapter 3.
More sophisticated decision-making schemes can also be investigated. For example, a sliding window approach to the observed data can be useful in keeping under control contingent changes of $\mathbb{P}_\Delta$ with time, as well as allowing an interesting frequentist interpretation of the coverage properties, in the line of results for transductive predictors studied in [53].
The properties of decisions made according to iterative schemes are at present subject of research: an iterative scheme allows to update *on-line* a decision when new observations suggest that the decision can be improved. Finally, the study of more general cost functions (non-quadratic in the average approach and non-convex in the worst-case approach) and, from a more radical point of view, relaxations of the i.i.d. assumption constitute open and stimulating research areas. In this regard, the introduction of weak dependence assumptions, like β-mixing (see e.g. [43], Chapter 2, Section 5), is certainly worth being considered.

Appendix A

The scenario approach to constrained convex optimization problems

While, in our work, we have only considered convex min-max optimization problems, the theory of the scenario approach studied in [29, 30, 56, 92, 67, 44, 93, 57] is set-up for general convex constrained optimization problems. In this appendix, a brief overview of the main results of this theory is given. In Appendix B the FAST algorithm presented in Chapter 4 is extended to this general context.

A.1 General problem statement

Given a constant vector $r \in \mathbb{R}^{d+1}$, a convex and closed set $\mathcal{Z} \subseteq \mathbb{R}^{d+1}$ and a family of convex and closed sets $\mathcal{Z}_\delta$, parameterized in the uncertainty variable δ, consider the following constrained convex scenario program

$$\begin{aligned} &\min_{z \in \mathcal{Z} \subseteq \mathbb{R}^{d+1}} r^T z \\ &\text{subject to: } z \in \bigcap_{i=1,\dots,N} \mathcal{Z}_{\delta^{(i)}}, \end{aligned} \tag{A.1}$$

where $\delta^{(1)}, \dots, \delta^{(N)}$ are instances of δ independently sampled according to probability measure $\mathbb{P}_\Delta$. In a convex setting, linearity of the objective function is without loss of generality, since every convex program can be rewritten with linear objective, see e.g. [62]. Also, note that (A.1) generalizes the min-max problem:

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \max_{i=1,\dots,N} \ell(x, \delta^{(i)}). \tag{A.2}$$

In fact, (A.2) can be rewritten in epigraphic form as follows:

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

Hence, (A.2) is a particular case of (A.1) with $z = (x, c)$, $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$, $\mathcal{Z}_\delta = \{(x, c) : \ell(x, \delta) \leq c\}$, and $r^T = (0, 0, \ldots, 0, 1)$. In other words, a convex problem like (A.1) is an extension of the min-max problems discussed in our work, and it arises in modeling uncertain optimization where the feasible set $\mathcal{X}$ depends on δ, too. Note that (A.1) *is* more general than (A.2), *because* the cost function $\ell(x, \delta)$ is real-valued. If the cost function were defined as an extended real-valued function, i.e. $\ell : \mathcal{X} \times \Delta \to (\mathbb{R} \cup \{\pm\infty\})$, then (A.1) and (A.2) would be equivalent. Indeed, (A.1) can be formulated as a min-max problem by posing $x = z$, $\mathcal{X} = \mathcal{Z}$, $\ell(x, \delta) = +\infty$ for any $x \notin \mathcal{Z}_\delta$, and $\ell(x, \delta) = r^T x$ otherwise.

We are interested in quantifying the probability that z^*, the solution to (A.1), is violated by an unseen uncertainty instance δ, that is, we want to study $\mathbb{P}_\Delta\{\delta \in \Delta : \ z \notin \mathcal{Z}_\delta\}$. We give the following formal definition.

Definition 9 (violation probability)**.** *The* violation probability*, or just* violation*, of a given point $z \in \mathcal{Z}$ is defined as*

$$V(z) := \mathbb{P}_\Delta\{\delta \in \Delta : \ z \notin \mathcal{Z}_\delta\}.$$

$*$

Throughout, we will assume implicitly that, for any m and any choice of $\delta^{(1)}, \ldots, \delta^{(m)}$, any problem of the form

$$\begin{aligned} &\min_{z \in \mathcal{Z} \subseteq \mathbb{R}^{d+1}} r^T z \\ &\text{subject to: } z \in \bigcap_{i=1,\ldots,m} \mathcal{Z}_{\delta^{(i)}} \end{aligned} \tag{A.3}$$

is feasible and its feasibility domain has non-empty interior, and that the solution of (A.3) exists and is unique. This assumption is common in studying constrained convex problems. Relaxations of it are possible, in the line suggested in Section 2.1 of [56], but we will not consider them for simplicity. Note that in the min-max context, the condition of feasibility and non-empty interior is always satisfied.

A.2 Review of main results

We need a preliminary definition and a proposition.

Definition 10 (support scenario)**.** *For given scenarios $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, the scenario $\delta^{(r)}$, $r \in \{1, \ldots, N\}$, is called a support scenario for the optimization problem* (A.1) *if its removal changes the solution of* (A.1)*.*

$\star$

Proposition 2. *For every value of $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$, the number of support scenarios for* (A.1) *is at most $d + 1$, i.e. the number of optimization variables.*

$\star$

For a proof, see [94, 29].

A.2.1 Fundamental theorem

Following [56], we focus provisionally on situations where the following fully-supportedness assumption is satisfied.

Assumption 3 (fully-supportedness)**.** *Let consider* (A.1) *for all* $N \geq d+1$*. With probability one with respect to the extractions of samples* $\delta^{(1)}, \delta^{(2)}, \ldots, \delta^{(N)}$*, it holds that the optimization problem* (A.1) *has* exactly $d+1$ *support scenarios.*

$\star$

As shown in Chapter 3, the class of problems satisfying Assumption 3 is not empty, nor pathological. For this class, the following fundamental theorem holds true.

Theorem 12 ([56])**.** *Under Assumption 3, it holds that:*

$$\mathbb{P}_{\Delta}^{N}\{V(z^*) > \epsilon\} = \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}, \tag{A.4}$$

independently of $\mathbb{P}_{\Delta}$*.* $\star$

The equation (A.4) reads that for fully-supported problems the probability of seeing a "bad" sample $\mathsf{D}^N = \delta^{(1)}, \ldots, \delta^{(N)}$ such that $V(z^*) > \epsilon$ is exactly $\sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$. The right-hand side of (A.4) is the so-called incomplete Beta function ratio, see e.g. [70], that is, the violation $V(z^*)$ is a random variable having a *Beta distribution*, whatever $\mathbb{P}_{\Delta}$ is. When Assumption 3 is dropped, the distribution of $V(z^*)$ is still dominated by a Beta distribution, that is, the following theorem holds true.

Theorem 13 ([56])**.** *It holds that:*

$$\mathbb{P}_{\Delta}^{N}\{V(z^*) > \epsilon\} \leq \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}, \tag{A.5}$$

independently of $\mathbb{P}_{\Delta}$*.* $\star$

The bound in (A.5) is a bound valid irrespective of $\mathbb{P}_{\Delta}$, so that an application of Theorem 13 does not require knowledge of probability $\mathbb{P}_{\Delta}$. Moreover, result (A.5) is not improvable since the inequality $\leq$ in (A.5) becomes an equality $=$ for the class of the fully-supported problems.

When using (A.5), one can fix a value ϵ and an arbitrarily small confidence parameter β, and find the smallest integer N such that $\sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta$. Due to (A.5), this N entails that $\mathbb{P}_{\Delta}^{N}\{V(z^*) > \epsilon\} \leq \beta$, so that solving the optimization problem (A.1), based on N observed scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$, returns a solution such that $V(z^*) \leq \epsilon$ holds with (high) confidence $1-\beta$.

A.2.2 Explicit formulas

In [93], it is shown that every N satisfying

$$N \geq \frac{1}{\epsilon}\left(d + \ln\frac{1}{\beta} + \sqrt{2d\ln\frac{1}{\beta}}\right)$$

is such that $\sum_{i=0}^{d}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i} \leq \beta$, so that $\mathbb{P}_\Delta^N\{V(z^*) > \epsilon\} \leq \beta$. We here prefer to use the slightly less refined but more compact condition (also proved in [93])

$$N \geq \frac{e}{e-1}\frac{1}{\epsilon}\left(d + \ln\frac{1}{\beta}\right), \tag{A.6}$$

which still shows the fundamental fact that N has a logarithmic dependence on the confidence parameter β, and goes like $\frac{d}{\epsilon}$.

A.2.3 Expected value

Observe that for a problem with N scenarios and $d+1$ decision variables, the distribution of $V(z^*)$ has expected value equal to $\frac{d+1}{N+1}$, for fully-supported problems, while in general it holds that

$$\mathbb{E}_{\Delta^N}\left[V(z^*)\right] \leq \frac{d+1}{N+1}.$$

This result was first proved in [29] but can be derived from Theorem (A.5) as well.

A.2.4 Scenarios removal

Assume that, for any N scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$, we have a rule to remove k scenarios, say $\delta^{(i_1)}, \ldots, \delta^{(i_k)}$. Given N scenarios, we consider the solution $z^*_{N\setminus k}$ to the problem obtained by ignoring k scenarios, say $\delta^{(i_1)}, \ldots, \delta^{(i_k)}$, i.e. the solution to

$$\begin{aligned} &\min_{z\in\mathcal{Z}\subseteq\mathbb{R}^{d+1}} r^T z \\ &\text{subject to: } z \in \bigcap_{i\in\{1,\ldots,N\}\setminus\{i_1,\ldots,i_k\}} \mathcal{Z}_{\delta^{(i)}}. \end{aligned} \tag{A.7}$$

The rule to remove k scenarios is arbitrary, the only constraint we impose is that it must lead to a solution $z^*_{N\setminus k}$ that with probability 1 is violated by $\delta^{(i_1)}, \ldots, \delta^{(i_k)}$, i.e. it must hold true that $z^*_{N\setminus k} \notin \mathcal{Z}_{\delta^{(i_j)}}, j = 1, \ldots, k$. Under this condition, the following theorem, proved in [57], holds true.

Theorem 14 ([57])**.** *The violation probability of $z^*_{N\setminus k}$ can be bounded as follows:*

$$\mathbb{P}_\Delta^N\{V(z^*_{N\setminus k}) > \epsilon\} \leq \binom{k+d}{k}\sum_{i=0}^{k+d}\binom{N}{i}\epsilon^i(1-\epsilon)^{N-i},$$

independently of $\mathbb{P}_\Delta$.

This result allows the user for trading-off the probability that the solution is satisfied by the unseen uncertainty instances and the cost attained by the solution. Of no less importance is the fact that this result can be used to make the solutions more stable. Indeed, scenario solutions can be very different when different sets of scenarios are considered: removing "worst" scenarios reduces this variability.

A.3 Applications

The most important applications of the general theory here presented are in robust optimization (i.e. as probabilistic relaxations of robust problems) and chance-constrained optimization. See e.g. [29, 68, 79, 95] for applications of the theory here presented to robust control or input design, and [7] for an application to chance-constrained problems in finance. However, results here presented have found applications also in models for interval prediction, see [65], and a generalization of the mathematical machinery underlying Theorem 12 has been exploited in machine learning, to bound the probability of error of a classification algorithm presented in [96].

Appendix B

Generalized FAST algorithm

A generalized FAST algorithm can be applied to find a decision in the presence of uncertainty in the general constrained context of (A.1), see Appendix A. We will see that the generalized FAST algorithm produces, in two steps, a final decision z_F^* such that the distribution of the violation probability $V(z_F^*)$ can be kept under control, for relatively small N.

Before presenting the algorithm, a further assumption is needed: we have to assume that the user knows a "robustly feasible" point.

Assumption 4. *A point $\bar{z} \in \left(\bigcap_{\delta\in\Delta} \mathcal{Z}_\delta\right) \cap \mathcal{Z}$ is known to the user.* $*$

It is perhaps worth stating explicitly that there are no requirements on $\bar{z}$ other than it is robustly feasible, in particular there are no requirements on its performance value $r^T\bar{z}$. Assumption 4 is satisfied in many situations of interest. In particular, it is usually easy to check the feasibility of a "no action" solution: an example is robust feedback controller synthesis with bounded noise, as in [68], where one can take $\bar{z}$ corresponding to the zero controller set-up. Similarly, a suitable $\bar{z}$ can be found in applications as IPMs (Interval Predictor Models), see [65]. One way to search for a robustly feasible $\bar{z}$ in more general contexts is by sequential randomized algorithms, see e.g. [97, 98, 91].

In the following, the generalized FAST algorithm is given. The main theoretical result for the generalized FAST algorithm is presented in Section B.2, followed by a brief discussion in Section B.3. In Section B.4 a numerical example is given.

B.1 Generalized FAST algorithm

- INPUT:
  - $\epsilon \in]0, 1[$, violation parameter;
  - $\beta \in]0, 1[$, confidence parameter;
  - N_1, an integer such that $N_1 \geq d + 1$;
  - $\bar{z} \in \left(\bigcap_{\delta\in\Delta} \mathcal{Z}_\delta\right) \cap \mathcal{Z}$, a robustly feasible point.

1. Compute the smallest integer N_2 such that

$$N_2 \geq \frac{\ln \beta - \ln B_\epsilon^{N_1,d}}{\ln (1-\epsilon)}, \tag{B.1}$$

   where $B_\epsilon^{N,d} = \sum_{i=0}^{d} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}$.

2. Sample N_1+N_2 independent constraints $\delta^{(1)}, \ldots, \delta^{(N_1)}, \delta^{(N_1+1)}, \ldots, \delta^{(N_1+N_2)}$, according to $\mathbb{P}_\Delta$.

3. Solve problem (A.1) with $N = N_1$; let $z^*_{|N_1}$ be the solution.

4. (Detuning step) Let $\hat{z}[\alpha] := (1-\alpha) z^*_{N_1} + \alpha \bar{z}$, $\alpha \in [0,1]$, i.e. $\hat{z}[\alpha]$ describes the line segment
   connecting $z^*_{|N_1}$ with $\bar{z}$. Compute the solution α^* to the problem

$$\begin{aligned} &\min_{\alpha \in [0,1]} r^T \hat{z}[\alpha] \\ &\text{subject to: } \hat{z}[\alpha] \in \bigcap_{i=N_1+1}^{N_1+N_2} \mathcal{Z}_{\delta^{(i)}}. \end{aligned} \tag{B.2}$$

- OUTPUT:

  · $z^*_F := \hat{z}[\alpha^*]$.

B.2 Theoretical results

The violation of the solution z^*_F obtained with the generalized FAST algorithm is characterized by the following theorem.

Theorem 15 (main theorem on the generalized FAST algorithm)**.** *In the current set-up, it holds that*

$$\mathbb{P}_\Delta^{N_1+N_2}\{V(z^*_F) > \epsilon\} \leq \beta. \tag{B.3}$$

Moreover, N_2 given in point 1 of the generalized FAST algorithm cannot be improved in the sense that there are problems for which no N_2 smaller than that given in point 1 of the generalized FAST algorithm makes (B.3) *true.*

∗

A proof is given in Section B.5.

B.3 Discussion

The essential difference between the FAST algorithm of Section 4.2 and the generalized FAST algorithm of this section is the detuning step: the idea of raising $c^*_{|N_1}$ in the FAST algorithm is replaced in the generalized FAST algorithm by the idea of moving $z^*_{|N_1}$ towards $\bar{z}$. This operation can be performed at low computational effort since (B.2) is an optimization problem with a scalar decision variable α, so that (B.2) can be solved e.g. by means of bisection. Moreover, all observations in the discussion Section 4.2.2 can be carried over *mutatis mutandis* to the context of the present section.

Remark 5 (interpretation)**.** *Though mathematically analogue to the FAST algorithm of Chapter 4, the usage and interpretation of the generalized FAST algorithm here presented may differ substantially. Indeed, in the present general constrained context, we cannot always interpret a candidate solution z as a decision-cost pair: in general, a point $z \in \mathbb{R}^{d+1}$ represents a decision, with an associated cost $r^T z$ that may depend on all the $d+1$ components of z. Being the vector r deterministic, the question is not the uncertainty of the cost corresponding to z^*_F, but the fact that z^*_F may be unfeasible for a new instance of δ, so that, if $z^*_F \notin \mathcal{Z}_\delta$, the cost $r^T z^*_F$ looses its significance. Hence, $r^T z^*_F$ is still a guaranteed cost, but only because z^*_F is guaranteed to be ϵ-feasible.*

B.4 Numerical example

In this section, the classical scenario approach to convex problems is compared with FAST on an example.

B.4.1 Constrained convex scenario program

The following constrained convex problem with 200 optimization variables and uncertain LMI (Linear Matrix Inequality) constraints has no specific interpretation but resembles problems arising in robust control, see [99].

$$\begin{aligned} &\min_{z\in\mathbb{R}^{200}} \sum_{j=1}^{200} z_j \\ &\text{subject to: } \sum_{j=1}^{200} R_j(\delta^{(i)})B(\delta^{(i)})R_j(\delta^{(i)})^T z_j \preceq I, \quad i=1,\dots,N, \end{aligned} \tag{B.4}$$

where

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix};\ B(\delta^{(i)}) = \begin{bmatrix} \delta_1^{(i)} & \delta_2^{(i)} \\ \delta_2^{(i)} & \delta_3^{(i)} \end{bmatrix};$$

$$R_j(\delta^{(i)}) = \begin{bmatrix} \cos\left(2\pi\frac{j-1}{T(\delta^{(i)})}\right) & -\sin\left(2\pi\frac{j-1}{T(\delta^{(i)})}\right) \\ \sin\left(2\pi\frac{j-1}{T(\delta^{(i)})}\right) & \cos\left(2\pi\frac{j-1}{T(\delta^{(i)})}\right) \end{bmatrix},\ j = 1, ..., 200,$$

with $T(\delta^{(i)}) = 200 + 200^{2\delta_4^{(i)}}$, and $\delta^{(i)} = (\delta_1^{(i)}, \delta_2^{(i)}, \delta_3^{(i)}, \delta_4^{(i)})$ are sampled from $[0, 1]^4$ with uniform probability. $B(\delta^{(i)})$ is a stochastic matrix and $R_j(\delta^{(i)})$ is a rotation matrix whose period $T(\delta^{(i)})$ is also stochastic.

B.4.2 Classical approach vs FAST

Take $\epsilon = 0.01$ and $\beta = 10^{-9}$, i.e. we are interested in a scenario solution with a violation probability no more than 1%, with confidence $1 - 10^{-9}$.

In the classical approach, using (A.5) in Appendix A, we write

$$\sum_{i=0}^{199} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq 10^{-9},$$

which yields $N = 29631$. Solving (B.4) with $N = 29631$ yielded a cost value $\sum_{j=1}^{200} z_j^* = -1.052$. Turning to FAST, we took $N_1 = 4000$, and, according to (B.1), we obtained $N_2 = 2062$. Running (B.4) with $N = N_1 = 4000$ we obtained a solution $z^*_{|N_1}$ with cost value $\sum_{j=1}^{200} z^*_{|N_1,j} = -1.076$. Next, we selected $\bar{z} = 0$, so that $\hat{z}[\alpha] = (1-\alpha)z^*_{|N_1}$, and solved the detuning step with N_2 scenarios:

$$\begin{aligned} &\min_{\alpha\in[0,1]} (1-\alpha) \sum_{j=1}^{200} z^*_{|N_1,j} \\ &\text{subject to: } (1-\alpha) \sum_{j=1}^{200} R_j(\delta^{(i)}) B(\delta^{(i)}) R_j(\delta^{(i)})^T z^*_{|N_1,j} \preceq I, \\ &\qquad\qquad i = N_1 + 1, ..., N_1 + N_2. \end{aligned} \tag{B.5}$$

The optimal detuning value was $\alpha^* = 0.048$, yielding the final solution $z_F^* = (1-\alpha^*)z^*_{|N_1} = 0.952 z^*_{|N_1}$ with cost value $0.952 \cdot (-1.076) = -1.024$. Solving the problems by using cvx, [100], the total execution time with FAST was 20 times faster than with the classical scenario approach. With smaller values of ϵ, the comparison between the execution times is further unbalanced in favour of FAST, and FAST continues to offer a viable approach even for values of ϵ as small as 0.001 while the classical scenario approach becomes rapidly impractical as ϵ is let decrease.

B.5 Proof of Theorem 15

The proof of Theorem 15 follows the same line of reasoning as that of Theorems 9 and 10 in Chapter 4, see Section 4.3.1. For brevity, $\boldsymbol{\delta}_m^n := (\delta^{(m)}, \delta^{(m+1)}, ..., \delta^{(n)})$,

so that $\boldsymbol{\delta}_m^n \in \Delta^{n-m+1}$.
We want to compute the probability of set

$$H := \left\{ \boldsymbol{\delta}_1^{N_1+N_2} : V(z_F^*) > \epsilon \right\}. \tag{B.6}$$

Given $z^*_{|N_1}$, the solution z_F^* obtained by the generalized FAST algorithm lies on the half-line defined as $\hat{z}[\alpha] := (1-\alpha)z^*_{|N_1} + \alpha\bar{z}, \alpha \in]-\infty, 1]$: this half-line extends the line segment at point 4 of the generalized FAST algorithm in Section B.1 beyond point $z^*_{|N_1}$. The set Z of points on this half-line with a violation probability bigger than ϵ is formally defined as:

$$Z := \{\hat{z}[\alpha] : \alpha \in]-\infty, 1] \text{ and } V(\hat{z}[\alpha]) > \epsilon\}.$$

Since sets $\mathcal{Z}_\delta$ are convex and closed, $V(\hat{z}[\alpha])$ is right-continuous and nonincreasing in $\alpha \in]-\infty, 1]$. Hence, Z is an open half-line. In formulas, by defining

$$\bar{\alpha} := \sup_{\alpha \in]-\infty, 1]} \{\alpha : V(\hat{z}[\alpha]) > \epsilon\}, \tag{B.7}$$

Z can then be rewritten as

$$Z = \{\hat{z}[\alpha] : \alpha \in]-\infty, \bar{\alpha}[\}.$$

The following property provides a useful characterization of set H.

Property 2. *$\boldsymbol{\delta}_1^{N_1+N_2} \in H$ if and only if $V(z^*_{|N_1}) > \epsilon$ and $Z \cap \mathcal{Z}_{\delta^{(i)}} \neq \emptyset$,$\forall i \in \{N_1+1, ..., N_1+N_2\}$.*

*

This Property 2 can be proved similarly to Property 1 in Section 4.3.1, by observing that Z has here the same role as L in Section 4.3.1. Refer to Figure B.1 for a geometrical visualization of the various objects involved.

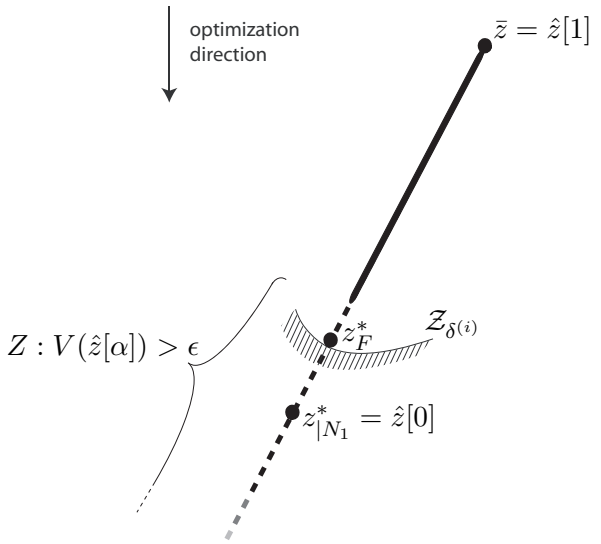


Figure B.1. Optimization domain for problem (B.2) in step 4 of the generalized FAST algorithm. The algorithm returns the point z_F^* closest to $z^*_{|N_1}$ and such that $z_F^* \in \mathcal{Z}_{\delta^{(i)}}, \forall i \in \{N_1+1, ..., N_1+N_2\}$. In this figure, set $\mathcal{Z}_{\delta^{(i)}}$ is the region above the shaded area, and $Z \cap \mathcal{Z}_{\delta^{(i)}} \neq \emptyset$.

Based on Property 2 and mimicking (4.8), the probability of H can be written as

$$\begin{aligned}
&\mathbb{P}_{\Delta}^{N_1+N_2}\{H\} \\
&\quad = \int_{\Delta^{N_1}} \mathbb{1}\{V(z^*_{|N_1}) > \epsilon\} \left[\int_{\Delta^{N_2}} \mathbb{1}\{Z \cap \mathcal{Z}_{\delta^{(i)}} \neq \emptyset, \forall i \in \{N_1+1, \ldots, N_1+N_2\}\} \right. \\
&\qquad\qquad \left. \mathbb{P}_{\Delta}^{N_2}\{\mathrm{d}\boldsymbol{\delta}_{N_1+1}^{N_1+N_2}\} \right] \mathbb{P}_{\Delta}^{N_1}\{\mathrm{d}\boldsymbol{\delta}_1^{N_1}\}.
\end{aligned}$$

By the independence of the samples, the inner integral in this latter equation can be written as

$$\begin{aligned}
&\int_{\Delta^{N_2}} \mathbb{1}\{Z \cap \mathcal{Z}_{\delta^{(i)}} \neq \emptyset, \forall i \in \{N_1+1, \ldots, N_1+N_2\}\} \mathbb{P}_{\Delta}^{N_2}\{\mathrm{d}\boldsymbol{\delta}_{N_1+1}^{N_1+N_2}\} \\
&\qquad = \left(\mathbb{P}_{\Delta}\{Z \cap \mathcal{Z}_{\delta} \neq \emptyset\}\right)^{N_2},
\end{aligned}$$

which, as we shall show below in this proof, is upper-bounded by $(1-\epsilon)^{N_2}$ for every $\boldsymbol{\delta}_1^{N_1}$ (it can also be proved that it is exactly $(1-\epsilon)^{N_2}$ whenever the sets $\mathcal{Z}_{\delta}$ satisfy a non-degeneracy condition). Thus, we conclude that

$$\begin{aligned}
&\mathbb{P}_{\Delta}^{N_1+N_2}\{H\} \\
&\qquad \leq (1-\epsilon)^{N_2} \int_{\Delta_{N_1}} \mathbb{1}\{V(z^*_{N_1}) > \epsilon\} \mathbb{P}_{\Delta}^{N_1}\{\mathrm{d}\boldsymbol{\delta}_1^{N_1}\} \\
&\qquad \leq (1-\epsilon)^{N_2} B_{\epsilon}^{N_1,d}, \qquad \text{(B.8)}
\end{aligned}$$

where the last inequality follows from Theorem 13 in Appendix A (for fully-supported problems it is an equality in light of Theorem 12). Theorem 15 follows by substituting in (B.8) the expression for N_2 given in (B.1).

The fact that $(\mathbb{P}_\Delta\{Z \cap \mathcal{Z}_\delta \neq \emptyset\})^{N_2} \leq (1-\epsilon)^{N_2}$ is now proved by working conditionally on a fixed $\boldsymbol{\delta}_1^{N_1}$, so that $\hat{z}[\alpha]$, $\alpha \in]-\infty, 1]$ has to be thought of as a fixed half-line. Define the sets

$$Z_n := \left\{\hat{z}[\alpha] : \alpha \in \left]-\infty, \bar{\alpha} - \tfrac{1}{n}\right]\right\},$$

for $n > 1$. Clearly, $\{\delta \in \Delta : Z_n \cap \mathcal{Z}_\delta \neq \emptyset\} = \{\delta \in \Delta : \hat{z}[\bar{\alpha} - \frac{1}{n}] \in \mathcal{Z}_\delta\}$, that is for $Z_n \cap \mathcal{Z}_\delta$ to be non empty, the extreme point $\hat{z}[\bar{\alpha} - \frac{1}{n}]$ of Z_n must be in $\mathcal{Z}_\delta$. Now, by the Definition 9 of violation probability, $\mathbb{P}_\Delta\{\delta \in \Delta : \hat{z}[\bar{\alpha} - \frac{1}{n}] \in \mathcal{Z}_\delta\} = 1 - V(\hat{z}[\bar{\alpha} - \frac{1}{n}])$, and by the σ-additivity of $\mathbb{P}_\Delta$, we have that

$$\begin{aligned}
&\mathbb{P}_\Delta\{Z \cap \mathcal{Z}_\delta \neq \emptyset\} \\
&\quad = \mathbb{P}_\Delta\left\{\bigcup_{n=1}^{\infty}\{Z_n \cap \mathcal{Z}_\delta \neq \emptyset\}\right\} \\
&\quad = \lim_{n\to\infty}\left[1 - V\left(\hat{z}\left[\bar{\alpha} - \tfrac{1}{n}\right]\right)\right] \\
&\quad \leq 1 - \epsilon,
\end{aligned}$$

where the last inequality follows from the fact that $V(\hat{z}[\bar{\alpha} - \frac{1}{n}]) > \epsilon, \forall n$, see (B.7). Thus, $(\mathbb{P}_\Delta\{Z \cap \mathcal{Z}_\delta \neq \emptyset\})^{N_2} \leq (1-\epsilon)^{N_2}$, and the theorem is proved. □

Bibliography

[1] A. Carè, S. Garatti, and M.C. Campi, "Randomized min-max optimization: The exact risk of multiple cost levels," in *Proceedings of the IEEE Conference on Decision and Control*, Orlando, Florida, USA, 2011.

[2] A. Carè, S. Garatti, and M.C. Campi, "FAST: An algorithm for the scenario approach with reduced sample complexity," in *Proceedings of IFAC 2011 World Congress*, Milano, Italy, 2011.

[3] A.N. Shiryaev, *Probability (2nd ed.)*, Springer-Verlag New York, Inc., Secaucus, New Jersey, USA, 1995.

[4] S.S. Wilks, "Order statistics," *Bulletin of the American Mathematical Society*, vol. 54, no. 1, pp. 6–50, 1948.

[5] D.A.S. Fraser and I. Guttman, "Tolerance regions," *The Annals of Mathematical Statistics*, vol. 27, no. 1, pp. 162–179, 1956.

[6] D. Bertsimas and A. Thiele, "Robust and data-driven optimization: Modern decision-making under uncertainty," in *Tutorials on Operations Research*. INFORMS, 2006.

[7] B. Pagnoncelli, D. Reich, and M. Campi, "Risk-return trade-off with the scenario approach in practice: A case study in portfolio selection," *Journal of Optimization Theory and Applications*, vol. 155, no. 2, pp. 707–722, 2012.

[8] M.C. Campi, "Why is resorting to fate wise? A critical look at randomized algorithms in systems and control," *European Journal of Control*, vol. 16, no. 5, pp. 419–430, 2010.

[9] S.K. Mitter, A. Nemirovski, and J. C. Willems, "Discussion on: "Why is resorting to fate wise? A critical look at randomized algorithms in systems and control"," *European Journal of Control*, vol. 16, no. 5, pp. 431–441, 2010.

[10] R.E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME-Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[11] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, no. 3, pp. 637–654, 1973.

[12] L. Ljung, *System Identification - Theory For the User*, Prentice Hall, Upper Saddle River, New Jersey, USA, 1999.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag New York, LLC, New York, USA, 2009.

[14] Z. Drezner, "Bounds on the optimal location to the Weber problem under conditions of uncertainty," *Journal of the Operational Research Society*, vol. 30, pp. 923–931, 1979.

[15] Z. Drezner and J. Guyse, "Application of decision analysis techniques to the Weber facility location problem," *European Journal of Operational Research*, vol. 116, no. 1, pp. 69–79, 1999.

[16] Z. Drezner and C.H. Scott, "On the feasible set for the squared euclidean Weber problem and applications," *European Journal of Operational Research*, vol. 118, no. 3, pp. 620–630, 1999.

[17] Z. Drezner, K. Klamroth, A. Schöbel, and G.O. Wesolowsky, "The Weber problem," in *Facility location - applications and theory*, Z. Drezner and H.W. Hamacher, Eds. Springer-Verlag New York, LLC, New York, USA, 2002.

[18] L.V. Snyder, "Facility location under uncertainty: A review," *IIE Transactions*, vol. 38, no. 7, pp. 547–564, 2004.

[19] B.D.O. Anderson and J.B. Moore, *Optimal Control: Linear Quadratic Methods*, Prentice Hall, Englewood Cliffs, New Jersey, USA, 1990.

[20] G.C. Calafiore and F. Dabbene, "Near optimal solutions to least-squares problems with stochastic uncertainty," *Systems and Control Letters*, vol. 54, no. 12, pp. 1219–1232, 2005.

[21] R.L. Plackett, "Studies in the history of probability and statistics. XXIX: The discovery of the method of least squares," *Biometrika*, vol. 59, no. 2, pp. 239–251, 1972.

[22] A.L. Soyster, "Convex programming with set-inclusive constraints and applications to inexact linear programming," *Operations Research*, vol. 21, no. 5, pp. 1154–1157, 1973.

[23] A. Ben-Tal and A. Nemirovski, "Robust convex optimization," *Mathematics of Operations Research*, vol. 23, no. 4, pp. 769–805, 1998.

[24] L. El Ghaoui and H. Lebret, "Robust solutions to uncertain semidefinite programs," *SIAM Journal on Optimization*, vol. 9, no. 1, pp. 33–52, 1998.

[25] A. Ben-Tal and A. Nemirovski, "Robust solutions of uncertain linear programs," *Operations Research Letters*, vol. 25, no. 1, pp. 1–13, 1999.

[26] L. El Ghaoui and S.-I. Niculescu, "Robust decision problems in engineering: a Linear Matrix Inequality approach," in *Advances in Linear Matrix Inequality Methods in Control*, L. El Ghaoui and S.-I. Niculescu, Eds. SIAM, 2000.

[27] D. Bertsimas and M. Sim, "The price of robustness," *Operations Research*, vol. 52, no. 1, pp. 35–53, 2004.

[28] D. Bertsimas and D.B. Brown, "Constructing uncertainty sets for robust linear optimization," *Operations Research*, vol. 57, no. 6, pp. 1483–1495, 2009.

[29] G.C. Calafiore and M.C. Campi, "Uncertain convex programs: randomized solutions and confidence levels," *Mathematical Programming*, vol. 102, no. 1, pp. 25–46, 2005.

[30] G.C. Calafiore and M.C. Campi, "The scenario approach to robust control design," *IEEE Transactions on Automatic Control*, vol. 51, no. 5, pp. 742–753, 2006.

[31] G.B. Dantzig, "Linear programming under uncertainty," *Management Science*, vol. 1, no. 3-4, pp. 197–206, 1955.

[32] A. Shapiro, "Monte Carlo sampling methods," in *Stochastic Programming*, A. Ruszczyński and A. Shapiro, Eds., vol. 10 of *Handbooks in Operations Research and Management Science*. Elsevier, London, New York and Amsterdam, 2003.

[33] J. Luedtke and S. Ahmed, "A sample approximation approach for optimization with probabilistic constraints," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 674–699, 2008.

[34] A. Charnes, W.W. Cooper, and G.H. Symonds, "Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil," *Management Science*, vol. 4, no. 3, pp. 235–263, 1958.

[35] A. Charnes and W.W. Cooper, "Chance constrained programming," *Management Science*, vol. 6, no. 1, pp. 73–79, 1959.

[36] A. Prékopa, *Stochastic Programming*, Kluwer, Boston, Massachusetts, USA, 1995.

[37] A. Prékopa, "Probabilistic programming," in *Stochastic Programming*, A. Ruszczyński and A. Shapiro, Eds., vol. 10 of *Handbooks in Operations Research and Management Science*. Elsevier, London, New York and Amsterdam, 2003.

[38] D. Dentcheva, "Optimization models with probabilistic constraints," in *Probabilistic and Randomized Methods for Design under Uncertainty*, G. Calafiore and F. Dabbene, Eds. Springer-Verlag, London, UK, 2006.

[39] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*, MPS-SIAM, Philadelphia, Pennsylvania, USA, 2009.

[40] H. Henrion and C. Strugarek, "Convexity of chance constraints with independent random variables," *Computational Optimization and Applications*, vol. 41, no. 2, pp. 263–276, 2008.

[41] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, USA, 1996.

[42] M. Vidyasagar, "Statistical learning theory and randomized algorithms for control," *IEEE Control Systems Magazine*, vol. 18, no. 6, pp. 69–85, 1998.

[43] M. Vidyasagar, *A Theory of Learning and Generalization*, Springer-Verlag New York, Inc., Secaucus, New Jersey, USA, 2nd edition, 2002.

[44] T. Alamo, R. Tempo, and E.F. Camacho, "Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2545–2559, 2009.

[45] V. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, 2001.

[46] T. Kanamori and A. Takeda, "Worst-case violation of sampled convex programs for optimization with uncertainty," *Journal of Optimization Theory and Applications*, vol. 152, no. 1, pp. 171–197, 2012.

[47] S. Boucheron, G. Lugosi, and O. Bousquet, "Concentration inequalities," in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds., vol. 3176 of *Lecture Notes in Computer Science*. Springer-Verlag Berlin and Heidelberg, Germany, 2003.

[48] L. El Ghaoui and H. Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM Journal on Matrix Analysis and Applications*, vol. 18, no. 4, pp. 1035–1064, 1997.

[49] H. Hindi and S. Boyd, "Robust solutions to l1, l2, and l-infinity uncertain linear approximation problems using convex optimization," in *Proceedings of the American Control Conference*, Philadelphia, Pennsylvania, USA, 1998.

[50] G.C. Calafiore, U. Topcu, and L. El Ghaoui, "Parameter estimation with expected and residual-at-risk criteria," *Systems and Control Letters*, vol. 58, no. 1, pp. 39–46, 2009.

[51] S.S. Wilks, *Mathematical Statistics*, Wiley, New York, USA, 1962.

[52] H.N. Nagaraja H.A. David, *Order Statistics, 3rd Edition*, Wiley, New York, USA, 2003.

[53] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, Springer-Verlag New York, Inc., Secaucus, New Jersey, USA, 2005.

[54] J.G. Saw, M.C.K. Yang, and T.C. Mo, "Chebyshev inequality with estimated mean and variance," *The American Statistician*, vol. 38, no. 2, pp. 130–132, 1984.

[55] U. Köyluoglu, A.S. Cakmak, and S.R.K. Nielsen, "Interval algebra to deal with pattern loading and structural uncertainty," *Journal of Engineering Mechanics*, vol. 121, no. 11, pp. 1149–1157, 1995.

[56] M.C. Campi and S. Garatti, "The exact feasibility of randomized solutions of uncertain convex programs," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1211–1230, 2008.

[57] M.C. Campi and S. Garatti, "A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality," *Journal of Optimization Theory and Applications*, vol. 148, no. 2, pp. 257–280, 2011.

[58] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[59] S.L. Campbell and C.D. Meyer, "The Moore-Penrose or generalized inverse," in *Generalized Inverses of Linear Transformations*. SIAM, Philadelphia, Pennsylvania, USA, 2009.

[60] T. Drezner and Z. Drezner, "The Weber location problem: the threshold objective," *INFOR*, vol. 49, no. 3, pp. 212–220, 2011.

[61] W.W. Hager, "Updating the inverse of a matrix," *SIAM Review*, vol. 31, no. 2, pp. 221–239, 1989.

[62] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[63] H.L. Harter, "Minimax methods," in *Encyclopedia of Statistical Sciences*, vol. 4, pp. 514–516. John Wiley & Sons, 1982.

[64] S. Garatti and M.C. Campi, "L-infinity layers and the probability of false prediction," in *Proceedings of the 15th IFAC Symposium on System Identification*, Saint Malo, France, 2009.

[65] M.C. Campi, G. Calafiore, and S. Garatti, "Interval predictor models: identification and reliability," *Automatica*, vol. 45, no. 2, pp. 382–392, 2009.

[66] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[67] B.K. Pagnoncelli, S. Ahmed, and A. Shapiro, "Sample average approximation method for chance constrained programming: theory and applications," *Journal of Optimization Theory and Applications*, vol. 142, no. 2, pp. 399–416, 2009.

[68] M.C. Campi, S. Garatti, and M. Prandini, "The scenario approach for systems and control design," *Annual Reviews in Control*, vol. 33, no. 2, pp. 149–157, 2009.

[69] R. Hochreiter, "An evolutionary computation approach to scenario-based risk-return portfolio optimization for general risk measures," in *Applications of Evolutionary Computing*, M. Giacobini, Ed., vol. 4448 of *Lecture Notes in Computer Science*. Springer-Verlag Berlin and Heidelberg, Germany, 2007.

[70] A.K. Gupta and S. Nadarajah, Eds., *Handbook of Beta Distribution and Its Applications*, Mercel Dekker, Inc., New York, USA, 2004.

[71] C. Baudrit and D. Dubois, "Practical representations of incomplete probabilistic knowledge," *Computational Statistics & Data Analysis*, vol. 51, no. 1, pp. 86–108, 2006.

[72] R.D. Gupta and D.S.P. Richards, "The history of the Dirichlet and Liouville distributions," *International Statistical Review*, vol. 69, pp. 433–446, 2001.

[73] H. Finner and M. Roters, "Multiple hypotheses testing and expected number of type I errors," *The Annals of Statistics*, vol. 30, no. 1, pp. 220–238, 2002.

[74] A. Gouda and T. Szántai, "New sampling techniques for calculation of Dirichlet probabilities," *Central European Journal of Operations Research*, vol. 12, no. 4, pp. 389–403, 2004.

[75] K.S. Kwong and Y.M. Chan, "On the evaluation of the joint distribution of order statistics," *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5091–5099, 2008.

[76] A. Gouda and T. Szántai, "On numerical calculation of probabilities according to Dirichlet distribution," *Annals of Operations Research*, vol. 177, no. 1, pp. 185–200, 2010.

[77] MATLAB, *version 7.10.0 (R2010a)*, The MathWorks Inc., Natick, Massachusetts, 2010.

[78] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, `http://www.R-project.org`.

[79] M.C. Campi and S. Garatti, "Variable robustness control: principles and algorithms," in *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems*, Budapest, Hungary, 2010.

[80] G. Blanchard, T. Dickhaus, N. Hack, F. Konietschke, K. Rohmeyer, J. Rosenblatt, M. Scheer, and W. Werft, "Mutoss - multiple hypothesis testing in an open software system," in *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 2010, vol. 11.

[81] MuToss Coding Team, "Package mutoss - unified multiple testing procedures, version 0.1-7," `http://cran.r-project.org/web/packages/mutoss/`, May 2012.

[82] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

[83] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized Algorithms for Analysis and Control of Uncertain Systems*, Springer, London, UK, 2005.

[84] A. Mutapcic, S.J. Kim, and S.P. Boyd, "Robust Chebyshev FIR equalization," in *Proceedings of the 50th IEEE Global Communication Conference (GLOBECOM '07)*, Washington, DC, USA, 2007.

[85] J.G. Proakis and M. Salehi, *Digital Communications*, McGraw-Hill, 2008.

[86] A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, Upper Saddle River, New Jersey, USA, 2010.

[87] D.R. Mazur, *Combinatorics: A Guided Tour*, MAA textbooks. Mathematical Association of America, Washington, DC, USA, 2009.

[88] H. Cramér and H. Wold, "Some theorems on distribution functions," *Journal of the London Mathematical Society*, vol. s1-11, no. 4, pp. 290–294, 1936.

[89] Z. Drezner, "On minimax optimization problems," *Mathematical Programming*, vol. 22, pp. 227–230, 1982.

[90] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 969–996, 2006.

[91] Y. Oishi, "Polynomial-time algorithms for probabilistic solutions of parameter-dependent Linear Matrix Inequalities," *Automatica*, vol. 43, no. 3, pp. 538–545, 2007.

[92] J. Luedtke and S. Ahmed, "A sample approximation approach for optimization with probabilistic constraints," *SIAM Journal on Optimization*, vol. 19, pp. 674–699, 2008.

[93] T. Alamo, R. Tempo, and A. Luque, "On the sample complexity of randomized approaches to the analysis and design under uncertainty," in *Proceedings of the American Control Conference (ACC)*, Baltimore, Maryland, USA, 2010.

[94] V.L. Levin, "Application of E. Helly's theorem to convex programming, problems of best approximation and related questions," *Sbornik: Mathematics*, vol. 8, pp. 235–247, 1969.

[95] J. Welsh and H. Kong, "Robust experiment design through randomisation with chance constraints," in *Proceedings of IFAC 2011 World Congress*, Milano, Italy, 2011.

[96] M.C. Campi, "Classification with guaranteed probability of error," *Machine Learning*, vol. 80, no. 1, pp. 63–84, 2010.

[97] B.T. Polyak and R. Tempo, "Probabilistic robust design with linear quadratic regulators," *Systems and Control Letters*, vol. 43, pp. 343–353, 2001.

[98] Y. Fujisaki, F. Dabbene, and R. Tempo, "Probabilistic design of LPV control systems," *Automatica*, vol. 39, no. 8, pp. 1323–1337, 2003.

[99] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, vol. 15 of *Studies in Applied Mathematics*, SIAM, Philadelphia, Pennsylvania, USA, 1994.

[100] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," `http://cvxr.com/cvx`, October 2010.