Distributionally robust control and optimization

Author(s):
Parys, Bart Paul Gerard Van

Publication Date:
2015

Permanent Link:
https://doi.org/10.3929/ethz-a-010553920

DISS. ETH NO. 22911

Distributionally robust control and optimization

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Ir. BART PAUL GERARD VAN PARYS
MSc, Katholieke Universiteit Leuven

born on 25.04.1988
citizen of Belgium

accepted on the recommendation of

Prof. Dr. Manfred Morari (examiner)
Prof. Dr. Dimitris Bertsimas (co-examiner)
Prof. Dr. Erick Delage (co-examiner)

2015

Extreme times call for extreme measures

Acknowledgements

Here lays before you the fruits of almost four years of work at the Control Laboratory at ETH Zurich. The successful completion of this four year journey hinged upon the interaction with many wonderful people too numerous to list exhaustively here. I would nevertheless like to make an effort to name those special individuals deserving a word of acknowledgement. It might be appropriate to start with naming Moritz Diehl in Leuven who, thanks to his contagious enthusiasm for optimization, motivated me to take an internship at IWR in Heidelberg and ultimately to pursue a Ph.D. in Zurich.

I am further very much indebted to my thesis supervisor Manfred Morari for his unfaltering believe in me. From our first meeting in the Irish college in Leuven to long after my defense, Manfred has entrusted me with invaluable advise and given me unprecedented academic freedom. I really appreciate his, sometimes underappreciated, ability to state the facts as they are. Due to his generous nature, the last four years were a wonderful experience for which I will be him for ever grateful.

If anyone is deserving to be denoted as my mentor it is surely Paul Goulart. Without his patient supervision and motivation throughout the years, this dissertation would not have been the same. During his time at the Control Laboratory, Paul furthermore helped me transform from a young undergraduate student to an experienced academic researcher. Also my time in Oxford visiting Paul will not be forgotten any time soon. I wish that one day I could mean the same to a young ambitious student what Paul meant to me. Paul, you have set the bar very high!

I am furthermore grateful to Daniel Kuhn whom I met for the first time at Imperial College London. He must for sure be the best host I have come across during my academic career. Both my visit to Imperial College London and EPF Lausanne will go down as extremely valuable from both an academical and from a personal perspective. I am sure that our collaboration will not stop with this dissertation.

Cheers to Bartolomeo Stellato for being an amazing master thesis student. I hope that I was able to pass over at least some of the world-class supervision I have received to him. I also sincerely want to pay tribute to Bartolomeo and his colleagues Dhruva Raman, Goran Banjac and Moritz Schulze Darup of the department of Engineering Science at Oxford for my wonderful time there.

These last four years have been made unforgettable by a great many colleagues at the Control Laboratory. I wish to thank Peyman Mohajerin Esfahani for our many vibrant discussions on the most diverse topics. Thanks are due to Claudia Fischer for organizing the weekly German lunches on Wednesdays. I am also grateful to Frauke Oldewurtel and Line Roald for pointing out that this dissertation has some practical use in power systems. Robert Nguyen and Joseph Warrington will always be remembered for the pleasant cycling trips. I am deeply indebted to Riccardo Porreca for introducing me to cross country skiing at Club Schwedentritt in Einsiedeln. Thanks to Marcello Colombino for organizing our ascent to the top of the Gran Paradiso and many trips to Chamonix.

Having lived the entirety of my Swiss adventure above the Sternen Apotheke in Oerlikon, I want to thank the amazing people I had the good fortune to have come across over the years. I am deeply grateful to my flat mate and Energy SmartOps collaborator Giampoalo Torrisi. I hope that after almost three years together he still feels the same way about me. The most sincere thanks are due to my neighbor Aurélie Eisenberg for helping me at the defense of this dissertation and her continuing patience with my German. The employees at Bar Cheyenne are recognized for their excellent service at what has become my safe haven in Oerlikon.

As with most acknowledgements, the most important ones are saved for last. I would like to thank my parents and family for their unconditional support. My special thanks goes to Julia, the only person who managed to read this dissertation more than once. I thank her for her love, patience and unwavering cheerfulness. Julia, you are my hero!

Bart P.G. Van Parys
Zurich,
November 2015

Abstract

The principal aim of this dissertation is to discuss and advance the use of distributionally robust constraints for decision-making in an uncertain environment. Distributionally robust constraints are the robust counterpart of uncertain constraints subject to a random outcome of which the distribution is only partially known. This dissertation will study in particular control and optimization problems for which these type of constraints constitute sensible design objectives. The contributions of this thesis fall into three categories each of which support the principal aim by taking down a hurdle which prevents its attainment. We briefly discuss the aforementioned categories and the contributions of this dissertation therein.

First, we argue that distributionally robust constraints present sound design objectives which are often more practically relevant than the classical worst-case or chance constrained alternatives. Distributionally robust constraints are often less pessimistic and do not require the distribution of the disturbances involved to be known exactly. In practice distributions are indeed never observed directly, but rather need to be estimated from noisy historical data. We consider two types of distributionally robust constraints in this dissertation. In the first type, we require that constraints hold with a given probability for all disturbance distributions consistent with the known partial distributional information. These constraints are referred to as distributionally robust chance constraints. In a second type of constraints, referred to as distributionally robust CVaR constraints, we additionally require the expected constraint violation to be small for all relevant disturbance distributions. Either constraint type is discussed and promoted as sound and sensible design objectives for both static optimization and dynamic control problems.

Second, in many interesting situations distributionally robust constraints are amendable to practical computation. The need for computational tools applicable to distributionally robust constraints naturally leads to the study of uncertainty quantification problems in which a probabilistic question needs to be answered using only limited statistical information. Uncertainty quantification problems find their roots in the classical univariate probability inequalities advanced by the Russian school of probability (Chebyshev, Markov, Lyapunov & Bernstein) and their origins can be traced back to the middle of the 19th century. In this dissertation these classical Chebyshev type inequalities are generalized to bounds on the probability of events in arbitrary dimensions based solely on second-order moment information. Instead of a closed form solution, these bounds are stated in terms of a tractable convex optimization problem. We discuss why Chebyshev type bound are achieved by pathological discrete distributions which render the corresponding inequalities overly pessimistic. In an attempt to exclude these practically irrelevant distributions, Gauss type probability inequalities and uncertainty quantification problems will be at the center of attention. In aforementioned Gauss bounds, the considered distributions are required to enjoy further structural properties which many practical distributions possess such as unimodality or monotonicity.

We indicate lastly that all discussed problems can be treated in a unified fashion and stated in the language of convex optimization. This dissertation indeed brings together and merges many relevant results in probability theory by unveiling their innate convex nature. By presenting a deep analogy between vectors in R^n and probability distributions on R^n, it will be argued that the same mathematical tools used in the analysis of classical worst-case robust constraints can be wielded in our distributionally robust setting equally well. Many results found in this dissertation concerning probability theory and uncertainty quantification problems have indeed a direct counterpart in either convex analysis or optimization, respectively.

Zusammenfassung

Das Hauptziel dieser Dissertation ist das Erörtern und Weiterentwickeln des Gebrauchs von verteilungsrobusten Nebenbedingungen in Hinsicht auf das Fällen von Entscheidungen in einem ungewissen Umfeld. Verteilungsrobuste Nebenbedingungen sind das robuste Gegenstück zu ungewissen Nebenbedingungen, welche entstehen, wenn die Verteilung nur teilweise bekannt ist. Diese Dissertation behandelt insbesondere Steuerungs- und Optimierungsaufgaben, bei welchen solide Zielvorgaben fr diese Art von Nebenbedingungen dargestellt werden. Die Beiträge dieser These fallen in drei Kategorien, jede einzelne mit dem Hintergrund die Hürde zu überwinden, welche das Ergebnis verhindert. Wir werden die genannten Kategorien folgend kurz erläutern.

Zu Beginn argumentieren wir, dass verteilungsrobuste Nebenbedingungen fundierte Zielvorgaben präsentieren, die oft praktisch relevanter sind als klassische worst-case- bzw. wahrscheinlichkeitsbedingte Alternativen. Verteilungsrobuste Nebenbedingungen sind oft weniger pessimistisch und benötigen keine genaue Strungsverteilung. Exakte Verteilungen werden in der Praxis allerdings nie direkt observiert, sondern müssen aus korrupten historischen Aufzeichnungen abgeschätzt werden. In dieser Dissertation untersuchen wir zwei Arten von verteilungsrobusten Nebenbedingungen. In der ersten Art fordern wir, dass die Nebenbedingungen mit einer gewissen Wahrscheinlichkeit standhalten fr alle Störungsverteilungen, übereinstimmend mit den zum Teil vorhandenen Informationen. Diese Nebenbedingungen werden bezeichnet als verteilungsrobuste Wahrscheinlichkeitsbedingungen. In einer zweiten Art, bezeichnet als verteilungsrobuste CVaR-bedingungen, benötigen wir zusätzlich, dass die voraussehende Nebenbedingungsverletzung gering ist für alle relevanten Störungsverteilungen. Beide Nebenbedingungen werden erörtert und weiterentwickelt als solide Zielvorgaben für statische Optimierung, als auch dynamische Steuerungsaufgaben.

Zweitens sind in vielen interessanten Situationen verteilungsrobuste Nebenbedingungen berechenbar. Die Notwendigkeit für Berechnungstools anwendbar für verteilungsrobuste Nebenbedingungen führt gewissermassen zur Studie über Unsicherheitsquantifizierungsaufgaben, in der eine wahrscheinliche Frage mit Anwendung von Teilinformationen beantwortet werden muss. Unsicherheitsquantifizierungsaufgaben finden ihre Wurzeln in den univariablen Wahrscheinlichkeitsungleichheiten, gefördert durch die Russische Schule der Wahrscheinlichkeit (Chebyshev, Markov, Lyapunov & Bernstein). In dieser Dissertation werden diese klassischen Wahrscheinlichkeitsungleichheiten auf Wahrscheinlichkeitsgrenzen in mehreren Dimensionen verallgemeinert, basierend auf lediglich zweitrangigen Moment Informationen. Anstelle von analytischen Lösungen sind diese Grenzen festgelegt in Form einer lenkbaren konvexen Optimierungsaufgabe. Wir diskutieren, warum Chebyshev-artige Grenzen durch pathologisch diskrete Verteilungen erzielt werden, was die dazugehörigen Ungleichheiten übermässig pessimistisch erweisen. In einem Versuch diese irrelevanten Verteilungen wegzulassen, werden Gauss-artige Wahrscheinlichkeitsungleichheiten und Unsicherheitsquantifizierungsaufgaben hier im Mittelpunkt stehen. In diesen Gauss-artigen Ungleichheiten geniessen die betrachteten Verteilungen weitere strukturelle Eigenschaften, welche viele praktische Verteilungen besitzen, darunter Unimodalität oder Monotonie.

Schliesslich deuten wir an, dass alle erörterten Probleme in einer einheitlichen Art und Weise behandelt, sowie in der Sprache der konvexen Optimierung angegeben werden können. Diese Dissertation bringt und verschmelzt viele relevante Resultate in der Wahrscheinlichkeitstheorie zusammen durch das enthüllen ihrer angeborenen konvexen Natur. Durch das Präsentieren einer tiefen Analogie zwischen Vektoren in R^n und Verteilungen auf R^n, wird argumentiert, dass dieselben mathematischen Tools, welche auch in der Analyse von klassischen worst-case robusten Nebenbedingungen eingesetzt werden, im verteilungsrobusten Umfeld gleichermassen gut gehandhabt werden können. Viele Resultate in dieser Dissertation betreffend Wahrscheinlichkeits-

theorie und Unsicherheitsquantifizierungsaufgaben haben ein direktes Spiegelbild in entweder konvexer Analyse oder Optimierung.

Notation

Scalar Sets

N	the natural numbers
R	the real numbers
R_+	the nonnegative real numbers
$\bar{\mathrm{R}}$	the extended real numbers: $\bar{\mathrm{R}} := \mathrm{R} \cup \{-\infty, \infty\}$

Vectors and Matrices

R^n	the space of vectors of length n
R^n_+	the set of element-wise positive vectors in R^n
S_n	the canonical simplex in R^n
$\mathrm{R}^{n\times m}$	the space of matrices of size n rows by m columns
S^n	the space of symmetric matrices in $\mathrm{R}^{n\times n}$
S^n_+	the set of symmetric positive definite matrices in $\mathrm{R}^{n\times n}$
$\langle y, x\rangle$	standard inner product $y^\top x$ of vectors x and y
$\mathrm{Tr}\,\{A\}$	trace of the matrix A
$A^\top$	transpose of the matrix A
$A^\dagger$	pseudo-inverse of matrix A
A_i	i-th column of the matrix A
$A \otimes B$	Kronecker product of matrices A and B
$\mathbb{I}_n$	identity matrix in $\mathrm{R}^{n\times n}$

Distributions and Random Variables

$\mathcal{E}_n$	the space of signed measures on R^n
$\mathcal{E}^\star_n$	the space of measurable functions on R^n
$\mathcal{P}_n$	the set of probability distributions in $\mathcal{E}_n$
$\mathcal{U}_\alpha$	the set of α-unimodal distributions in $\mathcal{P}_n$
$\mathcal{M}_\gamma$	the set of γ-monotone distributions in $\mathcal{P}_n$
$\mathbb{P}(B)$	the probability of the measurable event B
$\mathbf{E}_\mathbb{P}[L(\xi)]$	expectation of a function L with ξ distributed as $\mathbb{P}$: $\mathbf{E}_\mathbb{P}[L(\xi)] := \int L(x)\, \mathbb{P}(\mathrm{d}x)$
$\mathrm{supp}\, \mathbb{P}$	the support of a probability distribution $\mathbb{P}$

Definitions and Inequalities

$A := B$	A is defined by B
$A \leq B$	element-wise inequality between A and B
$A < B$	strict element-wise inequality between A and B
$A \preceq B$	matrix inequality between symmetric matrices: $B - A$ is positive semidefinite
$A \prec B$	strict matrix inequality between symmetric matrices: $B - A$ is positive definite

Topology and Sets

$\operatorname{conv} C$	convex hull of the set C
$\operatorname{ex} C$	extreme points of the convex set C
$K^\star$	dual cone of the set K
$\operatorname{int} C$	interior of the set C
$\operatorname{rint} C$	relative interior of the set C
$\operatorname{cl} C$	closure of the set C
$\operatorname{bd} C$	boundary of the set C
$\operatorname{dom} f$	effective domain of the function f
$\operatorname{epi} f$	epigraph of the convex function f

Set Operations

$A \cup B$	union of sets A and B
$A \cap B$	intersection of sets A and B
$A \setminus B$	difference of the set A with B

Elementary Functions

$\mathbf{1}\{B\}$	indicator function of the set B
κ_B	gauge function of the set $0 \in B$
$B(u,v)$	Euler integral of the first kind: $B(u,v) := \int_0^1 \lambda^{u-1} \cdot (1-\lambda)^{v-1} \, \mathrm{d}\lambda$
$\Gamma(t)$	Euler integral of the second kind: $\Gamma(t) := \int_0^\infty \lambda^{t-1} \cdot \mathrm{e}^{-\lambda} \, \mathrm{d}\lambda$
$\binom{n}{k}$	binomial coefficient

Acronyms

CVaR	Conditional Value-at-Risk
DLTI	Discrete Linear Time Invariant
LMI	Linear Matrix Inequality
LP	Linear Program
LQR	Linear Quadratic Regulator
MPC	Model Predictive Control
QP	Quadratic Program
SDP	Semidefinite Program
SOC	Second-Order Cone
SOCP	Second-Order Cone Program
SOS	Sum-Of-Squares
VaR	Value-at-Risk

Contents

1 Motivation

1.1 Uncertainty and robustness

The problem of taking decisions as to optimally influence an outcome of interest, despite the uncertain nature of the environment in which the decision is taken, is an important and well studied problem. In many practical problems the influence of exogenous inputs can not simply be neglected and naturally gives rise to questions concerning safety and robustness of decisions in the face of extreme events.

This dissertation is concerned with the problem of robust constraint satisfaction when faced with limited information regarding uncertain input data. The aforementioned problem is motivated by the fact that in many real-world problems constraint satisfaction is paramount. Constraint satisfaction might mean keeping budget in planning problems, or ensuring safe operation in process control. Depending on how critical a constraint is, and the nature of its uncertain influence, a robust formulation must be chosen. We will briefly discuss here which types of robust formulations are available and appropriate in what situations. Furthermore, we will motivate the approach taken in this dissertation which is particularly well suited when faced with ambiguity regarding the distribution of the uncertain influence. The notation in this chapter is intentionally informal to start, with a more rigorous treatment deferred to later in the dissertation.

The problem of taking a decision u such that an outcome $x(u, \xi)$ remains within a given constraint set X, despite an uncertain influence ξ, is a fundamental problem in many research areas and practical applications. Depending on his or her background, the reader can think of x and u as the states and inputs of an uncertain control system or alternatively as the returns and investments in a planning problem. As suggested by the title of this work, the exposition will be biased towards the optimization and control interpretation although the aforementioned problem comes about in a much wider variety of research areas.

We will motivate the road taken in this dissertation by focussing on the following canonical uncertain constraint

$$x(u, \xi) \in X \tag{1.1}$$

influenced by both the decision u and uncertainty ξ, i.e. the outcome should remain in the constraint set *in some sense*. Indeed, the uncertain constraint (1.1) is not made mathematically precise just yet. We first describe two standard methods for modeling such a constraint; the now classical worst-case formulation and the more recent chance-constrained formulation. We will discuss the alternative approach taken in this dissertation and argue that it alleviates some of the shortcomings inherent to the two standard methods. This alternative distributionally robust formulation will be particularly well suited when faced with only a limited amount of information regarding the distribution of the exogenous influence ξ.

1.1.1 Worst-case constraints

The worst-case formulation starts by assuming that the support of the uncertain influence ξ is bounded and known, i.e. that the uncertainty ξ is restricted to realize within a compact set C. The uncertain constraint (1.1) is then interpreted as a condition that the uncertain outcome $x(u, \xi)$ should be an element of the constraint set X for all realizations of the uncertain influence within its support C. The uncertain constraint (1.1) thus translates to

$$\forall \xi \in C : \quad x(u, \xi) \in X. \tag{1.2}$$

Taking the decision u such that the uncertain outcome $x(u, \xi)$ remains in the given constraint set X, for all possible realizations of the uncertain influence ξ, is historically the most prevalent formulation of the uncertain constraint (1.1).

In the optimization literature the robust formulation considered here goes back to 1973 with the work of Soyster [120] on robust linear programming. This formulation was however mostly neglected in the optimization community for several decades thereafter due to a lack of numerical tools able to deal with the resulting robust constraints. Around the second millennium, the worst-case formulation has witnessed an explosive interest in both its theory and practical applications by virtue of the seminal works of Ben-Tal and Nemirovski [7] and El Ghaoui et al. [46] which provided tractable reformulations for constraint (1.2) for many interesting situations. The reader is referred to further works of Ben-Tal and Nemirovski [9, 10] and Bertsimas et al. [15] plus the many references therein to get an idea of the staggering research activity surrounding this worst-case formulation.

In the field of control, worst-case formulations of the type (1.2) are of long standing interest as well as illustrated by the early work of Witsenhausen [135] and Bertsekas [14]. With the advent of optimization based control strategies such as model predictive control (MPC) surveyed by for instance Garcia et al. [53] and approximate dynamic programming (ADP) for which Powell [104] is a standard reference, a similar surge of interest in the worst-case reformulation (1.2) can be noted in the control community as well. As the constraints in control applications are often safety critical, worst-case constraints of the type (1.2) indeed often present an appropriate reformulation. Stability constraints match particularly well with the worst-case formulation as loss of stability should be avoided at all costs. The reader is referred to the small selection of works by Bertsekas and Rhodes [14], Blanchini [25], Kerrigan [67], De Farias and Van Roy [42] and Mayne et al. [84] for a necessarily incomplete perspective on the field of worst-case robust control.

The popularity of the worst-case formulation draws largely from the fact that in many interesting situations the constraint (1.2) admits an exact and tractable reformulation in terms of a convex optimization problem as shown by Ben-Tal et al. [6]. The applicability of the powerful results in convex analysis and optimization to the worst-case constraint (1.2) can in fact be regarded as the key to its popularity.

The worst-case formulation (1.2) almost invariably requires that the support of the uncertain influence ξ is completely known and compact. The boundedness assumption may be quite restrictive, e.g. in cases where the uncertain influence follows a normal distribution and hence has unbounded support. Furthermore, Bertsimas and Sim [19] argued that the worst-case formulation (1.2) is often quite pessimistic and usually comes at the high price of significant loss of optimality when compared to the nominal problem. Where for safety critical or hard constraints this optimality loss can be justified, in others we may benefit from a more forgiving formulation.

1.1.2 Chance constraints

Chance constraints require that the uncertain outcome $x(u, \xi)$ realizes within the constraint set X only up to a specified probability level. The uncertain constraint (1.1) is then modeled as the requirement

$$\mathbb{P}(x(u, \xi) \in X) \geq 1 - \epsilon, \tag{1.3}$$

with $\epsilon \in (0, 1)$ the prescribed safety level and where the distribution $\mathbb{P}$ of the uncertain influence ξ is assumed known. The chance constraint formulation is primarily aimed at soft constraints for which a small number of violations might be regarded as acceptable. Chance constraints are in those situations often more practical than their worst-case counterpart (1.2) which can be seen as a degenerate chance constraint with $\epsilon = 0$ and which tends to encourage overly pessimistic decisions. Chance constraints can in sharp contrast to the worst-case formulation (1.2) furthermore readily deal with unbounded support of the uncertain influence ξ.

The concept of chance constraints was introduced already in 1955 in Dantzig's original publication [40] and hence predates the worst-case formulation by almost two decades. Chance constrained optimization has received significant attention in the optimization community ever since; see for instance the work by Charnes et al. [35], Miller and Wagner [87] and Prekopa [105]. In control applications the chance constrained formulation (1.3) has received attention as well in the works of Schwarm and Nikolaou [114], Cannon et al. [33] and Oldewurtel et al. [93] in particular in the context of stochastic MPC.

Although no boundedness assumption is required on the support of the probability distribution $\mathbb{P}$, chance constraints are arguably worse from a practical perspective since they require the availability of a probability distribution over the disturbances. In practice one usually resorts to some simplifying assumption regarding the disturbance ξ, e.g. that it is normal distributed with known mean and variance. Unfortunately, verifying a chance constraint in the form (1.3) is intractable under generic distributions, i.e. checking (1.3) even for a fixed decision u and given distribution $\mathbb{P}$ is intractable. Indeed, Shapiro and Nemirovski [88] point out that computing the probability of a weighted sum of uniformly distributed variables being non-positive is already NP-hard. As a consequence, recently the attention in the works by Calafiore [28, 29] and Campi [32] has shifted towards stochastic sampling methods, for which only probabilistic guarantees can typically be provided, e.g. that the chance constraint condition (1.3) holds only with some level of confidence.

In this dissertation we will take an approach intermediate to the two extremes presented up so far. Our goal is to provide a framework that addresses the uncertain constraint (1.1) using only partial information about the distribution $\mathbb{P}$ of the uncertain influence ξ, and without recourse to sampling.

1.1.3 Distributionally robust constraints

In many situations the distribution $\mathbb{P}$ of the uncertain influence ξ is unknown and must be estimated from historical data, and hence is ambiguous. We therefore assume only that the distribution $\mathbb{P}$ belongs to an ambiguity set $\mathcal{C}$ of distributions. The ambiguity set $\mathcal{C}$ should ideally be composed of all distributions consistent with the available information regarding the uncertain influence ξ. The distributionally robust counterpart of the chance constraint (1.3) hence becomes

$$\forall \mathbb{P} \in \mathcal{C} : \quad \mathbb{P}\left(x(u, \xi) \in X\right) \geq 1 - \epsilon. \tag{1.4}$$

The constraint (1.4) is referred to as a distributionally robust chance constraint on the uncertain outcome $x(u, \xi)$ following Calafiore and El Ghaoui [30]. Such a constraint is a robust version of the classical chance constraint (1.3) in that it is immunized to any distribution $\mathbb{P}$ from within the ambiguity set $\mathcal{C}$.

The classical worst-case (1.2) and chance constrained formulation (1.3) can be seen as special instances of the distributionally robust chance constraint (1.4) for an ambiguity set $\mathcal{C}$ consisting of all probability distributions supported on the set C or where the ambiguity set $\mathcal{C} = \{\mathbb{P}\}$ reduces to a singleton, respectively. The distributionally robust constraint (1.4) thus covers the entire spectrum between the two extreme interpretations given to the uncertain constraint (1.1) in either the classical worst-case or chance constrained formulation.

We will take it as an objective in the remainder of the work to show that this distributionally robust interpretation of the uncertain constraint (1.1) constitutes a mathematically sound constraint specification amenable to practical computation. We stress that all of the numerical methods we present for dealing with such constraints are deterministic. This in contrast to the stochastic methods presented by Calafiore [28, 29] and Campi [32], for which only probabilistic admissibility guarantees can be provided.

Distributionally robust constraints were considered in the optimization community only recently by Calafiore et al. [30] and Zymler et al. [142], but have nevertheless already received considerable interest ever since. One of the main advantages of the distributionally robust formulation over the classical chance constrained formulation is the fact that only partial information on the distribution $\mathbb{P}$ is required. Furthermore as shown by Zymler et al. [142], the resulting robust formulation (1.4) is in many interesting situations computationally favorable over its nominal chance constrained counterpart (1.3). In the control community some early work on chance constraints with ambiguous distributions has been done by Lagoa and Barmish [3, 73, 72] in which the ambiguity set $\mathcal{C}$ considered consisted of all symmetric distributions sharing a unimodal structural property with known rectangular support. Despite the aforementioned early work the distributional robust formulation lay dormant for more than a decade and was only reconsidered very recently again by this author in [131].

Before we can hope to present any tractable reformulations of distributionally robust chance constraints, we must first be able to answer more fundamental questions such as the problem of feasibility in the distributionally robust constraint (1.4). The problem of deciding feasibility of a fixed decision u in the distributionally robust chance constraint (1.4) is equivalent to the condition

$$(1.4) \iff \sup_{\mathbb{P} \in \mathcal{C}} \ \mathbb{P}\left(x(u, \xi) \notin X\right) \leq \epsilon. \tag{1.5}$$

The left hand side of the inequality in equivalence (1.5) consists of the supremum of the probability $\mathbb{P}\left(x(u, \xi) \notin X\right)$ over an ambiguity set $\mathcal{C}$ of the distributions of the uncertain influence ξ. This type of optimization problem over a set of distributions will be referred to in this dissertation as an uncertainty quantification problem. Indeed, the optimization problem in (1.5) is recognized to provide the best upper bound on the probability of the event $x(u, \xi) \notin X$ given merely partial information on the distribution of ξ represented through the ambiguity set $\mathcal{C}$. From condition (1.5) it is clear that if one cherishes any hope of providing tractable reformulations of the distributionally robust chance constraint (1.4) the uncertainty quantification problem (1.5) must be amendable to tractable computation as well.

Uncertainty quantification problems such as (1.5) are unfortunately not known to admit closed form expressions in general. An exception worth mentioning are the univariate probability bounds discussed in the subsequent section. A primary aim of this dissertation is to generalize these so called classical probability bounds to problems in which the event space is of arbitrary dimension. We will furthermore indicate that for many interesting uncertainty quantification problems, the exact same tools used in the reformulation of the classical worst-case formulation (1.4) are applicable in the context of distributionally robust constraints equally well.

1.2 Classical probability inequalities

Distributionally robust constraints are thus preceded, by virtue of equivalence (1.5), by uncertainty quantification problems $\sup_{\mathbb{P}\in\mathcal{C}} \mathbb{P}(\xi \notin \Xi)$ quantifying the worst-case probability of events corresponding to those disturbances such that the uncertain outcome $x(u,\xi)$ realizes outside the constraint set X merely using the fact that the ξ is distributed within the ambiguity set $\mathcal{C}$. The uncertainty quantification problem $\sup_{\mathbb{P}\in\mathcal{C}} \mathbb{P}(\xi \notin \Xi)$ must hence be answered before its corresponding distributionally robust constraint (1.4) can be approached.

Observe that uncertainty quantification problems are intimately related to probability inequalities. Indeed, based on the partial information ξ distributed in $\mathcal{C}$ we can say at most that

$$\mathbb{P}(\xi \notin \Xi) \leq p \quad \text{for any} \quad p \geq \sup_{\mathbb{P}\in\mathcal{C}} \mathbb{P}(\xi \notin \Xi).$$

The probability inequality is denoted as tight if p is taken to be $\sup_{\mathbb{P}\in\mathcal{C}} \mathbb{P}(\xi \notin \Xi)$. The former type of tight probability inequalities with univariate uncertainty quantification problems in which ξ realizes in R have been studied since at least the 19th century and were advanced predominantly by the Russian school (Chebyshev, Markov, Lyapunov and Bernstein) of probability. We state these 19th century probability inequalities and bounds here with the promise to generalize them to events in arbitrary dimensions later on.

The Chebyshev inequality provides an upper bound on the tail probability of a univariate random variable based on limited moment information. The most common formulation of this inequality asserts that the probability that a random variable ξ valued in R with distribution $\mathbb{P}$ differs from its mean by more than κ standard deviations is bounded by

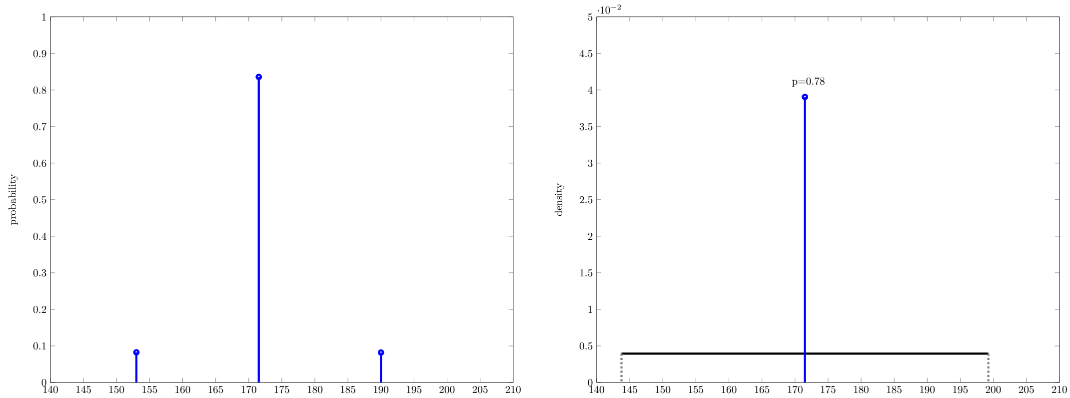
$$\sup_{\mathbb{P}\in\mathcal{H}(\mu,S)\cap\mathcal{P}_1} \mathbb{P}(|\xi-\mu| \geq \kappa\sigma) = \begin{cases} \frac{1}{\kappa^2} & \text{if } \kappa > 1, \\ 1 & \text{otherwise,} \end{cases} \tag{1.6}$$

where κ is a strictly positive constant, while μ and $S = \mu^2 + \sigma^2$ denote the mean and second moment of the random variable ξ distributed as $\mathbb{P}$, respectively. The Chebyshev bound is recognized as the solution to a particular uncertainty quantification problem in which the ambiguity set $\mathcal{H}(\mu,S)\cap\mathcal{P}_1$ consists of all univariate distributions $\mathbb{P} \in \mathcal{P}_1$ sharing a given mean μ and second moment S. In this special case a closed form solution for the optimization problem over distributions in (1.6) can be found.

The worst-case probability bound (1.6) was discovered by Bienaymé [20] in 1853 and proved by Chebyshev [36] in 1867. An alternative proof was offered by Chebyshev's student Markov [82] in 1884. The popularity of the Chebyshev inequality arises largely from its distribution-free nature. It holds for any distribution $\mathbb{P}$ under which ξ has mean μ and variance σ^2, and therefore can be used to construct robust confidence intervals for ξ relying exclusively on first and second-order moment information. Moreover, the inequality is sharp in the sense that, for any fixed κ, there exists a distribution $\mathbb{P}$ with given mean and variance achieving the worst-case bound (1.6). Unfortunately, the Chebyshev inequality may be quite pessimistic as the worst-case distributions achieving the bound (1.6) are of a degenerate nature and thus often practically irrelevant. The previous statement is best understood with the help of the fictitious example given in Figure 1.1(a).

In order to alleviate the pessimism inherent to the Chebyshev bound (1.6) an attempt can be made to exclude the pathological degenerate distributions from the ambiguity set $\mathcal{C}$. A property common to almost all practically relevant distributions is unimodality. Informally, a unimodal random variable has the intuitive property that small deviations from its mode should occur more frequently than large ones.

Definition 1.1 (Univariate unimodality [44])**.** *A univariate distribution $\mathbb{P}$ is called unimodal with mode c if the mapping $t \mapsto \mathbb{P}(\xi \leq t)$ is convex for $t < c$ and concave for $t > c$.*



(a) Chebyshev (b) Gauss

Figure 1.1: Consider a random variable ξ with mean $\mu = 171.5$ and standard deviation $\sigma = 7.5$. The Chebyshev bound (1.6) states that the probability of the event $|\xi - \mu| \geq 18.5$ is at most 16%. The Chebyshev inequality is tight in that it is achieved by the degenerate distribution shown in Figure 1.1(a) supported on a finite number of points. When unimodality around the mean μ is assumed, the Gauss bound (1.7) reduces the probability of the event $|\xi - \mu| \geq 18.5$ to at most 7.3%. The Gauss inequality is tight too in that it is achieved by the piece-wise uniform distribution in Figure 1.1(b). Both classical bounds can be compared to the probability obtained by assuming ξ to be distributed normally in which case the probability of the event of interest is 1.4%

In 1821 Gauss [54] proved that the classical Chebyshev bound (1.6) can be improved by a factor of 4/9 when the considered distributions are restricted to be unimodal $\mathbb{P} \in \mathcal{U}_1$ with mode $c = \mu$, that is,

$$\sup_{\mathbb{P} \in \mathcal{H}(\mu,S) \cap \mathcal{U}_1} \mathbb{P}(|\xi - \mu| \geq \kappa\sigma) = \begin{cases} \frac{4}{9\kappa^2} & \text{if } \kappa > \frac{2}{\sqrt{3}}, \\ 1 - \frac{\kappa}{\sqrt{3}} & \text{otherwise.} \end{cases} \tag{1.7}$$

The Gauss bound is thus recognized as the solution to an uncertainty quantification problem in which the ambiguity set $\mathcal{H}(\mu, S) \cap \mathcal{U}_1$ consists of all univariate unimodal distributions $\mathbb{P} \in \mathcal{U}_1$ sharing a given mean μ and second moment S. The Gauss bound (1.7) is again sharp and furthermore provides a much less pessimistic bound on the probability of the tail event $|\xi - \mu| \geq \kappa\sigma$ than its Chebyshev counterpart (1.6) when the random variable ξ is known to have a unimodal distribution; see Figure 1.1(b).

Since the 19th century several other probability inequalities have been discovered most of which try to include information other than merely the mean and standard deviation in an attempt to reduce the pessimism innate to the classical Chebyshev bound (1.6). The Pearson inequality [98] for instance states that the Chebyshev inequality can be adapted to include absolute moment information $\beta_r = \mathbf{E}_{\mathbb{P}}[|\xi - \mu|^r]$, that is

$$\mathbb{P}(|\xi - \mu| \geq \kappa\sigma) \leq \frac{\beta_r}{\sigma^r \kappa^r}. \tag{1.8}$$

Similarly, the Berge inequality [11] generalizes the Chebyshev inequality to bivariate random variables and reads

$$\mathbb{P}(|\xi_1 - \mu_1| \geq \kappa\sigma_1 \text{ or } |\xi_2 - \mu_2| \geq \kappa\sigma_2) \leq \frac{1 + \sqrt{1 - \rho_{12}^2}}{\kappa^2}, \tag{1.9}$$

where the correlation is in this context defined as $\rho_{12} = \sigma_{12}/(\sigma_1\sigma_2)$ for the covariance measure $\sigma_{12} = \mathbf{E}_{\mathbb{P}}[(\xi_1 - \mu_1)(\xi_2 - \mu_2)]$. Both probability inequalities (1.8) and (1.9) are tight when the right hand side is less than or equal to one in the same sense as discussed before. Some other probability type inequalities can be found in the literature too. An excellent starting point is a survey by Savage [112] on inequalities in probability theory in which many more inequalities are collected than stated here.

In all these exceptional cases a closed form expression for the corresponding uncertainty quantification problem was available. In general an uncertainty quantification problem does unfortunately not admit a closed form expression. Nevertheless, in this dissertation we will generalize all the aforementioned classical univariate probability bounds to arbitrary dimensions in terms of a tractable convex optimization problem.

1.3 Organization and highlights

A first important objective of this dissertation is hence to generalize the 19th century probability bounds of Chebyshev (1.6) and Gauss (1.7) to worst-case probabilities of events in arbitrary dimensions. The ultimate goal of the previous generalization being to facilitate the tractable reformulation of distributionally robust constraints of the type (1.4) by virtue of equivalence (1.5). Lastly, we will try to convince the reader in this dissertation that distributionally robust constraints and uncertainty quantification problems can be approached using the exact same tools as those used in case of the classical worst-case formulation (1.2).

This dissertation is consequently divided into three parts, each corresponding to one of the aforementioned points. In brief, the first part of this dissertation provides the necessary mathematical tools used to approach uncertainty quantification problems such as those appearing in (1.5). It will be argued that many concepts in probability theory enjoy an underlying convex structure and are thus amendable to the same tools used in standard robust optimization. The structure of the first part of the dissertation reflects the fact that uncertainty quantification problems can be recognized as optimization problems over convex sets of distributions. In Part II we will use convex analysis and optimization to study uncertainty quantification problems and by doing so generalize the classical probability inequalities discussed in the preceding section. Finally, in Part III we direct attention back to distributionally robust constraints and discuss their application in both optimization and control problems. The overall structure of this dissertation is pictorially represented in Figure 1.2.

In the remainder of this chapter we outline the structure and main contributions of this dissertation in greater detail. We also indicate which parts of the dissertation have been published before by the author, possibly in collaboration with others. In each of those works this author was however the principal investigator.

1.3.1 Mathematical tools

Convexity plays the protagonist role throughout the entire dissertation. Many sets of distributions in probability theory possess an underlying convex structure. However, sets of distributions reside in vector spaces more general than the finite dimensional space R^n. In the first part of this dissertation, we intend to show that the same tools can be used as for finite dimensional spaces nevertheless. Many results found in this dissertation concerning probability theory and uncertainty quantification problems have indeed a direct counterpart in either convex analysis or optimization in R^n.

The close similarity between vectors in R^n and distributions on R^n is made explicit by the intentional analogy between Chapters 2 and 3 dealing with convex analysis and Chapters 4 and 5 on optimization over convex sets of vectors and distributions, respectively. The notation throughout Chapters 2 to 5 is kept uniform as to facilitate the direct comparison between the results of convex analysis and convex optimization over either vectors in R^n or measures on R^n; see also Figure 1.2.

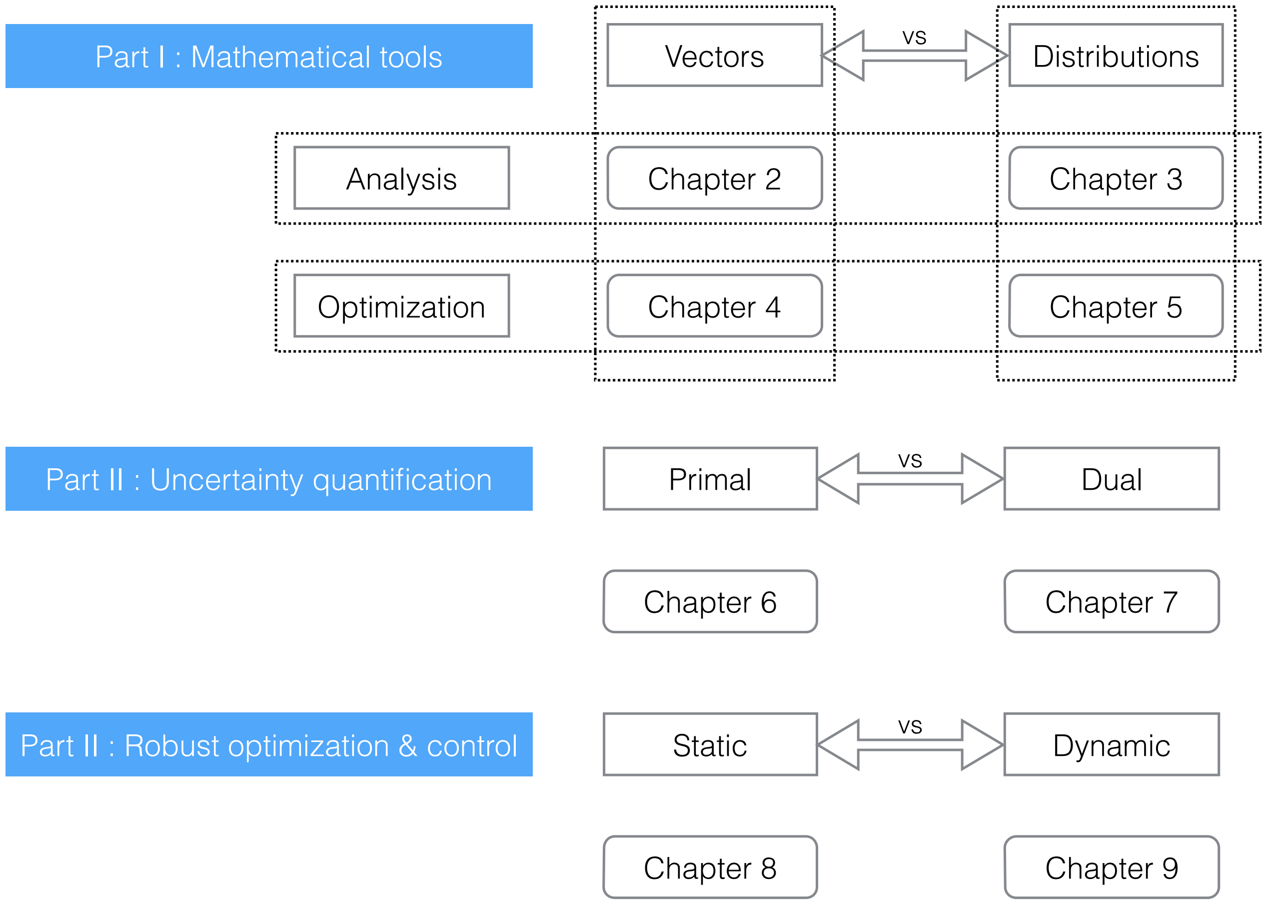


Figure 1.2: The structure of the dissertation. The first part intends to show that many convex sets of distributions in probability theory possess a convex structure and are very similar to convex sets of vectors. This close similarity between vectors and distributions is made explicit by the intentional analogy between Chapters 2 and 3 dealing with convex analysis and Chapters 4 and 5 on optimization over convex sets of vectors and distributions, respectively. Part II will discuss uncertainty quantification problems from either a primal perspective in Chapter 6 or a dual perspective in Chapter 7. The final part of the dissertation then investigates the use of distributionally robust constraints in either a static optimization context or in dynamic control problems.

By doing so we intend to convince the reader that essentially the same mathematical tools used to reformulate the standard worst-case robust constraint (1.2) apply to the distributionally robust formulation (1.4) as well. The only difference being that in the former we exploit the convexity of the set of possible realizations C of ξ in R^n and in the later the convexity of the set of possible distributions $\mathcal{C}$ of ξ on R^n. While this part of the dissertation contains almost no novel results, it facilitates the exposition of the remaining parts greatly and fixes the main ideas of the novel results found in all subsequent parts.

Chapter 2 The purpose of this chapter is twofold. We introduce the definitions of convex sets and functions in R^n both of which are fundamental to the dissertation. Particular emphasis is put on extreme points and Choquet representations of convex sets which will leave their mark throughout the remainder of the work. Additionally, this chapter will allude to the results of convex analysis for sets of distributions as discussed in the subsequent chapter. Indeed, many results in probability theory shall find their direct finite dimensional counterpart in this chapter.

Chapter 3 Probability distributions and ambiguity sets $\mathcal{C}$ are of critical importance to any discussion concerning distributionally robust constraints. This chapter brings together the necessary material from probability theory as required in the remainder of the dissertation. Many concepts and results in probability theory are shown to be intimately related to convexity and thus amendable to the same analysis put forward in Chapter 2. The notation in this chapter is intentionally chosen to parallel the notation introduced in Chapter 2 in order to emphasize the similarity between both chapters.

We will come across essentially two types of convex sets of distributions in this dissertation. The first type of convex sets $\mathcal{H}$ consists of measures sharing a finite number of given moments. The sets of measures $\mathcal{H}$ can be thought of as generalized hyperplanes in the space of measures on R^n. The prototypical moment set considered in this dissertation is the ambiguity set $\mathcal{H}(\mu, S)$ already encountered in the classical Chebyshev bound (1.6) consisting of all measures sharing first and second moments. Secondly, we will consider sets of probability distributions $\mathcal{K}$ enjoying a specific structural property such as unimodality or monotonicity. The set of all unimodal distributions $\mathcal{U}_n$ on R^n encountered in the classical Gauss bound (1.7) serves as an illustrative example to the latter kind of convex sets.

Extreme point or Choquet representations turn out to be of crucial importance in this dissertation. The sets $\mathcal{K}$ of the structural type are shown to admit explicit Choquet representations in terms of their radial extreme distributions in Propositions 3.2 and 3.3. Via Choquet star representable sets defined in Definition 3.14, we are able to put many different types of seemingly distinct structural requirements such as unimodality and monotonicity on an equal footing. Choquet star representable sets will serve us very well when analyzing uncertainty quantification problems over structured sets of probability distributions.

Lastly, we also introduce the value-at-risk (VaR) and conditional value-at-risk (CVaR) measure. The CVaR measure will come into play in the last part of the dissertation when trying to alleviate some of the problems inherent to the distributionally robust chance constrained formulation (1.4).

Chapter 4 This chapter will serve a dual purpose as well. The current chapter introduces the hierarchy of optimization problems ranging from linear programs (LPs) to semi-definite programs (SDPs). We point out that all classes of convex optimization problems within this optimization hierarchy can be solved efficiently and can thus be considered as de facto closed form expressions. Indeed, the solution to many problems in this dissertation will be stated in terms of a tractable convex optimization problem rather than as a closed form expression. This chapter will furthermore discuss the finite dimensional counterparts of the two central results discussed in the subsequent chapter dealing with optimization problems over sets of distributions.

Chapter 5 Uncertainty quantification problems of the type encountered in (1.5) can be regarded as specific instances of linear optimization problems over sets of distributions. The last chapter of this part will culminate in a discussion on how linear optimization problems over convex sets of distributions can be analyzed. The two main tools put forward in this chapter are (i) the fundamental theorem of linear programming and (ii) conic duality, both of which will play a key role in the remainder of this dissertation.

The fundamental theorem of linear programming 5.1 relates the geometry of the optimal distributions $\mathbb{P}^\star \in \mathcal{C}$ attaining the worst-case bound in (1.5) to the extreme points of the feasible set $\mathcal{C}$. The observation made in Figure 1.1 concerning the nature of the worst-case distribution achieving either the Chebyshev or Gauss bound is argued to be a direct consequence of the fundamental theorem.

A comprehensive duality theory can be developed based on pairing the space of measures on R^n with a dual space of measurable functions on R^n. Dual feasibility can be given a nice interpretation in terms of the positivity of dual functions. Strong duality is guaranteed by Theorem 5.2 under a very mild constraint qualification condition. To illustrate the power of the ideas presented in the first part of the dissertation, we show that the classical Gauss and Chebyshev inequalities can be proven and generalized easily within the presented framework using merely elementary manipulations.

1.3.2 Uncertainty quantification

In the second part of the dissertation we will, among other things, generalize the classical probability inequalities of Chebyshev (1.6) and Gauss (1.7) to worst-case probabilities of events in arbitrary dimensions. The resulting generalized Gauss type inequalities are an original contribution of this thesis. We will do so by considering the uncertainty quantification problem

$$\begin{aligned} B(L, \mathcal{K}, \mu, S) = \sup \quad & \int L(x)\,\mathbb{P}(\mathrm{d}x) \\ \text{s.t.} \quad & \mathbb{P} \in \mathcal{H}(\mu, S), \\ & \mathbb{P} \in \mathcal{K} \end{aligned} \tag{1.10}$$

for which the feasible set consists of all probability distributions in the structured set of distributions $\mathbb{P} \in \mathcal{K}$ sharing known second-order moment information $\mathbb{P} \in \mathcal{H}(\mu, S)$. The classical Chebyshev (1.6) and Gauss bound (1.6) can readily be seen to constitute special cases of the worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$ for a judicious choice of loss function L and structure $\mathcal{K}$.

As argued in Part I, the uncertainty quantification problem (1.10) can either be approached in a primal or an equivalent dual formulation. The structure of this part will follow this difference in perspective closely as Chapter 6 will take the primal perspective and Chapter 7 considers its dual. Nevertheless, a result central to both chapters is the fact that an uncertainty quantification problem over a structured set of distributions $\mathcal{K}$ can be transformed to an equivalent uncertainty quantification problem over the standard probability simplex $\mathcal{P}_n$ consisting of all distributions on R^n, i.e.

$$B(L, \mathcal{K}, \mu, S) = B(L_s, \mathcal{P}_n, \mu_s, S_s) \tag{1.11}$$

for a judiciously transformed loss function L_s, mean μ_s and second moment S. Theorems 6.1 and 7.1 prove the previous equivalence from both a primal and a dual perspective, respectively. The previous reduction is extremely beneficial to the exposition of both chapters in this part of the dissertation as only unstructured uncertainty quantification problems need be considered initially.

Although our interest in uncertainty quantification problems is mainly motivated by their close connection to distributionally robust constraints, the worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$

is of interest on its own. Recall that the structured set of distributions $\mathcal{K}$ can be used to represent additional structure enjoyed by many distributions in practice. We will illustrate the relevance of the bound $B(L, \mathcal{K}, \mu, S)$ to practical applications in the several numerical examples discussed in Chapters 6 and 7.

Chapter 6 This chapter analyzes the uncertainty quantification problem (1.10) starting from its primal formulation as a maximization problem over a set of distributions $\mathcal{H}(\mu, S) \cap \mathcal{K}$ on R^n. We initially derive exact tractable reformulations for uncertainty quantification problems when no structural assumptions are made ($\mathcal{K} = \mathcal{P}_n$). The central result in this chapter is stated in Theorem 6.2 which provides an exact tractable reformulation for the worst-case bound $B(L, \mathcal{P}_n, \mu, S)$ when the loss function is in the form

$$L(x) = \max_i \ell_i(x) \tag{1.12}$$

where the functions ℓ_i are understood to be all concave.

The reduction (1.11) can be used to deal with structured distributions ($\mathcal{K} \subset \mathcal{P}_n$) as well. In fact the class of worst-case bounds $B(L, \mathcal{P}_n, \mu, S)$, with L in the form (1.12), is indeed rich enough to generalize the classical Gauss bound ($\mathcal{K} = \mathcal{U}_1$) into arbitrary dimensions as shown in Theorem 6.5. Although attempts have been made before by Vandenberghe et al. [133] and Popescu [103], we are the first to obtain an exact and tractable representation of the Gauss bound in arbitrary dimensions. Furthermore, using a more flexible notion of unimodality we define in Theorem 6.4 a novel hierarchy of Gauss type bounds, all of which have a tractable representation, in which the Chebyshev and Gauss bounds are recognized as two extreme entities. This chapter is largely based on the results which appeared in the publication [128] of this author.

Chapter 7 This chapter approaches the uncertainty quantification problem (1.10) via its dual formulation as a minimization problem over the coefficients of positive functions on R^n. Again we initially derive exact tractable reformulations for uncertainty quantification problems when no structural assumptions are made ($\mathcal{K} = \mathcal{P}_n$). The central result in this chapter is found in Theorem 7.2 which provides a novel advantageous reformulation of the worst-case expectation bound $B(L, \mathcal{P}_n, \mu, S)$ when the loss function is in the form

$$L(x) = \max_i \ell_i(A_i x) \tag{1.13}$$

for arbitrary functions $\ell_i : \mathrm{R}^d \to \mathrm{R}$ in terms of a minimization problem over the coefficients of positive functions on R^d.

In the first part of the chapter, we indicate that many known worst-case probability and expectation bounds for unstructured sets of distributions ($\mathcal{K} = \mathcal{P}_n$) in the literature can be cast as corollaries of our Theorem 7.2. We then generalize the aforementioned results to structured sets of distributions ($\mathcal{K} \subset \mathcal{P}_n$) using again the reduction (1.11). The results in this chapter are largely based on the publication [130] by this author.

The main contributions, from a practitioners point of view, which can be found in this part of the thesis are collected in Table 1.1. We study essentially three types of uncertainty quantification problems and their corresponding bounds: (i) worst-case probability bounds, (ii) worst-case expectation bounds and (iii) worst-case CVaR bounds. Each mentioned problem type is approached through studying the uncertainty quantification problem (1.10) for a particular class of loss functions L in conjunction with a distinct structural assumption made through choice of the ambiguity set $\mathcal{K}$.

In order to examining the relative merits between the primal approach taken in Chapter 6 and the dual approach followed in this chapter, we could consider the classes of loss functions L either in the form (1.12) or (1.13) which can be dealt with effectively in the primal or dual

Structure	Worst-case probability	Worst-case expectation	Worst-case CVaR
Unstructured	Vandenberghe et al. [133]	Zymler et al. [142]	Zymler et al. [142]
Unimodal	Theorem 6.4	Corollary 7.4	Corollary 7.4
Montone	Corollary 7.3	Corollary 7.5	Corollary 7.5

Table 1.1: Listening of the worst-case bounds discussed in Part II organized by objective and distribution type. The results indicated in blue are novel contributions found in this dissertation.

approach, respectively. As both classes are non overlapping, no one approach is strictly stronger than the other. Nevertheless for the practically relevant loss functions we will encounter in this dissertation, the dual requirement (1.13) seems to offer more flexibility. That being said however, the dual approach does require technical conditions to guarantee strong duality and does seems to result in tractable but slightly more involved SDPs than its primal counterpart.

1.3.3 Robust optimization and control

In the final part of this dissertation the discussion turns back to distributionally robust constraints of the type (1.4) with the ambiguity set $\mathcal{C} = \mathcal{H}(\mu, S) \cap \mathcal{K}$ consisting of structured distributions sharing second-order moment information. We discuss their use in the static context of optimization problems in Chapter 8. Optimal control problems which can be regarded to represent a dynamic counterpart to the optimization problems discussed in Chapter 8 are considered in Chapter 9.

Chapter 8 Distributionally robust chance constraints (1.4) have the limitation that they are blind to severe constraint violations in which the uncertain outcome $x(u, \xi)$ strays far outside the constrained set X. Although the constraint (1.4) guarantees that the uncertain outcome $x(u, \xi)$ realizes within the constraint set X with probability at least $1 - \epsilon$ for all distributions within the ambiguity set $\mathcal{C}$, there is in general no bound on the severity of constraint violation in the remaining ϵ fraction of realizations.

In this chapter, we will consider distributionally robust CVaR constraints too as they provide a mechanism to control the level of constraint violation. We analyze both types of distributionally robust constraints in the context of polytopic constraint sets

$$X = \left\{x \ : \ a_i^\top x < b_i, \quad \forall i \in [1, \ldots, k]\right\}.$$

The chapter is divided into a first part discussing single uncertain constraints ($k = 1$) and a second part dealing with the general case ($k > 1$) of joint uncertain constraints.

In case of single uncertain constraints, both the chance and CVaR formulation are shown in Propositions 8.1 and 8.2 to admit an exact tractable reformulation in terms of a second-order cone (SOC) constraint. To the best of our knowledge, both propositions are novel. Joint distributionally robust constraints prove more challenging. In Section 8.3, we will outline when exact tractable reformulations are available, and when not, what type of approximation can be used instead.

Chapter 9 Where the previous chapter dealt with static optimization problems in which a single decision u needs to be taken in the face of a single uncertain realization ξ, this chapter will deal with control problems in which a sequence of feasible decisions has to be taken over time in the face of a sequence of disturbances in a causal manner. The main difference between the static optimization setting discussed in Chapter 8 is that the decisions $u_t(\xi)$ need to be taken adaptively in face of the disturbances ξ to ensure feedback.

We focus in this chapter on distributionally robust CVaR constraints with second-order moment information in the context of quadratically representable constraint sets

$$X = \left\{ x \;:\; x^\top E_i x + 2e_i^\top x + e_i^0 < 0, \quad \forall i \in [1, \ldots, k] \right\}.$$

The first half of the chapter deals with the finite horizon control of discrete-time linear time-invariant (DLTI) systems. The control decisions $u_t(\xi)$ are taken according to causal affine decision rules following Goulart et al. [55] and Ben-Tal et al. [6]. Our main contribution here is Theorem 9.1 which establishes that the best affine control policy can be characterized in terms of a tractable optimization problem. Theorem 9.3 extends this previous observation to infinite horizon control problems in the second half of the chapter. We furthermore show that for the constraint set X a single ellipsoid, the best linear control policy separates into a Kalman filter and a state feedback policy which can be found through the solution of a tractable SDP. The results in this chapter are illustrated on a wind blade control design case study for which distributionally robust constraints constitute sensible design objectives. The results in this final chapter are largely based on this author his publication [128]. The wind blade control design case study is presented in more detail in a conference contribution [132] again by this author.

Part I

Mathematical tools

2 Convex analysis

In this chapter we state various results concerning the convexity of both sets and functions in R^n. The purpose of this section is however twofold. As convexity plays a protagonist role throughout this dissertation, this section on convex analysis in R^n is meant to make the discussion self-contained. At the same time however, this section will try to prelude the corresponding analysis of convexity in more general vector spaces as discussed in Chapter 3.

The presentation of the results in this chapter is by no means exhaustive and entirely determined by their use in the remainder of this work. All results in this chapter are well known and are stated without proof. The reader is referred to the excellent works by Rockafellar [109] or Boyd and Vandenberghe [27] for a more complete treatment of convexity.

2.1 Convex sets and functions

Definition 2.1 (Convex set)**.** *A subset C of R^n is convex if it includes for every pair of points x, $y \in C$ the line segment that joins them, i.e.*

$$\forall x, y \in C : \quad tx + (1-t)y \in C, \ \textit{for all } t \in [0,1].$$

Convex sets play the protagonist role in this dissertation as they are involved in almost all subsequent results in one form or another. Various convex and non-convex sets are shown in Figure 2.1.

So many connections between convex sets and convex functions exist that it is best to introduce both objects at the same time.

Definition 2.2 (Convex function)**.** *A function $f : \mathrm{R}^n \to \bar{\mathrm{R}}$ is convex with respect to the convex set C in R^n if the following relationship holds for all points x and y in C,*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y), \ \textit{for all } t \in [0,1].$$

We denote with the set $\bar{\mathrm{R}} := \mathrm{R} \cup \{\infty, -\infty\}$ the extended real numbers on which the usual extended arithmetic is defined as found in Rockafellar and Wets [109, Section 1.E].

The main advantage of working with the extended real numbers $\bar{\mathrm{R}}$ is that the domain of a convex function f can be related to those points in R^n having a value which is bounded from above. The (effective) domain of f is then defined as

$$\operatorname{dom} f := \{x \in \mathrm{R}^n \ : \ f(x) < \infty\}.$$

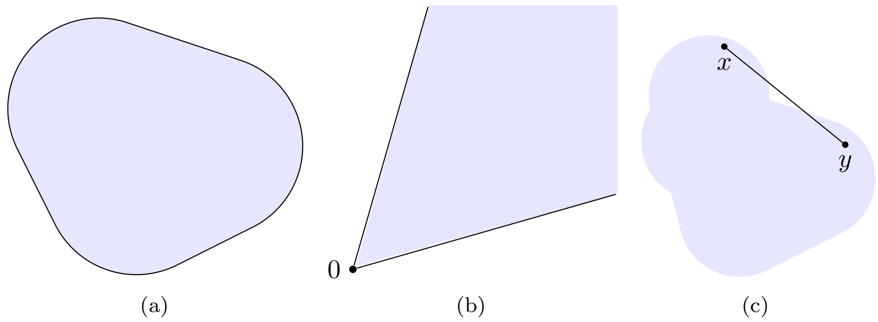


Figure 2.1: Various convex and non-convex sets in the plane. All but the set on the right are convex sets. The set in the middle is a proper convex cone, while the set on the left is a compact convex set. The open set 2.1(c) is not convex as the line segment $[x, y]$ is not entirely contained in it.

Note that if a convex function $g : C \to \mathrm{R}$ is only defined on a convex set $C \subset \mathrm{R}^n$ then it can be identified with a convex function f on R^n with $\operatorname{dom} f = C$ through

$$f(x) := \begin{cases} g(x) & x \in C, \\ \infty & \text{Otherwise.} \end{cases}$$

For most purposes, the study of convex functions can therefore be reduced to the framework of Definition 2.2 in which functions are defined everywhere but extended valued.

We say that the function f is proper if its effective domain is non-empty, i.e. $\operatorname{dom} f \neq \emptyset$, and $f(x) > -\infty$ for all $x \in \mathrm{R}^n$. A function f is denoted as concave if its negative $-f$ is a convex function. Furthermore, a function is called affine if it is both convex and concave.

The epigraph of a convex function f is defined as the set

$$\operatorname{epi} f := \{(x, s) \in \mathrm{R}^n \times \mathrm{R} \;:\; f(x) \leq s\}\,.$$

Convex functions are intimately related to convex sets as a function f is convex if and only if its epigraph $\operatorname{epi} f$ is a convex set. As the epigraph of a convex function is a convex set, convex functions are very well behaved as indicated by the following proposition.

Proposition 2.1 (Continuity of convex functions)**.** *A convex function $f : \mathrm{R}^n \to \bar{\mathrm{R}}$ is continuous on the interior of its domain* $\operatorname{int} \operatorname{dom} f$*.*

Although much of the analysis presented here holds for arbitrary convex sets, some convex sets will prove to be of particular interest in this dissertation and are discussed here further in more detail.

Canonical convex sets

Arguable the most primitive convex set is the simplex. Despite its simplicity the canonical simplex will be of central importance to this dissertation.

Example 2.1 (The canonical simplex)**.** *We denote with S_n the canonical simplex in R^n defined as*

$$S_n := \{x \in \mathrm{R}^n \;:\; x_i \geq 0, \quad \textstyle\sum_{i=1}^n x_i = 1\}\,.$$

The canonical simplex S_n is a compact convex set.

The class of convex sets carrying the most historical importance are the polyhedral sets which are defined through a finite number of linear inequalities.

Example 2.2 (Polyhedral sets)**.** *A set C in R^n is said to be a polyhedral set if it can be expressed as the intersection of $k \in \mathrm{N}$ linear constraints, i.e.*

$$C = \left\{x \in \mathrm{R}^n \ : \ a_i^\top x \leq b_i, \quad \forall i \in [1, \ldots, k]\right\}.$$

The simplex S_n is a polyhedral set as is clear from its definition.

The sets introduced in the preceding examples can be recognized as convex sets by checking the condition in Definition 2.2 directly. Direct inspection of the convexity condition in Definition 2.2 can however be quite tedious. Fortunately, convexity can often be established indirectly as discussed in the following section.

2.2 Operations that preserve convexity

Of interest are those operations which preserve the convexity of convex sets and functions. These operations are useful to establish the convexity of the various sets introduced throughout this dissertation without having to resort to the tedious convexity condition in Definition 2.2 directly. Indeed, an extensive algebra of convex sets and functions exists to verify the convexity of convex sets. Sets in R^n which can be recognized as convex sets using this convexity algebra are referred to as disciplined convex sets by Grant et al. [56]. We have listed here those operations most relevant to this dissertation.

Proposition 2.2 (Intersection)**.** *Suppose $C_i \subseteq \mathrm{R}^n$ for $i \in I$ is an arbitrary collection of convex sets, then their intersection $C := \cap_{i\in I} C_i$ is a convex set as well.*

Proposition 2.3 (Point-wise supremum)**.** *Suppose $f_i : \mathrm{R}^n \to \bar{\mathrm{R}}$ for $i \in I$ is an arbitrary collection of convex functions, then their point-wise supremum $f(x) := \sup_{i\in I} f_i(x)$ is a convex function as well.*

Note that Proposition 2.3 can be seen as a corollary of Proposition 2.2 using the fact that $\operatorname{epi} f = \cap_{i\in I} \operatorname{epi} f_i$. We remark here that the collection I need not be finite or even countable for Propositions 2.2 and 2.3 to hold.

Convexity of both sets and functions is preserved under linear, and even more general affine, transformations.

Proposition 2.4 (Affine transformations of sets)**.** *Let $C_1 \subseteq \mathrm{R}^n$, $C_2 \subseteq \mathrm{R}^m$ be convex sets and $L : \mathrm{R}^n \to \mathrm{R}^m$ an affine transformation. Then both the image set of C_1 under L defined as*

$$L(C_1) := \{y \in \mathrm{R}^m \ : \ \exists x \in C_1, \ y = L(x)\},$$

and the pre-image set of C_2 under L defined as

$$L^{-1}(C_2) := \{x \in \mathrm{R}^n \ : \ \exists y \in C_2, \ y = L(x)\}$$

are convex sets.

Proposition 2.5 (Affine transformations of functions)**.** *Let $f : \mathrm{R}^m \to \bar{\mathrm{R}}$ be a convex function and $L : \mathrm{R}^n \to \mathrm{R}^m$ an affine transformation. Then the composition $f \circ L : \mathrm{R}^n \to \bar{\mathrm{R}}$ is a convex function as well.*

Another transformation which preserves convexity is the perspective transformation for which we refer to Boyd and Vandenberghe [27]. We define the perspective transformation $P : \mathrm{R}^{n+1} \to \mathrm{R}^n$ with domain $\mathrm{R}^n \times \mathrm{R}_{++}$ as the transformation $P(z,t) := z/t$. The perspective transformation scales or normalizes a vector $(z,t) \in \mathrm{R}^{n+1}$ so that its last component is rescaled to one, and then drops this last component to yield a vector in R^n.

Proposition 2.6 (Perspectives of sets)**.** *Let $C_1 \subseteq \mathrm{R}^{n+1}$ and $C_2 \subseteq \mathrm{R}^n$ be convex sets and $P : \mathrm{R}^{n+1} \to \mathrm{R}^n$ a perspective transformation. Then both the image set of C_1 under P and the pre-image set of C_2 under P are convex sets.*

Closely related to the perspective transformation for convex sets is the perspective function of convex functions.

Proposition 2.7 (Perspective functions)**.** *Let $f : \mathrm{R}^n \to \bar{\mathrm{R}}$ be a convex function. Then its corresponding perspective function $g : \mathrm{R}^{n+1} \to \bar{\mathrm{R}}$ defined as $g(x,t) := tf(x/t)$ with effective domain*

$$\{(x,t) \in \mathrm{R}^n \times \mathrm{R}_{++} \ : \ x/t \in \operatorname{dom} f\}$$

is convex as well.

Notice that Proposition 2.7 can be seen as a corollary of Proposition 2.6 using the fact that $\operatorname{epi} g$ is the pre-image of $\operatorname{epi} f$ under a perspective transformation.

2.3 Convex hulls and extreme points

A non-convex set S can be "convexified" by considering its convex hull $\operatorname{conv} S$.

Definition 2.3 (Convex hull)**.** *The convex hull of a set $S \subseteq \mathrm{R}^n$ is defined as the intersection of all convex sets in R^n containing S.*

The convex hull of any set is a convex set because convexity is preserved under arbitrary many intersection; see Proposition 2.2. The convex hull and convex sets in R^n are closely related through the notion of convex combination.

Definition 2.4 (Convex combination)**.** *We call a point of the form $\sum_{i=1}^k x_i p_i$, where $\sum_{i=1}^k p_i = 1$ and $p_i \geq 0$ for $i \in [1,\ldots,k]$, a convex combination of the points $x_1,\ldots,x_k$. A convex combination $\sum_{i=1}^k x_i p_i$ is called strict if it is a convex combination for which all $p_i > 0$.*

It can be shown that a set $C \subseteq \mathrm{R}^n$ is convex if and only if it contains all convex combinations of its own elements. Furthermore, the convex hull of a set S can alternatively, and equivalently, be defined as

$$\operatorname{conv} S := \bigcup_{k \in \mathrm{N}} \left\{ \sum_{i=1}^k p_i x_i \ : \ x_i \in S,\ p_i \geq 0,\ i \in [1,\ldots,k],\ \sum_{i=1}^k p_i = 1 \right\}. \tag{2.1}$$

as shown by Rockafellar and Wets [109, Theorem 2.27]. Equation (2.1) establishes that the convex hull of S is the union of all convex combinations of elements in S. A useful result relating points in $\operatorname{conv} S$ to points in S is Carathéodory's Theorem.

Theorem 2.1 (Carathéodory's Theorem)**.** *If a set S in R^n is nonempty, then every point $x \in \operatorname{conv} S$ can be written as a convex combination of at most $n+1$ points in S.*

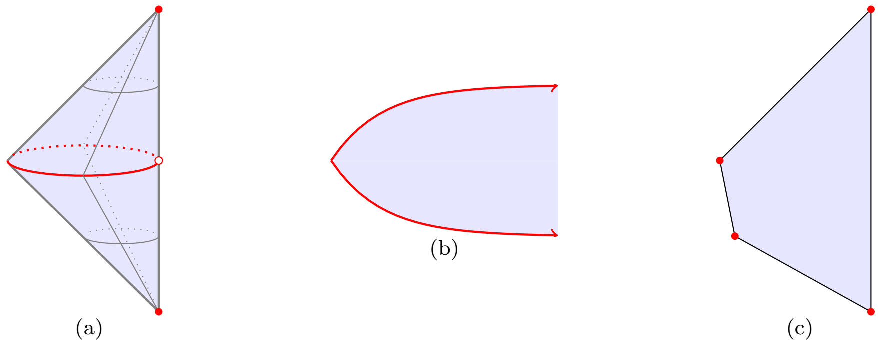


Figure 2.2: Closed convex sets with their extreme points shown in red. Observe that the set of extreme points is not necessarily closed as is the case for the convex set on the left. The convex set in the middle is not compact, but nevertheless admits a Choquet representation. The set on the right is polyhedral and hence has only finitely many extreme points.

From Carathéodory's theorem it must hence follow that k in equation (2.1) can, in fact, be limited to $n+1$, i.e.

$$\operatorname{conv} S = \left\{ \sum_{i=1}^{n+1} p_i x_i \ : \ x_i \in S, \ p_i \geq 0, \ i \in [1, \dots, n+1], \ \sum_{i=1}^{n+1} p_i = 1 \right\}.$$

Definition 2.5 (Extreme points)**.** *A point x in a convex set C is said to be an extreme point of C if it is not representable as a strict convex combination of two distinct points in C.*

The set of all extreme points of a convex set C is denoted as $\operatorname{ex} C$. In other words, an extreme point of C is a point that is not an interior point of any line segment contained in C. It is clear that the inclusion $\operatorname{ex} C \subseteq \operatorname{bd} C$ must hold. The converse is easily shown to be false, and some counter examples are shown in Figure 2.3.

One of the most important results on convex sets is the Krein-Milman Theorem [69] which takes on the following form in R^n.

Theorem 2.2 (Krein-Milman)**.** *Let C be a compact convex subset of R^n. Then C is the convex hull of its extreme points $\operatorname{ex} C$.*

Compactness, although sufficient, is not a necessary condition for a convex set $C \subset \mathrm{R}^n$ to be the convex hull of its extreme points, see Figure 2.2(b).

Definition 2.6 (Choquet representation)**.** *A convex set C in R^n is said to admit a Choquet representation if for every $x \in C$ there exists a number $k \in \mathrm{N}$ so that*

$$x = \sum_{i=1}^{k} p_i x_i, \quad x_i \in \operatorname{ex} C,$$

with $p \in S_k$.

From Carathéodory's Theorem, a necessary and sufficient condition for a convex set C in R^n to admit a Choquet representation is $C = \operatorname{conv} \operatorname{ex} C$. We emphasize this fact as we will encounter in the next Chapter convex sets in more general vector spaces for which the previous statement fails to hold.

We can illustrate the definitions and results stated in this section by revisiting the canonical simplex S_n introduced in Example 2.1.

Example 2.3 (Extreme points of S_n.)**.** *The extreme points of the canonical simplex S_n are the canonical vectors in R^n. As all polyhedral sets, the canonical simplex S_n has only finitely many extreme points. Furthermore, the canonical simplex S_n admits a unique Choquet representation as*

$$\forall x \in S_n,\ \exists! p \in S_n : \quad x = \sum_{i=1}^{n} e_i p_i.$$

where $e_i \in \operatorname{ex} S_n$ is the i^{th} the canonical vector.

2.4 Cones and dual cones

A prominent class of convex sets in this dissertation are the convex cones. Cones are subsets of R^n which are not necessarily convex although almost all cones encountered in this dissertation are.

Definition 2.7 (Cone)**.** *A set $K \subseteq \mathrm{R}^n$ is a cone if $0 \in K$ and $tx \in K$ for all $x \in K$ and $t \geq 0$.*

A cone contains rays emanating from the origin, i.e. sets of the particular form $\{tx \ : \ t \geq 0\}$ for some $x \in \mathrm{R}^n$. We say that the cone K is *proper* if it does not contain any lines, i.e. a proper cone satisfies $K \cap -K = \{0\}$. A non-conic set S can be "conified" by considering its conic hull co S.

Definition 2.8 (Conic hull)**.** *The conic hull of a set $S \subseteq \mathrm{R}^n$ is defined as the intersection of all convex conic sets in R^n containing S.*

We can relate to any cone $K \subseteq \mathrm{R}^n$ a corresponding dual cone $K^\star$ using a bilinear product $\langle \cdot, \cdot \rangle : \mathrm{R}^n \times \mathrm{R}^n \to \mathrm{R}$. Dual cones will plays an important role in developing duality for convex optimization problems; see Chapters 4 and 5.

Definition 2.9 (Dual cone)**.** *The dual cone $K^\star$ of any set K in R^n is defined as*

$$K^\star := \{ y \in \mathrm{R}^n \ : \ \langle y, x \rangle \geq 0, \quad \forall x \in K \}.$$

It can immediately be seen from its definition that the dual cone of any set is itself indeed always a cone in R^n. Furthermore, the dual cone of any set and its conic hull must necessarily coincide, i.e. $K^\star = (\text{co } K)^\star$. A more significant observation is that the dual cone of any, potentially non-convex, cone K is a closed convex cone. The most important properties concerning dual cones relevant to this dissertation are stated in the following proposition.

Proposition 2.8 (Properties of dual cones)**.** *Given a cone K in R^n, then the following properties hold:*

1. *The dual cone $K^\star$ is a closed convex cone in R^n.*
2. *If the cone K is closed and convex then $K = K^{\star\star}$.*
3. *Given two cones $K_1 \subseteq K_2$ then $K_2^\star \subseteq K_1^\star$.*

From Definition 2.9 it can be remarked that the dual cone $K^\star$ depends on the bilinear product $\langle \cdot, \cdot \rangle$ considered. Throughout the dissertation, we will take the standard inner product as our bilinear product, i.e. $\langle x, y \rangle := x^\top y$. In Figure 2.3 we have depicted a cone and dual cone pair in R^3 for the standard inner product $\langle a, b \rangle = a^\top b$.

In the remainder of this section we consider cones which are of particular interest in subsequent chapters of this dissertation.

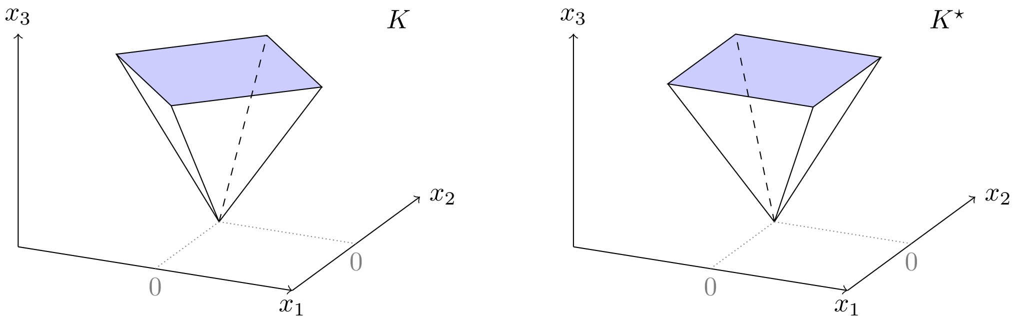


Figure 2.3: A convex cone $K = \left\{x \in \mathrm{R}^3 \ : \ \|(x_1, x_2)\|_1 \leq x_3\right\}$ and its dual $K^\star = \left\{x \in \mathrm{R}^3 \ : \ \|(x_1, x_2)\|_\infty \leq x_3\right\}$ for the standard inner product $\langle a, b\rangle = a^\top b$.

Example 2.4 (Positive orthant)**.** *The positive orthant* R^n_+ *defined as the set of vectors in* R^n *which are component-wise positive, i.e.*

$$\mathrm{R}^n_+ = \{x \in \mathrm{R}^n \ : \ x_i \geq 0, \quad \forall i \in [1, \dots, n]\},$$

is a proper convex cone.

Example 2.5 (Norm cones)**.** *Any norm* $\|\cdot\| : \mathrm{R}^n \to \mathrm{R}$ *is a convex function and its epigraph a proper convex cone. For the p-norms with* $1 \leq p \leq \infty$*, we define the norm cones* L^n_p *as follows*

$$L^n_p := \operatorname{epi} \|\cdot\|_p = \left\{(x, s) \in \mathrm{R}^n \times \mathrm{R} \ : \ \|x\|_p \leq s\right\}.$$

The norm cone L^n_2 *is sometimes referred to as the Lorentz cone. The cones* L^n_p *and* L^n_q *are dual pairs whenever* $1/p + 1/q = 1$*. In Figure 2.3 the cone* L^2_1 *and its dual* L^2_∞ *are depicted.*

Example 2.6 (The positive definite cone)**.** *The set of symmetric positive definite matrices in* $\mathrm{R}^{n\times n}$

$$\left\{Q \in \mathrm{S}^n \ : \ x^\top Q x \geq 0, \quad \forall x \in \mathrm{R}^n\right\}$$

is denoted as the positive semidefinite cone S^n_+*.*

Lastly, a cone K is referred to as self-dual if it coincides with its dual cone, i.e. we have that $K = K^\star$. The positive orthant R^n_+, the Lorentz cone L^n_2 and the semi-definite cone S^n_+ are all known to be self dual cones. Self dual cones play a prominent role in the development of computational methods for convex optimization problems as discussed in Chapter 4.

3 Probability Theory

In this chapter we try to communicate the fact that the concepts introduced in Chapter 2 concerning convex sets in R^n are relevant to many structures encountered in probability theory as well.

However the convex structures encountered in this chapter will reside in vector spaces more general then R^n. Many of the results stated in Chapter 2 can be extended directly to more general vector spaces; see for instance Zalinescu [140]. We chose to stick to a finite-dimensional analysis so as not to cloud the picture with the many complications that a treatment in general vector spaces inevitably would bring. Instead, we treat here convex sets in measure spaces on R^n and use the finite dimensional case discussed in Chapter 2 as a useful metaphor. Many of the results stated here can indeed be recognized as generalizations or direct counterparts of statements made in Chapter 2. Nevertheless, great care must be taken when working in vector spaces more general than R^n. We refer to Barvinok [5, Chapter 4] for an excellent discussion on what can go wrong when working in infinitely dimensional vector spaces.

Readers who are unacquainted with measure spaces and probability theory are referred to Appendix A.1 in which we discuss the measure space $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$ and random vectors valued in R^n to the extent relevant to this work. The standard reference by Billengsley [22] can be consulted for a more comprehensive treatment of general measure spaces and probability theory.

3.1 Convexity in measure spaces

As mentioned earlier, one might generalize the convex analysis of Chapter 2 almost directly to more general vector spaces as for instance done by Zalinescu [140] or Barvinok [5]. In doing so convex sets are required to be subsets of a locally convex topological vector space instead of R^n. The more general vector space should be topological as to have appropriately defined open and closed sets. Local convexity of the more general vector space ensures a comprehensive duality theory. In this dissertation we will mainly deal with probabilistic problems on R^n. The aforementioned generalization is thus wholly unnecessary for our purposes and would yield certain mathematical difficulties which could complicate the exposition considerably.

In the sequel therefore we will consider exclusively the measure space $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$ and the vector space of signed measures thereupon. The reader who is not familiar with measure spaces and signed measures is referred to Appendix A.1 for a terse, but self-contained, treatment of both concepts. Denote with $\mathcal{E}_n$ the vector space of finite signed measures on R^n. It is indeed a vector space for the vector addition and scalar multiplication defined in the natural way. We further assume that $\mathcal{E}_n$ is endowed with the topology of weak convergence of measures.

Definition 3.1 (Convex set)**.** *A subset $\mathcal{C}$ of $\mathcal{E}_n$ is convex if it includes for every pair of measures $\mathbb{x}$, $\mathbb{y} \in \mathcal{C}$ the line segment that joins them, i.e.*

$$\forall \mathbb{x}, \mathbb{y} \in C : \quad t\mathbb{x} + (1-t)\mathbb{y} \in \mathcal{C}, \text{ for all } t \in [0,1].$$

Previous definition extends its counterpart 2.1 in Chapter 2 concerning convex sets residing in R^n in a natural way. Some convex sets will proof to be of particular interest in this dissertation and are discussed here further in more detail. We let $\mathcal{E}_n^+$ represent the convex cone of all positive measures in $\mathcal{E}_n$. A measure $\mathbb{m}$ is denoted as positive if it assigns a positive measure to all measurable sets. We say $\mathbb{m} \geq 0$ using shorthand notation.

Definition 3.2 (Probability simplex)**.** *The set of all probability distributions in $\mathcal{E}_n$, i.e. all positive measures $\mathbb{P}$ for which $\mathbb{P}(\mathrm{R}^n) = 1$, shall be denoted as $\mathcal{P}_n$. It is conventional to refer to the convex set $\mathcal{P}_n$ as the standard probability simplex on R^n.*

It will proof convenient to refer to a subset of the probability simplex $\mathcal{P}_n$ as an ambiguity set. Any ambiguity set hence consists of measures in $\mathcal{E}_n$, but not every subset of $\mathcal{E}_n$ is an ambiguity set. Furthermore, we will use $\mathcal{P}(B)$ to refer to the convex set of probability distributions supported on a measurable subset B of R^n, i.e.

$$\mathcal{P}(B) := \{\mathbb{m} \in \mathcal{E}_n \ : \ \mathbb{m} \geq 0, \ \mathbb{m}(B) = 1\}.$$

Using previous notation it can be seen that $\mathcal{P}_n = \mathcal{P}(\mathrm{R}^n)$. In what follows we discuss various types of convex sets which will be used frequently throughout the dissertation. Particular emphasis will be put on convex ambiguity sets as they form the backbone of this dissertation.

Hyperplanes and moment sets

We say that a subset $\mathcal{H}$ of the measure space $\mathcal{E}_n$ is a hyperplane of codimension k if it is in the canonical form

$$\mathcal{H} := \left\{ \mathbb{m} \in \mathcal{E}_n : \int g_i(x)\, \mathbb{m}(\mathrm{d}x) = m_i, \quad \forall i \in [0, \ldots, k-1] \right\}, \tag{3.1}$$

where the moment functions $g_0, \ldots, g_{k-1}$ are real valued measurable functions on R^n. The vector $m \in \mathrm{R}^k$ will be referred to as the moment vector. The subset $\mathcal{H}$ is an affine subset of the measure space $\mathcal{E}_n$. We remark that the moment set $\mathcal{C}$ can be interpreted as a generalized hyperplane in $\mathcal{P}_n$. Likewise, a subset $\mathcal{C}$ of the standard probability simplex $\mathcal{P}_n$ is a moment set if it is the intersection of a hyperplane with the standard probability simplex, i.e.

$$\mathcal{C} := \left\{ \mathbb{P} \in \mathcal{E}_n^+ : \int g_i(x)\, \mathbb{P}(\mathrm{d}x) = m_i, \quad \forall i \in [0, \ldots, k-1] \right\}, \tag{3.2}$$

where it is assumed that $g_0 = 1$ and $m_0 = 1$ as to guarantee that the moment set $\mathcal{C}$ contains exclusively probability distributions $\mathbb{P}$. The moment set $\mathcal{C}$ is thus a convex subset of the standard probability simplex $\mathcal{P}_n$ as can be seen directly from its definition. We remark that the moment set $\mathcal{C}$ can be interpreted as a generalized polytope in $\mathcal{P}_n$.

Unlike hyperplanes in R^n, the affine set $\mathcal{H}$ in (3.1) is not necessarily a closed set. A similar remark can be made concerning the generalized polytope $\mathcal{C}$ in (3.2) of probability distributions in $\mathcal{P}_n$. When the moment functions g_i are bounded and continuous the related hyperplane $\mathcal{H}$ and moment set $\mathcal{C}$ are closed as seen immediately from Definition A.4. However, in this thesis we shall be primarily interested in moment functions which are neither bounded nor continuous and thus both sets $\mathcal{H}$ and $\mathcal{C}$ are not required to be closed.

The classical moment problem discussed by for instance Landau [75] occupies itself with determining whether the ambiguity set $\mathcal{C}$ is non-empty for a given moment function $g : \mathrm{R}^n \to \mathrm{R}^k$

and moment vector m. Most of the literature on moment problems is concerned with univariate problems and thus focusses on the existence of probability distributions with specific moments on the real line. Historically, the Hamburger moment problem with polynomial power moment functions $g_i(x) = x^i$ has attracted the most attention. A brief history of the classical moment problem is given in Section 3.3. Nevertheless, there is a significant body of work that concerns itself with the multivariate setting $n \neq 1$, see for instance Karlin and Studden [65], Krein and Nudelmann [70] or Gantmacher and Krein [52] plus the various references therein.

Moment sets as defined in (3.2) are immediately seen as convex sets as they are the finite intersection of k affine sets in the standard probability simplex $\mathcal{P}_n$. On the other hand various well known and common sets of probability distributions are convex as well, but are not immediately recognized to be so. We now introduce the concepts of unimodal and monotone distributions and indicate that both concepts lead in fact to convex sets.

Unimodal distributions

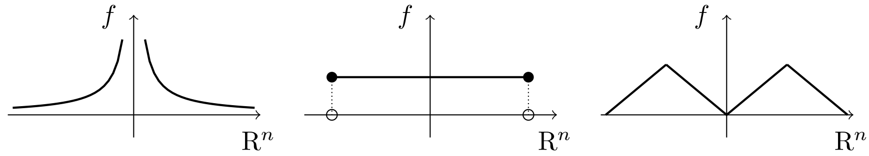
A common structural property enjoyed by many probability distributions encountered in practical situations is unimodality. Informally, a continuous probability distribution is unimodal if it has a centre c, referred to as the mode, such that its density function is non-increasing with increasing distance from the mode. Note that most probability distributions commonly studied in probability theory are unimodal. A huge variety of named and popular probability distributions are indeed unimodal. Even when sticking to the first few letters of the alphabet, we have that the Bates, Beta (α, $\beta > 1$), Birnbaum-Saunders, Burr, Cauchy, Chi and Chi-squared probability distributions are all unimodal. So, too, are all stable probability distributions, which are ubiquitous in statistics as they represent the attractors for properly normed sums of independent and identically distributed random variables.

Definition 3.3 (Star-shaped sets). *A set $B \subseteq \mathbb{R}^n$ is said to be star-shaped with center 0 if for every $x \in B$ the line segment $[0, x]$ is a subset of B.*

Definition 3.4 (Star unimodal distributions [44]). *A probability distribution $\mathbb{P} \in \mathcal{P}_n$ is called star-unimodal if it belongs to the weak closure of the convex hull of all uniform probability distributions on star-shaped sets with center 0. The set of all star-unimodal distributions is denoted as $\mathcal{U}_n$.*

Definition 3.4 assumes without loss of generality that the mode of a star-unimodal distribution is located at the origin, which can always be enforced by applying a suitable coordinate translation. We remark that for multivariate probability distributions there exist several other notions of unimodality such as linear, convex or log-concave unimodality etc. While not equivalent for $n > 1$, all customary notions of unimodality, including the star unimodality of Definition 3.4, coincide with Definition 1.1 in the univariate case; see for instance Dharmadhikari and Joag-Dev [44]. We also remark that the definition of star-unimodality is in line with our intuitive idea of unimodality when $\mathbb{P}$ has a continuous density function f. In this case Dharmadhikari and Joag-Dev [44] prove that $\mathbb{P}$ is star-unimodal if and only if $f(tx)$ is non-increasing in $t \in (0, \infty)$ for all $x \neq 0$, which means that the density function is non-increasing along any ray emanating from the origin. Definition 3.4 extends this intuitive idea to a broader class of probability distributions that may have no density functions.

We remark that the definition of star unimodality 3.4 is quite rigid as it provides no indication on how unimodal a probability distribution exactly is. In this dissertation we will consider the slightly more flexible notion of α-unimodality first introduced by Olshen and Savage [94]. An excellent treatment of the theory concerning α-unimodal distributions and their application in a wide range of problems is given by Dharmadhikari and Joag-Dev [44].



(a) $\alpha < n$ (b) $\alpha = n$ (c) $\alpha > n$

Figure 3.1: Univariate α-unimodal distributions and their density functions. We remark that in case $\alpha = n$, the notion of α-unimodality coincides with the intuitive notion of star unimodality where a probability distribution is unimodal if its density function is decreasing or at worst non-increasing.

Definition 3.5 (α-Unimodal distributions [94, 44])**.** *For any fixed $\alpha \in \mathrm{R}_+$, a probability distribution $\mathbb{P} \in \mathcal{P}_n$ is called α-unimodal with mode 0 if $t^\alpha \mathbb{P}(B/t)$ is non-decreasing in $t \in \mathrm{R}_{++}$ for every measurable set B. The set of all α-unimodal distributions with mode 0 is denoted as $\mathcal{U}_\alpha$.*

To develop an intuitive understanding of Definition 3.5, it is instructive to study again the special case of continuous probability distributions. We have that a continuous probability distribution $\mathbb{P} \in \mathcal{P}_n$ with a continuous density function $f(x)$ is α-unimodal about 0 iff $t^{n-\alpha} f(tx)$ is non-increasing in $t \in \mathrm{R}_{++}$ for every fixed $x \neq 0$. This implies that if an α-unimodal distribution on R^n has a continuous density function f, then $f(x)$ does not grow faster than $\|x\|_2^{\alpha-n}$. In particular, for α equal to the dimension n the density is non-increasing along rays emanating from the origin. In this case, the notion of α-unimodality coincides with the notion of star unimodality defined in 3.4.

The density function of a continuous α-unimodal distribution may in fact increase along rays, but the rate of increase is controlled by the parameter α. Hence, the parameter α can be seen as a characterization of how unimodal a probability distribution is; see also Figure 3.1.

The sets $\mathcal{U}_\alpha$ enjoy the nesting property $\mathcal{U}_\alpha \subseteq \mathcal{U}_\beta$ whenever $\alpha \leq \beta$. It can be shown that the ambiguity sets $\mathcal{U}_\alpha$ are closed convex sets and that the closure of $\mathcal{U}_\infty := \cup_{\alpha \geq 0} \mathcal{U}_\alpha$ coincides with the standard probability simplex $\mathcal{P}_n$. Hence although $\mathcal{U}_\alpha \subset \mathcal{P}_n$ for all α, the standard simplex $\mathcal{P}_n$ is included in the limit of the hierarchy of α-unimodal ambiguity sets $\mathcal{U}_\alpha$ for the unimodality parameter α tending to infinity. At the other end of the spectrum, we have that $\mathcal{U}_0 = \{\delta_0\}$ reduces to a singleton containing exclusively the Dirac distribution in the origin. The Dirac distribution in the origin can thus be regarded as the most unimodal distribution. On the other hand, it can be seen that $\delta_x \notin \mathcal{U}_\alpha$ for any α if $x \neq 0$. We can thus claim that a Dirac distribution at a point distinct from the mode is completely not unimodal. Hence as the closure of $\mathcal{U}_\infty$ coincides with the standard probability simplex $\mathcal{P}_n$ the Dirac distributions δ_x must be boundary elements of $\mathcal{P}_n$. The present discussion is visually illustrated in Figure 3.2.

We remark here that it is not so clear from Definition 3.5 that the sets of α-unimodal distributions $\mathcal{U}_\alpha$ are convex. In this chapter we will argue that in fact the set of star and α-unimodal distributions are convex sets admitting a so called Choquet representation; see Section 3.2.

Monotone distributions

A property which is closely related to unimodality is monotonicity. Where unimodality requires intuitively that the density function of a continuous probability distribution should be decreasing with increasing distance from its mode, monotonicity additionally requires that this decrease is smooth; see Figure 3.3. Citing Pestana and Mendonça [99], monotonicity is indeed often used in mathematics to express the vague notion of *smoothness* of a probability distribution. Similar to unimodality, monotonicity is a reoccurring property of probability distributions in both theory

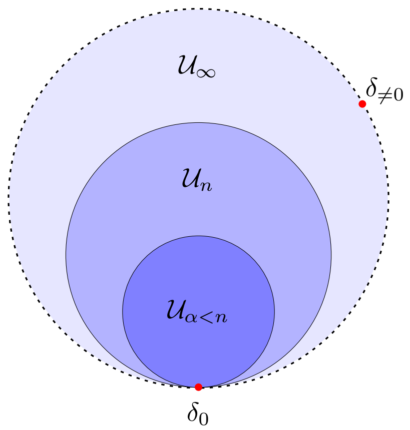


Figure 3.2: The hierarchy of the ambiguity sets $\mathcal{U}_\alpha$. The Dirac distribution δ_0 is contained in $\mathcal{U}_\alpha$ for all α. The set $\mathcal{U}_n$ consists of all star unimodal distributions. The sets $\mathcal{U}_\alpha$ are closed convex sets in $\mathcal{P}_n$. The closure of the set $\mathcal{U}_\infty$ coincides with the standard probability simplex $\mathcal{P}_n$. All Dirac distributions δ_x are boundary elements of the standard probability simplex and, with the exception of δ_0, are not in any set $\mathcal{U}_\alpha$.

and practice. Lifetime distributions, hazard rates and network performance are all commonly modeled using monotone distributions as done for instance in Jewell [64], Harris and Singpurwalla [60] and Feldmann and Whitt [49], respectively. Furthermore, monotonicity has deep connections to infinite divisibility as expressed in the celebrated Lévy-Kinchine transformation as outlined in Barndorff-Nielsen et al. [4]. In the remainder we adopt the standard definition of monotonicity of probability distributions found in for instance Pestana and Mendonça [99]. The definition of a monotone distribution is inspired on the notion of monotone functions as originally discussed by Lévy [77].

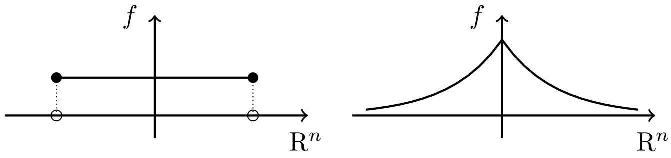
Definition 3.6 (γ-monotone functions)**.** *A univariate function $f : \mathrm{R}_+ \to \mathrm{R}$ is denoted as γ-monotone if it is γ times differentiable and*

$$(-1)^k f^{(k)}(t) \geq 0, \quad \forall t > 0,\ k \in \{1, \ldots, \gamma\}.$$

Definition 3.7 (γ-monotone distributions [13])**.** *For any $1 \leq \gamma \in \mathrm{N}$, a probability distribution $\mathbb{P}$ is called γ-monotone with mode 0 if $t^{\gamma+n-1}\mathbb{P}(B/t)$ is γ-monotone in $t \in (0, \infty)$ for every measurable set B. The set of all γ-monotone distributions with mode 0 is denoted as $\mathcal{M}_\gamma$.*

The class of γ-monotone distributions defined here can be identified with the class of (n, γ)-unimodal distributions discussed by Bertin et al. [13, Theorem 3.1.14]. Again it is instructive to consider the case of continuous probability distributions. We have that a continuous probability distribution $\mathbb{P}$ is γ-monotone if its continuous density function $f(t\xi)$ is γ-monotone in $t \in [0, \infty)$ for every fixed ξ. This means that if a γ-monotone distribution $\mathbb{P}$ admits a continuous density f, then f is γ-monotone along rays emanating from the mode. Hence γ can be seen as a characterization of how monotone a probability distribution is.

The ambiguity sets $\mathcal{M}_\gamma$ enjoy the nesting property $\mathcal{M}_\delta \subseteq \mathcal{M}_\gamma$ whenever $\gamma \leq \delta$. It can be seen that the sets $\mathcal{M}_\gamma$ are convex and closed subsets of $\mathcal{P}_n$. Historically, a probability distribution in $\mathcal{M}_\infty := \cap_{\gamma \geq 1} \mathcal{M}_\gamma$ has been denoted as a completely monotone distribution by Bernstein [12]; see also Figure 3.4. We remark here that $\mathcal{M}_\infty$ is a closed and convex subset of $\mathcal{P}_n$ as it is the intersection of a collection of closed sets. At the other end of the extreme, we have that the set of all 1-monotone distributions reduces to the set of star unimodal distributions $\mathcal{U}_n$.



(a) Unimodal (b) Montone

Figure 3.3: Comparison between unimodality and monotonicity. Where unimodality requires that the density function of a continuous probability distribution should be non-increasing with increasing distance from the mode, monotonicity additionally requires that the density function is smooth.

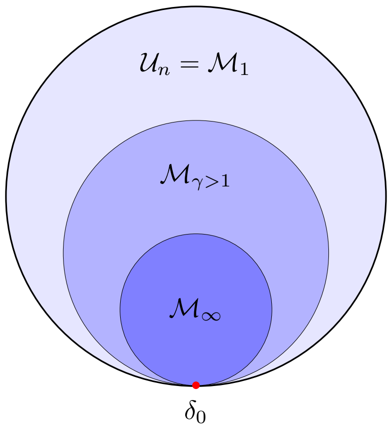


Figure 3.4: The hierarchy of the ambiguity sets $\mathcal{M}_\gamma$ in $\mathcal{U}_n$. The ambiguity set $\mathcal{M}_1$ coincides with the set of all star unimodal distributions $\mathcal{U}_n$. The Bernstein [12] or completely monotone distributions are included in the limit of the hierarchy for γ tending to infinity as $\mathcal{M}_\infty := \cap_{\gamma \geq 1} \mathcal{M}_\gamma$.

We remark here that it is not so clear from Definition 3.7 that the sets of γ-monotone distributions $\mathcal{M}_\gamma$ are convex. In this chapter we will argue that, in fact, the set completely and γ-monotone distributions are convex sets admitting a so called Choquet representation; see Section 3.2.

Jensen's inequality

The first part of this section dealt mainly with convex sets of probability distributions in the vector space of measures $\mathcal{E}_n$. In the last part of the section we will discuss integration and expectation with respect to probability distributions. Emphasis is put on the intimate relation between integration and convex functions known as Jensen's inequality.

Integration and expectation are well known to represent linear functionals of probability distributions and random variables, respectively, having deep links with convexity. A rigorous definition of the integral of a function or the expectation of a random variable requires a certain familiarity with measure theory. The unacquainted reader is referred to the standard reference by Billingsley [22] for a comprehensive treatment of the topic. For the sake of the exposition we intend to keep the measure theoretical technicalities to a minimum here.

Given a measurable function $f : \mathrm{R}^n \to \mathrm{R}$ and probability distribution $\mathbb{P}$, we denote the integral of f with respect to $\mathbb{P}$ as

$$\int f(x)\,\mathbb{P}(\mathrm{d}x) \in \mathrm{R}.$$

The function f is denoted as integrable with respect to $\mathbb{P}$ if the integral of $|f|$ with respect to $\mathbb{P}$ is well defined and finite. It can be remarked that integration is well known to be a linear operation as for any two integrable functions f_1 and f_2 we have the relation

$$\int f_1(x) + f_2(x)\,\mathbb{P}(\mathrm{d}x) = \int f_1(x)\,\mathbb{P}(\mathrm{d}x) + \int f_2(x)\,\mathbb{P}(\mathrm{d}x).$$

In fact we will discuss in Section 3.3 how a comprehensive duality theory between functions and probability distributions can be developed using the bilinear pairing $\langle f, \mathbb{P}\rangle := \int f(x)\,\mathbb{P}(\mathrm{d}x)$ defined through integration.

The expectation of a random variable ξ valued in R^n can be defined with the help of integration as

$$\mathbf{E}_{\mathbb{P}}[\xi] := \int x\,\mathbb{P}(\mathrm{d}x),$$

where $\mathbb{P}$ is the distribution of ξ. It can be remarked that the expectation $\mathbf{E}_{\mathbb{P}}[\xi]$ of a random variable ξ does not depend on the random variable directly but rather on its distribution. The expectation of the random variable $f(\xi)$ with f a measurable function is well known to be given as the integral $\mathbf{E}_{\mathbb{P}}[f(\xi)] = \int f(x)\,\mathbb{P}(\mathrm{d}x)$. A very important connection between integration and convex functions is expressed in Jensen's inequality.

Theorem 3.1 (Jensen's inequality)**.** *Given a convex function f and a random variable ξ distributed as $\mathbb{P}$ then the following relation always holds*

$$f\left(\mathbf{E}_{\mathbb{P}}[\xi]\right) \leq \mathbf{E}_{\mathbb{P}}\left[f\left(\xi\right)\right].$$

Jensen's inequality can be seen as a generalization of the secant condition in Definition 2.1 for convex functions. By virtue of its generality Jensen's inequality will appear throughout the remainder of this work.

3.2 Extreme point representations

This section mirrors the corresponding discussion held in Section 2.3 concerning extreme points of convex sets in R^n. There are however some important differences that need to be addressed when working in the more general vector space $\mathcal{E}_n$ of signed measures on $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$. Extreme points of convex sets in more general vector spaces are characterized in the same fashion as in Definition 2.5.

Definition 3.8 (Extreme points)**.** *A measure $\mathbb{m}$ in a convex set $\mathcal{C}$ in $\mathcal{E}_n$ is said to be an extreme measure of $\mathcal{C}$ if $\mathbb{m}$ is not representable as a strict convex combination of two distinct measures in $\mathcal{C}$.*

The extreme measures of a convex set $\mathcal{C}$ are those measures which can not be represented as the convex combination of other measures in $\mathcal{C}$. We would like to recover a direct counterpart of Theorem 2.2 (Krein-Milman) and the closely related Choquet representation of convex sets. There are however a number of caveats present in vector spaces more general than R^n which prevent a verbatim restatement of Theorem 2.2 in the current setting. The following theorem states the classic counterpart to Theorem 2.2 in general topological vector spaces.

Theorem 3.2 (Krein-Milman)**.** *Let $\mathcal{C}$ be a compact convex subset of $\mathcal{E}_n$. Then $\mathcal{C}$ is the closed convex hull of* $\mathrm{ex}\,\mathcal{C}$.

Compared to Theorem 2.2, the Krein-Milman Theorem 3.2 must be adapted slightly to include the closure of the convex hull of the extreme measures of $\mathcal{C}$. To show that this closure is indeed necessary, let us consider the probability simplex $\mathcal{P}(B)$ on a compact set B in R^n. The set $\mathcal{P}(B)$ is a compact subset of $\mathcal{E}_n$ as can be established with the help of Prokhorov's Theorem A.1. The set of all Dirac distributions supported on a measurable set B is denoted as $\mathcal{D}(B)$. It can be seen that the Dirac distributions δ_x with $x \in B$ are all extreme in $\mathcal{P}(B)$. The following proposition considering the extreme points of $\mathcal{P}(B)$ is well known.

Definition 3.9 (The Dirac measure δ_x). *We define the Dirac distribution $\delta_x : \mathcal{B}(\mathrm{R}^n) \to [0,1]$ as the probability distribution which assigns one to any measurable set B containing $x \in \mathrm{R}^n$ and zero otherwise.*

Proposition 3.1 (Phelps [101]). *The extreme points of $\mathcal{P}(B)$ are given as*

$$\operatorname{ex} \mathcal{P}(B) = \{\delta_x \ : \ x \in B\} = \mathcal{D}(B)$$

for any measurable set B.

The convex hull of all Dirac distributions $\mathcal{D}(B)$ is nevertheless not the standard probability simplex $\mathcal{P}(B)$. Indeed, the convex hull of $\mathcal{D}(B)$ is the set of all probability distributions in $\mathcal{P}(B)$ having a countable support on B which we denote as $\mathcal{F}(B)$. It is not difficult to show that $\mathcal{F}(B)$ is dense in $\mathcal{P}(B)$ under the weak topology. We thus obtain that the standard probability simplex is the closed convex hull of its extreme distributions, i.e. $\mathcal{P}(B) = \operatorname{cl}\operatorname{conv}\mathcal{D}(B)$, and that the closure in Theorem 3.2 is necessary.

Compactness, although sufficient, is not a necessary condition for a convex set $\mathcal{C}$ to be the closed convex hull of its extreme measures. Indeed, while $\mathcal{P}_n$ is not compact it still is the closed convex hull of its extreme distributions $\mathcal{D}(\mathrm{R}^n)$.

It remains to develop a counterpart to the Choquet representation discussed for convex sets in R^n in Definition 2.6. In order to arrive at a natural counterpart of Choquet representable sets in $\mathcal{E}_n$, we must extend Definition 2.4 of convex combinations to include a continuum of points.

Definition 3.10 (Mixing combination). *We call a point $\int \mathbb{x}\ \mathbb{m}(\mathrm{d}\mathbb{x})$, with mixing distribution $\mathbb{m} : \mathcal{B}(\mathcal{S}) \to [0,1]$, a mixing combination of the set $\mathcal{S}$.*

It can be shown that all mixing combinations of a set $\mathcal{S}$ are elements of its closed convex hull $\operatorname{cl}\operatorname{conv}\mathcal{S}$. Furthermore, the mixing hull of a set $\mathcal{S}$ can be defined as

$$\operatorname{mix}\mathcal{S} := \left\{ \int \mathbb{x}\ \mathbb{m}(\mathrm{d}\mathbb{x}) \ : \ \mathbb{m} : \mathcal{B}(\mathcal{E}_n) \to [0,1], \quad \mathbb{m}(\mathcal{S}) = 1 \right\}. \tag{3.3}$$

Using the notion of mixing combination, we can define in a similar fashion as in Definition 2.6 Choquet representable sets in $\mathcal{E}_n$.

Definition 3.11 (Choquet representation). *A convex set $\mathcal{C}$ in $\mathcal{E}_n$ is said to admit a Choquet representation if for every measure $\mathbb{P} \in \mathcal{C}$ there exists a mixing distribution $\mathbb{m} : \mathcal{B}(\mathcal{C}) \to [0,1]$ supported on $\operatorname{ex}\mathcal{C}$ such that*

$$\mathbb{P} = \int \mathbb{x}\ \mathbb{m}(\mathrm{d}\mathbb{x}). \tag{3.4}$$

We remark here that the mixing distribution $\mathbb{m}$ in equation (3.3) and Definition 3.11 is not a measure on R^n. Rather, the probability distribution $\mathbb{m}$ is an element of the vector space of measures on the space of measures $\mathcal{E}_n$. Unfortunately as shown by Bishop and De Leeuw [24], the extreme measures of a set of measures $\mathcal{C}$ need not form a measurable set. Thus, the statement that the probability distribution $\mathbb{m}$ is supported by the extreme points of $\mathcal{C}$ in Definition 3.11 is in that case meaningless under our present definitions.

The existence of a Choquet representation for a convex set $\mathcal{C}$ in a general topological vector space is the topic of Choquet theory. A standard reference on Choquet theory are the lecture notes of Phelps [101]. It can be shown that under the relatively mild assumption that $\operatorname{ex}\mathcal{C}$ is metrizable, convex and compact such Choquet representation is always well posed and exists. In this work however, we will take the simplifying assumption that the set of extreme measures $\operatorname{ex}\mathcal{C}$ admits a spatial parametrization and circumvent many technicalities by doing so.

Definition 3.12 (Spatial parameterization [103])**.** *We say that the set of extreme measures of a convex set $\mathcal{C}$ admits a spatial parameterization if*

$$\operatorname{ex}\mathcal{C} = \{\mathbb{e}_x \ : \ x \in X\},$$

where $x \in \mathrm{R}^\ell$ parameterizes the extreme measures of $\mathcal{C}$ and ranges over a closed convex set $X \subseteq \mathrm{R}^\ell$, while the mapping $x \mapsto \mathbb{e}_x(B)$ is a measurable function for any fixed measurable set B.

The spatial parametrization has as a main benefit that it eliminates the need of having to deal with a mixing distribution $\mathbb{m}$ in a space more general then $\mathcal{E}_n$ as discussed in the previous paragraph. Indeed using the spatial parametrization, the integral (3.4) can be written as

$$\mathbb{P} = \int \mathbb{x} \ \mathbb{m}(\mathrm{d}\mathbb{x}) = \int \mathbb{e}_x \ \tilde{\mathbb{m}}(dx),$$

with $\tilde{\mathbb{m}}$ a mixing distribution in $\mathcal{P}_\ell(X)$. As we only deal with spatial parametrizable extreme point sets $\operatorname{ex}\mathcal{C}$ in the remainder of this work, we do not have to deal with the more general mixing distribution $\mathbb{m}$ but deal with the mixing distribution $\tilde{\mathbb{m}} \in \mathcal{P}_\ell(X)$ instead.

The next theorem unveils a deep connection between Choquet representations and the set of extreme distributions. We will indicate that a Choquet representation can be used to argue that a certain set $\partial\mathcal{K}$ of distributions contains all extreme distributions of a given set $\mathcal{K}$.

Theorem 3.3 (Extreme representations)**.** *Suppose we have a set of probability distributions $\mathcal{K}$ in $\mathcal{E}_n$ and a set $\partial\mathcal{K} = \{\mathbb{e}_x \ : \ x \in \mathrm{R}^n\} \subseteq \mathcal{K}$ that satisfies the condition* $\operatorname{mix}\partial\mathcal{K} \supseteq \mathcal{K}$*, then $\partial\mathcal{K} \supseteq \operatorname{ex}\mathcal{K}$.*

Proof. For the sake of contradiction, assume that there exists a probability distribution $\mathbb{P} \in \operatorname{ex}\mathcal{K} \setminus \partial\mathcal{K}$. The probability distribution $\mathbb{P}$ is hence extreme in the set $\mathcal{K}$, but not contained in the set $\partial\mathcal{K}$. From the premise $\operatorname{mix}\partial\mathcal{K} \supseteq \mathcal{K}$ it follows that there exists a probability distribution $\mathbb{m}$ such that

$$\mathbb{P} = \int \mathbb{e}_x \, \mathbb{m}(\mathrm{d}x). \tag{3.5}$$

Observe that the condition $\mathbb{P} \notin \partial\mathcal{K}$ implies that the probability distribution $\mathbb{m} \notin \mathcal{D}(\mathrm{R}^n)$ can not be a Dirac distribution. Therefore there exists a measurable set B such that $\mathbb{m}(B) \notin \{0,1\}$. Consider now the restricted distributions $\mathbb{m}[B]$ and $\mathbb{m}[\mathrm{R}^n \setminus B]$ defined through $\mathbb{m}[X] : B \mapsto \mathbb{m}(B \cap X)/\mathbb{m}(X)$. It is clear that we have the following chain of equalities from equation (3.5)

$$\begin{aligned}
\mathbb{P} &= \int_B \mathbb{e}_x \, \mathbb{m}(\mathrm{d}x) + \int_{\mathrm{R}^n \setminus B} \mathbb{e}_x \, \mathbb{m}(\mathrm{d}x) \\
&= \mathbb{m}(B) \int_B \frac{\mathbb{e}_x}{\mathbb{m}(B)} \, \mathbb{m}(\mathrm{d}x) + \mathbb{m}(\mathrm{R}^n \setminus B) \int_{\mathrm{R}^n \setminus B} \frac{\mathbb{e}_x}{\mathbb{m}(\mathrm{R}^n \setminus B)} \, \mathbb{m}(\mathrm{d}x) \\
&= \mathbb{m}(B) \int \mathbb{e}_x \, \mathbb{m}[B](\mathrm{d}x) + (1 - \mathbb{m}(B)) \int \mathbb{e}_x \, \mathbb{m}[\mathrm{R}^n \setminus B](\mathrm{d}x)
\end{aligned}$$

which contradicts the fact that $\mathbb{P}$ is extreme in $\mathcal{K}$. Indeed, the preceding equality reads that $\mathbb{P}$ is the strict convex combination of two distinct probability distributions $\mathbb{m}[B]$ and $\mathbb{m}[\mathrm{R}^n \setminus B]$. We must thus conclude that $\operatorname{ex}\mathcal{K} \subseteq \partial\mathcal{K}$. $\square$

The previous theorem thus guarantees that if a set of distributions $\mathcal{K}$ can be represented as the mixing combination of another set $\partial\mathcal{K}$, then the set of distributions $\partial\mathcal{K}$ necessarily contains all the extreme points of the set $\mathcal{K}$. The reverse statement does obviously not hold.

We will show in what remains of this section that both the set of all α-unimodal distributions $\mathcal{U}_\alpha$ and the set of all γ-monotone distributions $\mathcal{M}_\gamma$ admit Choquet representations despite not being compact. We first discuss what the extreme distributions of $\mathcal{U}_\alpha$ and $\mathcal{M}_\gamma$ look like. In order to do so we start with defining radial probability distributions.

Definition 3.13 (Radial probability distributions)**.** *We will refer to a probability distribution $\mathbb{P}$ supported on a ray emanating from the origin as a radial probability distribution, i.e. $\mathbb{P}(\{\lambda x \ : \ \lambda \geq 0\}) = 1$ for some $x \in \mathrm{R}^n$.*

Where the Dirac distributions $\mathcal{D}(\mathrm{R}^n)$ were the extreme distributions of the standard probability simplex $\mathcal{P}_n$, both sets $\mathcal{U}_\alpha$ and $\mathcal{M}_\gamma$ have (distinct) sets of radial probability distributions as their extreme distributions. The fact that both sets of structured probability distributions have known extreme distributions will be instrumental in Chapters 6 and 7. In the last part of this section we indicate that both the set of α-unimodal and γ-monotone distributions can be put on the same footing and even generalized further.

Unimodal distributions

For any $\alpha > 0$ and $x \in \mathrm{R}^n$ we denote by $\mathbb{u}_x^\alpha$ the radial distribution supported on the line segment $[0, x]$ in R^n with the property that

$$\mathbb{u}_x^\alpha\left([0, tx]\right) = \alpha \cdot \int_0^t \lambda^{\alpha-1} \, \mathrm{d}\lambda \quad \forall t \in [0, 1]. \tag{3.6}$$

The next proposition establishes that the probability distributions $\mathbb{u}_x^\alpha$ are extreme in $\mathcal{U}_\alpha$.

Proposition 3.2 (Extreme distributions of $\mathcal{U}_\alpha$ [44])**.** *The set of all α-unimodal distributions $\mathcal{U}_\alpha$ has as extreme distributions $\mathrm{ex}\,\mathcal{U}_\alpha = \{\mathbb{u}_x^\alpha \ : \ x \in \mathrm{R}^n\}$ and admits a Choquet representation of the form*

$$\forall \mathbb{P} \in \mathcal{U}_\alpha, \ \exists! \mathbb{m} \in \mathcal{P}_n : \quad \mathbb{P} = \int \mathbb{u}_x^\alpha \ \mathbb{m}(\mathrm{d}x).$$

One can confirm that the radial probability distribution $\mathbb{u}_x^\alpha$ is an element of $\mathcal{U}_\alpha$ by direct application of Definition 3.5. Indeed, we have

$$\begin{aligned} t^\alpha \mathbb{u}_x^\alpha(B/t) &= t^\alpha \int_{\mathrm{R}^n} \mathbf{1}_B(yt) \, \mathbb{u}_x^\alpha(\mathrm{d}y) \\ &= t^\alpha \int_0^1 \mathbf{1}_B(xt\lambda) \, \alpha\lambda^{\alpha-1} \mathrm{d}\lambda \\ &= \int_0^t \mathbf{1}_B(x\lambda) \, \alpha\lambda^{\alpha-1} \mathrm{d}\lambda \, , \end{aligned}$$

and the last expression is manifestly non-decreasing in $t \in \mathrm{R}_{++}$. Alternatively, one can express the radial probability distributions $\mathbb{u}_x^\alpha$ as weak limits of continuous probability distributions that are readily identified as members of $\mathcal{U}_\alpha$. As $\mathcal{U}_\alpha$ is closed under weak convergence, one can again conclude that $\mathbb{u}_x^\alpha \in \mathcal{U}_\alpha$. For example, denote by $\mathbb{P}_\theta$ the uniform probability distribution on the intersection of the closed ball $B(\|x\|_2) = \{y \in \mathrm{R}^n \ : \ \|y\|_2 \leq \|x\|_2\}$ and the second-order cone

$$K(x, \theta) = \left\{ y \in \mathrm{R}^n : \frac{x^\top y}{\|x\|} \leq \tan(\theta) \left\| \left(\mathbb{I} - \frac{x x^\top}{\|x\|^2} \right) y \right\| \right\}$$

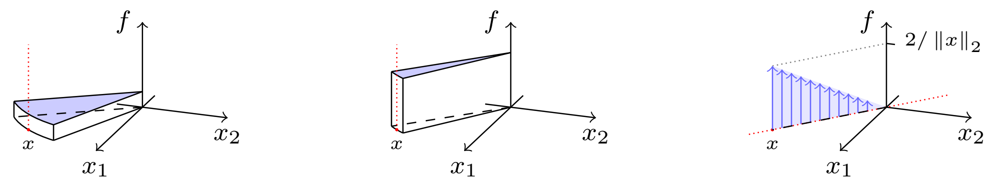


Figure 3.5: The radial measure $\mathbb{u}_x^n \in \mathcal{U}_n$ depicted on the right can be seen as the limit of uniform probability distributions on the wedge $K(x,\theta)$ for θ tending to zero. Remark that $\mathbb{u}_x^n$ is not a continuous probability distribution and hence admits no density function on R^n. The density of $\mathbb{u}_x^n$ depicted is with respect to the Lebesque measure on the affine hull of its support; see also Example A.1.

with principal axis x and opening angle $\theta \in (0, \pi/2)$. As both $B(\|x\|_2)$ and $K(x,\theta)$ are star-shaped, $\mathbb{P}_\theta$ is star-unimodal. Using standard arguments, one can show that $\mathbb{P}_\theta$ converges weakly to $\mathbb{u}_x^n$ as θ tends to 0. This confirms the (maybe surprising) result that $\mathbb{u}_x^n$ is star-unimodal as $\mathcal{U}_n$ is closed. The weak convergence of $\mathbb{P}_\theta$ to $\mathbb{u}_x^n$ for θ tending to zero is depicted in Figure 3.5.

Monotone distributions

For any natural number $\gamma \geq 1$ and $x \in \mathrm{R}^n$ we denote by $\mathbb{m}_x^\gamma$ the unique radial distribution supported on the line segment $[0, x] \subset \mathrm{R}^n$ with the property that

$$\mathbb{m}_x^\gamma([0, tx]) = \frac{1}{B(n,\gamma)} \cdot \int_0^t \lambda^{n-1} \cdot (1-\lambda)^{\gamma-1} \, \mathrm{d}\lambda \quad \forall t \in [0,1], \tag{3.7}$$

using the beta function or Euler integral of the first kind $B : \mathrm{R}_{++}^2 \to \mathrm{R}_{++}$. The next proposition establishes that the probability distributions $\mathbb{m}_x^\gamma$ are extreme in $\mathcal{M}_\gamma$.

Proposition 3.3 (Extreme distributions of $\mathcal{M}_\gamma$ [13])**.** *The set of all γ-monotone distributions $\mathcal{M}_\gamma$ has as extreme distributions* $\operatorname{ex} \mathcal{M}_\gamma = \{\mathbb{m}_x^\gamma \ : \ x \in \mathrm{R}^n\}$ *and admits a Choquet representation of the form*

$$\forall \mathbb{P} \in \mathcal{M}_\gamma, \ \exists ! \mathbb{m} \in \mathcal{P}_n : \quad \mathbb{P} = \int \mathbb{m}_x^\gamma \, \mathbb{m}(\mathrm{d}x).$$

It is not very hard to verify that the sequence $\mathbb{m}_{x\gamma}^\gamma$ of radial monotone distributions converges weakly, when γ tends to infinity, to a radial probability distribution $\mathbb{m}_x^\infty$ supported on the ray $\{\lambda x \ : \ \lambda \in \mathrm{R}_+\}$ with the property

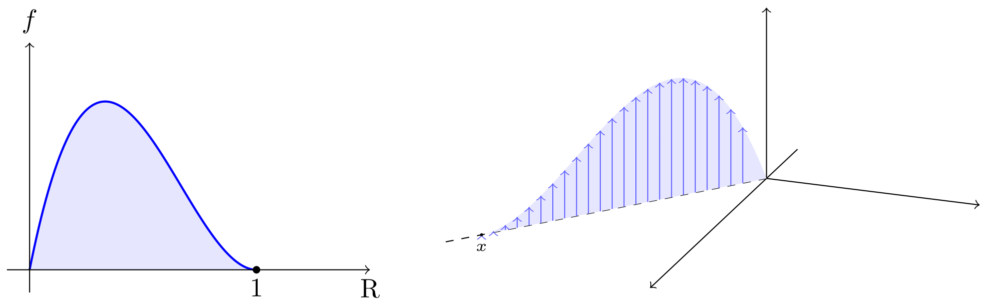
$$\mathbb{m}_x^\infty([0, tx]) = \frac{1}{\Gamma(n)} \int_0^t \lambda^{n-1} \mathrm{e}^{-\lambda} \, \mathrm{d}\lambda \quad \forall t \in [0, \infty),$$

using the gamma function or Euler integral of the second kind $\Gamma : \mathrm{R} \to \mathrm{R}$. Note that for positive integers, the gamma function reduces to $\Gamma(n) = (n-1)!$.

Proposition 3.4 (Bernstein's theorem)**.** *The set of all completely monotone distributions $\mathcal{M}_\infty$ has as extreme distributions* $\operatorname{ex} \mathcal{M}_\infty = \{\mathbb{m}_x^\infty \ : \ x \in \mathrm{R}^n\}$ *and admits a Choquet representation of the form*

$$\forall \mathbb{P} \in \mathcal{M}_\infty, \ \exists ! \mathbb{m} \in \mathcal{P}_n : \quad \mathbb{P} = \int \mathbb{m}_x^\infty \, \mathbb{m}(\mathrm{d}x).$$

In view of the Proposition 3.4, the set of completely monotone distributions is included as the limit of the hierarchy of γ-monotone ambiguity sets $\mathcal{M}_\gamma$ for γ tending to infinity.



(a) Generating distribution $\mathbb{T}$ (b) Radial distributions $\mathbb{T}_x$

Figure 3.6: Visual illustration of Definition 3.14. Consider the univariate distribution $\mathbb{T}$ which generates a family of distributions $\{\mathbb{T}_x \ : \ x \in \mathrm{R}^n\}$ as illustrated in the two figures above. The univariate distribution $\mathbb{T}$ dictates the shape of $\mathbb{T}_x$ along any direction x in R^n. A convex set $\mathcal{K}$ is a Choquet star simplex if there exists a distribution $\mathbb{T}$ such that all extreme distributions $\mathrm{ex}\,\mathcal{K} = \{\mathbb{T}_x \ : \ x \in \mathrm{R}^n\}$ are generated by $\mathbb{T}$.

Choquet star simplices

Propositions 3.2 and 3.3 indicate that both the set of α-unimodal distributions $\mathcal{U}_\alpha$ and γ-monotone distributions $\mathcal{M}_\gamma$ admit unique Choquet representations in terms of radial extreme distributions. As the Choquet representation is unique in either case, the sets $\mathcal{U}_\alpha$ and $\mathcal{M}_\gamma$ can thus be regarded as simplicial sets in $\mathcal{E}_n$. Indeed, a characteristic property of simplices in R^n is that any element can be expressed as a unique convex combination of its (finitely many) extreme points.

It will proof advantageous to the exposition of this dissertation to put the sets of α-unimodal and γ-monotone distributions on equal footing via the notion of Choquet star representable set.

Definition 3.14 (Choquet star representation)**.** *Suppose that $\mathbb{T}$ is a univariate measure on R_+, and define a derived family of radial measures $\mathbb{T}_x$ on R^n such that, for every $x \in \mathrm{R}^n$ and every set B,*

$$\mathbb{T}_x(C) = \mathbb{T}(\{\lambda \geq 0 \ : \ \lambda x \in B\}).$$

We say that the a closed convex set of measures $\mathcal{K}$ admits a Choquet star representation if it admits a unique Choquet representation over

$$\mathrm{ex}\,\mathcal{K} = \{\mathbb{T}_x \ : \ \forall x \in \mathrm{R}^n\}\,. \tag{3.8}$$

In this case we say that $\mathcal{K}$ is generated by the univariate measure $\mathbb{T}$.

A Choquet star simplex $\mathcal{K}$ is hence a convex set of measures for which the extreme measures $\mathrm{ex}\,\mathcal{K} = \{\mathbb{T}_x \ : \ x \in \mathrm{R}^n\}$ are radially scaled versions of a unique generating measure $\mathbb{T}$ on R_+. In Figure 3.6 the previous statement is visually illustrated. We will now make Definition 3.14 concrete by showing that many convex sets encountered up to this point are, in fact, Choquet star simplices. The results are condensed in Table 3.1.

Example 3.1 (The standard probability simplex $\mathcal{P}_n$)**.** *As stated in Proposition 3.1, the extreme distributions of the standard probability simplex are the Dirac distributions, i.e. $\mathrm{ex}\,\mathcal{P}_n = \{\delta_x \ : \ x \in \mathrm{R}^n\}$. The set of all Dirac distributions is recognized to be generated by the univariate Dirac distribution at unity δ_1. The standard probability simplex $\mathcal{P}_n$ is thus a Choquet star simplex generated by $\mathbb{T} = \delta_1$.*

Choquet simplex $\mathcal{K}$	Generator $\mathbb{T}$	$\int_0^\infty t\,\mathbb{T}(\mathrm{d}t)$	$\int_0^\infty t^2\,\mathbb{T}(\mathrm{d}t)$	$\operatorname{supp}\mathbb{T} \subseteq \mathrm{R}_+$
Standard simplex $\mathcal{P}_n$	δ_1	1	1	$\{1\}$
α-Unimodal $\mathcal{U}_\alpha$	$\mathbb{u}_1^\alpha$	$\frac{\alpha}{\alpha+1}$	$\frac{\alpha}{\alpha+2}$	$[0,1]$
γ-Monotone $\mathcal{M}_\gamma$	$\mathbb{m}_1^\gamma$	$\frac{n}{n+\gamma}$	$\frac{n}{n+\gamma}\,\frac{n+1}{n+\gamma+1}$	$[0,1]$
Completely monotone $\mathcal{M}_\infty$	$\mathbb{m}_1^\infty$	$n!$	$(n+1)!$	R_+

Table 3.1: The generating distributions $\mathbb{T}$ and their properties for the Choquet simplices most relevant to this dissertation.

Example 3.2 (The α-unimodal distributions $\mathcal{U}_\alpha$)**.** *The extreme distributions of the set of all α-unimodal distributions* $\operatorname{ex}\mathcal{U}_\alpha = \{\mathbb{u}_x^\alpha \,:\, x \in \mathrm{R}^n\}$ *are given in Proposition 3.2. The set of all radial unimodal distributions $\mathbb{u}_x^\alpha$ is recognized to be generated by the univariate extreme radial distribution*

$$\mathbb{u}_1^\alpha(B) = \alpha \int_B t^{\alpha-1}\mathrm{d}t, \quad \forall B \subseteq [0,1].$$

The set of all α-unimodal distributions $\mathcal{U}_\alpha$ is a Choquet star simplex generated by $\mathbb{T} = \mathbb{u}_1^\alpha$.

Example 3.3 (The γ-monotone distributions $\mathcal{M}_\gamma$)**.** *The extreme distributions of the set of all γ-monotone distributions* $\operatorname{ex}\mathcal{M}_\gamma = \{\mathbb{m}_x^\gamma \,:\, x \in \mathrm{R}^n\}$ *are given in Proposition 3.3. The set of all radial monotone distributions $\mathbb{m}_x^\gamma$ is recognized to be generated by the univariate extreme radial distribution*

$$\mathbb{m}_1^\gamma(B) = \frac{1}{B(n,\gamma)} \int_B t^{n-1}(1-t)^{\gamma-1}\mathrm{d}t, \quad \forall B \subseteq [0,1].$$

The set of all γ-monotone distributions $\mathcal{M}_\gamma$ is a Choquet star simplex generated by $\mathbb{T} = \mathbb{m}_1^\gamma$.

Example 3.4 (The completely monotone distributions $\mathcal{M}_\infty$)**.** *The extreme distributions of the set of all completely monotone distributions* $\operatorname{ex}\mathcal{M}_\gamma = \{\mathbb{m}_x^\infty \,:\, x \in \mathrm{R}^n\}$ *are given in Proposition 3.4. The set of all radial completely monotone distributions $\mathbb{m}_x^\infty$ is recognized to be generated by the univariate extreme radial distribution*

$$\mathbb{m}_1^\infty(B) = \frac{1}{\Gamma(n)} \int_B t^{n-1}\mathrm{e}^{-t}\ \mathrm{d}t, \quad \forall B \subseteq \mathrm{R}_+.$$

The set of all completely monotone distributions $\mathcal{M}_\infty$ is a Choquet star simplex generated by $\mathbb{T} = \mathbb{m}_1^\infty$.

From Definition 3.14 it is evident that if a convex set $\mathcal{C}$ admits a Choquet star representation, then it is isomorphic to the standard probability simplex $\mathcal{P}_n$. We might therefore also refer to a Choquet star representable set as a Choquet star simplex. The power of Choquet star representable sets will become clear already in subsequent section. Choquet star representable sets are of particular importance when we will discuss optimization over convex ambiguity sets.

In what remains of this section, we will indicate that Choquet star simplices are preserved under linear projection. Define the projection operator P_A as the function which maps a probability distribution $\mathbb{P}$ in $\mathcal{P}_n$ to a probability distribution $\mathbb{Q}$ in $\mathcal{P}_m$ in accordance to

$$\mathbb{Q}(B) = \mathbb{P}(\{x \in \mathrm{R}^n \,:\, Ax \in B\}), \quad \forall B \in \mathcal{B}(\mathrm{R}^m)$$

where $A \in \mathrm{R}^{m\times n}$. The projection operator P_A is best understood by considering its relation with the projection of random variables. Indeed if a random variable ξ valued in R^n is distributed as $\mathbb{P}$ then its projection $A\xi$ will be distributed as $P_A(\mathbb{P})$. Slightly abusing notation, we will further denote with $P_A(\mathcal{C})$ the projection of an ambiguity set $\mathcal{C}$, i.e.

$$P_A(\mathcal{C}) = \{\mathbb{Q} \in \mathcal{P}_m \,:\, \exists \mathbb{P} \in \mathcal{C},\ \ \mathbb{Q} = P_A(\mathbb{P})\}\,.$$

The following theorem indicates that Choquet star simplices are preserved under linear projection and will find its application in Chapters 8 and 9.

Theorem 3.4 (Projection theorem)**.** *Consider two Choquet star simplices generated by the same generating measure* $\mathbb{T}$*, i.e.* $\mathcal{K}_n = \operatorname{mix}\{\mathbb{T}_x \,:\, x \in \mathrm{R}^n\}$ *and* $\mathcal{K}_m = \operatorname{mix}\{\mathbb{T}_x \,:\, x \in \mathrm{R}^m\}$*, then we have that*

$$\mathcal{K}_m = P_A(\mathcal{K}_n)$$

for any full row rank matrix A *in* $\mathrm{R}^{m\times n}$*.*

Proof. We need only consider the extreme points $\operatorname{ex}\mathcal{K}_n = \{\mathbb{T}_x \,:\, x \in \mathrm{R}^n\}$ as the projection operator is linear and thus $P_A(\mathcal{K}_n) = \operatorname{mix} P_A(\operatorname{ex}\mathcal{K}_n)$. For any set B we have $P_A(\mathbb{T}_x)(B) = \mathbb{T}_x(\{y \in \mathrm{R}^n \,:\, Ay \in B\}) = \mathbb{T}(\{\lambda \geq 0 \,:\, A\lambda x \in B\}) = \mathbb{T}_{Ax}(B)$. Hence, $P_A(\mathbb{T}_x) = \mathbb{T}_{Ax}$ for all $x \in \mathrm{R}^n$ and thus $P_A(\{\mathbb{T}_x \,:\, x \in \mathrm{R}^n\}) = \{\mathbb{T}_{Ax} \,:\, x \in \mathrm{R}^n\} = \{\mathbb{T}_x \,:\, x \in \mathrm{R}^m\}$, where the last equality requires A to be of full row rank. ☐

3.3 Moment problems

In 1884 Stieltjes published his now classical paper [122] containing a wealth of new ideas. In this paper Stieltjes introduced what is now known as the problem of moments already briefly referred to in Section 3.1. In its most classical form the problem of moments can be stated as finding a probability distribution $\mathbb{P}$ element of the moment set

$$\mathcal{C}(m) := \left\{\mathbb{P} \in \mathcal{E}_n^+ : \int g_i(x)\,\mathbb{P}(\mathrm{d}x) = m_i,\; i \in [0, \dots, k-1]\right\},$$

and thus is recognized as a feasibility problem. In the original investigations made by Stieltjes concerning the problem of moments [122] continuous fractions play the predominant role. Modern approaches as in Isii [63] or Lasserre [76] however all draw from the intimate relationship between the problem of moments and positive functions. In this section we will show how Section 2.4 on cones and dual cones resonates to more general vector spaces as well.

The problem of moments concerns itself with the question whether or not we can find a probability distribution $\mathbb{P}$ in the moment set $\mathcal{C}(m)$ and hence whether it is non-empty. In this section we will indicate how the moment problem can be posed in terms of dual cones. Denote with K the cone of moments m in R^k for which the moment set $\mathcal{C}(m)$ is non-empty, i.e.

$$K := \left\{m \in \mathrm{R}^k \;:\; \exists \mathbb{P} \in \mathcal{E}_n^+ \text{ s.t. } \mathbb{P} \in \mathcal{C}(m)\right\}. \tag{3.9}$$

The set K identifies those moments m for which the moment set $\mathcal{C}(m)$ is non-empty and hence the moment problem admits a solution. As the first moment function is taken to be $g_0(x) = 1$, the set $M = K \cap \left\{m \in \mathrm{R}^k \,:\, m_0 = 1\right\}$ then represents all moments for which a probability distribution can be found in $\mathcal{C}(m)$. The dual cone $K^\star$ can be determined explicitly and has a nice geometrical interpretation. From Proposition 2.8 it follows that both cones K and $K^\star$ determine each other up to a closure operation uniquely, i.e. $\operatorname{cl} K = K^{\star\star}$, and thus both can essentially be regarded as a solution to the moment problem.

Proposition 3.5 (Dual positive functions)**.** *The dual cone* $K^\star$ *of the cone* K *defined in equation* (3.9) *is*

$$K^\star = \left\{a \in \mathrm{R}^k \;:\; \sum_{i=0}^{k-1} a_i g_i(x) \geq 0, \quad \forall x \in \mathrm{R}^n\right\},$$

and can be stated in terms a semi-infinite constraint.

The cone $K^\star$ admits a nice geometrical interpretation. Indeed, the cone $K^\star$ consists of all linear combinations a_i of the moment functions g_i for which the function $\sum_{i=0}^{k-1} a_i g_i$ is positive. We remark that the cone $K^\star$ is given in terms of a semi-infinite constraint discussed further in Chapter 4.

Proof. The following identities are all standard and are included merely for the sake of completeness.

$$\begin{aligned} K^\star &:= \left\{a \in \mathrm{R}^k \ : \ a^\top m \geq 0, \quad \forall m \in K\right\}, \\ &= \left\{a \in \mathrm{R}^k \ : \ a^\top m \geq 0, \quad \exists \mathbb{P} \in \mathcal{E}_n^+ \text{ s.t. } \textstyle\int g_i(x)\, \mathbb{P}(\mathrm{d}x) = m_i \ \ \forall i\right\}. \end{aligned}$$

We can now restate the inner product as the sum $\sum_i a_i m_i$ where $m_i = \int g_i(x)\, \mathbb{P}(\mathrm{d}x)$ after which we obtain

$$K^\star = \left\{a \in \mathrm{R}^k \ : \ \textstyle\int \left[\sum_i a_i g_i(x)\right] \mathbb{P}(\mathrm{d}x) \geq 0, \quad \forall \mathbb{P} \in \mathcal{E}_n^+\right\}$$

As the integral of $\sum_i a_i g_i$ needs to be non-negative with respect to all measures $\mathbb{P}$ in $\mathcal{E}_n^+$ the integrand must be a positive function. $\square$

Second-order moment problems

In what remains of this section we will illustrate the power of previous analysis for the moment set $\mathcal{C}(\mu, S) := \mathcal{H}(\mu, S) \cap \mathcal{P}_n$ defined as the collection of all probability distributions in $\mathcal{P}_n$ sharing a known mean vector $\mu \in \mathrm{R}^n$ and second moment matrix $S \in \mathrm{S}^n$, i.e.

$$\mathcal{H}(\mu, S) := \left\{\mathbb{P} \in \mathcal{E}_n \ : \ \int \mathbb{P}(dx) = 1, \ \int x\, \mathbb{P}(\mathrm{d}x) = \mu, \ \int xx^\top\, \mathbb{P}(\mathrm{d}x) = S\right\}.$$

The ambiguity set $\mathcal{C}(\mu, S)$ will play a protagonist role throughout the entire dissertation. The ambiguity set $\mathcal{C}(\mu, S)$ is recognized immediately as a moment set in the form (3.2) for $(n+1)(n+2)/2$ unique quadratic moment functions.

The following well known proposition solves the moment problem related to the ambiguity set $\mathcal{C}(\mu, S)$ with the help of Proposition 3.5 and will be of interest in Chapters 6 and 7. As we only have that cl $K = K^{\star\star}$, we will only proof the following proposition up to a closure. The proof of the complete proposition can be found for instance in other work of this author [128] and is quite standard.

Proposition 3.6 (Second-moment information)**.** *The set of all means μ and second moments S for which $\mathcal{C}(\mu, S)$ is non-empty can be represented as*

$$M_\delta := \{(1, \mu, S) \ : \ \exists \mathbb{P} \in \mathcal{C}(\mu, S)\} = \left\{(1, \mu, S) \in \mathrm{R} \times \mathrm{R}^n \times \mathrm{S}^n \ : \ \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} \succeq 0\right\}$$

which is a linear matrix inequality (LMI).

Proof. The proposition is an almost immediate corollary of Proposition 3.5 and the S-Lemma. Indeed Proposition 3.5 states that the dual cone $K_\delta^\star$ of $\mathcal{C}(\mu, S)$ can be found as

$$K_\delta^\star = \left\{(a_0, a_1, A_2) \in \mathrm{R} \times \mathrm{R}^n \times \mathrm{S}^n \ : \ a_0 + a_1^\top x + \mathrm{Tr}\left\{A_2 \cdot xx^\top\right\} \geq 0, \quad \forall x \in \mathrm{R}^n\right\}$$

for the particular moment functions $g_0(x) = 1 \in \mathrm{R}$, $g_1(x) = x \in \mathrm{R}^n$ and $g_2(x) = xx^\top \in \mathrm{S}^n$ with corresponding moments $m_0 = 1$, $m_1 = \mu$ and $M_2 = S$. With the help of the S-Lemma we can reexpress the previous semi-infinite constraint as an LMI yielding

$$K_\delta^\star = \left\{(a_0, a_1, A_2) \in \mathrm{R} \times \mathrm{R}^n \times \mathrm{S}^n \ : \ \begin{pmatrix} A_2 & \frac{1}{2}a_1 \\ \frac{1}{2}a_1^\top & a_0 \end{pmatrix} \succeq 0\right\}$$

The dual cone of $K_\delta^\star$ can be explicitly calculated and is given as the cone

$$K_\delta^{\star\star} = \left\{ (m_0, m_1, M_2) \ : \ \begin{pmatrix} M_2 & m_1 \\ m_1^\top & m_0 \end{pmatrix} \succeq 0 \right\}.$$

The proposition then follows from the observation that the set M_δ is determined up to a closure operation as cl $M_\delta = K_\delta^{\star\star} \cap \{m \in \mathrm{R}^k \ : \ m_0 = 1\}$. □

In the Chapters 6 and 7 we will come across sets of probability distributions which, besides sharing a mean vector and second moment matrix, are more richly structured. More specifically, we will frequently encounter sets of probability distributions in the general form

$$\mathcal{H}(\mu, S) \cap \mathcal{K},$$

where the ambiguity set $\mathcal{K}$ admits a Choquet star representation with generating measure $\mathbb{T}$. The set $\mathcal{H}(\mu, S) \cap \mathcal{K}$ of probability distributions is not a moment set unless $\mathcal{K}$ is taken to be the standard probability simplex $\mathcal{P}_n$. Nevertheless, Proposition 3.6 can be used to derive the following proposition.

Proposition 3.7 (Second-moment information)**.** *The set of all means μ and second moments S for which $\mathcal{H}(\mu, S) \cap \mathcal{K}$ is non-empty can be represented as $M_\mathbb{T} = \{(1, \mu, S) \ : \ \exists \mathbb{P} \in \mathcal{H}(\mu, S) \cap \mathcal{K}\} =$*

$$\left\{ (1, \mu, S) \ : \ \begin{pmatrix} S_s & \mu_s \\ \mu_s^\top & 1 \end{pmatrix} \succeq 0, \ S_s \cdot \int_0^\infty t^2 \, \mathbb{T}(\mathrm{d}t) = S, \ \mu_s \cdot \int_0^\infty t \, \mathbb{T}(\mathrm{d}t) = \mu \right\}$$

which is an LMI.

Proof. As the set $\mathcal{K}$ admits a Choquet representation in terms of the generating distribution $\mathbb{T}$, we have

$$M_\mathbb{T} = \left\{ (1, \mu, S) \ : \ \exists \mathbb{m} \in \mathcal{P}_n, \ \int \mathbb{T}_x \, \mathbb{m}(\mathrm{d}x) \in \mathcal{H}(\mu, S) \right\}.$$

With the help of the substitution $\mathbb{P} = \int \mathbb{T}_x \, \mathbb{m}(\mathrm{d}x)$ we obtain the following two identities concerning the moment functions

$$\begin{aligned} \int y \, \mathbb{P}(dy) &= \int \left[\int y \mathbb{T}_x(\mathrm{d}y) \right] \mathbb{m}(\mathrm{d}x) \\ &= \int t \, \mathbb{T}(\mathrm{d}t) \cdot \int x \, \mathbb{m}(\mathrm{d}x) \end{aligned} \quad \text{and} \quad \begin{aligned} \int y \cdot y^\top \, \mathbb{P}(dy) &= \int \left[\int y \cdot y^\top \mathbb{T}_x(\mathrm{d}y) \right] \mathbb{m}(\mathrm{d}x) \\ &= \int t^2 \, \mathbb{T}(\mathrm{d}t) \cdot \int x x^\top \, \mathbb{m}(\mathrm{d}x) \end{aligned}$$

The final result is obtained with Proposition 3.6 by taking $\mu_s = \int x \, \mathbb{m}(\mathrm{d}x)$ and $S_s = \int x x^\top \, \mathbb{m}(\mathrm{d}x)$ with which we arrive at $M_\mathbb{T} = \{(1, \mu, S) \ : \ \exists \mathbb{m} \in \mathcal{C}(\mu_s, S_s)\}$. □

3.4 Risk measures

In many applications it is of interest to quantify the risk carried by a real random variable ξ representing an unknown outcome. Depending on the application the random variable ξ may express an economic loss or simply quantify how undesirable the outcomes of ξ are. A risk measure ρ assigns to random variables a number in $\bar{\mathrm{R}}$ and satisfies certain properties:

Definition 3.15 (Risk measure [2])**.** *A function ρ mapping a random variable ξ to $\bar{\mathrm{R}}$ which satisfies*

1. *Normalized*

$$\rho(0) = 0.$$

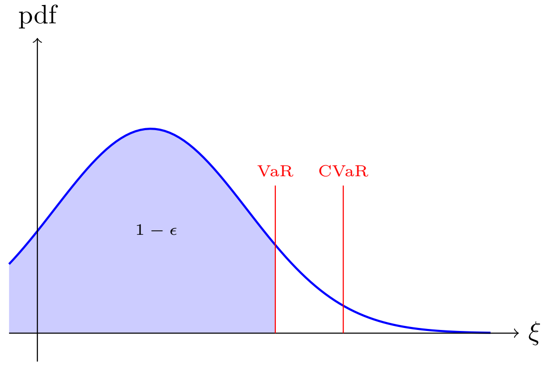


Figure 3.7: The VaR measure $\mathrm{VaR}_\epsilon(\xi)$ of the random variable ξ at risk level $\epsilon \in (0,1)$ quantifies the largest value of ξ which occurs with frequency at least ϵ. The CVaR measure $\mathrm{CVaR}_\epsilon(\xi)$ at level ϵ coincides for continuous probability distributions with the conditional expectation of the distribution of ξ above $\mathrm{VaR}_\epsilon(\xi)$.

2. *Translative*
$$\text{If } a \in \mathbb{R} \text{ then } \rho(\xi + a) = \rho(\xi) + a$$

3. *Monotone*
$$\text{If } \xi_1 \leq \xi_2 \text{ then } \rho(Z_2) \leq \rho(Z_1)$$

is denoted as a risk measure.

In this work we will only consider law invariant risk measures. A risk measure ρ is said to be law invariant if $\rho(\xi)$ depends only on the distribution of ξ; i.e. if ξ and ξ' are two random variables sharing the same distribution, then $\rho(\xi) = \rho(\xi')$. Informally, the law invariant assumption on the risk measure ρ has the benefit that all necessary information of a random variable ξ to compute its risk is subsumed in its distribution. We discuss in what remains of this chapter the three risk measures which are of relevance to this dissertation.

Value-at-Risk measure The value-at-risk (VaR) measure of a random variable ξ quantifies the largest loss occurring with odds at least $\epsilon \in (0,1)$; see Figure 3.7.

Definition 3.16 (Value-at-Risk)**.** *The value-at-risk of a univariate random variable ξ distributed as $\mathbb{P}$ is defined as*
$$\mathrm{VaR}_\epsilon(\xi) := \inf \{\beta \in \mathrm{R} \ : \ \mathbb{P}(\xi \geq \beta) \leq \epsilon\}\,. \tag{3.10}$$

The VaR measure is the predominant risk measure in for instance the financial and insurance risk quantification literature. The VaR is easily recognized as a quantile function of the distribution of random variable ξ and its use is encouragement in the Basel II banking accords, see Danielsson et al. [39], which most banks abide by. While the VaR is likely the most commonly used measure to quantify the risk carried by the random outcomes ξ in the literature, it comes with a few shortcomings. Indeed, the VaR measure is blind to the severity of the risk taken and hence encourages large but remote risks to be taken.

Conditional Value-at-Risk measure A risk measure which has been proposed as an alternative to the VaR measure, is its convex counterpart; the CVaR measure.

Definition 3.17 (Conditional Value-at-Risk)**.** *The conditional value-at-risk of a univariate random variable ξ distributed as $\mathbb{P}$ is defined as*

$$\mathrm{CVaR}_\epsilon(\xi) := \inf_{\beta \in \mathrm{R}} \left\{ \beta + \frac{1}{\epsilon} \mathbf{E}_{\mathbb{P}} \left[(\xi - \beta)^+ \right] \right\}. \tag{3.11}$$

Rockafellar and Uryasev [108] have shown that the set of optimal solutions for β in (3.11) is a closed interval whose left endpoint is given by $\mathrm{VaR}_\epsilon(\xi)$. Moreover, it can be shown that if the random variable ξ follows a continuous probability distribution, then its CVaR at risk level ϵ coincides with the conditional expectation above the $\mathrm{VaR}_\epsilon(\xi)$-quantile; see Figure 3.7. This observation originally motivated the term *conditional value-at-risk*.

The CVaR measure enjoys a number of practical advantages over VaR measure since it is monotone, homogeneous and convex in the sense that:

$$t\mathrm{CVaR}_\epsilon(L_1(\xi)) + (1-t)\mathrm{CVaR}_\epsilon(L_2(\xi)) \geq \mathrm{CVaR}_\epsilon([tL_1 + (1-t)L_2](\xi))$$

for all measurable real functions L_1 and L_2 as shown in Pflug [100]. The CVaR measure assigns a higher risk to larger realizations of ξ. In contrast, the VaR measure assigns a uniform risk to all realizations irrespective of their size above a given quantile.

Expectation The most elementary but nevertheless versatile risk measure is the expectation of ξ as measured by a to be chosen loss function L,

$$\rho(\xi) := \mathbf{E}_{\mathbb{P}}[L(\xi)] \tag{3.12}$$

The measurable function L is most often referred to as a loss function as it quantifies the severity of the realized outcome of the random variable ξ. The risk measure ρ as defined in equation (3.12) assigns the expected severity $L(\xi)$ as risk to the random variable ξ. As the loss function L can be chosen in accordance to the severity quantification most suitable to the practitioner, the expectation risk measure ρ as presented in equation (3.12) is quite flexible. In Chapter 5 we will indicate that many risk quantification problems with VaR or CVaR measure can in fact be reduced to the expectation risk measure defined in (3.12) for a judiciously chosen loss function L.

4 Convex optimization

In this chapter we state various results concerning convex optimization problems which are used throughout this dissertation. Yet again the purpose of the current chapter is twofold. Firstly, it is meant to state a selection of results concerning convex optimization problems based on their relevance to this dissertation. Yet at the same time this chapter will also provide an analogue to the corresponding results in more general vector spaces as discussed in Chapter 5.

A convex optimization program in R^n is a problem in the canonical form

$$\begin{aligned} \sup \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \geq 0, \quad \forall i \in [1, \ldots, k] \\ & g_j(x) = 0, \quad \forall j \in [1, \ldots, \ell] \end{aligned} \tag{4.1}$$

where any of the functions $f_i : \mathrm{R}^n \to \bar{\mathrm{R}}$ is concave and the functions $g_j : \mathrm{R}^n \to \mathrm{R}$ are all affine. The function f_0 is denoted as the objective function, while the functions f_i and g_j are referred to as the inequality and equality constraint functions, respectively. The feasible set of the convex optimization problem (4.1)

$$C := \{x \in \mathrm{R}^n \ : \ f_i(x) \geq 0, \ \ \forall i \in [1, \ldots, k] \quad \text{and} \quad g_i(x) = 0, \ \ \forall j \in [1, \ldots, \ell]\}$$

is always a convex set.

The canonical form (4.1) is slightly unorthodox in that it defines a convex optimization problem in terms of the supremum of a concave objective function, instead of the more customary infimum of a convex objective function. Please note that any infimum of a convex function can be recast as the supremum of a concave objective function with the help of a sign change. We justify our use of the slightly unorthodox canonical form (4.1) so that the current chapter bears a closer resemblance to Chapter 5 in which the supremum is the more conventional form.

We remark that a huge variety of optimization problems in practice can be phrased in the canonical form offered in (4.1). During the last decade, a wide spectrum of problems from different fields have been recognized as convex optimization problems. An excellent selection of these problems can be found in Boyd and Vandenberghe [27]. The selection of the results presented here is entirely based on their further relevance to the dissertation. The reader may be referred again to the standard work by Boyd and Vandenberghe [27] for a more complete treatment of convex optimization.

4.1 The optimization hierarchy

In this section we discuss the classes of convex optimization problems most relevant for our purpose. These classes of optimization problems will be recognized as particular instances of the general class of convex optimization problems of the type (4.1) and will be introduced in order of increasing descriptive power. For historical reasons these classes of convex optimization problems are referred to alternatively as convex programming problems too.

Linear programming A linear program (LP) is a convex optimization problem which can be brought in the canonical form

$$\begin{aligned} \sup \quad & c^\top x \\ \text{s.t.} \quad & a_i^\top x \leq b_i, \quad i \in [1, \ldots, k]. \end{aligned} \tag{4.2}$$

The feasible set of an LP is a polyhedral set as defined in Example 2.2. Linear programs are the most classical examples of convex optimization problems and have been studied since at least the 1930s. Much of the theory on linear programming and many ideas in convex optimization in general were put forward by Kantorovich in 1939 motivated by his interest in transportation problems. Nowadays, the applications of linear programming to practice are so many and diverse that it would be impossible to list them here.

Quadratic programming A quadratic program (QP) is a convex optimization problem, see Lobo et al. [79], which can be brought in the canonical form

$$\begin{aligned} \sup \quad & x^\top Q x + c^\top x \\ \text{s.t.} \quad & a_i^\top x \leq b_i, \quad i \in [1, \ldots, k]. \end{aligned} \tag{4.3}$$

where $Q \in \mathrm{S}^n_+$ is a positive definite matrix. Quadratic programming can be seen as a generalization of the classical least-square problem as discussed by Stigler [123]. It should hence not come as a surprise that quadratic programming is an indispensable tool in regression analysis; see for instance Suykens and Vandewalle [125]. Research on quadratic programming took off in the 1950th and was partially motivated by the portfolio optimization problem first posed by Markowitz [83]. Quadratic programming is also frequently encountered in optimization driven solution strategies to constrained control problems such as model predictive control (MPC). Note that when the matrix Q is zero then the QP (4.3) reduces to the LP (4.2).

Second-order cone programming A second-order cone program (SOCP) is a convex optimization problem, see again Lobo et al. [79], which can be brought in the canonical form

$$\begin{aligned} \sup \quad & c^\top x \\ \text{s.t.} \quad & \|C_i x + d_i\|_2 \leq a_i^\top x + b_i, \quad i \in [1, \ldots, k]. \end{aligned} \tag{4.4}$$

Second-order cone programming is intimately related to robust optimization. Worst-case robust optimization problems with linear constraints indeed do give rise to second-order cone (SOC) constraints. Furthermore, in Chapter 8 it will be argued that many stochastic robust optimization problems share a similar affinity to second-order cone programming as well. Note that if the matrices C_i are chosen to be zero in (4.4) then the SOCP reduces to an LP. It can be shown that SOCP problems are more expressive than QPs too. For a more exhaustive discussion on second-order cone programming and its applications the reader is referred to Lobo et al. [79].

Semi-definite programming A semi-definite program (SDP) is a convex optimization problem, see Boyd and Vandenberghe [27], which can be brought in the canonical form

$$\begin{aligned} \sup \quad & c^\top x \\ \text{s.t.} \quad & F_0 + \textstyle\sum_{i=1}^n F_i x_i \succeq 0. \end{aligned} \tag{4.5}$$

The generalized conic inequality in the optimization problem (4.5) is referred to as a LMI. For any two symmetric matrices $X, Y \in \mathrm{S}^d$ the relation $X \succeq Y$ ($X \preceq Y$) is short hand for

$X - Y \in \mathrm{S}^d_+$ ($Y - X \in \mathrm{S}^d_+$). The LMI in the optimization problem (4.5) hence represents a convex feasible set. Semidefinite programming had a tremendous impact on a wide range of disciplines as illustrated by the explosive growth of publications on the topic. The expressive power of SDP has for instance enabled the unified treatment of many problems and methods in linear control as pioneered by Boyd et al. [26]. Many results in this dissertation will involve LMI inequalities or will be stated in terms of an SDP. Please note that any SOCP can be rephrased as an SDP with an appropriate choice for the matrices F_i.

Semi-infinite programming A semi-infinite program is a convex optimization problem, see Hettich and Kortanek [61], which can be brought in the canonical form

$$\begin{array}{ll} \sup & c^\top x \\ \text{s.t.} & h_0(y) + \sum_{i=1}^n h_i(y) x_i \geq 0, \quad \forall y \in Y. \end{array} \tag{4.6}$$

with the help of the dual functions $h_i : \mathrm{R}^d \to \mathrm{R}$ for all $i \in [0, \ldots, n]$. Note that the feasible set of the semi-infinite program

$$C = \{x \in \mathrm{R}^n \ : \ h_0(y) + \textstyle\sum_{i=1}^n h_i(y) x_i \geq 0, \quad \forall y \in Y\} \tag{4.7}$$

is indeed a convex set as it is the intersection of possibly infinitely many half-spaces. Indeed, the feasible set C is represented through the intersection of as many half-spaces as there are elements in the set Y in R^d. For certain choices of the set Y and dual functions h_i the semi-infinite constraint (4.7) can be recast as an LMI. Two particular cases are of importance in the remainder of the work and are stated here for the sake of completeness.

Theorem 4.1 (The S-Lemma [102]). *Assume that the dual functions h_i for $i \in [0, \ldots, n]$ and the set Y are all quadratically representable, i.e. the functions h_i and set Y are in the form*

$$\begin{aligned} h_i(y) &= y^\top E_i y + 2 e_i^\top y + e_i^0, \quad \forall i \in [0, \ldots, n] \\ Y &= \left\{ y \in \mathrm{R}^d \ : \ y^\top F y + 2 f^\top y + f^0 \geq 0 \right\} \end{aligned}$$

If there exists now a Slater point $\bar{y} \in \operatorname{int} Y$, then the feasible set (4.7) *can be represented as an LMI.*

We remark that the Slater condition $\bar{y} \in \operatorname{int} Y = \left\{ y \in \mathrm{R}^d \ : \ y^\top F_i y + 2 f_i^\top y + f_i^0 > 0 \right\}$ is essential for Theorem 4.1 to hold. The Slater condition $\bar{y} \in \operatorname{int} Y$ is a constraint qualification condition on the set Y of a type which we shall encounter frequently throughout this dissertation.

Theorem 4.2 (Sum-of-squares representation [89]). *Assume that the dual functions h_i for $i \in [0, \ldots, n]$ are univariate polynomials, i.e. the functions $h_i : \mathrm{R} \to \mathrm{R}$ are in the form*

$$h_i(y) = \sum_{r=0}^d c_{i,r} y^r, \quad \forall i \in [0, \ldots, n],$$

and the set Y is a (possibly infinite) closed interval, then the feasible set (4.7) *can be represented as an LMI.*

Conic programming The convex optimization problems discussed up to this point can be recast as linear optimization problems over the semi-definite cone. We can take this observation one step further. It is well known that any convex optimization problem can be recast into a linear optimization problem over a convex cone K. An optimization problem over a general convex cone is denoted here as a conic optimization problem or conic program.

A conic program is a convex optimization problem in the canonical form

$$\begin{aligned} \sup \quad & c^\top x \\ \text{s.t.} \quad & Ax = b, \\ & x \in K, \end{aligned} \tag{4.8}$$

where we assume $K \subseteq \mathrm{R}^n$ to be a convex cone. As mentioned any convex optimization problem can be brought in the conic form (4.8). In what remains of this chapter we will deal with all encountered convex optimization problems directly in their corresponding conic programming canonical forms (4.8). This will allow us to keep the exposition uncluttered and in line with the presentation of the results in the subsequent Chapter 5 on optimization problems in more general vector spaces.

4.2 Computational aspects

The main reason for the focus on the specific classes of convex optimization problems introduced in Section 4.1 stems from the fact that they all admit, with the exception of the class of semi-infinite programs, efficient solution methods in both theory and practice. In this section we briefly discuss the computational aspects of the aforementioned classes of optimization problems. The interested reader can be referred to the lectures of Ben-Tal and Nemirovski [8, Chapter 5] for a more in depth discussion on the efficient solvability of convex optimization problems.

A solution method for a class of optimization problems is an algorithm that computes the optimum and a feasible solution achieving that optimum (up to some given accuracy), for a given particular problem instance from that class. Hence whether the supremum in (4.8) is attained or not shall not be of major concern in this dissertation as only approximate solutions are of interest here.

Where it is hard to solve convex optimization problems (4.1) in general, the classes of convex optimization problems discussed in Section 4.1 have the major advantage of being tractable in theory. Recent decades have seen a large effort in developing algorithms to solve convex optimization problems efficiently with spectacular success. Indeed, modern interior point methods as discussed by Ye [138] can be proven to solve SDPs problems in polynomial time.

In practice, mature software exists that implements modern convex optimization methods efficiently. So much so that although SDPs do not admit a closed form solution, stating a problem in the form (4.5) is for many practical purposes excepted as a de facto closed form solution. Many of the central results in this work will in fact be stated as an SDP.

We remark here that although any LP can be formulated as an SOCP and any SOCP can further be recast as a SDP it might not be advisable to do so from a practical perspective. Indeed, the convex optimization algorithms dealing specifically with SOCPs come with better complexity certificates and their software implementation may be more mature than the corresponding algorithms for SDPs. Analogous remarks can be made between the other classes in the optimization hierarchy discussed in Section 4.1. Fortunately, software such as `YALMIP` made available to the public by Löfberg [80] nowadays makes the previously mentioned issue mostly transparent to the optimization practitioner.

4.3 The fundamental theorem

One of the most important results concerning convex optimization is the fundamental theorem. The fundamental theorem relates convex optimization problems of the type (4.8) with the extreme points of their respective feasible sets.

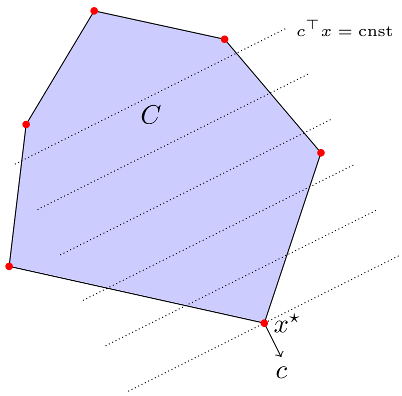


Figure 4.1: Geometric visualization of the fundamental Theorem 4.1. The compact feasible set C is a polyhedral set. The objective function $c^\top x$ is linear, so its level curves are hyperplanes orthogonal to c (depicted as dotted lines). The point $x^\star$ is optimal; it can always be found as an extreme point of C.

Theorem 4.3 (Fundamental theorem)**.** *Let C be a compact convex set in R^n, then we have the equivalence* $\sup\left\{c^\top x \;:\; x \in C\right\} = \sup\left\{c^\top x \;:\; x \in \operatorname{ex} C\right\}$.

Proof. The proof of this theorem is well known and hence omitted. See for instance Barvinok [5, Corollary II.3.4] for a proof. ☐

The fundamental theorem can be read informally as the guarantee that the minimum over the extreme points of the compact feasible set C coincides with the minimum over the set C itself. The fundamental theorem can be understood in a geometrical fashion; see also Figure 4.1.

Despite its conceptual simplicity the fundamental theorem carries a great historical importance in particular to linear programming. When the set C is polyhedral the optimization problem

$$\sup\left\{c^\top x \;:\; x \in \operatorname{ex} C\right\}$$

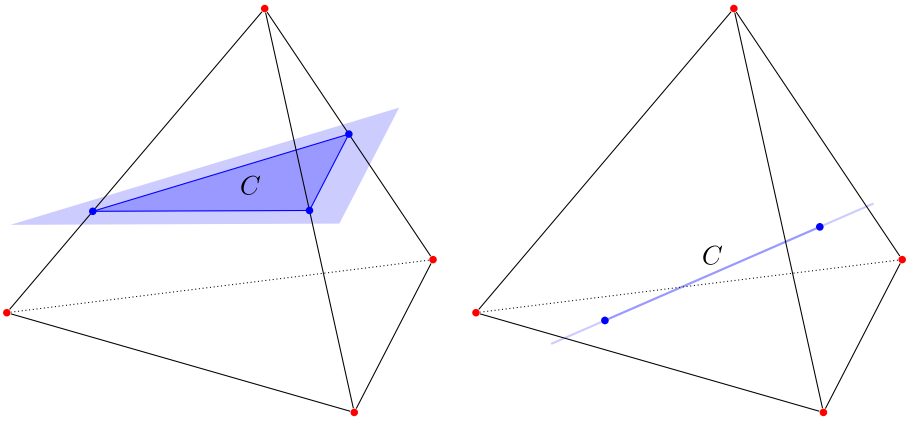
reduces to a maximization problem over its finitely many extreme points. This observation inspired the development of the simplex method contributed to Dantzig [41]. Direct maximization over the extreme points of C to solve a convex optimization problem $\sup\left\{c^\top x \;:\; x \in C\right\}$ might not always be a good idea as the number of extreme points may be vast even when the convex set C is a polyhedral set.

In the remainder of this section we state a theorem which is specific to LPs. The following proposition characterizes the sparseness properties of maximizers of linear optimization problems over the intersection of a hyperplane with a compact convex set. This proposition will be of great importance in the next chapter which will deal with convex optimization problems in more general vector spaces.

Proposition 4.1 ([5, Theorem III.9.2])**.** *Let $C = H \cap S$ be the intersection of an affine subspace of codimension k, i.e. $H := \left\{x \;:\; a_i^\top x = b_i,\ \forall i \in [1, \dots, k]\right\}$, and a compact convex set S, then*

$$\sup\left\{c^\top x \;:\; x \in C\right\} = \sup\left\{c^\top x \;:\; x \in \operatorname{conv}_{k+1}\{\operatorname{ex} S\}\right\},$$

where $\operatorname{conv}_{k+1}\{S\} := \left\{\sum_{i=0}^k p_i x_i \;:\; x_i \in \operatorname{ex} S,\ p_i \geq 0,\ \sum_{i=0}^k p_i = 1\right\}$.



(a) Codimension $k = 1$ (b) Codimension $k = 2$

Figure 4.2: Visual illustration of Proposition 4.1. Let C the intersection of an affine subspace of codimension k and a compact convex set S. The extreme points of the convex set C indicated in blue are recognized as the linear combination of at most $k + 1$ extreme points of the convex set S shown in red.

The previous proposition, informally stated, guarantees that the optimal solutions $x^\star$ of the convex optimization problem $\sup \left\{ c^\top x \ : \ x \in C = H \cap S \right\}$ can be found as the convex combination of at most $k+1$ extreme points of the compact convex set S. Proposition 4.1 can be understood in a geometrical fashion as well; see Figure 4.2.

4.4 Duality

From its earliest beginnings duality theory has played a preeminent role in the theory of convex optimization. Duality theory can indeed be traced back all the way to the early work by Kantorovich and von Neumann on linear programming.

There are many ways in which a comprehensive duality theory for convex optimization problems can be developed. Lagrangian duality is for instance used by Boyd and Vandenberghe [27] to establish a duality theory for convex optimization problems in the form (4.1). Alternatively, the Fenchel duality discussed in Rockafellar [107] provides a means to the same end as well.

We will state in this dissertation the conic duality theory as presented by Shapiro [116] for the convex conic optimization problem (4.8). Conic duality has the benefit that it is ideally suited to provide a duality theory for convex conic optimization problems in the form (4.8) and can furthermore be extended without much effort to the convex optimization problems in more general vector spaces encountered in Chapter 5.

With every primal conic program (4.8) we will associate a conic dual optimization problem, i.e.

$$\begin{array}{ll} \inf & \lambda^\top b \\ \text{s.t.} & A^\top \lambda - c \in K^\star, \end{array} \tag{4.9}$$

where the dual cone $K^\star$ is defined as in Definition 2.9. The dual problem (4.9) can thus be recognized as a convex optimization problem as well. Both convex optimization problems satisfy the classical weak duality relationship

$$\lambda^\top b \geq c^\top x$$

for every feasible x and y in the primal problem (4.8) and dual problem (4.9), respectively. Hence, the dual problem bounds the primal from above. In fact, under a very mild constraint qualification condition both problems share the same optimal value.

Theorem 4.4 (Strong duality)**.** *If K is a closed convex cone and there exists a point $\bar{x} \in \operatorname{rint} K$ such that $A\bar{x} = b$, then strong duality holds, i.e.*

$$\begin{array}{llcll} \sup & c^\top x & = & \inf & \lambda^\top b \\ \text{s.t.} & Ax = b, & & \text{s.t.} & A^\top \lambda - c \in K^\star. \\ & x \in K. & & & \end{array}$$

The existence of a point $\bar{x}$ such that $\bar{x} \in \operatorname{rint} K$ and $A\bar{x} = b$ is commonly referred to as Slater's constraint qualification condition [118]. The point $\bar{x}$ itself is referred to as a Slater point of the feasible set. Slater's condition is a specific example of a constraint qualification condition to establish strong duality between the primal problem (4.8) and its dual problem (4.9). While weaker type of constraint qualification conditions exists, Slater's constraint qualification condition will proof sufficient for the discussion in this work. In fact, Slater's condition will be encountered again in a slightly modified shape in the subsequent Chapter 5 when discussing convex optimization problems in more general vector spaces.

If strong duality holds the convex optimization problem (4.8) can alternatively be approached via its dual formulation (4.9). The dual of a convex optimization problem belonging to one of the classes discussed in Section 4.1 belongs again to that same class of optimization problems. That means that solving either primal or dual formulation in case of LPs, SOCPs or SDPs can be regarded as equally difficult.

Nevertheless, in many applications one is interested in what the primal maximizer $x^\star$ looks like, rather then the maximum itself. When instead the maximum was determined using a dual approach, the relationship between primal maximizers $x^\star$ and dual minimizers $\lambda^\star$ is of interest. Because of strong duality it follows that the primal maximizer $x^\star$ and dual minimizer $\lambda^\star$ are related as $\lambda^{\star\top} b = c^\top x^\star$. As the primal maximizer $x^\star$ is feasible we have by definition that $b = Ax^\star$. The primal and dual extrema are thus related as

$$(A^\top \lambda^\star - c)^\top x^\star = 0. \tag{4.10}$$

Condition (4.10) is commonly referred to as a complementarity condition between the primal maximum $x^\star$ and dual minimum $\lambda^\star$. Indeed, when the cone K is taken to be the positive orthant R^n_+, the complementarity condition (4.10) implies the standard relationships

$$A_i^\top \lambda_i^\star - c_i > 0 \implies x_i^\star = 0 \quad \text{or equivalently} \quad x_i^\star > 0 \implies A_i^\top \lambda_i^\star - c_i = 0$$

between primal and dual slack. The previous complementarity slackness implications will be given a nice geometrical interpretation in the convex optimization problems discussed in the subsequent chapter.

5 Convex optimization over probability distributions

In this chapter, we will consider the following class of optimization problems over convex sets of probability distributions in the following canonical form

$$\begin{aligned} \sup \quad & \int L(x)\,\mathbb{P}(\mathrm{d}x) \\ \text{s.t.} \quad & \int g_i(x)\,\mathbb{P}(\mathrm{d}x) = m_i, \quad i \in [0,\ldots,k-1] \\ & \mathbb{P} \in \mathcal{K} \end{aligned} \tag{5.1}$$

with L, $g_i : \mathrm{R}^n \to \mathrm{R}_+$ measurable functions and $\mathcal{K}$ a closed convex set of probability distributions. The function L will be called the loss function and the functions g_i will be denoted further as moment functions. We will further refer to the set of measures

$$\mathcal{C} := \mathcal{H} \cap \mathcal{K} \tag{5.2}$$

as the feasible set of problem (5.1) where the hyperplane $\mathcal{H}$ is defined as in (3.1). For the sake of exposition, assume that $g_0(x) = 1$ and the corresponding moment $m_0 = 1$. Note that the previous assumption is indeed redundant as $\mathcal{K}$ is assumed to be a subset of $\mathcal{P}_n$ and hence $\int g_0(x)\,\mathbb{P}(\mathrm{d}x) = 1$ is automatically satisfied.

Observe that the optimization problem (5.1) is a linear optimization problem over the convex set of probability distributions $\mathcal{C}$. The optimization problem (5.1) bears close resemblance to the convex optimization problems (4.8) in R^n discussed in Chapter 4. Many of the corresponding results from Chapter 4 on convex optimization in R^n will carry over immediately to the more general optimization problem (5.1) in the vector space of measures $\mathcal{E}_n$. We will seek to show in this chapter that the fundamental theorem of linear programming can be generalized and has something to say about the properties of the maximizers of the optimization problem (5.1). Furthermore, a comprehensive duality theory for the optimization problem (5.1) can be developed based on pairing the primal space of measures $\mathcal{E}_n$ with the space of positive functions $\mathcal{E}_n^\star$.

5.1 The fundamental theorem

We would like to recover a direct counterpart to the fundamental theorem stated in Theorem 4.3 of convex optimization in R^n. There are however a number of caveats present in vector spaces more general than R^n which prevent a verbatim restatement of Theorem 4.3 in the current setting.

Theorem 5.1 (Fundamental theorem)**.** *Let $\mathcal{C}$ be a compact convex set of probability distributions and L a bounded and continuous measurable function, then we have the equivalence* $\sup\left\{\int L(x)\,\mathbb{P}(\mathrm{d}x) \ : \ \mathbb{P}\in\mathcal{C}\right\} = \sup\left\{\int L(x)\,\mathbb{P}(\mathrm{d}x) \ : \ \mathbb{P}\in\mathrm{ex}\,\mathcal{C}\right\}$.

Proof. The proof of this theorem is well known and analogous to the proof of its counterpart Theorem 4.3. A proof can be found in for instance Barvinok [5, Corollary III.4.2]. □

The difference with the finite dimensional setting of Chapter 4 is that in general a linear cost function $\int L(x)\,\mathbb{P}(\mathrm{d}x)$ need not be a continuous function in $\mathbb{P}$. The proof of Theorem 5.1 nevertheless requires the affine contours $\left\{\mathbb{P}\in\mathcal{E}_n \ : \ \int L(x)\,\mathbb{P}(\mathrm{d}x)=\alpha\right\}$ to be closed sets or equivalently $\int L(x)\,\mathbb{P}(\mathrm{d}x)$ to be continuous. The requirement on the loss function L in Theorem 5.1 can thus be explained immediately in view of Definition A.4.

We have that the minimum of a continuous linear function over the extreme distributions of a compact convex set $\mathcal{C}$ recovers the minimum over the set $\mathcal{C}$ itself. We would like to apply Theorem 5.1 in order to derive geometric properties of the extrema of the convex optimization problem (5.1) in case $\mathcal{C}$ is the intersection of a hyperplane and a closed convex ambiguity set. However, in the remainder of the work we will consider in many instances loss functions L and moment functions g_i which are neither bounded nor continuous resulting in a non-compact feasible set $\mathcal{C}$ and non-continuous linear function $\int L(x)\,\mathbb{P}(\mathrm{d}x)$. Nevertheless, we can present the following counterpart to Proposition 4.1.

Proposition 5.1 ([110])**.** *Let $\mathcal{C}=\mathcal{H}\cap\mathcal{K}$ be the intersection of a hyperplane $\mathcal{H}$ of codimension k and a closed convex set $\mathcal{K}$ of probability distributions, then*

$$\sup\left\{\int L(x)\,\mathbb{P}(\mathrm{d}x) \ : \ \mathbb{P}\in\mathcal{C}\right\} = \sup\left\{\int L(x)\,\mathbb{P}(\mathrm{d}x) \ : \ \mathbb{P}\in\mathrm{conv}_{k+1}\{\mathrm{ex}\,\mathcal{K}\}\right\}.$$

Proof. It is remarked that the ambiguity set $\mathcal{K}\subseteq\mathcal{P}_n$ can not contain any lines. Indeed, suppose $\mathbb{P}_1$ and $\mathbb{P}_2$ are distinct elements in $\mathcal{K}$. As $\mathbb{P}_1$ and $\mathbb{P}_2$ are distinct, there exists a measurable set B such that $\mathbb{P}_1(B)\neq\mathbb{P}_2(B)$. The line $\{t\mathbb{P}_1+(1-t)\mathbb{P}_2 \ : \ t\in\mathrm{R}\}$ can not be contained in $\mathcal{P}_n$ as follows immediately from the contradiction

$$0\leq t\mathbb{P}_1(B)+(1-t)\mathbb{P}_2(B)\leq 1, \quad \forall t\in\mathrm{R}.$$

A modern proof of the proposition for closed convex sets $\mathcal{K}$ not containing lines is provided by Barvinok [5, Theorem III.9.2]. □

The previous proposition, informally stated, guarantees that the optimal solutions $\mathbb{P}^\star$ of the convex optimization problem (5.1) can be found as the convex combination of at most $k+1$ extreme distributions of the convex set $\mathcal{K}$. Please note that no further assumptions on the loss function L and moment functions g_i are required in Proposition 5.1 besides of course measurability.

When the ambiguity set $\mathcal{K}$ in the convex optimization problem (5.1) admits a Choquet star representation generated by the univariate probability distribution $\mathbb{T}$, i.e.

$$\mathrm{ex}\,\mathcal{K}=\{\mathbb{T}_x \ : \ x\in\mathrm{R}^n\},$$

then Proposition 5.1 provides directly the general structure of worst-case distributions $\mathbb{P}^\star$ in optimization problem (5.1). As discussed in Chapter 3, both the set of α-unimodal distributions $\mathcal{U}_\alpha$ and the set of γ-monotone distributions $\mathcal{M}_\gamma$ admit such Choquet star representations.

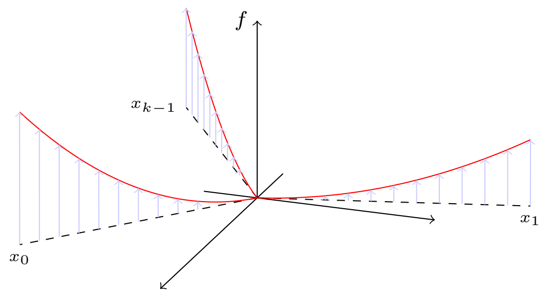


Figure 5.1: A worst-case distribution $\mathbb{P}^\star$ in the plane R^2 of the uncertainty quantification problem (5.1) for α-unimodal distributions ($\mathcal{K} = \mathcal{U}_\alpha$) can always be found as the convex combination of at most k extreme distributions $\mathbb{u}_x^\alpha$ by merit of Proposition 5.1.

Proposition 5.1 now guarantees that the worst-case probability distributions $\mathbb{P}^\star$ in optimization problem (5.1) will be of the general form

$$\mathbb{P}^\star = \sum_{i=0}^{k-1} p_i \mathbb{T}_{x_i}.$$

A worst-case distribution $\mathbb{P}^\star$ can thus be found as the convex combination of at most k extreme distributions $\mathbb{T}_x$ in the Choquet star simplex $\mathcal{K}$. This previous observation is made visual in Figure 5.1.

5.2 Duality

A comprehensive duality theory for the convex optimization problem (5.1) can be developed akin to the duality theory discussed in Section 4.4 for optimization problems in R^n. A duality theory for the optimization problem (5.1) was already developed in the early work of Isii [63]. In this dissertation however, we will present the modern duality theory outlined by Shapiro [116] which nicely mirrors its counterpart discussed in Section 4.4. The reader interested in the technicalities of the results stated in this section is hence referred to the work of Shapiro [116], or references therein, for a more in depth discussion on duality in $\mathcal{E}_n$.

Let $\mathcal{E}_n^\star$ be the vector space of real measurable functions f on R^n. The space $\mathcal{E}_n$ and $\mathcal{E}_n^\star$ are paired by a bilinear product $\langle \cdot, \cdot \rangle : \mathcal{E}_n^\star \times \mathcal{E}_n \to \mathrm{R}$ defined through the integral

$$\langle f, \mathbb{m} \rangle := \int f(x)\, \mathbb{m}(\mathrm{d}x). \tag{5.3}$$

More information on the pairing between both spaces is provided in Appendix A.2. The bilinear product in equation (5.3) will serve the same role in this chapter as the standard inner product between vectors in R^n in the development of the duality theory between the vector spaces of measures $\mathcal{E}_n$ and positive functions $\mathcal{E}_n^\star$ as can be seen from the next definition.

Definition 5.1 (Dual cone)**.** *The dual cone of functions $\mathcal{K}^\star$ in $\mathcal{E}_n^\star$ of any set of measures $\mathcal{K}$ is defined as*

$$\mathcal{K}^\star := \{ f \in \mathcal{E}_n^\star \;:\; \langle f, \mathbb{m} \rangle \geq 0, \quad \forall \mathbb{m} \in \mathcal{K} \}.$$

It can immediately be seen from its definition that the dual cone of any set of measures is itself indeed always a convex cone of functions in $\mathcal{E}_n^\star$.

As in the finite dimensional case, we can relate to the primal optimization problem (5.1) over measures in the vector space $\mathcal{E}_n$ a dual optimization problem over measurable functions in the dual space $\mathcal{E}_n^\star$ of functions, i.e.

$$\begin{aligned} \inf \quad & \sum_{i=0}^{k-1} \lambda_i \cdot m_i \\ \text{s.t.} \quad & \sum_{i=0}^{k-1} \lambda_i \cdot g_i - L \in \mathcal{K}^\star \end{aligned} \tag{5.4}$$

were $\mathcal{K}^\star$ is the dual cone of the convex set of distributions $\mathcal{K}$. We remark that the dual problem is always convex and is recognized as the counterpart of the dual (4.9) for convex optimization problems in R^n. At the end of this section we discuss several interesting sets of probability distributions $\mathcal{K}$ of particular interest to this dissertation for which the dual cone $\mathcal{K}^\star$ is known explicitly and comes with a nice geometrical interpretation.

The primal optimization problem over probability distributions (5.1) and the dual optimization problem over functions (5.4) satisfy the weak duality relationship

$$\sum_{i=0}^{k-1} \lambda_i \cdot m_i \geq \int L(x)\, \mathbb{P}(\mathrm{d}x)$$

for every feasible probability distribution $\mathbb{P}$ and dual vector λ in the primal problem (5.1) and dual problem (5.4), respectively. The dual problem (5.4) thus bounds the primal problem (5.1) from above. Under a mild constraint qualification condition both problems in fact share the same optimal value.

Theorem 5.2 (Strong duality [116]). *Denote with M the set of moments for which the feasible set is non-empty, i.e. $M := \left\{m \in \mathrm{R}^k \ : \ \exists \mathbb{P} \in \mathcal{K} : \int g_i(x)\, \mathbb{P}(\mathrm{d}x) = m_i,\ \forall i \in [0, \dots, k-1]\right\}$. If $m \in \operatorname{rint} M$ then strong duality holds, i.e.*

$$\begin{aligned} \sup_{\mathbb{P}\in\mathcal{K}} \quad & \int L(x)\, \mathbb{P}(\mathrm{d}x) & = \quad \inf \quad & \sum_i \lambda_i \cdot m_i \\ \text{s.t.} \quad & \int g_i(x)\, \mathbb{P}(\mathrm{d}x) = m_i, \quad \forall i & \text{s.t.} \quad & \sum_i \lambda_i \cdot g_i - L \in \mathcal{K}^\star \end{aligned} \tag{5.5}$$

Proof. The theorem is proven in a slightly different form by Shapiro [116, Proposition 3.4]. □

The condition $m \in \operatorname{rint} M$ can be seen as a generalized Slater condition. The previous generalized Slater constraint qualification condition is relatively mild as illustrated by the following example.

Example 5.1 (Constraint qualification). *In Chapters 6 and 7 the set $\mathcal{H}(\mu, S) \cap \mathcal{K}$ of all distributions sharing a mean μ and second moment S in a Choquet star simplex $\mathcal{K} \subseteq \mathcal{P}_n$ will be of particular interest. As stated in Proposition 3.6 the ambiguity set $\mathcal{H}(\mu, S) \cap \mathcal{P}_n$ is non-empty when $S \succeq \mu\mu^\top$. In that case, the Slater constraint qualification condition is fulfilled when the given moments satisfy $S \succ \mu\mu^\top$. Because of Proposition 3.7, the ambiguity set $\mathcal{H}(\mu, S) \cap \mathcal{K}$ is non-empty when $S_s \succeq \mu_s\mu_s^\top$ for $S_s \cdot \int_0^\infty t^2\, \mathbb{T}(\mathrm{d}t) = S$ and $\mu_s \cdot \int_0^\infty t\, \mathbb{T}(\mathrm{d}t) = \mu$ with $\mathbb{T}$ the generator of the Choquet star simplex $\mathcal{K}$. The set $\mathcal{H}(\mu, S) \cap \mathcal{K}$ thus satisfies the Slater constraint qualification condition if $S_s \succ \mu_s\mu_s^\top$.*

In the last part of the section we make the dual optimization problem (5.4) concrete in case the set of probability distributions $\mathcal{K}$ admits a Choquet star representation in terms of a generating univariate probability distribution $\mathbb{T}$. The case in which the set of probability distributions $\mathcal{K}$

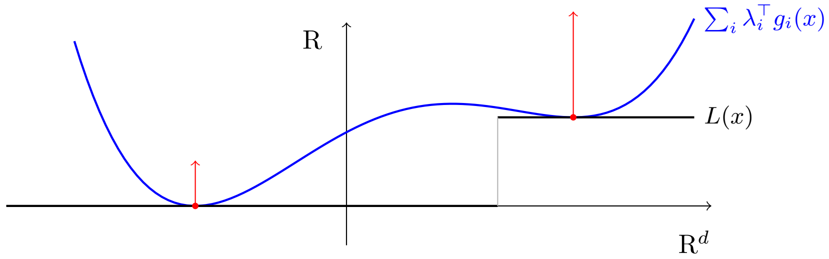


Figure 5.2: A feasible dual solution $\lambda \in \mathrm{R}^k$ in the dual constraint set (5.7) results in a linear combination of moment functions $\sum_i \lambda_i \cdot g_i$ which majorizes the function L. The complementarity condition (5.8) guarantees further that the worst-case distributions $\mathbb{P}^\star$ of the primal problem (5.1) are supported on those points where the optimal dual function $\sum_i \lambda_i^\star \cdot g_i$ in the dual constraint (5.7) kisses the function L.

coincides with the standard probability simplex $\mathcal{P}_n$ will proof of particular interest and admits a nice geometric interpretation of the dual constraint in problem (5.4).

Recall that the dual cone $\mathcal{K}^\star$ is defined as in Definition 5.1 and hence the dual constraint in problem (5.4) can be stated as

$$\sum_{i=0}^{k-1} \lambda_i \cdot g_i - L \in \mathcal{K}^\star \iff \int \textstyle\sum_{i=0}^{k-1} \lambda_i \cdot g_i(x) - L(x)\, \mathbb{m}(\mathrm{d}x) \geq 0, \quad \forall \mathbb{m} \in \mathcal{K}.$$

The previous constraint does not immediately admit a favorable representation as it consists of as many linear constraints as there are probability distributions in $\mathcal{K}$ which may be formidable. However, when the set of probability distributions $\mathcal{K}$ admits a Choquet star representation the dual set $\mathcal{K}^\star$ can be represented as a standard semi-infinite constraint over R^n instead.

Proposition 5.2 (Semi-infinite representation [103])**.** *If the set of probability distributions $\mathcal{K}$ admits a Choquet star representation with generating distribution $\mathbb{T}$ then the dual constraint in* (5.4) *can be represented as the semi-infinite constraint*

$$\int_0^\infty \left[\sum_{i=0}^{k-1} \lambda_i \cdot g_i(tx) - L(tx)\right] \mathbb{T}(\mathrm{d}t) \geq 0, \quad \forall x \in \mathrm{R}^n. \tag{5.6}$$

When in the primal optimization problem the set of probability distributions $\mathcal{K}$ is taken to be the standard probability simplex, then the corresponding dual cone $\mathcal{K}^\star$ consists of all measurable positive functions. Indeed, in this case $\mathcal{K} = \mathcal{P}_n = \mathrm{mix}\,\{\delta_x \,:\, x \in \mathrm{R}^n\}$ which admits a Choquet star representation with generating probability distribution given in Example 3.1. As the particular case $\mathcal{K} = \mathcal{P}_n$ will occur frequently in the remainder of this work, we make the dual constraint (5.6) explicit for this situation:

$$\sum_{i=0}^{k-1} \lambda_i \cdot g_i(x) - L(x) \geq 0, \quad \forall x \in \mathrm{R}^n. \tag{5.7}$$

The previous dual set (5.7) admits a nice geometrical interpretation. Indeed, the dual set (5.7) consists of all linear combinations λ_i of the moment functions g_i majorizing the loss function L in R^n; see also Figure 5.2.

5.3 Worst-case probability distributions

As discussed in the previous section the convex optimization problem over probability distributions (5.1) can be approached via its dual formulation (5.4) under the mild constraint qualification conditions mentioned in Theorem 5.2. Instead of maximizing over a convex set of probability distributions, the dual characterization consists of minimizing over a set of functions.

Be this as it may, in some cases it might be of interest to know what the distributions $\mathbb{P}^\star \in \mathcal{C}$ achieving the supremum $\int L(x)\,\mathrm{d}\mathbb{P}^\star$ looks like. In view of subsequent Chapters 8 and 9 we will also refer to the optima $\mathbb{P}^\star$ of the optimization problem (5.1) as worst-case distributions. As we approached the convex optimization problem (5.1) here via its dual characterization (5.4), we are now interested in the relation between the worst-case probability distributions $\mathbb{P}^\star \in \mathcal{C}$ for the primal problem (5.1) and the optimal solutions $\lambda^\star$ in its dual characterization (5.4).

Because of strong duality it follows that the primal maximizer $\mathbb{P}^\star$ and dual minimizer $\lambda^\star$ are related as $\int L(x)\,\mathrm{d}\mathbb{P}^\star = \sum_{i=0}^{k} \lambda_i^\star m_i$. As the primal maximizer $\mathbb{P}^\star$ is feasible we have by definition that $m_i = \int g_i\,\mathrm{d}\mathbb{P}^\star$. The primal and dual extrema are thus related as

$$\int L(x)\,\mathrm{d}\mathbb{P}^\star = \int \textstyle\sum_{i=0}^{k-1} \lambda_i^{\star\top} g_i(x)\,\mathrm{d}\mathbb{P}^\star \tag{5.8}$$

making use of the linearity of integration. Condition (5.8) is commonly referred to as a complementarity condition between the primal maximum $\mathbb{P}^\star$ and dual minimum $\lambda^\star$. In what remains of this section we will show that when the set of probability distributions $\mathcal{K}$ coincides with the standard probability simplex $\mathcal{P}_n$ then we can again provide a nice geometric interpretation to the complementarity condition (5.8).

When $\mathcal{K} = \mathcal{P}_n$ we must have according to (5.7) that $\sum_{i=0}^{k-1} \lambda_i^{\star\top} g_i(x) - L(x) \geq 0$ point-wise in R^n. A direct consequence of previous inequality in combination with the complementarity condition (5.8) is that the optimal probability distribution $\mathbb{P}^\star$ must be supported on the points at which the dual function $\sum_{i=0}^{k-1} \lambda_i^{\star\top} g_i(x)$ kisses the loss function $L(x)$, i.e.

$$\operatorname{supp} \mathbb{P}^\star \subseteq S^\star = \left\{ x \in \mathrm{R}^n \;:\; \sum_{i=0}^{k-1} \lambda_i^{\star\top} g_i(x) = L(x) \right\}. \tag{5.9}$$

The previous statement is illustrated visually in Figure 5.2.

It is clear that the worst-case distribution $\mathbb{P}^\star$ is not uniquely characterized by condition (7.14). In fact, any feasible probability distribution $\mathbb{P}$ supported on the set $S^\star$ is necessarily optimal in the convex optimization problem (5.1). The set $S^\star$ has in many interesting situations a finite cardinality. Indeed, if the functions g_i and L are (piece-wise) polynomial then the fundamental theorem of algebra guarantees that $S^\star$ has finite cardinality. For $S^\star$ a finite set of points $x_j^\star$ constructing a worst-case probability distribution

$$\mathbb{P}^\star = \textstyle\sum_j p_j^\star \delta_{x_j^\star}, \quad \text{with } x_j^\star \in S^\star$$

requires only the solution of the following system of linear inequalities

$$\begin{aligned} \textstyle\sum_j p_j g_i(x_j^\star) &= m_i, \quad \forall i \in [0, \ldots, k-1] \\ p_j &\geq 0, \qquad \forall j \end{aligned}$$

When $S^\star$ has finite cardinality, the condition (7.14) thus allows for the efficient extraction of a worst-case probability distribution $\mathbb{P}^\star$ from a dual optimal solution $\lambda^\star$ at virtually no additional computation cost.

Properties of $\mathbb{P}$	$\mathbb{P}(\lvert\xi-\mu\rvert \geq \kappa\sigma)$	$\mathbb{P}(\xi-\mu \geq \kappa\sigma)$
None, $\mathbb{P} \in \mathcal{P}_1$	Chebyshev [36]	Cantelli [34]
Unimodal, $\mathbb{P} \in \mathcal{U}_1$	Gauss [54]	Proposition 5.6

Table 5.1: Table organizing the probability inequalities presented in Section 5.4.1.

5.4 Optimal expectation inequalities

We will first indicate how the convex optimization problem (5.1) can be related to optimal inequalities in probability theory such as the classical Chebyshev (1.6) and Gauss (1.7) inequalities discussed in the introduction of this dissertation. In the second part of this section we will show that the convex optimization problem (5.1) is very relevant to optimal CVaR inequalities as well.

5.4.1 Classical probability inequalities

The classical Chebyshev inequality (1.6) discussed in the introduction presents a tight upper bound on the probability

$$\mathbb{P}(|\xi-\mu| \geq \kappa\sigma)$$

for $\kappa > 0$ given only the mean μ and standard deviation $\sigma > 0$ of the univariate random variable ξ. The Chebyshev inequality is denoted as a bilateral inequality as it bounds the probability of a two-sided tail event $|\xi - \mu| \geq \kappa\sigma$. A unilateral inequality bounds the probability of a single tail event $\xi - \mu \geq \kappa\sigma$. The unilateral counterpart to the Chebyshev inequality (1.6) was discovered by Cantelli [34] in 1910. The classical Cantelli inequality presents a tight upper bound on the probability

$$\mathbb{P}(\xi-\mu \geq \kappa\sigma)$$

for $\kappa > 0$ given only the mean μ and standard deviation $\sigma > 0$ of the univariate random variable ξ. On the other hand, the classical Gauss inequality (1.7) gives a counterpart to the classical Chebyshev inequality (1.6) for unimodal distributions. This naturally begs the question whether also a unimodal counterpart to the unilateral Cantelli inequality can be found. Somewhat surprisingly, no such probability inequality can be found in the literature.

The convex optimization problem (5.1) is ideally suited to discuss and generalize the classical Chebyshev and Gauss inequalities in the direction hinted upon. We will show that using only the results stated so far in this dissertation, we are already in a position to put many classical probability inequalities on an equal footing. In doing so, we will provide a unimodal counterpart to the Cantelli inequality which we believe is novel. The results found in this section are organized in the Table 5.1.

What distinguishes the probability inequalities from one another is the additional structure assumed, e.g. unimodality or monotonicity, on the probability distribution $\mathbb{P}$. Hence, the convex optimization problems

$$\begin{aligned} \sup \quad & \mathbb{P}(|\xi-\mu| \geq \kappa\sigma) \\ \text{s.t.} \quad & \int x\,\mathbb{P}(\mathrm{d}x) = \mu, \\ & \int x^2\,\mathbb{P}(\mathrm{d}x) = \mu^2+\sigma^2 \\ & \mathbb{P} \in \mathcal{K} \end{aligned} \tag{5.10}$$

$$\begin{aligned} \sup \quad & \mathbb{P}(\xi-\mu \geq \kappa\sigma) \\ \text{s.t.} \quad & \int x\,\mathbb{P}(\mathrm{d}x) = \mu, \\ & \int x^2\,\mathbb{P}(\mathrm{d}x) = \mu^2+\sigma^2 \\ & \mathbb{P} \in \mathcal{K} \end{aligned} \tag{5.11}$$

can be recognized to generalize the classical probability bounds for a judicious choice of ambiguity set $\mathcal{K}$. The classical Gauss bound for instance can be related to the optimization problem (5.10) in which the distribution $\mathbb{P} \in \mathcal{U}_1$ is required to be unimodal. Similarly, the classical Cantelli bound can be related to the convex optimization problem (5.11) where $\mathcal{K}$ imposes no additional structure.

As we have the equivalences $\mathbb{P}(|\xi-\mu| \geq \kappa\sigma) = \mathbb{P}(|z| \geq \kappa)$ and $\mathbb{P}(\xi-\mu \geq \kappa\sigma) = \mathbb{P}(z \geq \kappa)$ with $z = (\xi-\mu)/\sigma$ a univariate standardized random variable, we need only to consider the situation $\mu = 0$ and $\sigma = 1$. A random variable z is denoted as standardized if it has zero mean and unit variance. In what follows we will show that for unimodal ($\mathcal{K} = \mathcal{U}_1$) and for unstructured probability distributions ($\mathcal{K} = \mathcal{P}_1$), the convex optimization problems (5.10) and (5.11) can be solved in closed form. Before we make our results specific though, we will indicate first that most of the analysis can be done merely by assuming that the ambiguity set $\mathcal{K}$ admits a Choquet star representation in terms of the generating univariate distribution $\mathbb{T}$. The set $\mathcal{K}$ has then as extreme distributions the following set of radial probability distributions $\operatorname{ex}\mathcal{K} = \{\mathbb{T}_x \ : \ x \in \mathrm{R}\}$.

Proposition 5.1 as applied to the particular convex optimization problems (5.10) and (5.11) guarantees now that we can restrict attention to probability distributions $\mathbb{P}^\star \in \operatorname{conv}_3\{\mathbb{T}_x \ : \ x \in \mathrm{R}\}$ consisting of the convex combination of at most three extreme distributions. From the symmetry of the problem (5.10) with $\mu = 0$, it follows that $\mathbb{P}^\star$ must be symmetric around the origin. That is ξ and $-\xi$ have the same variance and assign the same probability to the event of interest. Attention can thus be restricted to probability distributions in the canonical form

$$\mathbb{P}^\star = (1-p)\mathbb{T}_0 + \frac{1}{2}p\,\mathbb{T}_x + \frac{1}{2}p\,\mathbb{T}_{-x}, \tag{5.12}$$

for some $x \in \mathrm{R}_+$ and $p \in [0,1]$. Similarly, a moment of reflexion learns that in the optimization problem (5.11) attention can be restricted to probability distributions in the canonical form

$$\mathbb{P}^\star = (1-p)\mathbb{T}_{-px/(1-p)} + p\,\mathbb{T}_x, \tag{5.13}$$

for some $x \in \mathrm{R}_+$ and $p \in [0,1]$. The previous results are extremely powerful as they reduce the convex optimization problems (5.10) and (5.11) over probability distributions with zero mean $\mu = 0$ and unit variance $\sigma = 1$ to standard (non-convex) optimization problem in only two variables

$$\begin{array}{rl} (5.10) & = \\ \max & p\mathbb{T}_x([\kappa,\infty)) \\ \text{s.t.} & p \in [0,1], \ x \in \mathrm{R}_+ \\ & \int_0^\infty t^2\,\mathbb{T}(\mathrm{d}t)\cdot px^2 = 1. \end{array} \tag{5.14}$$

$$\begin{array}{rl} (5.11) & = \\ \max & p\mathbb{T}_x([\kappa,\infty)) \\ \text{s.t.} & p \in [0,1], \ x \in \mathrm{R}_+ \\ & \int_0^\infty t^2\,\mathbb{T}(\mathrm{d}t)\cdot px^2 = 1-p. \end{array} \tag{5.15}$$

The non-convex optimization problems (5.14) and (5.15) admit in many interesting situations explicit solutions. In what follows we will illustrate the previous statement for both bilateral and unilateral probability inequalities thereby completing Table 5.1.

Bilateral inequalities

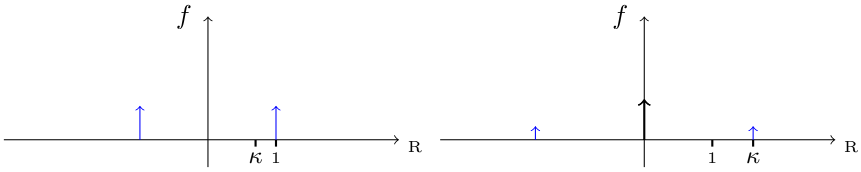
Proposition 5.3 (Classical Chebyshev inequality [36]). *Let ξ be a real random variable, then we have the bilateral probability inequality*

$$\mathbb{P}(|\xi-\mu| \geq \kappa\sigma) \leq \begin{cases} \frac{1}{\kappa^2} & \text{if} \quad \kappa > 1, \\ 1 & \text{otherwise.} \end{cases} \tag{5.16}$$

Proof. The quantity $p\mathbb{T}_x^\star([\kappa,\infty))$ for $\mathbb{T}_x = \delta_x$ can be explicitly calculated to be $p\mathbf{1}\{x \geq \kappa\}$. The variable x can be eliminated from problem (5.14) yielding the equivalent problem

$$p_{\mathrm{cheb}}(\kappa) = \max\left\{p \ : \ 0 \leq p \leq 1, \ p \leq \frac{1}{\kappa^2}\right\}. \tag{5.17}$$

The result in (5.16) follows now from standard manipulations of (5.17). □



(a) $0 \leq \kappa < 1$ (b) $\kappa \geq 1$

Figure 5.3: Worst-case unimodal probability distributions achieving the bound in the probability inequality (5.16) for $\mu = 0$ and $\sigma = 1$. The colored part depicts the size of the event corresponding to the Chebyshev inequality (5.16).

The worst-case distributions attaining the classical Chebyshev inequality (5.16) can be explicitly constructed as well. The density $f_{\text{cheb}}(\kappa)$ of the worst-case distribution $\mathbb{P}^\star$ attaining the bound (5.16) for $\mu = 0$ and $\sigma = 1$ is given in closed form as

$$f_{\text{cheb}}(\kappa) = \begin{cases} (1 - \frac{1}{\kappa^2})\delta_0 + \frac{1}{2\kappa^2}\delta_\kappa + \frac{1}{2\kappa^2}\delta_{-\kappa} & \text{if} \quad \kappa \geq 1, \\ \frac{1}{2}\delta_1 + \frac{1}{2}\delta_{-1} & \text{otherwise.} \end{cases}$$

The distribution $f_{\text{cheb}}(\kappa)$ is graphically depicted in Figure 5.3.

Proposition 5.4 (Classical Gauss inequality [54])**.** *Let ξ be a real unimodal random variable with centre $c = \mu$, then we have the tight bilateral probability inequality*

$$\mathbb{P}(|\xi - \mu| \geq \kappa\sigma) \leq \begin{cases} \frac{4}{9\kappa^2} & \text{if} \quad \kappa > \frac{2}{\sqrt{3}}, \\ 1 - \frac{\kappa}{\sqrt{3}} & \text{otherwise.} \end{cases} \tag{5.18}$$

Proof. The quantity $p\mathbb{T}_x^\star([\kappa, \infty))$ for $\mathbb{T}_x = \mathbb{u}_x^1$ can be explicitly calculated to be $p\left(1 - \frac{\kappa}{x}\right) \mathbf{1}\left\{x \geq \kappa\right\}$. The variable x can be eliminated from problem (5.14) yielding the equivalent problem

$$p_{\text{gauss}}(\kappa) = \max\left\{ p\left(1 - \frac{\kappa\sqrt{p}}{\sqrt{3}}\right) \ : \ 0 \leq p \leq 1,\ p \leq \frac{3}{\kappa^2} \right\}. \tag{5.19}$$

Using standard arguments, the maximum must be attained at a critical point of the objective function or at the boundary of the feasible domain of the non-convex optimization problem (5.19). The boundary points $p^\star = 0$ and $p^\star = \frac{3}{\kappa^2}$ establish that (5.14) is bounded from below by zero. The last boundary point $p^\star = 1$ yields $p_{\text{gauss}}(\kappa) \geq 1 - \frac{\kappa}{\sqrt{3}}$. The critical points are located at

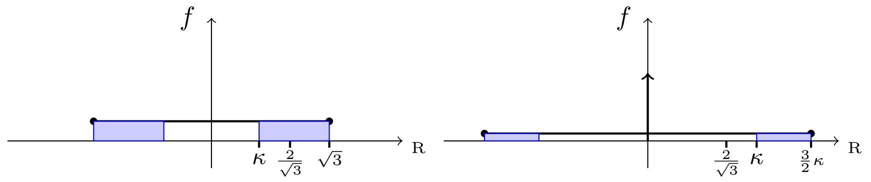
$$\frac{\mathrm{d}}{\mathrm{d}p}\left[p\left(1 - \frac{\kappa\sqrt{p}}{\sqrt{3}}\right)\right] = 0 \iff p = \frac{4}{3\kappa^2}$$

and belong to the feasible region whenever $\kappa \geq \frac{2}{\sqrt{3}}$ which results in $p_{\text{gauss}}(\kappa) \geq 4/(9\kappa^2)$. $\square$

The worst-case distributions attaining the classical Gauss inequality (5.18) can be explicitly constructed as well. The density $f_{\text{gauss}}(\kappa)$ of the worst-case distribution $\mathbb{P}^\star$ attaining the bound (5.18) for $\mu = 0$ and $\sigma = 1$ is given in closed form as

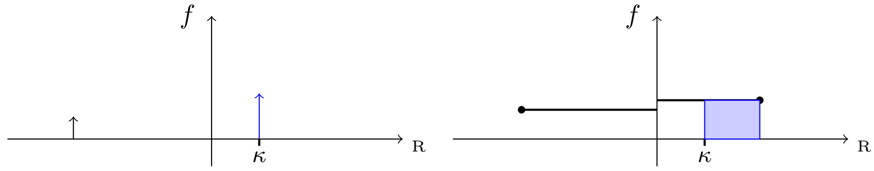
$$f_{\text{gauss}}(\kappa) = \begin{cases} (1 - \frac{4}{3\kappa^2})\delta_0 + \frac{4}{9\kappa^3}\mathbf{1}\left\{-\frac{3}{2}\kappa \leq t \leq \frac{3}{2}\kappa\right\} & \text{if} \quad \kappa \geq \frac{2}{\sqrt{3}}, \\ \frac{1}{2\sqrt{3}}\mathbf{1}\left\{-\sqrt{3} \leq t \leq \sqrt{3}\right\} & \text{if} \quad 0 \leq \kappa < \frac{2}{\sqrt{3}}. \end{cases}$$

The distribution $f_{\text{gauss}}(\kappa)$ is graphically depicted in Figure 5.4.



(a) $0 \leq \kappa < \frac{2}{\sqrt{3}}$ (b) $\kappa \geq \frac{2}{\sqrt{3}}$

Figure 5.4: Worst-case unimodal probability distributions achieving the bound in the Gauss inequality (5.18) for $\mu = 0$ and $\sigma = 1$. The colored part depicts the size of the event corresponding to the Gauss inequality (5.18).



(a) The classical Cantelli inequality (b) The unimodal Cantelli inequality

Figure 5.5: Worst-case probability distributions achieving the bounds in the Cantelli inequality (5.20) and the probability inequality (5.22) and for $\mu = 0$ and $\sigma = 1$. The colored part depicts the size of the event corresponding to the probability inequalities.

Unilateral inequalities

Proposition 5.5 (Classical Cantelli inequality [34])**.** *Let ξ be a real random variable, then we have the tight unilateral probability inequality*

$$\mathbb{P}(\xi - \mu \geq \kappa\sigma) \leq \frac{1}{1+\kappa^2}. \tag{5.20}$$

Proof. The quantity $p\mathbb{T}_x^\star([\kappa, \infty))$ for $\mathbb{T}_x = \delta_x$ can be explicitly calculated to be $p\mathbf{1}\{x \geq \kappa\}$. The variable x can be eliminated from problem (5.15) yielding the equivalent problem

$$p_{\text{cant}}(\kappa) = \max\left\{p \ : \ 0 \leq p \leq \frac{1}{\kappa^2+1}\right\}. \tag{5.21}$$

The result in (5.20) follows now from standard manipulations of (5.21). $\square$

The worst-case distributions attaining the classical Cantelli inequality (5.18) can be explicitly constructed as well. The density $f_{\text{cant}}(\kappa)$ of the worst-case distribution $\mathbb{P}^\star$ attaining the bound (5.18) for $\mu = 0$ and $\sigma = 1$ is given in closed form as

$$f_{\text{cant}}(\kappa) = \frac{\kappa^2}{\kappa^2+1}\delta_{-\frac{1}{\kappa}} + \frac{1}{\kappa^2+1}\delta_\kappa$$

The distribution $f_{\text{cant}}(\kappa)$ is graphically depicted in Figure 5.5(a).

Proposition 5.6 (Unimodal Cantelli inequality)**.** *Let ξ be a real unimodal random variable with centre $c = \mu$, then we have the tight unilateral probability inequality*

$$\mathbb{P}(\xi - \mu \geq \kappa\sigma) \leq \frac{1}{2}\left(\frac{3\kappa^2}{\sqrt[3]{3\kappa^2\left(3+\sqrt{3(3+\kappa^2)}\right)}} - \sqrt[3]{3\kappa^2\left(3+\sqrt{3(3+\kappa^2)}\right)} + 2\right) \tag{5.22}$$

Proof. The quantity $p\mathbb{T}_x([\kappa,\infty))$ for $\mathbb{T}_x = \mathbb{u}_x^1$ can be explicitly calculated to be $p\left(1-\frac{\kappa}{x}\right)\mathbf{1}\{x \geq \kappa\}$. The variable x can be eliminated from problem (5.15) yielding the equivalent problem

$$p_{\text{ucant}}(\kappa) = \max\left\{p\left(1-\frac{\kappa\sqrt{p}}{\sqrt{3(1-p)}}\right) \;:\; 0 \leq p \leq \frac{3}{3+\kappa^2}\right\}. \tag{5.23}$$

Using standard arguments, the maximum must be attained at a critical point of the objective function or at the boundary of the feasible domain of the non-convex optimization problem (5.23). The boundary points $p^\star = 0$ and $p^\star = \frac{3}{3+\kappa^2}$ establish that (5.15) is bounded from below by zero. The critical points are located at

$$\frac{\mathrm{d}}{\mathrm{d}p}\left[p\left(1-\frac{\kappa\sqrt{p}}{\sqrt{3(1-p)}}\right)\right] = 0 \iff \sqrt{3p}\cdot\kappa(3-2p) - 6(1-p)^{3/2} = 0.$$

The previous equation has only one real root, i.e.

$$p = -\frac{-36\kappa^4 - 108\kappa^2}{72(\kappa^2+3)\sqrt[3]{-\kappa^6 - 6\kappa^4 - 9\kappa^2 + \sqrt{3}\sqrt{\kappa^{10}+9\kappa^8+27\kappa^6+27\kappa^4}}} +$$
$$\frac{\sqrt[3]{-\kappa^6 - 6\kappa^4 - 9\kappa^2 + \sqrt{3}\sqrt{\kappa^{10}+9\kappa^8+27\kappa^6+27\kappa^4}}}{2(\kappa^2+3)} + 1.$$

Elementary manipulations now lead to the desired result. □

The worst-case distributions attaining the unimodal Cantelli inequality (5.22) can be explicitly constructed as well using the characterization of $\mathbb{P}^\star$ in equation (5.12) where $x = \sqrt{3(1-p)}/\sqrt{p}$ and p is given in the last equation in the proof of Proposition 5.6. Their explicit expression is however quite complex and is therefore omitted. Nevertheless, the worst-case distribution is graphically depicted in Figure 5.5(b).

5.4.2 Conditional value-at-risk inequalities

As in the worst-case expectation problem (5.1), we also want to consider the following worst-case CVaR problem:

$$\begin{aligned} B_{\text{CVaR}} := \sup \quad & \mathbb{P}\text{-CVaR}_\alpha\left(L(\xi)\right) \\ \text{s.t.} \quad & \mathbb{P} \in \mathcal{C}. \end{aligned}$$

Unfortunately, the previous worst-case CVaR problem is in general not a convex optimization problem in the form (5.1) discussed throughout this chapter. However, from a computational point of view the CVaR problem can be reduced to a worst-case expectation problem. Defining

$$\mathcal{L}(\beta,\mathbb{P}) := \beta + \frac{1}{\epsilon}\mathbf{E}_{\mathbb{P}}\left[(L(\xi)-\beta)^+\right]$$

and recalling the definition (3.11), our worst-case CVaR problem becomes

$$\begin{aligned} B_{\text{CVaR}} &= \sup_{\mathbb{P}\in\mathcal{C}} \inf_{\beta} \; \mathcal{L}(\beta,\mathbb{P}) = \inf_{\beta} \sup_{\mathbb{P}\in\mathcal{C}} \; \mathcal{L}(\beta,\mathbb{P}) \\ &= \inf_{\beta} \left\{ \beta + \sup_{\mathbb{P}\in\mathcal{C}} \frac{1}{\epsilon} \mathbf{E}_{\mathbb{P}} \left[(L(\xi)-\beta)^{+} \right] \right\}. \end{aligned}$$

Since $\mathcal{L}(\beta,\mathbb{P})$ is convex in β and linear in $\mathbb{P}$, the interchange of the supremum and infimum operations is justified by virtue of a stochastic saddle point theorem due to Shapiro [117]. The worst-case expectation problem (5.1) can now be seen to constitute an inner problem in the worst-case CVaR problem. Since the optimal $\beta^{\star}$ is shown to lie in a closed interval by Rockafellar [108] and $\sup_{\mathbb{P}\in\mathcal{C}} \; \mathcal{L}(\beta,\mathbb{P})$ is convex in β, computing a solution to the worst-case CVaR problem reduces to solving a sequence of worst-case expectation problems. For instance, the golden section search discussed in Kiefer [68] can be used to optimize $\sup_{\mathbb{P}\in\mathcal{C}} \; \mathcal{L}(\beta,\mathbb{P})$ only requiring a polynomial number of evaluations of $\sup_{\mathbb{P}\in\mathcal{C}} \; \mathbf{E}_{\mathbb{P}}[(L(\xi)-\beta)^{+}]$.

Hence in this dissertation we will deal with the more general worst-case expectation problem (5.1) directly. Nevertheless, we will make our results for worst-case CVaR bounds explicit at various places in this dissertation using the results derived for the worst-case expectation problem (5.1) and the discussion presented above.

Part II

Uncertainty quantification with second-order moment information

6 Primal uncertainty quantification with second-order moment information

In a wide range of applications, one is faced with the problem of quantifying the expected cost $L(\xi)$ of a random variable ξ with distribution $\mathbb{P}$. Common problems include determining the expected profit of a stock portfolio with uncertain stock returns as in Lo [78] or Bertsimas and Popescu [17], or quantifying the symbol error rate in a noisy communication channel as discussed by Vandenberghe et al. [133]. When the probability distribution $\mathbb{P}$ of the random vector ξ is known, computing $\mathbf{E}_{\mathbb{P}}[L(\xi)]$ typically reduces to the evaluation of a (high dimensional) integral. The evaluation of a high dimensional integral is shown by Nemirovski and Shapiro [88] to be in general however a computationally formidable task.

Furthermore, in practice it is often the case that the information available concerning the probability distribution $\mathbb{P}$ is limited. This means that the distribution of ξ is ambiguous and only known to belong to some ambiguity set $\mathcal{C}$ containing all probability distributions consistent with the known partial information concerning the distribution $\mathbb{P}$. We are thus limited to providing an upper bound on the expected cost $\mathbf{E}_{\mathbb{P}}[L(\xi)]$ holding uniformly for all probability distributions $\mathbb{P}$ in the ambiguity set $\mathcal{C}$. Hence when faced with limited information on the distribution of ξ, the least upper bound on the expected cost is given as

$$\begin{array}{ll} \sup & \int L(x)\, \mathbb{P}(\mathrm{d}x) \\ \text{s.t.} & \mathbb{P} \in \mathcal{C} \end{array} \tag{6.1}$$

Unfortunately, such worst-case expectation bounds or inequalities are generally unavailable in closed form, except in special cases where one can resort to classical bounds such as the Chebyshev or Gauss bounds discussed in Chapter 1.

Recently several scientific communities have however made a renewed effort to develop worst-case expectation bounds using a computational approach. Depending on the community, the problem

(6.1) is either referred to as an uncertainty quantification problem as in the work of Owhadi and Han [95, 59] or as a generalized moment problem by Bertsimas and Popescu [18, 103]. To avoid confusion, we will from here on always refer to optimization problems over ambiguity sets, i.e. sets of probability distributions, as uncertainty quantification problems. As indicated in Chapter 3, we reserve the name generalized moment problem to denote the problem of deciding feasibility of a distribution $\mathbb{P}$ in an ambiguity set $\mathcal{C}$ defined through moment conditions.

It can be remarked that for convex ambiguity sets $\mathcal{C}$ the problem (6.1) can be recognized as a convex optimization problem over probability distributions which were discussed at length in the previous chapter. For convex ambiguity sets $\mathcal{C}$, the uncertainty quantification problem can be approached via either its primal formulation (6.1) or its corresponding dual characterization as discussed in Chapter 5. It is of interest to remark that the literature in which problem (6.1) is denoted as an "uncertainty quantification problem" usually focusses on the primal characterization. At the same time, the corresponding dual characterization is the starting point in the literature referring to problem (6.1) as a "generalized moment problem". This difference in perspective between both approaches is reflected in this dissertation as well. The current chapter will present the primal approach to the uncertainty quantification problem (6.1), while the next chapter will approach the problem via its dual characterization. Despite this difference in perspective, both approaches ultimately start from the characterization of the uncertainty quantification problem (6.1) as an optimization problem over probability distributions as discussed in Chapter 5.

6.1 Second-moment information and Choquet simplices

Unfortunately, uncertainty quantification problems such as (6.1) are shown by Bertsimas and Popescu [18] to be in general NP-hard to solve. On the other hand, several tractable reformulations based on convex programming are given by Vandenberghe et al. [133] and Zymler et al. [142] for the case where the ambiguity set $\mathcal{C} = \mathcal{C}(\mu, S)$ consists of all probability distributions sharing a known mean and second moment. Thanks to modern interior point algorithms, these convex programming reformulations provide a de facto closed form solution to the resulting worst-case expectation bounds. The resulting bounds are widely used across many different disciplines such as in distributionally robust optimization by Delage and Ye [43] and in control by Van Parys et al. [131, 132] or portfolio selection and hedging as done by Yamada and Primbs [137] or Zymler et al. [143].

The main downside of these inequalities stems from the fact that the ambiguity set $\mathcal{C}(\mu, S)$ contains probability distributions that are not realistic in many applications and that consequently render the inequalities overly pessimistic. As a direct consequence of Proposition 5.1 the probability distributions achieving the worst-case expectation bound generically have discrete support with a finite number of discretization points. Vandenberghe et al. [133] describe this shortcoming as follows: *"In practical applications, the worst-case probability distribution will often be unrealistic, and the corresponding bound overly conservative."* The same adverse effect is also reported in Section 1.2 in case of the classical Chebyshev bound (1.6).

In this chapter we will therefore consider uncertainty quantification problems in which the ambiguity set $\mathcal{C}$ consists of *structured* probability distributions sharing a known mean and second moment. We will from hereon now consider the following uncertainty quantification problem with second-order moment information:

$$\begin{aligned} B(L, \mathcal{K}, \mu, S) = \sup \quad & \int L(x)\, \mathbb{P}(\mathrm{d}x) \\ \text{s.t.} \quad & \mathbb{P} \in \mathcal{H}(\mu, S) \\ & \mathbb{P} \in \mathcal{K}, \end{aligned} \tag{6.2}$$

where the hyperplane $\mathcal{H}(\mu, S)$ is defined as the collection of all measures sharing a known mean

μ and second moment S, i.e.

$$\mathcal{H}(\mu, S) := \left\{ \mathbb{P} \in \mathcal{E}_n \ : \ \int \mathbb{P}(dx) = 1, \ \int x \, \mathbb{P}(\mathrm{d}x) = \mu, \ \int xx^\top \, \mathbb{P}(\mathrm{d}x) = S \right\}.$$

Apart from mean and second moment, we will enforce the ambiguity set $\mathcal{C}$ to only contain probability distributions enjoying additional structure. The set $\mathcal{K}$ will be used to characterize any further structural information about the probability distributions $\mathbb{P}$ considered, e.g. unimodality or monotonicity. When $\mathcal{K}$ is taken to be the standard probability simplex $\mathcal{P}_n$, then problem (6.2) reduces to a standard uncertainty quantification problem as discussed by for instance Bertsimas and Popescu [18], Zymler et al. [142] and Vandenberghe et al. [133].

The principal aim of this and the subsequent chapter is to provide a unified computational approach to the situations under which problem (6.2) admits a tractable reformulation in terms of a convex optimization problem. As mentioned already, we are primarily interested in the situation in which the ambiguity set in richly structured $\mathcal{K} \subset \mathcal{P}_n$ so as to exclude pathological distributions which can make the corresponding bound overly pessimistic. In the remainder of this section we will assume that the ambiguity set $\mathcal{K}$ admits a Choquet star representation in terms of the generating distribution $\mathbb{T}$. As argued in Section 3.2, many type of structures commonly imposed on distributions in practice such as unimodality and monotonicity in fact admit a Choquet star representation. The domain of the worst-case bound $B(L, \mathcal{K}, \mu, S)$ can now be determined explicitly using Proposition 3.7 as an LMI.

Fact 6.1 (Domain of $B(L, \mathcal{K}, \mu, S)$)**.** *The domain of the worst-case bound $B(L, \mathcal{K}, \mu, S)$ with respect to its last two variables is given as*

$$\left\{ (\mu, S) \in \mathrm{R}^n \times \mathrm{S}^n \ : \ \begin{pmatrix} S_s & \mu_s \\ \mu_s^\top & 1 \end{pmatrix} \succeq 0, \ S_s \cdot \int_0^\infty t^2 \, \mathbb{T}(\mathrm{d}t) = S, \ \mu_s \cdot \int_0^\infty t \, \mathbb{T}(\mathrm{d}t) = \mu \right\}.$$

The first and second moment of the generating distribution $\mathbb{T}$ in case of α-unimodal and γ-monotone distributions are given explicitly in Table 3.1.

The worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$ over $\mathcal{K}$ a choquet star representable set can be reduced to a related worst-case expectation bound over the standard probability simplex $\mathcal{P}_n$, i.e.

$$B(L, \mathcal{K}, \mu, S) = B(L_s, \mathcal{P}_n, \mu_s, S_s)$$

for judiciously chosen loss function L_s mean μ_s and second moment S_s. With this in mind, the power of Choquet star representable ambiguity sets becomes clear. Indeed, the Choquet star structure of the set $\mathcal{K}$ allows us to restrict attention to uncertainty quantification problems over the standard probability simplex. This reduction will greatly benefit the unified presentation of the computational results on the worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$ in this and the subsequent Chapter.

We now state how a Choquet star representation of $\mathcal{K}$ can be utilized to remodel a structured problem in the form (6.2) as an equivalent unstructured problem (i.e. one with ambiguity set $\mathcal{K} = \mathcal{P}_n$) via an appropriate transformation of the loss function and moments.

Theorem 6.1 (Reduction to the standard simplex $\mathcal{P}_n$)**.** *Assume that the ambiguity set $\mathcal{K}$ admits a Choquet star representation with generating distribution $\mathbb{T}$, then*

$$B\left(L, \mathcal{K}, \mu, S\right) = B\left(L_s, \mathcal{P}_n, \mu_s, S_s\right) \tag{6.3}$$

for $L_s(x) := \int_0^\infty L(tx) \, \mathbb{T}(\mathrm{d}t)$, $S_s \cdot \int_0^\infty t^2 \, \mathbb{T}(\mathrm{d}t) = S$ and $\mu_s \cdot \int_0^\infty t \, \mathbb{T}(\mathrm{d}t) = \mu$.

Proof. Since the ambiguity set $\mathcal{K}$ admits a Choquet star representation, we can optimize over the mixture representations $\mathbb{m}$ instead of $\mathbb{P}$. Using the reparametrization $\mathbb{P} = \int \mathbb{T}_y \, \mathbb{m}(\mathrm{d}y)$ we obtain

$$\begin{array}{llcll} \sup\limits_{\mathbb{P}} & \displaystyle\int L(x)\,\mathbb{P}(\mathrm{d}x) & = & \sup\limits_{\mathbb{m}} & \displaystyle\int \left[\int L(x)\,\mathbb{T}_y(\mathrm{d}x)\right] \mathbb{m}(\mathrm{d}y) \\ \text{s.t.} & \mathbb{P} \in \mathcal{K} \cap \mathcal{H}(\mu, S) & & \text{s.t.} & \displaystyle\int \mathbb{T}_y \, \mathbb{m}(\mathrm{d}y) \in \mathcal{H}(\mu, S). \end{array}$$

Indeed, we have that the condition $\mathbb{P} = \int \mathbb{T}_y \, \mathbb{m}(\mathrm{d}y) \in \mathcal{K} \cap \mathcal{H}(\mu, S)$ is equivalent to the requirement that $\int \mathbb{T}_y \, \mathbb{m}(\mathrm{d}y) \in \mathcal{C}(\mu, S)$. Furthermore, we have the identity

$$\int [x^\top, 1]^\top \cdot [x^\top, 1] \; \mathbb{P}(\mathrm{d}x) = \int \left[\int [x^\top, 1]^\top \cdot [x^\top, 1] \, \mathbb{T}_y(\mathrm{d}x)\right] \; \mathbb{m}(\mathrm{d}y),$$

which equals using Fubini's Theorem and the Choquet star property of $\mathbb{T}_y$

$$\int [x^\top, 1]^\top \cdot [x^\top, 1] \; \mathbb{P}(\mathrm{d}x) = \int \begin{pmatrix} \int_0^\infty t^2 \, \mathbb{T}(\mathrm{d}t) \; y \cdot y^\top & \int_0^\infty t \, \mathbb{T}(\mathrm{d}t) \; y \\ \int_0^\infty t \, T(\mathrm{d}t) \; y^\top & 1 \end{pmatrix} \mathbb{m}(\mathrm{d}y).$$

Hence the condition $\mathbb{P} = \int \mathbb{T}_y \, \mathbb{m}(\mathrm{d}y) \in \mathcal{C}(\mu, S)$ is equivalently stated as the requirement $\mathbb{m} \in \mathcal{C}(\mu_s, S_s)$. We have because of Fubini's Theorem that the integral $\int L(x)\,\mathbb{P}(\mathrm{d}x)$ for $\mathbb{P} = \int \mathbb{T}_y \, \mathbb{m}(\mathrm{d}y)$ equals the integral $\int L_s(x)\,\mathbb{m}(\mathrm{d}x)$ where both loss functions are related via the transformation $L_s(y) = \int_0^\infty L(ty)\,\mathbb{T}(\mathrm{d}t)$. □

Hence, an uncertainty quantification problem over a Choquet star simplex $\mathcal{K}$ can be reduced to an equivalent problem over the standard probability simplex $\mathcal{P}_n$; see also Figure 6.1. Both uncertainty quantification problems are related in terms of their loss functions as

$$L_s(y) = \int L(x)\,\mathbb{T}_y(\mathrm{d}x) \tag{6.4}$$

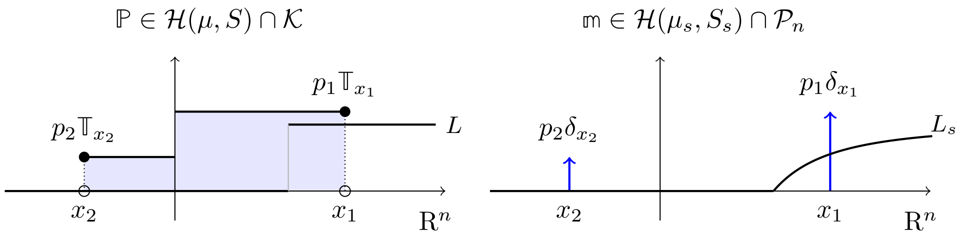
according to the result presented in Theorem 6.1. The equivalent problems $B\,(L, \mathcal{K}, \mu, S) = \int L(x)\,\mathbb{P}^\star(\mathrm{d}x)$ and $B\,(L_s, \mathcal{P}_n, \mu_s, S_s) = \int L_s(x)\,\mathbb{m}^\star(\mathrm{d}x)$ are not only related in terms of the worst-case expectation bound but also in terms of the distributions in their corresponding convex optimization problems. As indicated in the proof of Theorem 6.1, the relationship

$$\mathbb{P} = \int \mathbb{T}_x \, \mathbb{m}(\mathrm{d}x)$$

always holds. Because of the fundamental theorem stated in Chapter 5, the worst-case distributions $\mathbb{P}^\star$ and $\mathbb{m}^\star$ can always be found as a convex combination of at most $(n+1)(n+2)/2$ extreme distributions of $\mathcal{K}$ or $\mathcal{P}_n$, respectively. We thus have that the worst-case distributions attaining either $B\,(L, \mathcal{K}, \mu, S)$ or $B\,(L_s, \mathcal{P}_n, \mu_s, S_s)$ are in the canonical forms

$$\mathbb{P}^\star = \sum_i p_i \cdot \mathbb{T}_{x_i} \quad \text{and} \quad \mathbb{m}^\star = \sum_i p_i \cdot \delta_{x_i}.$$

Both in the current chapter and in subsequent Chapter 7, we will derive computational reformulations for uncertainty quantification problems over the standard probability simplex $\mathcal{P}_n$. As the previous discussion indicated taking $\mathcal{K} = \mathcal{P}_n$ is without loss of generality when the ambiguity set $\mathcal{K}$ admits a Choquet star representation. In a next step we will then show how uncertainty quantification problems over a more richly structured Choquet star simplex $\mathcal{K}$ can be treated equally well via the transformation (6.4) and the reduction Theorem 6.1.



(a) $B(L, \mathcal{K}, \mu, S)$ (b) $B(L_s, \mathcal{P}_n, \mu_s, S_s)$

Figure 6.1: The uncertainty quantification problem (6.2) for mean μ and second moment S over the Choquet star simplex $\mathcal{K}$ is equivalent to an uncertainty quantification problem over the standard probability simplex for judiciously chosen loss function L_s, mean μ_s, and second moment S_s. The fundamental theorem guarantees that the worst-case distribution $\mathbb{P}^\star$ or $\mathbb{m}^\star$ is a convex combination of a finite number of extreme distributions in either $\mathcal{K}$ or $\mathcal{P}_n$, respectively. Furthermore, these worst-case distributions are in both situations related according to $\mathbb{P} = \int \mathbb{T}_x \, \mathbb{m}(\mathrm{d}x)$ as depicted in blue for star unimodal distributions ($\mathcal{K} = \mathcal{U}_n$).

6.2 Primal uncertainty quantification via perspective functions

In this chapter we will approach the uncertainty quantification problem (6.2) via its primal reformulation. Initially, we will only consider unstructured probability distributions $\mathcal{K} = \mathcal{P}_n$. We focus on the uncertainty quantification problem (6.2) for unstructured probability distributions

$$B(\max_{i\in I_0} \ell_i(x), \mathcal{P}_n, \mu, S) = \begin{array}{ll} \sup\limits_{\mathbb{P}} & \displaystyle\int [\max_{i\in I_0} \ell_i(x)] \, \mathbb{P}(\mathrm{d}x) \\ \text{s.t.} & \mathbb{P} \in \mathcal{H}(\mu, S) \cap \mathcal{P}_n, \end{array} \tag{6.5}$$

where the functions ℓ_i are understood to be concave for all i in the index set $I_0 = I \cup \{0\} = [0, \ldots, k]$. We assume further that the loss function L is positive. Previous assumption is without loss of generality when L is bounded from below and is enforced by taking $\ell_0(x) = 0$. In Section 6.3, we will then illustrate that the uncertainty quantification problem (6.5) is in fact rich enough to handle several interesting uncertainty quantification problems (6.2) in which $\mathcal{K} \subset \mathcal{P}_n$, following the discussion at the end of the preceding section.

Under aforementioned conditions, an uncertainty quantification problem is most naturally treated in its primal form. Indeed, we will show that problem (6.5) can be restated as a convex optimization problem in terms of perspective functions of the functions ℓ_i. Subsequently, we will discuss in Section 6.2.1 how the variables of this optimization problem can be related to the worst-case probability distribution in the uncertainty quantification problem (6.2). It will be shown in Section 6.2.2 that the generalized Chebyshev bound discovered by Vandenberghe et al. [133] can be recognized as a corollary of Theorem 6.2.

Theorem 6.2 (Perspective functions)**.** *The bound $B(\max_{i\in I_0} \ell_i(x), \mathcal{P}_n, \mu, S)$ for concave functions ℓ_i can be reformulated as*

$$\begin{array}{ll} \sup & \displaystyle\sum_{i\in I} p_i \, \ell_i\!\left(\frac{z_i}{p_i}\right) \\ \text{s.t.} & p_i \in \mathrm{R}, \;\; z_i \in \mathrm{R}^n, \;\; Z_i \in \mathrm{R}^{n\times n}, \quad \forall i \in I \\ & \displaystyle\sum_{i\in I} \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \preceq \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} \\ & \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \succeq 0, \quad \forall i \in I \end{array} \tag{6.6}$$

which is a convex optimization problem.

Proof. We first show that any feasible solution for the optimization problem (6.6) can be used to construct a feasible probability distribution $\mathbb{P} \in \mathcal{C}(\mu, S)$ achieving the same objective value in (6.5). Let $\{p_i, z_i, Z_i, \}_{i\in I}$ be feasible in (6.6) and set $x_i = z_i/p_i$ if $p_i > 0$; $= 0$ otherwise. Moreover, assume without loss of generality that $Z_i = z_i z_i^\top / p_i$ if $p_i > 0$; $= 0$ if $p_i = 0$. This choice preserves feasibility of $\{p_i, z_i, Z_i\}_{i\in I}$ and has no effect on its objective value in (6.6). Next, define

$$\begin{pmatrix} Z_0 & z_0 \\ z_0^\top & p_0 \end{pmatrix} := \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} - \sum_{i\in I} \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \succeq 0\,, \tag{6.7}$$

which is positive semidefinite due to the first constraint in (6.6). Assume now that $p_0 > 0$ and define $\mu_0 = z_0/p_0$ and $S_0 = Z_0/p_0$. Proposition 3.6 then guarantees the existence of a probability distribution $\mathbb{P}_0 \in \mathcal{C}(\mu_0, S_0)$, which allows us to construct

$$\mathbb{P} = p_0\,\mathbb{P}_0 + \sum_{i\in I} p_i\,\delta_{x_i}.$$

The first and second moments of $\mathbb{P}$ are given by

$$\begin{aligned} \int \begin{pmatrix} x \\ 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}^\top \mathbb{P}(\mathrm{d}x) &= p_0 \begin{pmatrix} S_0 & \mu_0 \\ \mu_0^\top & 1 \end{pmatrix} + \sum_{i\in I} p_i \begin{pmatrix} x_i x_i^\top & x_i \\ x_i^\top & 1 \end{pmatrix} \\ &= \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix}, \end{aligned}$$

where the second equation follows from equation (6.7). We hence conclude that $\mathbb{P} \in \mathcal{C}(\mu, S)$. Moreover, elementary calculation shows that

$$\begin{aligned} \int \max_{i\in I_0} \ell_i(x)\,\mathbb{P}(\mathrm{d}x) &= p_0 \int \max_{i\in I_0} \ell_i(x)\,\mathbb{P}_0(\mathrm{d}x) + \sum_{j\in I} \max_{i\in I_0} \ell_i(x_j) p_j, \\ &\geq \sum_{i\in I} \ell_i(x_i) p_i = \sum_{i\in I} p_i \ell_i \left(\frac{z_i}{p_i}\right) \end{aligned}$$

In summary, $\mathbb{P}$ is feasible in the worst-case expectation problem (6.5) with an objective value that is at least as large as that of $\{p_i, z_i, Z_i\}_{i\in I}$ in (6.6). If $p_0 = 0$, we can set $\hat{z}_i = (1-\epsilon) z_i$, $\hat{Z}_i = (1-\epsilon) Z_i$ and $\hat{p}_i = (1-\epsilon) p_i$ for some $\epsilon \in (0, 1)$. By repeating the above arguments for $p_0 > 0$, we can use $\{\hat{p}_i, \hat{z}_i, \hat{Z}_i\}_{i\in I}$ to construct a feasible probability distribution of (6.5) with an objective value of at least $(1-\epsilon)\sum_{i\in I} p_i \ell_i\,(z_i/p_i)$. As ϵ tends to zero, we obtain a sequence of probability distributions feasible in (6.5) whose objective values asymptotically approach that of $\{p_i, z_i, Z_i\}_{i\in I}$ in (6.6). We conclude that (6.6) provides a lower bound on (6.5).

Next, we prove that (6.5) also provides a lower bound on (6.6) and that any feasible solution for (6.5) gives rise to a feasible solution for (6.6) with the same objective value. To this end, we define for all i in I_0 the set

$$\Xi_i = \{x \in \mathrm{R}^n \,:\, \ell_i(x) \geq \ell_j(x),\ \forall j < i \text{ and } \ell_i(x) > \ell_j(x),\ \forall j > i\}\,.$$

Note that the sets Ξ_i form a partition of R^n. Consider now a probability distribution $\mathbb{P}$ that is feasible in (6.5). Next, define

$$\begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} = \int_{\Xi_i} \begin{pmatrix} x \\ 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}^\top \mathbb{P}(\mathrm{d}x) \succeq 0 \tag{6.8}$$

for all i in I. By construction, we have

$$\sum_{i\in I_0}\begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i\end{pmatrix} = \sum_{i\in I_0}\int_{\Xi_i}\begin{pmatrix} x \\ 1\end{pmatrix}\begin{pmatrix} x \\ 1\end{pmatrix}^\top \mathbb{P}(\mathrm{d}x) = \int \begin{pmatrix} x \\ 1\end{pmatrix}\begin{pmatrix} x \\ 1\end{pmatrix}^\top \mathbb{P}(\mathrm{d}x)$$
$$= \begin{pmatrix} S & \mu \\ \mu^\top & 1\end{pmatrix}.$$

Thus, the $\{p_i, z_i, Z_i\}_{i\in I}$ constructed in (6.8) are feasible in (6.6). Their objective value in (6.6) can be represented as

$$\sum_{i\in I} p_i \ell_i\left(\frac{z_i}{p_i}\right) = \sum_{i\in I} p_i \ell_i(x_i)$$
$$\geq \sum_{i\in I_0}\int_{\Xi_i} \ell_i(x)\,\mathbb{P}(\mathrm{d}x) = \int \left[\max_{i\in I_0}\,\ell_i(x)\right]\,\mathbb{P}(\mathrm{d}x)$$

where the second line follows from Jensen's inequality and the last equality is a direct consequence of the definition of the sets Ξ_i. The objective value of $\{p_i, z_i, Z_i\}_{i\in I}$ thus coincides with the objective value of $\mathbb{P}$ in (6.5). $\square$

6.2.1 Worst-case probability distributions

In addition to identifying a tractable reformulation of the uncertainty quantification problem (6.5), it is also of interest to identify a worst-case probability distribution $\mathbb{P}^\star \in \mathcal{C}(\mu, S)$, i.e.

$$\int [\max_{i\in I_0}\,\ell_i(x)]\ \mathbb{P}^\star(\mathrm{d}x) = B(\max_{i\in I_0}\,\ell_i(x), \mathcal{P}_n, \mu, S),$$

should it exist. The fundamental theorem of linear programming 5.1 states that if the supremum in problem (6.5) is attained, then it is attained in particular by a probability distribution $\mathbb{P}^\star$ consisting of at most $(n+2)(n+1)/2$ Dirac distributions. Indeed, the number of unique half-space constraints or co-dimension of the set $\mathcal{H}(\mu, S)$ is exactly n for the mean μ on top of the $(n+1)n/2$ constraints related to the second moment S.

The proof of Theorem 6.2 suggests an explicit construction of a worst-case probability distribution as the convex combination of at most $2n+k$ extreme distributions. With the help of any maximizer $\{p_i^\star, z_i^\star\}_{i\in I}$ of the convex optimization problem (6.6) satisfying $0 < p_0^\star = 1 - \sum_{i\in I} p_i^\star$, a worst-case probability distribution in the form

$$\mathbb{P}^\star = p_0^\star \mathbb{P}_0 + \sum_{i\in I} p_i^\star \delta_{x_i^\star}$$

can be constructed where $x_i^\star = z_i^\star/p_i^\star$ if $p_i^\star > 0$; $= 0$ otherwise, and where the probability distribution $\mathbb{P}_0 \in \mathcal{C}(\mu_0, S_0)$ with

$$p_0^\star\begin{pmatrix} S_0 & \mu_0 \\ \mu_0^\top & 1\end{pmatrix} = \begin{pmatrix} S & \mu \\ \mu^\top & 1\end{pmatrix} - \sum_{i\in I} p_i^\star \begin{pmatrix} x_i^\star x_i^{\star\top} & x_i^\star \\ x_i^{\star\top} & 1\end{pmatrix} \succeq 0.$$

Such a probability distribution $\mathbb{P}_0$ can always be found as the convex combination of at most $2n$ extreme distributions of $\mathcal{P}_n$. Indeed, the positive semidefinite variance matrix $\Sigma_0 := S_0 - \mu_0\mu_0^\top$ can be factored as $\Sigma_0 = \sum_{i=1}^r w_i w_i^\top$ where $r \leq n$ is the rank of the variance matrix Σ_0. It can now be readily verified that the probability distribution

$$\mathbb{P}_0 = \sum_{i=1}^r \frac{1}{2r}\delta_{\mu_0+\sqrt{r}w_i} + \sum_{i=1}^r \frac{1}{2r}\delta^\alpha_{\mu_0-\sqrt{r}w_i},$$

satisfies $\mathbb{P}_0 \in \mathcal{C}(\mu_0, S_0)$. When compared to the fundamental theorem, it is clear that the worst-case probability distribution $\mathbb{P}^\star$ constructed from a maximizer of the optimization problem (6.6) is not necessarily maximally sparse.

We remark that the reformulation offered in Theorem 6.2 is exact even though no worst-case probability distribution $\mathbb{P}$ may exist. The nonexistence of a worst-case probability distribution in problem (6.2) occurs only when $p_0^\star = 0$ in its reformulation (6.6). In that case, any maximizer of (6.6) can be used to construct a sequence of probability distributions $\{\mathbb{P}_t\}$, $\mathbb{P}_t \in \mathcal{C}(\mu, S)$ with the property

$$\lim_{t\to\infty} \mathbb{P}_t(\xi \notin \Xi) = B(\max_{i\in I_0} \ell_i(x), \mathcal{P}_n, \mu, S).$$

6.2.2 Generalized Chebyshev inequalities

Recent generalizations of the classical Chebyshev inequality (1.6) provide upper bounds on the probability of a multivariate random vector $\xi \in \mathrm{R}^n$ falling outside a prescribed confidence region $\Xi \subseteq \mathrm{R}^n$ if only the mean and second moment of ξ are known. The best upper bound of this kind is given by the optimal value of the worst-case probability problem

$$G_\infty(\mu, S) := \sup_{\mathbb{P}\in\mathcal{C}(\mu,S)} \mathbb{P}\left(\xi \notin \Xi\right). \tag{6.9}$$

The problem (6.9) has a natural interpretation as an uncertainty quantification problem using the standard identity

$$\mathbb{P}(\mathrm{R}^n \setminus \Xi) = \int \mathbf{1}\left\{\mathrm{R}^n \setminus \Xi\right\}(x)\, \mathbb{P}(\mathrm{d}x)$$

between measure and expectation of indicator functions. Vandenberghe et al. [133] showed that (6.9) admits an exact reformulation as a single SDP whenever Ξ is polytopic and described through finitely many half-space constraints. The resulting generalized Chebyshev bounds are widely used across many different application domains, ranging from distributionally robust optimization in Delage and Ye [43] to chance-constrained programming by Zymler et al. [142] and Xu et al. [136] and Cheng et al. [38], stochastic control applications by this author [131], machine learning techniques by Lanckriet et al. [74], for signal processing in Vorobyov et al. [134], in Lo [78] and Grundy [57] and Bertsimas and Popescu [17] for option pricing, portfolio selection and hedging applications are found in Yamada and Primbs [137], or finally in decision theory by Smith [119] etc.

Consider an open polytope Ξ representable as a finite intersection of open half spaces,

$$\Xi = \left\{x \in \mathrm{R}^n \ : \ a_i^\top x < b_i, \quad \forall i \in I\right\}, \tag{6.10}$$

where $a_i \in \mathrm{R}^n$, $a_i \neq 0$, and $b_i \in \mathrm{R}$ for all $i \in I$. The corresponding generalized Chebyshev bound (6.9) can be reformulated as an SDP as shown for instance by Vandenberghe et al. [133]. We now indicate that this result is readily obtained as a corollary of our Theorem 6.2.

Theorem 6.3 (Generalized Chebyshev bounds [133])**.** *If Ξ is a polytope of the form* (6.10)*, the worst-case probability problem* (6.9) *with ambiguity set $\mathcal{C}(\mu, S)$ is equivalent to a tractable SDP:*

$$\begin{array}{rcll}\sup\limits_{\mathbb{P}\in\mathcal{C}(\mu,S)} \mathbb{P}(\xi \notin \Xi) & = & \max & \sum\limits_{i\in I} p_i \\ & & \text{s.t.} & z_i \in \mathrm{R}^n,\ Z_i \in \mathrm{S}^n,\ p_i \in \mathrm{R} \quad \forall i \in I \\ & & & a_i^\top z_i \geq b_i p_i \quad \forall i \in I \\ & & & \sum\limits_{i\in I} \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \preceq \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} \\ & & & \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \succeq 0 \quad \forall i \in I. \end{array} \tag{6.11}$$

Proof. It is clear that the uncertainty quantification problem (6.9) is equivalent to the bound $B(\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}, \mathcal{P}_n, \mu, S)$ defined by (6.5). The loss function $L = \mathbf{1}\{\mathrm{R}^n \setminus \Xi\}$ can be written in the form required by Theorem 6.2, i.e. $L(x) = \max_{i\in I_0} \ell_i(x)$ where the functions

$$\ell_i(x) = \begin{cases} 1 & \text{if } a_i^\top x \geq b_i, \\ -\infty & \text{Otherwise} \end{cases}$$

are concave for all $i \in I$. The corresponding perspective functions are in this case given as $p_i \ell(z_i/p_i) = p_i$ if $a_i^\top z_i \geq b_i p_i$; $-\infty$ otherwise. $\square$

As illustrated in the preceding section, the solution of the convex optimization problem (6.11) by Vandenberghe et al. [133] can be used to construct a *discrete* worst-case probability distribution for the uncertainty quantification problem (6.9). The existence of optimal discrete distributions has distinct computational benefits and can be viewed as the key enabling property that facilitates the SDP reformulation of the uncertainty quantification problem (6.9). However, it also renders the corresponding Chebyshev bound rather pessimistic. Indeed, uncertainties encountered in real physical, technical or economic systems are unlikely to follow discrete distributions with few atoms. By accounting for such pathological distributions, problem (6.9) tends to overestimate the probability of the event $\xi \notin \Xi$ significantly as already remarked upon by Vandenberghe et al. [133].

6.3 Generalized Gauss inequalities

In order to mitigate the pessimism innate to the Chebyshev bound, one could impose additional restrictions on the ambiguity set $\mathcal{C}$ that complement the given moment information. A minimal structural property commonly encountered in practical situations is unimodality. Note that most probability distributions commonly studied in probability theory are unimodal. So too are all stable distributions, which are ubiquitous in statistics as they represent the attractors for properly normed sums of independent and identically distributed random variables.

The purpose of this section is to generalize the classical Gauss inequality (1.7) to multivariate probability distributions, providing a counterpart to the generalized Chebyshev inequality (6.11). Extensions of the univariate Gauss inequality involving generalized moments have previously been proposed by Sellke [115], while multivariate extensions have been investigated by Meaux et al. [85]. Popescu [103] uses ideas from Choquet theory similar to ours in conjunction with sums-of-squares polynomial techniques to derive *approximate* multivariate Gauss-type inequalities. However, to the best of our knowledge, until now no efficient algorithm is known to compute the underlying worst-case probabilities *exactly*.

6.3.1 The α-unimodal bound

We will now investigate the worst-case probability of the event $\xi \notin \Xi$ over all probability distributions from within $\mathcal{C}_\alpha(\mu, S) := \mathcal{H}(\mu, S) \cap \mathcal{U}_\alpha$,

$$G_\alpha(\mu, S) = \sup_{\mathbb{P}\in\mathcal{C}_\alpha(\mu,S)} \mathbb{P}(\xi \notin \Xi), \tag{6.12}$$

and we will prove that the worst-case probability $G_\alpha(\mu, S)$ can be computed efficiently by solving a tractable SDP. Following the discussion in Section 3.1, we will indicate afterwards that the classical Gauss bound is seen to constitute a special case of the hierarchy of bounds $G_\alpha(\mu, S)$ by letting $\alpha = n$. Furthermore, the generalized Chebyshev bound $G_\infty(\mu, S)$ will be proved to present the limit of $G_\alpha(\mu, S)$ for α tending to infinity. This should not come as a surprise as we have that $\mathcal{P}_n = \mathrm{cl} \cup_{\alpha\geq 0} \mathcal{U}_\alpha$.

It is clear that the worst-case probability bound $G_\alpha(\mu, S)$ is equivalent to the worst-case bound $B(\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}, \mathcal{U}_\alpha, \mu, S)$. For completeness we make Fact 6.1 explicit in case of α-unimodal distributions using the appropriate result in Table 3.1.

Fact 6.2 (Domain of $G_\alpha(\mu, S)$). *The domain of the worst-case probability bound $G_\alpha(\mu, S)$ is given as*

$$\operatorname{dom} G_\alpha = \left\{ (\mu, S) \ : \ \begin{pmatrix} \frac{\alpha+2}{\alpha} S & \frac{\alpha+1}{\alpha}\mu \\ \frac{\alpha+1}{\alpha}\mu^\top & 1 \end{pmatrix} \succeq 0 \right\}.$$

As stressed repeatedly, Theorem 6.2 deals only with uncertainty quantification problems over the standard probability simplex $\mathcal{P}_n$. Nevertheless, using the reduction Theorem 6.1 we will show that we can handle the uncertainty quantification problem (6.12) for α-unimodal probability distributions as well. The following theorem establishes that the uncertainty quantification problem (6.12) admits a convex reformulation and can be recognized as a corollary of Theorem 6.2 via the reduction Theorem 6.1.

Theorem 6.4 (α-Unimodal bound). *For $0 \in \Xi$ the problem (6.12) is equivalent to the convex optimization problem*

$$\begin{array}{rcll} G_\alpha(\mu, S) & = & \max & \sum_{i \in I} (p_i - \tau_i) \\ & & \text{s.t.} & z_i \in \mathrm{R}^n, \ Z_i \in \mathrm{S}^n, \ p_i \in \mathrm{R}, \ \tau_i \in \mathrm{R} \quad \forall i \in I \\ & & & \sum_{i \in I} \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \preceq \begin{pmatrix} \frac{\alpha+2}{\alpha} S & \frac{\alpha+1}{\alpha}\mu \\ \frac{\alpha+1}{\alpha}\mu^\top & 1 \end{pmatrix} \\ & & & \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \succeq 0 \quad \forall i \in I \\ & & & a^\top z_i \geq 0, \ \tau_i \geq 0, \ \tau_i (a_i^\top z_i)^\alpha \geq p_i^{\alpha+1} b_i^\alpha \quad \forall i \in I. \end{array} \tag{6.13}$$

Proof. The proof is very similar to the proof of Theorem 6.3. From the reduction Theorem 6.1 it follows that we have $B(\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}, \mathcal{U}_\alpha, \mu, S) = B(L_s, \mathcal{P}_n, \mu_s, S_s)$ where L_s is defined by transformation (6.4). In case of α-unimodal distributions, the generating probability distribution $\mathbb{T}$ is given in Example 3.2.

$$\begin{aligned} L_s(x) &= \int_0^1 \mathbf{1}\{\mathrm{R}^n \setminus \Xi\}(tx)\, \alpha t^{\alpha-1} \mathrm{d}t \\ &= \max_{i \in I} \int_0^1 \mathbf{1}\left\{a_i^\top x \geq b_i\right\}(tx) \alpha t^{\alpha-1} \mathrm{d}t \\ &= \max_{i \in I_0} \ell_i(x) \end{aligned}$$

where the functions ℓ_i are for all $i \in I$ given as

$$\ell_i(x) = \begin{cases} 1 - \left(\frac{b_i}{a_i^\top x}\right)^\alpha & \text{if } a_i^\top x \geq 0 \\ -\infty & \text{Otherwise} \end{cases}$$

and are all concave because $b_i > 0$ as $0 \in \Xi$. The corresponding perspective functions are in this case given as $p_i \ell(z_i/p_i) = p_i - b_i^\alpha p_i^{\alpha+1}/(a_i^\top z_i)^\alpha$ if $a_i^\top z \geq 0$; $-\infty$ otherwise. The concave perspective function can be represented in epigraph form using an auxiliary variable τ_i and a convex constraint $\tau_i \geq b_i^\alpha p_i^{\alpha+1}/(a_i^\top z_i)^\alpha$ as suggested in (6.13). □

Note that problem (6.13) fails to be an SDP in standard form due to the nonlinearity of its last constraint. It is hence unclear whether the convex reformulation (6.13) of the uncertainty quantification problem (6.12) is tractable. However, in Lemmas 6.1 and 6.2 below we will show that this constraint is in fact second-order cone representable under the mild additional assumption that α is rational and not smaller than 1. In this case, (6.13) is thus equivalent to a tractable SDP. We start with the case that the unimodality parameter α is a rational number not smaller than 1.

Lemma 6.1 (Second-order cone representation for rational α)**.** *Suppose that $b > 0$ and $\alpha = v/w$ for $(v, w) \in \mathrm{N}$ with $v \geq w$. If the linear constraints $p \geq 0$, $\tau \geq 0$ and $a^\top z \geq 0$ hold, then the nonlinear constraint $(a^\top z)^\alpha \, \tau \geq p^{\alpha+1} \, b^\alpha$ has an equivalent representation in terms of $\mathcal{O}(v)$ second-order constraints involving z, τ, p and $\mathcal{O}(v)$ auxiliary variables.*

Proof. We have

$$
\begin{aligned}
(a^\top z)^{\frac{v}{w}} \, \tau \geq p^{\frac{v+w}{w}} \, b^{\frac{v}{w}} \quad &\iff \quad (a^\top z)^v \, \tau^w \geq p^{v+w} \, b^v \\
&\iff \quad \exists t \geq 0 \ : \ (a^\top z)^v \, t^v \geq p^{2v} \, b^v, \ \ \tau^w \, p^{v-w} \geq t^v \\
&\iff \quad \exists t \geq 0 \ : \ (a^\top z) \, t \geq p^2 \, b, \ \ \tau^w \, p^{v-w} \, t^{2^\ell - v} \geq t^{2^\ell},
\end{aligned}
$$

where $\ell = \lceil \log_2 v \rceil$. Both constraints in the last line of the above expression are second-order cone representable. Indeed, the first (hyperbolic) constraint is equivalent to

$$
\left\| \begin{pmatrix} 2pb \\ tb - a^\top z \end{pmatrix} \right\|_2 \leq tb + a^\top z \, ,
$$

while the second constraint can be reformulated as $(\tau^w \, p^{v-w} \, t^{2^\ell - v})^{1/2^\ell} \geq t$ and thus requires the geometric mean of 2^ℓ nonnegative variables to be non-inferior to t. Using a result of Nesterov and Nemirovski [90, Section 6.2.3.5], this requirement can be re-expressed in terms of $\mathcal{O}(2^\ell)$ second-order cone constraints involving $\mathcal{O}(2^\ell)$ auxiliary variables. $\square$

Lemma 6.1 establishes that (6.13) has a tractable reformulation for any rational $\alpha \geq 1$ by exploiting a well-known second-order cone representation for geometric means. When α is integer, one can construct a more efficient reformulation involving far fewer second-order cone constraints and auxiliary variables. The following lemma derives this reformulation explicitly.

Lemma 6.2 (Second-order cone representation for integer α)**.** *Suppose that $b > 0$ and $\alpha \in \mathrm{N}$. If the linear constraints $p \geq 0$, $\tau \geq 0$ and $a^\top z \geq 0$ hold, then the nonlinear constraint $(a^\top z)^\alpha \, \tau \geq p^{\alpha+1} \, b^\alpha$ is equivalent to*

$$
\exists t \in \mathbb{R}^{\ell+1} \quad : \quad \begin{cases} t \geq 0, \quad t_0 = \tau, \quad \left\| \begin{pmatrix} 2pb \\ t_\ell b - a^\top z \end{pmatrix} \right\|_2 \leq t_\ell b + a^\top z, \\ \left\| \begin{pmatrix} 2t_{j+1} \\ t_j - p \end{pmatrix} \right\|_2 \leq t_j + p \quad \forall j \in E, \quad \left\| \begin{pmatrix} 2t_{j+1} \\ t_j - t_\ell \end{pmatrix} \right\|_2 \leq t_j + t_\ell \quad \forall j \in O, \end{cases}
$$

where $\ell = \lceil \log_2(\alpha) \rceil$, $E = \{j \in [0, \ldots, \ell - 1] : \lceil \alpha/2^j \rceil$ is even$\}$ and $O = \{j \in [0, \ldots, \ell - 1] : \lceil \alpha/2^j \rceil$ is odd$\}$.

Proof. We have

$$
\begin{aligned}
(a^\top z)^\alpha \, \tau \geq p^{\alpha+1} \, b^\alpha \quad &\iff \quad \exists s \geq 0 \ : \ (a^\top z)^\alpha \, s^\alpha \geq (p^2 \, b)^\alpha, \ \ \tau \, p^{\alpha-1} \geq s^\alpha, \\
&\iff \quad \exists s \geq 0 \ : \ a^\top z \, s \geq p^2 \, b, \ \ \tau \, p^{\alpha-1} \geq s^\alpha \, .
\end{aligned} \tag{6.14}
$$

The first inequality in (6.14) is a hyperbolic constraint equivalent to

$$
\left\| \begin{pmatrix} 2pb \\ s \, b - a^\top z \end{pmatrix} \right\|_2 \leq s \, b + a^\top z
$$

and implies that $p = 0$ whenever $s = 0$. Next, we show that the second inequality in (6.14) can be decomposed into ℓ hyperbolic constraints. To this end, we observe that

$$\begin{aligned} t_j\, p^{\lceil \alpha/2^j \rceil - 1} \geq s^{\lceil \alpha/2^j \rceil} \quad &\iff \quad \exists t_{j+1} \geq 0 \ : \ p^{\lceil \alpha/2^j \rceil - 2}\, t_{j+1}^2 \geq s^{\lceil \alpha/2^j \rceil}, \ t_j\, p \geq t_{j+1}^2 \\ &\iff \quad \exists t_{j+1} \geq 0 \ : \ p^{\lceil \alpha/2^{j+1} \rceil - 1}\, t_{j+1} \geq s^{\lceil \alpha/2^{j+1} \rceil}, \ t_j\, p \geq t_{j+1}^2 \end{aligned}$$

for all $j \in E$ and $t_j \geq 0$, while

$$\begin{aligned} t_j\, p^{\lceil \alpha/2^j \rceil - 1} \geq s^{\lceil \alpha/2^j \rceil} \quad &\iff \quad \exists t_{j+1} \geq 0 \ : \ p^{\lceil \alpha/2^j \rceil - 1}\, t_{j+1}^2 \geq s^{\lceil \alpha/2^j \rceil + 1}, \ t_j\, s \geq t_{j+1}^2 \\ &\iff \quad \exists t_{j+1} \geq 0 \ : \ p^{\lceil \alpha/2^{j+1} \rceil - 1}\, t_{j+1} \geq s^{\lceil \alpha/2^{j+1} \rceil}, \ t_j\, s \geq t_{j+1}^2 \end{aligned}$$

for all $j \in O$ and $t_j \geq 0$. Applying the above equivalences iteratively for $j = 0, \ldots, \ell - 1$, we find

$$\begin{aligned} \tau\, p^{\alpha - 1} \geq s^\alpha \quad &\iff \quad \exists t \in \mathbb{R}^{\ell+1} \ : \ t \geq 0, \quad t_0 = \tau, \quad t_j\, p \geq t_{j+1}^2 \ \ \forall j \in E, \quad t_j\, s \geq t_{j+1}^2 \ \ \forall j \in O \\ &\iff \quad \exists t \in \mathbb{R}^{\ell+1} \ : \ t \geq 0, \quad t_0 = \tau, \\ &\qquad \left\| \begin{pmatrix} 2t_{j+1} \\ t_j - p \end{pmatrix} \right\|_2 \leq t_j + p \ \ \forall j \in E, \quad \left\| \begin{pmatrix} 2t_{j+1} \\ t_j - s \end{pmatrix} \right\|_2 \leq t_j + s \ \ \forall j \in O. \end{aligned}$$

The claim now follows as we can set $s = t_\ell$ without loss of generality. □

Theorem 6.4 in combination with Lemma 6.1 or 6.2 establish that the worst-case probability bound $G_\alpha(\mu, S)$ for α-unimodal probability distributions is equivalent to a tractable SDP. As discussed at the end of Section 6.1 a worst-case probability distribution $\mathbb{P}^\star$ of $G_\alpha(\mu, S) = \mathbb{P}^\star(\xi \notin \Xi)$ can be found as

$$\mathbb{P}^\star = \sum_i p_i^\star \cdot \mathbb{u}_{x_i^\star}^\alpha$$

where $(x_i^\star, p_i^\star)$ can be derived from the solution of the tractable reformulation (6.13) with the help of the procedure discussed in Section 6.2.1. The worst-case distribution $\mathbb{P}^\star$ is hence recognized as the finite convex combination of extreme α-unimodal distributions.

In order for the worst-case probability problem (6.12) to be of practical value we need to establish that its optimal value depends continuously on the distributional parameters μ and S.

Proposition 6.1 (Well-posedness of problem (6.12))**.** *The optimal value function $G_\alpha(\mu, S)$ of problem (6.12) is concave and continuous on the set of all $\mu \in \mathrm{R}^n$ and $S \in \mathrm{S}^n$ with*

$$\begin{pmatrix} \frac{\alpha+2}{\alpha} S & \frac{\alpha+1}{\alpha} \mu \\ \frac{\alpha+1}{\alpha} \mu^\top & 1 \end{pmatrix} \succ 0\,.$$

Proof. Concavity of $G_\alpha(\mu, S)$ is a direct consequence of Rockafellar [109, Proposition 2.22] and Theorem 6.4. Because of Proposition 2.1 the bound $G_\alpha(\mu, S)$ is thus continuous on the interior of its domain. The claim now follows from the characterization of the domain of $G_\alpha(\mu, S)$ in Fact 6.2. □

6.3.2 The Chebyshev bound

To derive the classical Chebyshev bound (1.6), assume without loss of generality that $\mu = 0$, and define the confidence region Ξ as

$$\Xi = \{x \in \mathrm{R} \ : \ -x < \kappa\sigma, \quad x < \kappa\sigma\}\,.$$

The worst-case probability of the event $\xi \notin \Xi$ then coincides with the optimal value of the SDP of Theorem 6.3 and its dual, which are given by

$$
\begin{array}{llcll}
\max & \sum_{i=1}^{2} p_i & = & \min & \mathrm{Tr}\left\{ \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} \begin{pmatrix} P & q \\ q^\top & r \end{pmatrix} \right\} \\
\text{s.t.} & p_i, z_i, Z_i \in \mathrm{R} \;\; \forall i \in \{1,2\} & & \text{s.t.} & P, q, r \in \mathrm{R},\; \tau_i \in \mathrm{R} \;\; \forall i \in \{1,2\} \\
& a_i^\top z_i \geq b_i p_i \;\; \forall i \in \{1,2\} & & & \tau_i \geq 0 \;\; \forall i \in \{1,2\} \\
& \sum_{i=1}^{2} \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \preceq \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} & & & \begin{pmatrix} P & q \\ q^\top & r-1 \end{pmatrix} \succeq \tau_i \begin{pmatrix} 0 & \frac{a_i}{2} \\ \frac{a_i^\top}{2} & -b_i \end{pmatrix} \;\; \forall i \in \{1,2\} \\
& \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \succeq 0 \;\; \forall i \in \{1,2\} & & & \begin{pmatrix} P & q \\ q^\top & r \end{pmatrix} \succeq 0.
\end{array}
$$

A pair of optimal primal and dual solutions is provided in the following table. Note that the dual solution serves as a certificate of optimality for the primal solution.

Primal solution	**Dual solution**
$p_1 = p_2 = \begin{cases} \frac{1}{2\kappa^2} & \text{if } \kappa > 1 \\ \frac{1}{2} & \text{otherwise} \end{cases}$	$P = \begin{cases} \frac{1}{\sigma^2\kappa^2} & \text{if } \kappa > 1 \\ 0 & \text{otherwise} \end{cases}$
$z_1 = -z_2 = \begin{cases} \frac{\sigma}{2} & \text{if } \kappa > 1 \\ \frac{\kappa\sigma}{2} & \text{otherwise} \end{cases}$	$q = 0$ $r = \begin{cases} 0 & \text{if } \kappa > 1 \\ 1 & \text{otherwise} \end{cases}$
$Z_1 = Z_2 = \begin{cases} \frac{\sigma^2}{2} & \text{if } \kappa > 1 \\ \frac{\kappa^2\sigma^2}{2} & \text{otherwise} \end{cases}$	$\tau_1 = \tau_2 = \begin{cases} \frac{2}{\sigma\kappa} & \text{if } \kappa > 1 \\ 0 & \text{otherwise} \end{cases}$

The worst-case probability is thus given by $p_1 + p_2 = \min\left(\frac{1}{\kappa^2}, 1\right)$. Hence, Theorem 6.3 is a generalization of the classical Chebyshev bound (1.6).

In the remainder of this section we will formalize our intuition that the generalized Chebyshev bound $G_\infty(\mu, S)$ constitutes a special case of the α-unimodal bound $G_\alpha(\mu, S)$ when α tends to infinity. The next proposition establishes well-posedness of the Chebyshev bound, which is needed to prove this asymptotic result.

Proposition 6.2 (Well-posedness of the generalized Chebyshev bound)**.** *The value function $G_\infty(\mu, S)$ of problem (6.9) is concave and continuous on the set of all $\mu \in \mathrm{R}^n$ and $S \in \mathrm{S}^n$ with*

$$\begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} \succ 0\,.$$

Proof. The proof largely parallels that of Proposition 6.1 and is omitted. □

Note that the function $G_\infty(\mu, S)$ can be discontinuous on the boundary of its domain when the variance matrix $S - \mu\mu^\top$ is positive semidefinite but has at least one zero eigenvalue. Since the confidence region Ξ constitutes an open polytope, there exists a converging sequence $(x_i)_{i\in\mathrm{N}}$ in Ξ whose limit $x = \lim_{i\to\infty} x_i$ is not contained in Ξ. Defining $\mu_i = x_i$ and $S_i = x_i x_i^\top$ for all $i \in \mathrm{N}$, it is clear that $\lim_{i\to\infty}(\mu_i, S_i) = (x, xx^\top)$. Since $\mathcal{C}(\mu_i, S_i) = \{\delta_{x_i}\}$ and $x_i \in \Xi$ for all $i \in \mathrm{N}$, we conclude that $\lim_{i\to\infty} G_\infty(\mu_i, S_i) = 0$. However, we also have $G_\infty(x, xx^\top) = 1$ because $\mathcal{C}(\mu, S) = \{\delta_x\}$ and $x \notin \Xi$. Thus, $G_\infty(\mu, S)$ is discontinuous at $(x, xx^\top)$.

We are now ready to prove that the Chebyshev bound $G_\infty(\mu, S)$ is *de facto* embedded into the family of all α-unimodal bounds $G_\alpha(\mu, S)$ for $\alpha > 0$.

Proposition 6.3 (Embedding of the Chebyshev bound)**.** *For any $\mu \in \mathrm{R}^n$ and $S \in \mathrm{S}^n$ with $S \succ \mu\mu^\top$ we have $\lim_{\alpha\to\infty} G_\alpha(\mu, S) = G_\infty(\mu, S)$.*

Proof. Select any $\mu \in \mathrm{R}^n$ and $S \in \mathrm{S}^n$ with $S \succ \mu\mu^\top$. It is clear that $\lim_{\alpha\to\infty} G_\alpha(\mu, S) \leq G_\infty(\mu, S)$ as (6.12) constitutes a restriction of (6.9) for all $\alpha > 0$. In order to prove the converse inequality, we need the following relation between the extreme distributions of $\mathcal{U}_\alpha$ and $\mathcal{P}_n$:

$$\left(1 - \tfrac{1}{\alpha}\right) \delta_{x/\alpha^{\frac{1}{\alpha}}}(\xi \notin \Xi) \;\leq\; \mathbb{u}_x^\alpha(\xi \notin \Xi) \tag{6.15}$$

For $x/\alpha^{\frac{1}{\alpha}} \in \Xi$ the left hand side vanishes and (6.15) is trivially satisfied. For $x/\alpha^{\frac{1}{\alpha}} \notin \Xi$ and Ξ star-shaped, an elementary calculation shows that $\mathbb{u}_x^\alpha(\xi \notin \Xi) \geq \mathbb{u}_x^\alpha([x/\alpha^{\frac{1}{\alpha}}, x]) = 1 - \frac{1}{\alpha}$ for all $\alpha \geq 1$. Thus, (6.15) holds because $\delta_{x/\alpha^{\frac{1}{\alpha}}}(\xi \notin \Xi) \leq 1$. Taking mixtures with $\mathbb{m} \in \mathcal{C}\left(\frac{\alpha+1}{\alpha}\mu, \frac{\alpha+2}{\alpha}S\right)$ on both sides of (6.15) yields

$$\left(1 - \frac{1}{\alpha}\right) \int \delta_{x/\alpha^{\frac{1}{\alpha}}}(\xi \notin \Xi)\, \mathbb{m}(\mathrm{d}x) \;\leq\; \int \mathbb{u}_x^\alpha(\xi \notin \Xi)\, \mathbb{m}(\mathrm{d}x)\,. \tag{6.16}$$

By using elementary manipulations one can show that

$$\int \delta_{x/\alpha^{\frac{1}{\alpha}}}(\cdot)\, \mathbb{m}(\mathrm{d}x) \in \mathcal{C}\left(\alpha^{-\frac{1}{\alpha}} \tfrac{\alpha+1}{\alpha}\mu, \alpha^{-\frac{2}{\alpha}} \tfrac{\alpha+2}{\alpha} S\right) \quad \text{and} \quad \int \mathbb{u}_x^\alpha(\cdot)\, \mathbb{m}(\mathrm{d}x) \in \mathcal{U}_\alpha(\mu, S)\,.$$

Maximizing both sides of (6.16) over all mixture distributions $\mathbb{m} \in \mathcal{C}\left(\frac{\alpha+1}{\alpha}\mu, \frac{\alpha+2}{\alpha}S\right)$ thus yields

$$\left(1 - \tfrac{1}{\alpha}\right) G_\infty\left(\alpha^{-\frac{1}{\alpha}} \tfrac{\alpha+1}{\alpha}\mu, \alpha^{-\frac{2}{\alpha}} \tfrac{\alpha+2}{\alpha} S\right) \leq G_\alpha(\mu, S)\,.$$

Since $\lim_{\alpha\to\infty} \alpha^{-\frac{1}{\alpha}} \frac{\alpha+1}{\alpha} = \lim_{\alpha\to\infty} \alpha^{-\frac{2}{\alpha}} \frac{\alpha+1}{\alpha} = 1$ and $G_\infty(\mu, S)$ is continuous whenever $S \succ \mu\mu^\top$ (see Proposition 6.2), we conclude that $G_\infty(\mu, S) \leq \lim_{\alpha\to\infty} G_\alpha(\mu, S)$. □

In addition to generalizing the univariate Chebyshev bound (1.6), the multivariate Chebyshev bound $G_\infty(\mu, S)$ also generalizes Cantelli's classical one sided inequality and the bivariate Birnbaum-Raymond-Zuckerman inequality [23]. By virtue of Proposition 6.3, all of these classical inequalities can now be seen as special instances of the general problem (6.12).

6.3.3 The Gauss bound

Following the discussion in Section 3.1, the generalized Gauss bound for star-unimodal distributions can be defined as

$$G_n(\mu, S) = \sup_{\mathbb{P} \in \mathcal{C}_n(\mu, S)} \mathbb{P}(\xi \notin \Xi)\,. \tag{6.17}$$

From Definition 3.5 and the subsequent discussion we indeed know that the set of n-unimodal distributions $\mathcal{U}_n$ coincides with the set of star unimodal distributions; see also Dharmadhikari and Joag-Dev [44]. Theorem 6.5 is actually a straightforward corollary of Theorem 6.4 in conjunction with Lemma 6.2.

Theorem 6.5 (Generalized Gauss bounds)**.** *If Ξ is a polytope of the form* (6.10) *with $0 \in \Xi$, the worst-case probability problem* (6.17) *is equivalent to a tractable SDP,*

$$
\begin{array}{rcll}
\sup\limits_{\mathbb{P}\in\mathcal{C}_n(\mu,S)} \mathbb{P}(\xi \notin \Xi) & = & \max & \sum\limits_{i\in I}(p_i - t_{i,0}) \\
 & & \text{s.t.} & z_i \in \mathrm{R}^n,\ Z_i \in \mathrm{S}^n,\ p_i \in \mathrm{R},\ t_i \in \mathrm{R}^{\ell+1} \quad \forall i \in I \\
 & & & \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \succeq 0, \quad a_i^\top z_i \geq 0, \quad t_i \geq 0 \quad \forall i \in I \\
 & & & \sum\limits_{i\in I} \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \preceq \begin{pmatrix} \frac{n+2}{n}S & \frac{n+1}{n}\mu \\ \frac{n+1}{n}\mu^\top & 1 \end{pmatrix} \\
 & & & \left\| \begin{pmatrix} 2p_i b_i \\ t_{i,\ell} b_i - a_i^\top z_i \end{pmatrix} \right\|_2 \leq t_{i,\ell} b_i + a_i^\top z_i \quad \forall i \in I \\
 & & & \left\| \begin{pmatrix} 2t_{i,j+1} \\ t_{i,j} - p_i \end{pmatrix} \right\|_2 \leq t_{i,j} + p_i \quad \forall j \in E, \quad \forall i \in I \\
 & & & \left\| \begin{pmatrix} 2t_{i,j+1} \\ t_{i,j} - t_{i,\ell} \end{pmatrix} \right\|_2 \leq t_{i,j} + t_{i,\ell} \quad \forall j \in O, \quad \forall i \in I,
\end{array}
\tag{6.18}
$$

where $\ell = \lceil \log_2 n \rceil$, $E = \{j \in [0,\ldots,\ell-1] : \lceil n/2^j \rceil$ is even$\}$ and $O = \{j \in [0,\ldots,\ell-1] : \lceil n/2^j \rceil$ is odd$\}$.

Theorem 6.5 establishes that the Gauss bound $G_n(\mu, S)$ for star unimodal probability distributions is equivalent to a tractable SDP. As discussed at the end of Section 6.1, a worst-case probability distribution $\mathbb{P}^\star$ of $G_n(\mu, S) = \mathbb{P}^\star(\xi \notin \Xi)$ can be found as

$$\mathbb{P}^\star = \sum_i p_i^\star \cdot \mathbb{u}^n_{x_i^\star}$$

where $(x_i^\star, p_i^\star)$ can be derived from the solution of the tractable reformulation (6.18) at no additional computational cost with the help of the procedure discussed in Section 6.2.1. The worst-case distribution $\mathbb{P}^\star$ is hence recognized as the finite convex combination of extreme star unimodal distributions.

We demonstrate now that the classical Gauss bound (1.7) arises indeed as a special case of Theorem 6.5.

Example 6.1 (Classical Gauss bound)**.** *To derive the classical Gauss bound* (1.7)*, assume without loss of generality that $\mu = 0$, and define the confidence region Ξ as*

$$\Xi = \{x \in \mathrm{R} \ : \ -x < \kappa\sigma, \quad x < \kappa\sigma\}.$$

The worst-case probability of the event $\xi \notin \Xi$ then coincides with the optimal value of the SDP of Theorem 6.5 and its dual, which are given by

$$
\begin{array}{llcll}
\max & \sum\limits_{i=1}^{2} p_i - \tau_i & = & \min & \mathrm{Tr}\left\{ \begin{pmatrix} 3S & 2\mu \\ 2\mu^\top & 1 \end{pmatrix} \begin{pmatrix} P & q \\ q^\top & r \end{pmatrix} \right\} \\
\text{s.t.} & p_i, \tau_i, z_i, Z_i \in \mathrm{R} \ \ \forall i \in \{1,2\} & & \text{s.t.} & P, q, r \in \mathrm{R}, \ \Lambda_i \in \mathrm{S}^2 \ \ \forall i \in \{1,2\} \\
 & \begin{pmatrix} \tau_i b_i & p_i b_i \\ p_i b_i & a_i^\top z_i \end{pmatrix} \succeq 0 \ \ \forall i \in \{1,2\} & & & \begin{pmatrix} P & q \\ q^\top & r-1 \end{pmatrix} \succeq \begin{pmatrix} 0 & \frac{a_i}{2}\Lambda_{i,2,2} \\ \frac{a_i^\top}{2}\Lambda_{i,2,2} & 2b_i\Lambda_{i,1,2} \end{pmatrix} \\
 & \sum\limits_{i=1}^{2} \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \preceq \begin{pmatrix} 3S & 2\mu \\ 2\mu^\top & 1 \end{pmatrix} & & & \begin{pmatrix} P & q \\ q^\top & r \end{pmatrix} \succeq 0 \\
 & \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \succeq 0 \ \ \forall i \in \{1,2\} & & & b_i\Lambda_{i,1,1} \leq 1, \ \ \Lambda_i \succeq 0 \ \ \forall i \in \{1,2\}.
\end{array}
$$

A pair of optimal primal and dual solutions is provided in the following table. Note that the dual solution serves as a certificate of optimality for the primal solution.

Primal Solution	***Dual Solution***
$p_1 = p_2 = \begin{cases} \frac{2}{3\kappa^2} & \text{if } \kappa > \frac{2}{\sqrt{3}} \\ \frac{1}{2} & \text{otherwise} \end{cases}$	$P = \begin{cases} \frac{4}{27\sigma^2\kappa^2} & \text{if } \kappa > \frac{2}{\sqrt{3}} \\ \frac{\kappa}{6\sqrt{3}\sigma^2} & \text{otherwise} \end{cases}$
	$q = 0$
$\tau_1 = \tau_2 = \begin{cases} \frac{4}{9\kappa^2} & \text{if } \kappa > \frac{2}{\sqrt{3}} \\ \frac{\kappa}{2\sqrt{3}} & \text{otherwise} \end{cases}$	$r = \begin{cases} 0 & \text{if } \kappa > \frac{2}{\sqrt{3}} \\ 1 - \frac{\sqrt{3}\kappa}{2} & \text{otherwise} \end{cases}$
$z_1 = -z_2 = \begin{cases} \frac{\sigma}{\kappa} & \text{if } \kappa > \frac{2}{\sqrt{3}} \\ \frac{\sqrt{3}\sigma}{2} & \text{otherwise} \end{cases}$	$\Lambda_1 = \Lambda_2 = \begin{cases} \frac{1}{\sigma\kappa}\begin{pmatrix} 1 & -\frac{2}{3} \\ -\frac{2}{3} & \frac{4}{9} \end{pmatrix} & \text{if } \kappa > \frac{2}{\sqrt{3}} \\ \frac{1}{\sigma\kappa}\begin{pmatrix} 1 & -\frac{\kappa}{\sqrt{3}} \\ -\frac{\kappa}{\sqrt{3}} & \frac{\kappa^2}{3} \end{pmatrix} & \text{otherwise} \end{cases}$
$Z_1 = Z_2 = \frac{3\sigma^2}{2}$	

The worst-case probability is thus given by $(p_1 - \tau_1) + (p_2 - \tau_2) = \frac{4}{9\kappa^2}$ *when* $\kappa > \frac{2}{\sqrt{3}}$*;* $= 1 - \frac{\kappa}{\sqrt{3}}$ *otherwise. Hence, Theorem 6.5 is a generalization of the classical Gauss bound* (1.7)*.*

In the univariate case, the Gauss bound tightens the Chebyshev bound by a factor of 4/9. However, the tightening offered by unimodality is less pronounced in higher dimensions since $G_n(\mu, S)$ converges to the Chebyshev bound $G_\infty(\mu, S)$ when n tends to infinity. To make this point more concrete, we now provide *analytic* probability inequalities for the quantity of interest $\mathbb{P}(\max_i\ |\xi_i - \mu_i| \geq \kappa\sigma)$, where it is only known that $\mathbb{P} \in \mathcal{U}_\alpha$, $\mathbf{E}_\mathbb{P}[\xi] = \mu = c$ and $\mathbf{E}_\mathbb{P}\left[\xi\xi^\top\right] = \sigma^2/n\ \mathbb{I}_n + \mu\mu^\top$. This problem can be seen to constitute a multivariate generalization of the classical Chebyshev (1.6) and Gauss (1.7) inequalities, but is itself a particular case of the general problem (6.12). We have indeed the tight inequality $\mathbb{P}(\max_i\ |\xi_i - \mu_i| \geq \kappa\sigma) \leq \sup_{\mathbb{P}\in\mathcal{U}_\alpha(0,\mathbb{I}_n/n)} \mathbb{P}(\max_i\ |\xi_i| \geq \kappa)$. This last particular instance of the problem (6.12) however admits an analytic solution which can be proven analogously to the results in Section 5.4.1 as done by Stellato [121].

Lemma 6.3. *Let ξ be an α-unimodal random variable with centre $c = \mu$, then we have the tight bilateral probability inequality*

$$\mathbb{P}(\max_i\ |\xi_i - \mu_i| \geq \kappa\sigma) \leq \begin{cases} \left(\frac{2}{\alpha+2}\right)^{\frac{2}{\alpha}} \frac{1}{\kappa^2} & \text{if } \kappa > \left(\frac{2}{\alpha+2}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha+2}{\alpha}\right)^{\frac{1}{2}}, \\ 1 - \left(\frac{\alpha}{\alpha+2}\right)^{\frac{\alpha}{2}} \kappa^\alpha & \text{otherwise.} \end{cases}$$

The corresponding Chebyshev bound can be obtained by letting $\alpha \to \infty$ and yields $\mathbb{P}(\max_i\ |\xi_i - \mu_i| \geq \kappa\sigma) \leq 1/\kappa^2$ if $\kappa > 1$; $= 1$ otherwise. When compared to the corresponding Gauss bound found as a particular case $\alpha = n$ of Lemma 6.3, we see that the factor by which the Chebyshev bound is improved upon is indeed $4/9 = (2/(n+2))^{2/n}$, $n = 1$ for univariate problems. However, for higher dimensional problems the returns diminish as

$$\left(\frac{2}{n+2}\right)^{\frac{2}{n}} = 1 - \frac{2\log\left(\frac{n}{2}\right)}{n} + \mathcal{O}\left(\frac{1}{n^2}\right), \quad \text{and} \quad \lim_{n\to\infty}\left(\frac{2}{n+2}\right)^{\frac{2}{n}} = 1.$$

An intuitive explanation for this seemingly surprising result follows from the observation that most of the volume of a high-dimensional star-shaped set is concentrated in a thin layer near its surface. Thus, the radial distributions $\mathbb{u}_x^n$ converge weakly to the Dirac distributions δ_x as n grows, which implies that all probability distributions are approximately star-unimodal in high dimensions.

6.4 Immediate extensions

We now demonstrate that the results presented in this chapter can be used to solve a wider range of diverse worst-case probability problems as well. The corresponding ambiguity sets are more general than $\mathcal{C}_\alpha(\mu, S)$, which contains merely α-unimodal distributions with precisely known first and second-order moments.

Moment ambiguity The worst-case probability bound $G_\alpha(\mu, S)$ requires full and accurate information about the mean and second moment matrix of the random vector ξ. In practice, however, these statistics must typically be estimated from noisy historical data and are therefore themselves subject to ambiguity. Assume therefore that the first and second-order moments are known only to belong to an SDP-representable confidence set M, that is as in Nesterov and Nemirovski [90] we assume,

$$\begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} \in M \subseteq \mathrm{S}_+^{n+1}. \tag{6.19}$$

Then, the worst-case probability problem with ambiguity set $\mathcal{C} = \cup_{(\mu,S)} \mathcal{C}_\alpha(\mu, S)$, where the union is taken over all μ and S satisfying condition (6.19) can be reformulated as a tractable SDP. This is an immediate consequence of the identity

$$\sup_{\mathbb{P}\in\mathcal{C}} \ \mathbb{P}(\xi \notin \Xi) \ = \ \begin{array}{ll} \max\limits_{\mu,S} & G_\alpha(\mu, S) \\ \text{s.t.} & \left(\begin{smallmatrix} S & \mu \\ \mu^\top & 1 \end{smallmatrix}\right) \in M. \end{array}$$

Theorem 6.4 and its corollaries imply that the optimization problem on the right hand side of the above expression admits a tractable SDP reformulation. Hence, the α-unimodal bound $G_\alpha(\mu, S)$ can be generalized to handle ambiguity in both mean and the second-moment matrix.

Support information Suppose that, in addition to being α-unimodal, the distribution of ξ is known to be supported on a convex closed polytope representable as

$$B = \left\{x \in \mathrm{R}^n \ : \ c_i^\top x - d_i \leq 0 \quad \forall j \in J\right\}. \tag{6.20}$$

In this case we could use an ambiguity set of the form $\mathcal{C} = \mathcal{C}_\alpha(\mu, S) \cap \mathcal{P}(B)$, where $\mathcal{P}(B) = \{\mathbb{P} \in \mathcal{P}_n \ : \ \mathbb{P}(B) = 1\}$. Unfortunately, even checking whether the ambiguity set $\mathcal{C}$ is non-empty is NP-hard in general. Indeed, if S is given by the nonnegative orthant R_+^n, it can be seen from the Choquet representation of $\mathcal{U}_\alpha$ that checking whether $\mathcal{C}$ is non-empty is equivalent to checking whether the matrix

$$\begin{pmatrix} \frac{\alpha+2}{\alpha} S & \frac{\alpha+1}{\alpha}\mu \\ \frac{\alpha+1}{\alpha}\mu^\top & 1 \end{pmatrix}$$

is completely positive, which is hard for any $\alpha > 0$. A tractable alternative ambiguity set is given by $\mathcal{C} = \cup_{S' \preceq S} \mathcal{C}_\alpha(\mu, S') \cap \mathcal{P}(B)$. The resulting generalized moment problem treats S as an upper bound (in the positive semidefinite sense) on the second-order moment matrix of ξ, as suggested by Delage and Ye [43]. This relaxation is justified by the observation that the worst-case distribution in (6.9) tends to be maximally spread out and thus typically attains the upper bound imposed on its second-order moments. In all other cases, the relaxation results in a conservative estimate for the worst-case probability of the event $\xi \notin \Xi$. However, the relaxed problem always admits an exact reformulation as an SDP. This result is formalized in the following proposition. The proof is omitted because it requires no new ideas.

Proposition 6.4 (Support information). *The worst-case probability problem* (6.9) *with ambiguity set $\mathcal{C} = \cup_{S' \preceq S} \mathcal{C}_\alpha(\mu, S') \cap \mathcal{P}(B)$, where and Ξ and B are defined as in* (6.10) *and* (6.20), *respectively, can be reformulated as*

$$\begin{array}{ll} \max & \sum_{i \in I} (p_i - \tau_i) \\ \text{s.t.} & z_i \in \mathrm{R}^n,\ Z_i \in \mathrm{S}^n,\ p_i \in \mathrm{R},\ \tau_i \in \mathrm{R} \quad \forall i \in I \\ & \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \succeq 0 \quad \forall i \in I \\ & c_j^\top z_i \leq d_j p_i \quad \forall i \in I,\ \forall j \in J \\ & \sum_{i \in I} \begin{pmatrix} Z_i & z_i \\ z_i^\top & p_i \end{pmatrix} \preceq \begin{pmatrix} \frac{\alpha+2}{\alpha} S & \frac{\alpha+1}{\alpha} \mu \\ \frac{\alpha+1}{\alpha} \mu^\top & 1 \end{pmatrix} \\ & \left(a_i^\top z_i\right)^\alpha \tau_i \geq p_i^{\alpha+1} b_i^\alpha, \quad a_i^\top z_i \geq 0, \quad \tau_i \geq 0 \ \ \forall i \in I, \end{array}$$

which is equivalent to a tractable SDP for any rational $\alpha \geq 1$.

We remark that our techniques can be used to derive many other worst-case probability bounds involving α-unimodal distributions for instance by combining the results discussed in this section. These further generalizations are omitted for the sake of brevity, and we consider instead a practical application of the new worst-case probability bounds derived in this chapter.

6.5 Digital communication example

We use the generalized Gauss bounds presented in this chapter to estimate the probability of correct signal detection in a digital communication example inspired by Boyd and Vandenberghe [27]. All SDP problems are implemented in `Matlab` via the `YALMIP` interface and solved using `SDPT3`.

Consider a set of c possible symbols or signals $\mathfrak{S} = \{s_1, \ldots, s_c\} \subseteq \mathrm{R}^2$, which is termed the signal constellation. The signals are transmitted over a noisy communication channel and perturbed by additive noise. A transmitted signal s_t thus results in an output $s_\mathrm{o} = s_\mathrm{t} + \xi$, where ξ valued in R^2 follows a star-unimodal distribution with zero mean and variance matrix $\sigma^2 \mathbb{I}_2$. A *minimum distance detector*[1] then decodes the output, that is, it determines the symbol $s_\mathrm{r} \in \mathfrak{S}$ that is closest in Euclidian distance to the output s_o. Note that the detector is uniquely defined by the Voronoi diagram implied by the signal constellation $\mathfrak{S}$ as shown in Figure 6.3(a).

The quantity of interest is the average probability of correct symbol transmission

$$p = \frac{1}{c} \sum_{i=1}^{c} \mathbb{P}(s_i + \xi \in C_i) = 1 - \frac{1}{c} \sum_{i=1}^{c} \mathbb{P}(s_i + \xi \notin C_i),$$

where C_i is the (polytopic) set of outputs that are decoded as s_i. The generalized Chebyshev bound of Theorem 6.3 and the generalized Gauss bound of Theorem 6.5 both provide efficiently computable lower bounds on p, which are plotted in Figure 6.3(b) as a function of the Channel noise power σ. Note that the generalized Gauss bound is substantially tighter because the Chebyshev bound disregards the star-unimodality of the channel noise. For the sake of comparison, Figure 6.3(b) also shows the probability of correct detection when the noise ξ is assumed to be normal or block uniformly distributed. We say that a random variable is block uniformly distributed if it is distributed uniformly on a square in R^n. The both reference probabilities were computed using numerical integration. Using the procedure described in Section 6.2.1, we are able to explicitly construct a worst-case probability distribution for $\mathbb{P}(s_1 + \xi \notin C_1)$. In Figures 6.4(a) and 6.4(b) the support of these worst-case probability distributions for respectively the

[1]If the noise is Gaussian, then minimum distance decoding is the same as maximum likelihood decoding.

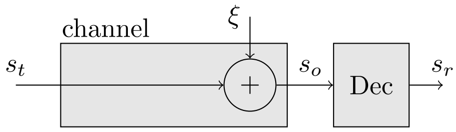
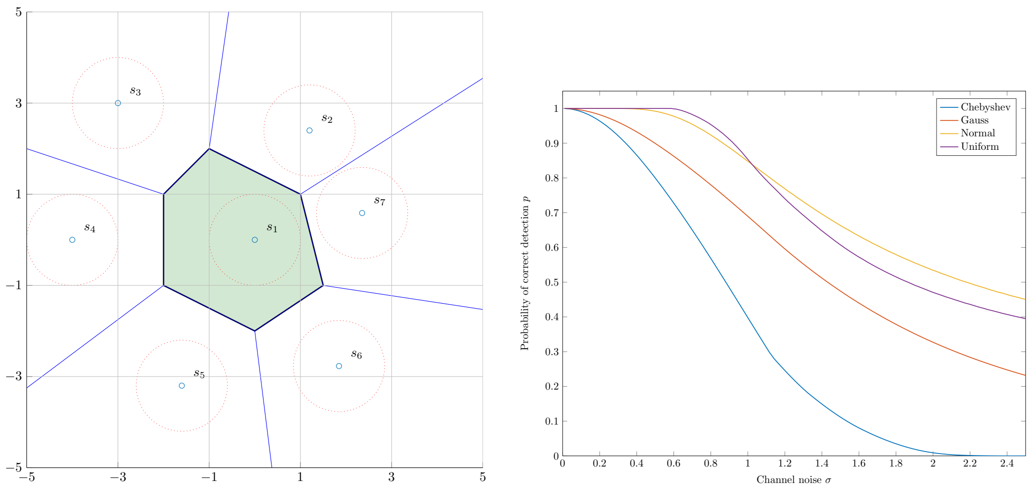


Figure 6.2: Upon transmitting the symbol s_{t}, a noisy output $s_{\mathrm{o}} = s_{\mathrm{t}} + \xi$ is received and decoded using a maximum likelihood decoder into the symbol s_{r}.

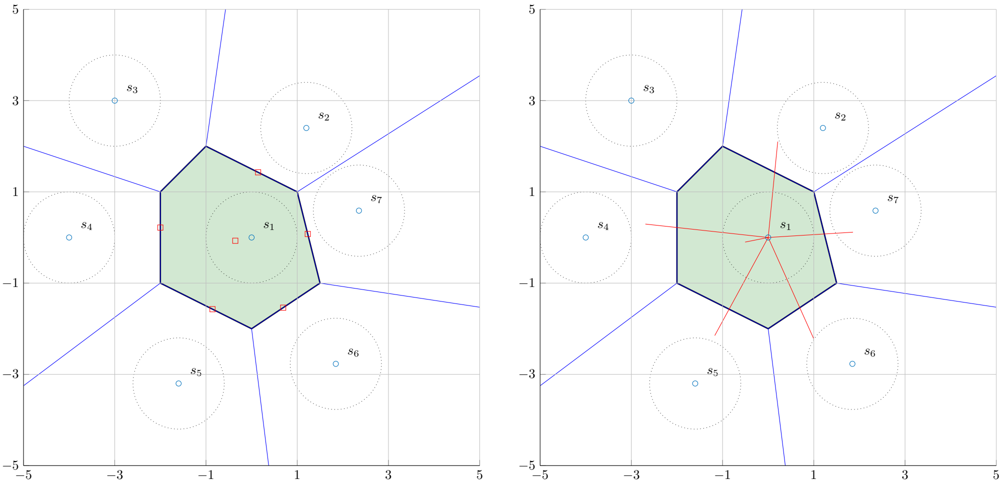


(a) (b)

Figure 6.3: Figure 6.3(a) depicts the signal constellation $\mathfrak{S}$. The probability distribution of the outputs is visualized by the dashed circles, while the detector is visualized by its Voronoi diagram. For example, the green polytope represents the set of outputs s_{o} which are decoded as s_1. Figure 6.3(b) shows the lower bounds on the correct detection probabilities as predicted by the Chebyshev and Gauss inequalities. The exact detection probability for normal and block uniform distributed noise is shown for the sake of comparison.

Chebyshev and Gauss bound are shown in case the channel noise is $\sigma = 1$. We used `Matlab` on a PC[2] operated by `Debian GNU/Linux 7 (wheezy)` in combination with the software `YALMIP` made available by Löfberg [80] and `SDPT3` described in Tütüncü [127] to solve the resulting convex SDPs. The Chebyshev and Gauss bounds depicted in Figure 6.3(b) each required the solution of 700 SDPs, with seven SDPs per channel noise level. On our computing hardware it took on average 1.0 s and 2.0 s to solve each of the 700 SDPs for the Chebyshev and Gauss bounds, respectively, to an accuracy of six significant digits.

[2]An `Intel(R) Core(TM) Xeon(R) CPU E5540 @ 2.53GHz` machine.



(a) (b)

Figure 6.4: Figures 6.4(a) and 6.4(b) depict in red the support of the worst-case probability distributions for $\mathbb{P}(s_1 + \xi \notin C_1)$ for the Chebyshev and Gauss bounds, respectively, in case of the channel noise power $\sigma = 1$.

7 Dual uncertainty quantification with second-order moment information

This chapter will deal with uncertainty quantification problems with second-order moment information of the type (6.2) as discussed in the previous chapter but now from the dual perspective. Instead of maximizing over probability distributions in an ambiguity set $\mathcal{C}$, the dual is stated in terms of minimization over dual functions.

We remark that the dual approach described in this chapter is the predominant approach in the literature concerned with the uncertainty quantification problem (6.2). Indeed, the early work by Isii [63] on uncertainty quantification problems already emphasizes the dual perspective. Furthermore, the recent pioneering results of Bertsimas and Popescu [18, 103] which spurred a renewed interest in the uncertainty quantification problem (6.2) are essentially all dual based. The problem (6.2) is in the aforementioned literature also referred to as a generalized moment problem. As discussed in the previous chapter we will not follow this convention and refer to problem (6.2) as an uncertainty quantification problem with second-order moment information.

The dual of the uncertainty quantification problem (6.2) defining the worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$ is described in Chapter 5. In case of second-order moment information the dual (5.4) specializes to a minimization problem over the coefficients of quadratic dual functions, i.e.

$$\begin{aligned}
\inf \quad & \mathrm{Tr}\left\{ \begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \cdot \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix} \right\} \\
\text{s.t.} \quad & Y \in \mathrm{S}^n,\ y \in \mathrm{R}^n,\ y_0 \in \mathrm{R} \\
& \int \left[x^\top Y x + 2x^\top y + y_0 - L(x) \right] \mathbb{P}(\mathrm{d}x) \geq 0, \quad \forall \mathbb{P} \in \mathcal{K}.
\end{aligned} \tag{7.1}$$

Note that the convex optimization problem (7.1) has finitely many dual variables (Y, y, y_0) but infinitely many constraints parameterized in the universal quantifier $\mathbb{P}$ in the ambiguity set $\mathcal{K}$. As in the previous chapter, we will show that if the ambiguity set $\mathcal{K}$ admits a Choquet star representation then the dual (7.1) can often be solved exactly.

7.1 Second-moment information and Choquet simplices

If the ambiguity set $\mathcal{K}$ admits a Choquet star representation generated by the univariate probability distribution $\mathbb{T}$, then the constraint of the dual problem (7.1) reduces to

$$\int_0^\infty \left[t^2 x^\top Y x + 2t x^\top y + y_0 - L(xt)\right] \mathbb{T}(\mathrm{d}t) \geq 0 \quad \forall x \in \mathrm{R}^n \tag{7.2}$$

as shown in Proposition 5.2. Notice that the constraint (7.2) is a semi-infinite constraint in the universal quantifier x in R^n. The representation (7.2) of the dual constraint is favorable over its counterpart in (7.1) which is parametrized in the probability distributions $\mathbb{P}$ in the ambiguity set $\mathcal{K}$. Indeed, in general there can be many more probability distributions in the ambiguity set $\mathcal{K}$ than that there are points in R^n.

It turns out that the parametric integral in (7.2) evaluates to a piecewise polynomial in x for many natural choices of the loss function L and the ambiguity set $\mathcal{K}$. In this case the constraint of the dual problem (7.1) requires a piecewise polynomial to be non-negative on R^n and can thus be reformulated as an LMI or approximated by a hierarchy of increasingly tight LMIs by using sum-of-squares techniques discussed by Lasserre [76]. The dual problem (7.1) can thus be approximated systematically with tractable SDPs. Popescu [103] has used this general approach to derive efficiently computable, albeit approximate, Chebyshev and Gauss-type bounds for several structured classes of probability distributions. In this work however, we will indicate that in many situations an exact reformulation can be obtained too.

It is worth mentioning that in case no additional structure is imposed ($\mathcal{K} = \mathcal{P}_n$), the constraint (7.2) reduces to the standard semi-infinite constraint

$$x^\top Y x + 2x^\top y + y_0 - L(x) \geq 0 \quad \forall x \in \mathrm{R}^n \tag{7.3}$$

as discussed before in Section 5.2. This observation was used by Vandenberghe et al. [133] and Zymler et al. [143] to derive exact and tractable reformulations of Chebyshev type bounds based on the S-Lemma 4.1. As Vandenberghe et al. [133] remarked themselves, the resulting bounds are rather pessimistic as many distributions in practice do enjoy additional structure. As mentioned already, we are primarily interested in the situation in which the ambiguity set $\mathcal{K}$ is more richly structured $\mathcal{K} \subset \mathcal{P}_n$ so as to exclude pathological distributions which can make the corresponding bound overly pessimistic.

In order to equate the worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$ with its dual characterization (7.1), strong duality needs to hold. As discussed in Theorem 5.2, strong duality requires that the feasible set $\mathcal{H}(\mu, S) \cap \mathcal{K}$ satisfies a Slater constraint qualification condition. In view of Example 5.1, we make the following standing assumption in this chapter.

Assumption 7.1 (Strong duality). *To guarantee strong duality between the optimization problems* (6.2) *and* (7.1)*, we assume that the mean vector* $\mu \in \mathrm{R}^n$ *and second moment matrix* $S \in \mathrm{S}^n_+$ *satisfy*

$$\begin{pmatrix} S_s & \mu_s \\ \mu_s^\top & 1 \end{pmatrix} \succ 0 \iff S_s \succ \mu_s \mu_s^\top$$

for $S_s \cdot \int_0^\infty t^2 \, \mathbb{T}(\mathrm{d}t) = S$ *and* $\mu_s \cdot \int_0^\infty t \, \mathbb{T}(\mathrm{d}t) = \mu$.

Fact 6.1 indicates that the feasible set $\mathcal{H}(\mu, S) \cap \mathcal{K}$ is non-trivial if $S_s \succeq \mu_s \mu_s^\top$. Strong duality thus holds on the entire interior of the domain of the worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$. Assumption 7.1 can hence be regarded as very mild.

We will argue now from a dual perspective that if the ambiguity set $\mathcal{K}$ admits a Choquet star representation then the worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$ over the Choquet star simplex $\mathcal{K}$ can be reduced to a related worst-case expectation bound this time over the standard probability simplex $\mathcal{P}_n$, i.e.

$$B(L, \mathcal{K}, \mu, S) = B(L_s, \mathcal{P}_n, \mu_s, S_s)$$

for judiciously chosen loss function L_s, mean μ_s, and second moment S_s. The power of Choquet star representable ambiguity sets $\mathcal{K}$ lies hence in the previous statement. Indeed, the Choquet star structure of the ambiguity set $\mathcal{K}$ will allow us to restrict attention to uncertainty quantification problems over the standard probability simplex $\mathcal{P}_n$. This reduction from an arbitrary Choquet star representable ambiguity set $\mathcal{K}$ to the standard probability simplex $\mathcal{P}_n$ greatly benefits the unified exposition of the computational results on the worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$ as already pointed out in the previous chapter. In this chapter however, the resulting worst-case expectation bounds over the standard probability simplex $\mathcal{P}_n$ will be approached via their dual characterization (7.1).

The representation (7.2) of the dual constraint in the uncertainty quantification problem (7.1) implies the following Theorem which is in fact the dual counterpart of Theorem 6.1:

Theorem 7.1 (Reduction to the standard simplex $\mathcal{P}_n$)**.** *Assume that the ambiguity set $\mathcal{K}$ admits a Choquet star representation with generating distribution $\mathbb{T}$, with Assumption 7.1 we have $B(L, \mathcal{K}, \mu, S) = B(L_s, \mathcal{P}_n, \mu_s, S_s) =$*

$$\begin{aligned} \inf \quad & \mathrm{Tr}\left\{\begin{pmatrix} Y_s & y_s \\ y_s^\top & y_{s0}\end{pmatrix} \cdot \begin{pmatrix} S_s & \mu_s \\ \mu_s^\top & 1\end{pmatrix}\right\} \\ \text{s.t.} \quad & Y_s \in \mathrm{S}^n,\ y_s \in \mathrm{R}^n,\ y_{s0} \in \mathrm{R} \\ & x^\top Y_s x + 2x^\top y_s + y_{s0} - L_s(x) \geq 0, \quad \forall x \in \mathrm{R}^n \end{aligned} \tag{7.4}$$

for $L_s(x) := \int_0^\infty L(tx)\, \mathbb{T}(\mathrm{d}t)$, $S_s \cdot \int_0^\infty t^2\, \mathbb{T}(\mathrm{d}t) = S$ and $\mu_s \cdot \int_0^\infty t\, \mathbb{T}(\mathrm{d}t) = \mu$.

Proof. With Assumption 7.1 strong duality holds and we can equate $B(L, \mathcal{K}, \mu, S)$ with its dual representation (7.1). We now exploit the Choquet star property of the ambiguity set $\mathcal{K}$ to restate the constraint (7.2) as

$$\int_0^\infty t^2\, \mathbb{T}(\mathrm{d}t) \cdot x^\top Y x + \int_0^\infty t\, \mathbb{T}(\mathrm{d}t) \cdot 2x^\top y + y_0 - \int_0^\infty L(xt)\, \mathbb{T}(\mathrm{d}t), \quad \forall x \in \mathrm{R}^n.$$

With $Y_s = Y \cdot \int t^2\, \mathbb{T}(\mathrm{d}t)$, $y_s = y \int t\, \mathbb{T}(\mathrm{d}t)$ and $y_{s0} = y_0$ the theorem follows immediately after reorganizing the terms in the dual (7.1) and comparing them term by term with constraint (7.3). $\square$

The worst-case expectation bound $B(L, \mathcal{K}, \mu, S)$ over the ambiguity set $\mathcal{K}$ can thus be reduced to the dual of the worst-case expectation bound $B(L_s, \mathcal{P}_n, \mu_s, S_s)$ over the standard probability simplex $\mathcal{P}_n$. Theorem 7.1 can hence be seen to provide a dual counterpart to Theorem 6.1. As Theorem 7.1 derives the equivalence between the worst-case expectation bounds $B(L, \mathcal{K}, \mu, S)$ and $B(L_s, \mathcal{P}_n, \mu_s, S_s)$ via a dual perspective, the condition $S_s \succ \mu_s \mu_s^\top$ in Assumption 7.1 is necessary to establish strong duality. Hence Theorem 7.1 is slightly weaker than its counterpart Theorem 6.1 and only establishes

$$B(L, \mathcal{K}, \mu, S) = B(L_s, \mathcal{P}_n, \mu_s, S_s)$$

on the interior of the domain of both functions.

In what follows, we will derive computational reformulations only for uncertainty quantification problems over the standard probability simplex $\mathcal{P}_n$. As the previous discussion indicated taking $\mathcal{K} = \mathcal{P}_n$ is without loss of generality when the ambiguity set $\mathcal{K}$ admits a Choquet star representation. In a next step we will then show how many uncertainty quantification problems over a more richly structured Choquet star simplex $\mathcal{K}$ can be treated equally well.

7.2 Dual representation via moment functions

Similar as in Chapter 6 we will start the discussion with assuming that the ambiguity set $\mathcal{K}$ is taken to be the standard probability simplex $\mathcal{P}_n$. As explained before, this is without loss of generality in virtue of the reduction Theorems 6.1 or its dual counterpart 7.1 when $\mathcal{K}$ admits a Choquet star representation. Subsequently, in Section 7.3 this will be made concrete in case of both α-unimodal and γ-monotone distributions.

The loss functions we will consider for the worst-case expectation bound $B(L, \mathcal{P}_n, \mu, S)$ will be in the following form

$$L(x) = \max_{i \in I_0} \; \ell_i(A_i x). \tag{7.5}$$

where each $\ell_i : \mathrm{R}^d \to \mathrm{R}$ and $A_i \in \mathrm{R}^{d\times n}$ for all $i \in I_0 := I \cup \{0\} = [0, \ldots, k]$. We assume further that the loss function L is positive, which is enforced by taking $\ell_0(x) = 0$ as in the preceding chapter. Please note that unlike in the preceding chapter, the functions ℓ_i are not required to be concave. By the end of this section we will have indicated that many interesting loss functions can be brought in the form (7.5).

Under aforementioned conditions, an uncertainty quantification problem is most naturally treated in its dual form. Indeed, we will show in this chapter that the dual problem (7.4) can be restated as a tractable convex optimization problem with semi-infinite constraints. The following theorem which does not provide a tractable reformulation of the worst-case expectation bound $B(L, \mathcal{P}_n, \mu, S)$ just yet, will be at the basis of many results found in this chapter.

Theorem 7.2. *The worst-case expectation problem with second-order moment information* (6.2) *can be reformulated as* $B\left(\max_{i\in I_0} \ell_i(A_i x), \mathcal{P}_n, \mu, S\right) =$

$$\begin{aligned}
\inf \quad & \operatorname{Tr}\left\{\begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \cdot \begin{pmatrix} S & \mu \\ \mu^\top & 1 \end{pmatrix}\right\} \\
\text{s.t.} \quad & \left.\begin{aligned} & \begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \in \mathrm{S}_+^{n+1}, \begin{pmatrix} T_{1,i} & T_{2,i} \\ T_{2,i}^\top & T_{3,i} \end{pmatrix} \in \mathrm{S}_+^{d+1}, \; \Lambda_{1,i} \in \mathrm{R}^{d\times d}, \; \Lambda_{2,i} \in \mathrm{R}^d \\ & \begin{pmatrix} \Lambda_{1,i} + \Lambda_{1,i}^\top - T_{1,i} & \Lambda_{2,i} - T_{2,i} & -\Lambda_{1,i}^\top A_i \\ \Lambda_{2,i}^\top - T_{2,i}^\top & y_0 - T_{3,i} & y^\top - \Lambda_{2,i}^\top A_i \\ -A_i^\top \Lambda_{1,i} & y - A_i^\top \Lambda_{2,i} & Y \end{pmatrix} \succeq 0, \quad \forall i \in I \end{aligned}\right\} \quad (C_1) \\
& T_{3,i} + 2q^\top T_{2,i} + q^\top T_{1,i} q \geq \ell_i(q), \quad \forall q \in \mathrm{R}^d, \quad \forall i \in I \qquad (C_2)
\end{aligned}$$

when Assumption 7.1 holds.

Proof. The constraint in the dual problem (7.4) can be reformulated as

$$\forall i \in I_0, \quad \forall q \in \mathrm{R}^d : \quad \inf_{A_i x = q} \; x^\top Y x + 2x^\top y + y_0 \; \geq \; \ell_i(q).$$

As we assume that the loss function L is positive, it must hence follow that the matrix $(Y, y; y^\top, y_0)$ is positive semidefinite. The claim now follows immediately from Theorem A.2 applied to the parametric optimization problem $\inf_{A_i x = q} \; x^\top Y x + 2x^\top y + y_0$. □

Note that this reformulation of the standard dual problem (7.4) into the more unconventional form in Theorem 7.2 is motivated by a desire to replace the semi-infinite constraint over R^n with one over R^d. Hence when $d \ll n$, the reformulation offered by Theorem 7.2 is preferable to the standard dual (7.4). Observe that bar for the semi-infinite constraint (C_2), the reformulation offered in Theorem 7.2 for the worst-case expectation bound $B\left(\max_{i\in I_0} \ell_i(A_i x), \mathcal{P}_n, \mu, S\right)$ is tractable. Hence the remainder of this chapter, we can focus on providing tractable reformulations of the semi-infinite constraint (C_2) for various functions ℓ_i.

The semi-infinite constraint (C_2) in R^d of Theorem 7.2 for piece-wise polynomial ℓ_i admits a tractable reformulation in the univariate case when $d = 1$ using sum-of-squares reformulations as indicated in Theorem 4.2. Similarly, when the functions ℓ_i are quadratically representable Theorem 7.2 provides a tractable reformulation of the worst-case expectation bound $B(L, \mathcal{P}_n, \mu, S)$ with help of the S-Lemma 4.1. In either case, the semi-infinite constraint (C_2) is represented through a tractable LMI and evaluating the worst-case expectation bound $B(L, \mathcal{P}_n, \mu, S)$ reduces to solving an SDP.

7.2.1 Unstructured uncertainty inequalities

Before considering structured classes of probability distributions, it is instructive to apply Theorem 7.2 to the unstructured case $\mathcal{K} = \mathcal{P}_n$, and to restate some well-known tractable reformulations of uncertainty quantification problems. The purpose of this section is thus not so much as to present novel results, but rather to indicate that many results in the literature can be recognized as corollaries of Theorem 7.2 for an appropriate choice of loss function L in the worst-case expectation bound

$$B(L, \mathcal{P}_n, \mu, S).$$

We reconsider the generalized Chebyshev bound $G_\infty(\mu, S)$ discussed in Chapter 6 on the worst-case probability of the event $\xi \notin \Xi$ based solely on second-moment information by letting $L = \mathbf{1}\{\mathrm{R}^n \setminus \Xi\}$. We will indicate here how Theorem 7.2 then provides a dual counterpart to the tractable reformulation (6.11). Additionally, we present tractable reformulations of worst-case CVaR and expectation problems which will come to fruition in Chapters 8 and 9. The main results presented in this section, from a practitioners point of view, are summarized in the table below. In Section 7.3 we will then present counterparts to the result presented here when considering more richly structured ambiguity sets $\mathcal{K} \subset \mathcal{P}_n$.

Structure	Probability inequalities	Expectation & CVaR inequalities
Unstructured $\mathcal{P}_n$	Example 7.1	Theorem 7.3

Probability inequalities

We revisit here the generalized Chebyshev inequalities as introduced in Section 6.2.2 of the previous chapter. Recall that we are interested in computing the worst-case probability of an event $\mathbb{P}(\xi \notin \Xi)$, merely using the fact that ξ has mean μ and second moment S. In Section 6.2.2 a primal reformulation was offered in terms of the tractable convex optimization problem (6.11) for polytopic event sets Ξ.

In light of Theorem 7.2, we can approach the generalized Chebyshev bound $G_\infty(\mu, S)$ alternatively via its dual characterization. There are in fact two ways in which we can apply Theorem 7.2 so as to obtain a tractable reformulation of the generalized Chebyshev bound

$$G_\infty(\mu, S) = B(\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}, \mathcal{P}_n, \mu, S)$$

using once more the equivalence between measure and expectation of indicator functions. We will present both approaches and indicate that a slight generalization of the results in Section 6.2.2 can be achieved.

(i) In a first approach, we again assume that the set Ξ is polytopic and represented as the finite intersection of half-spaces $\Xi = \left\{x \in \mathrm{R}^n \ : \ a_i^\top x < b_i, \quad \forall i \in I\right\}$. The indicator function of the set of interest $\mathrm{R}^n \setminus \Xi$ can be brought in the form required by Theorem 7.2 using the equivalence

$$\mathbf{1}\{\mathrm{R}^n \setminus \Xi\} = \max_{i \in I} \ell_i\left(a_i^\top x\right)$$

where each of the functions $\ell_i(q) = \mathbf{1}\{q \geq b_i\}$ for all $i \in I$ is an indicator function. Each of the functions $\ell_i\left(a_i^\top x\right)$ thus represents the indicator function of the half-space $\left\{x \in \mathrm{R}^n \ : \ a_i^\top x \geq b_i\right\}$.

The Chebyshev bound $G_\infty(\mu, S)$ can now be reformulated via Theorem 7.2 using the previous observation, i.e.

$$G_\infty(\mu, S) = B\left(\max_{i \in I} \ell_i(a_i^\top x), \mathcal{P}_n, \mu, S\right).$$

The following example makes the last remaining problematic constraint (C_2) in the reformulation offered in Theorem 7.2 explicit in terms of an LMI thus providing a tractable reformulation to the Chebyshev bound $G_\infty(\mu, S)$ for polytopic event sets Ξ.

Example 7.1 (Polytopes)**.** *The worst-case probability bound $G_\infty(\mu, S)$ for the event $\xi \notin \Xi$ as defined before can be modeled as in Theorem 7.2. The constraint (C_2) specializes to*

$$T_{3,i} - 1 + 2q^\top T_{2,i} + q^\top T_{1,i} q \geq 0, \quad \forall q \geq b_i, \quad \forall i \in I,$$

which can be rewritten as an LMI by virtue of the S-Lemma 4.1. The generalized Chebyshev bound $G_\infty(\mu, S)$ for polytopic sets Ξ is therefore equivalent to the following tractable SDP

$$\begin{aligned} G_\infty(\mu, S) = \inf \quad & \mathrm{Tr}\{YS\} + 2y^\top \mu + y_0 \\ \text{s.t.} \quad & (C_1), \\ & \exists \tau_i \in \mathrm{R}_+, \begin{pmatrix} T_{1,i} & T_{2,i} \\ T_{2,i} & T_{3,i} - 1 \end{pmatrix} \succeq \tau_i \begin{pmatrix} 0 & 1 \\ 1 & -2b_i \end{pmatrix}, \ \forall i \in I. \end{aligned} \tag{7.6}$$

It can be remarked that the tractable reformulation (7.6) represents a dual counterpart to Theorem 6.3.

(ii) An alternative approach to the generalized Chebyshev bound $G_\infty(\mu, S)$ based directly on the S-Lemma 4.1 can be found in the works of Vandenberghe et al. [133] or Zymler et al. [142] and in fact allows for a slight generalization of the result discussed in Example 7.1.

Let the set of interest $\Xi = \cap_{i \in I} \Xi_i$ be for the moment the intersection of finitely many generalized ellipsoids Ξ_i, i.e. $\Xi_i := \left\{x \in \mathrm{R}^n \ : \ x^\top E_i x + 2e_i^\top x + e_i^0 < 0\right\}$ with $E_i \in \mathrm{S}^n$ not necessarily positive definite matrices. The indicator function of the set of interest $\mathrm{R}^n \setminus \Xi$ can be brought again in the form required by Theorem 7.2 using

$$\mathbf{1}\{\mathrm{R}^n \setminus \Xi\} = \max_{i \in I} \ell_i(x)$$

where each of the functions $\ell_i(q) := \mathbf{1}\left\{q^\top E_i q + 2e_i^\top q + e_i^0 \geq 0\right\}$ is the indicator function corresponding to the complement of a single generalized ellipsoid Ξ_i.

An exact tractable reformulation of the corresponding Chebyshev bound $G_\infty(\mu, S)$ based directly on the S-Lemma 4.1 seems to have been found first by Vandenberghe et al. [133] and was later independently rediscovered by Zymler et al. [142]. We will now indicate that the Chebyshev bound $G_\infty(\mu, S)$ for generalized ellipsoids can again be reformulated immediately via Theorem 7.2 based on the observation that

$$G_\infty(\mu, S) = B\left(\max_{i \in I} \ell_i(x), \mathcal{P}_n, \mu, S\right).$$

The following example reformulates the last remaining problematic constraint (C_2) in the reformulation offered in Theorem 7.2 via the S-Lemma 4.1, thereby providing a tractable reformulation to the Chebyshev bound $G_\infty(\mu, S)$ for quadratically representable event sets Ξ.

Example 7.2 (Ellipsoids [133, 142]). *The worst-case probability $G_\infty(\mu, S)$ for the event $\xi \notin \Xi$ as defined before can be modeled as in Theorem 7.2. The constraint (C_2) becomes*

$$T_{3,i} + 2q^\top T_{2,i} + q^\top T_{1,i} q \geq \ell_i(q), \quad \forall q \in \mathrm{R}^n, \quad \forall i \in I,$$

which can be rewritten with the help of the $\mathcal{S}$-Lemma 4.1 as an LMI. The generalized Chebyshev bound $G_\infty(\mu, S)$ for ellipsoidal sets Ξ is therefore equivalent to the following tractable SDP

$$\begin{aligned} G_\infty(\mu, S) = \inf \quad & \mathrm{Tr}\,\{YS\} + 2y^\top \mu + y_0 \\ \text{s.t.} \quad & (C_1), \\ & \exists \tau_i \in \mathrm{R}_+, \begin{pmatrix} T_{1,i} & T_{2,i} \\ T_{2,i} & T_{3,i} - 1 \end{pmatrix} \succeq \tau_i \begin{pmatrix} E_i & e_i \\ e_i^\top & e_i^0 \end{pmatrix}, \ \forall i \in I. \end{aligned} \tag{7.7}$$

Expectation inequalities

We have indicated in Section 5.4.2 that any worst-case CVaR problem can be reduced to standard worst-case expectation problems. In this part we will make the previous claim concrete for worst-case CVaR problems of the general form

$$\begin{aligned} B_{\mathrm{CVaR}}(L, \mathcal{P}_n, \mu, S) := \sup \quad & \mathbb{P}\text{-CVaR}_\epsilon\,(L(\xi)) \\ \text{s.t.} \quad & \mathbb{P} \in \mathcal{H}(\mu, S) \\ & \mathbb{P} \in \mathcal{P}_n. \end{aligned} \tag{7.8}$$

where the loss function is in the form $L = \max_{i \in I} \ \ell_i(x)$ for quadratic functions $\ell_i(x) = x^\top E_i x + 2e_i^\top x + e_i^0$. Recall that the worst-case CVaR problem can be reduced to a standard worst-case expectation problem using

$$B_{\mathrm{CVaR}}(L, \mathcal{K}, \mu, S) = \min_\beta \ \beta + \frac{1}{\epsilon} B((L - \beta)_+, \mathcal{K}, \mu, S). \tag{7.9}$$

In what follows, we will present a tractable reformulation of the worst-case CVaR problem (7.8) for quadratically representable loss functions based on Theorem 7.2 and the reducation (7.9).

The following example reformulates the last remaining problematic constraint (C_2) in the reformulation offered in Theorem 7.2 via the S-Lemma 4.1, thereby providing a tractable reformulation to the worst-case expectation bound $B((L-\beta)_+, \mathcal{P}_n, \mu, S)$ for quadratically representable loss functions L and by virtue of the reduction (7.9) to the worst-case CVaR problem $B_{\mathrm{CVaR}}(L, \mathcal{K}, \mu, S)$ as well.

Example 7.3 (Zymler et al. [142]). *For a piecewise quadratic loss function $L(x) = \max_{i \in I} \ \ell_i(x)$, the constraint (C_2) in Theorem 7.2 for the worst-case expectation bound $B((L-\beta)_+, \mathcal{P}_n, \mu, S)$ becomes*

$$T_{3,i} + 2q^\top T_{2,i} + q^\top T_{1,i} q \geq q^\top E_i q + 2e_i^\top q + e_i^0 - \beta, \quad \forall q \in \mathrm{R}^n, \quad \forall i \in I,$$

which can be rewritten as an LMI using the S-Lemma 4.1. From equivalence (7.9) *it follows that the worst-case CVaR problem for the quadratic loss function L is therefore equivalent to the SDP*

$$\begin{aligned} B_{\mathrm{CVaR}}(L, \mathcal{P}_n, \mu, S) = \inf \quad & \beta + \tfrac{1}{\epsilon}\left[\mathrm{Tr}\,\{YS\} + 2y^\top \mu + y_0\right] \\ \text{s.t.} \quad & (C_1), \\ & \begin{pmatrix} T_{1,i} & T_{2,i} \\ T_{2,i} & T_{3,i} \end{pmatrix} \succeq \begin{pmatrix} E_i & e_i \\ e_i^\top & e_0 - \beta \end{pmatrix}, \ \forall i \in I. \end{aligned} \tag{7.10}$$

It is shown by Zymler et al. [142] that the tractable reformulation (7.10) offered in Example 7.3 for the worst-case CVaR bound $B_{\mathrm{CVaR}}(L, \mathcal{P}_n, \mu, S)$ can in fact be written more concisely as done in Theorem 7.3. The worst-case CVaR bound $B_{\text{CVaR}}(L, \mathcal{P}_n, 0, \Sigma)$, where the loss function L is a single centralized quadratic ($k = 1$), for a zero mean random variable ξ with given variance matrix $\Sigma \in \mathrm{S}^n_+$ can be even further simplified and admits the closed form expression stated in Corollary 7.1.

Theorem 7.3 (Zymler et al. [142]). *For a piecewise quadratic loss function $L(x) = \max_{i \in I}\ x^\top E_i x + 2e_i^\top x + e_i^0$ the worst-case CVaR bound $B_{\text{CVaR}}(L, \mathcal{P}_n, \mu, S)$ reduces to the tractable SDP*

$$\begin{aligned} B_{\text{CVaR}}(L, \mathcal{P}_n, \mu, S) = \inf \quad & \beta + \tfrac{1}{\epsilon}\left[\mathrm{Tr}\left\{YS\right\} + 2y^\top \mu + y_0\right] \\ \text{s.t.} \quad & \begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \succeq 0 \\ & \begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \succeq \begin{pmatrix} E_i & e_i \\ e_i^\top & e_i^0 - \beta \end{pmatrix}, \quad \forall i \in I \end{aligned} \tag{7.11}$$

when Assumption 7.1 holds.

Theorem 7.3 will come to use in the subsequent Chapters 8 and 9. The following corollary of Theorem 7.3 will play an important role in Chapter 9.

Corollary 7.1 (Concentric distributions and loss functions). *If $L(x) = x^\top E_1 x + e_1^0$ constitutes a single quadratic function centered at the origin, while the random vector ξ has mean $\mu = 0$ and variance Σ, then the worst-case CVaR bound*

$$B_{\text{CVaR}}(L, \mathcal{P}_n, 0, \Sigma) = e_1^0 + \frac{1}{\epsilon}\mathrm{Tr}\left\{\Sigma E_1\right\} \tag{7.12}$$

admits a closed form exact reformulation.

Proof. For the loss function $L(x) = x^\top E_1 x + e_1^0$ the previous Theorem 7.3 implies

$$\begin{aligned} B_{\text{CVaR}}(L, \mathcal{P}_n, 0, \Sigma) \quad = \quad \inf \quad & \beta + \frac{1}{\epsilon}\left(\mathrm{Tr}\left\{\Sigma Y\right\} + y_0\right) \\ \text{s.t.} \quad & Y \in \mathrm{S}^n_+,\ y \in \mathrm{R}^n,\ y_0 \in \mathrm{R}_+,\ \beta \in \mathrm{R} \\ & \begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \succeq 0, \quad \begin{pmatrix} Y - E_1 & y \\ y^\top & y_0 - e_1^0 + \beta \end{pmatrix} \succeq 0. \end{aligned} \tag{7.13}$$

As $Y = E_1$, $y = 0$, $y_0 = 0$ and $\beta = e_1^0$ is feasible in (7.13), it is clear that the worst-case CVaR is bounded above by $e_1^0 + \frac{1}{\epsilon}\mathrm{Tr}\left\{\Sigma E_1\right\}$. To prove the converse inequality, we let $(Y^\star, y^\star, y_0^\star, \beta^\star)$ be an optimal solution of (7.13). Then, we find

$$\begin{aligned} B_{\text{CVaR}}(L, \mathcal{P}_n, 0, \Sigma) \quad & = \quad \beta^\star + \frac{1}{\epsilon}\left(\mathrm{Tr}\left\{\Sigma Y^\star\right\} + y_0^\star\right) \\ & \geq \quad \beta^\star + \frac{1}{\epsilon}\left(\mathrm{Tr}\left\{\Sigma E_1\right\} + (e_1^0 - \beta^\star)^+\right) \ \geq \ e_1^0 + \frac{1}{\epsilon}\mathrm{Tr}\left\{\Sigma E_1\right\}, \end{aligned}$$

where the first inequality exploits the feasibility of $(Y^\star, y^\star, y_0^\star, \beta^\star)$ in (7.13), and the second inequality exploits the fact that $y_0^\star \geq (e_1^0 - \beta^\star)^+$ and $\epsilon \in (0, 1)$. □

7.2.2 Worst-case probability distributions

In many situations it is desirable to know what the worst-case probability distributions $\mathbb{P}^\star \in \mathcal{H}(\mu, S) \cap \mathcal{P}_n$ achieving the bound $B(L, \mathcal{K}, \mu, S) = \int L(x)\, \mathbb{P}^\star(\mathrm{d}x)$ look like. As we approach the uncertainty quantification problem (7.4) via its dual characterization, the dual variables do not relate directly to a worst-case probability distribution. As described in Section 5.2, we can nevertheless extract a worst-case probability distribution using the complementarity condition (5.8) between primal worst-case probability distribution $\mathbb{P}^\star$ and dual optima $(Y^\star, y^\star, y_0^\star)$ at virtually no additional cost.

We will illustrate the previous statement for the generalized Chebyshev bound $G_\infty(\mu, S)$ discussed in Sections 6.2.2 and 7.2.1. The complementarity condition (5.8) between the optima in the uncertainty quantification problem (6.2) and its dual (7.4) can be specialized to

$$\int \mathbf{1}\{\mathrm{R} \setminus \Xi\}(x)\, \mathbb{P}^\star(\mathrm{d}x) = \int x^\top Y^\star x + 2x^\top y^\star + y_0^\star \; \mathbb{P}^\star(\mathrm{d}x).$$

A direct consequence of dual feasibility in combination with the previous complementarity condition is that $\mathbb{P}^\star$ must be supported on the points at which the dual function kisses the indicator function $\mathbf{1}\{\mathrm{R} \setminus \Xi\}$, i.e.

$$\begin{aligned} \operatorname{supp} \mathbb{P}^\star \subseteq S^\star &= \left\{ x \in \mathrm{R}^n \;:\; x^\top Y^\star x + 2{y^\star}^\top x + y_0^\star = \mathbf{1}\{\mathrm{R} \setminus \Xi\}(x) \right\}. \qquad (7.14) \\ &\subseteq \left\{ x \in \mathrm{R}^n \;:\; x^\top Y^\star x + 2{y^\star}^\top x + y_0^\star = 0 \right\} \cup \\ &\qquad \left\{ x \in \mathrm{R}^n \;:\; x^\top Y^\star x + 2{y^\star}^\top x + y_0^\star = 1 \right\} \end{aligned}$$

The set $S^\star$ for the generalized Chebyshev bound $G_\infty(\mu, S)$ is hence recognized as a subset of the roots of two quadratic equations in R^n and can be computed at virtually no additional computational cost. This intimate relationship between worst-case probability distribution and dual optimal solution is visually illustrated in Figure 7.1. We remark that the same relationship between primal and dual optima $\mathbb{P}^\star$ and $(Y^\star, y^\star, y_0^\star)$ has been made by Vandenberghe et al. [133]. They obtained the support condition (7.14) however via direct complementarity between the tractable reformulations (6.11) and (7.6) instead of between the worst-case probability bound $G_\infty(\mu, S)$ and its dual formulation as done here.

It is of interest to remind the reader here that the previous discussion extends to more richly structured ambiguity sets $\mathcal{K} \subset \mathcal{P}_n$ as well. Indeed, with the help of the equivalence $B(L, \mathcal{K}, \mu, S) = B(L_s, \mathcal{P}_n, \mu_s, S_s)$ put forward in both Theorems 6.1 and its dual counterpart 7.1 we can determine the worst-case distribution $\mathbb{P}^\star \in \mathcal{H}(\mu, S) \cap \mathcal{K}$ achieving the bound $B(L, \mathcal{K}, \mu, S) = \int L(x)\, \mathbb{P}^\star(\mathrm{d}x)$ as

$$\mathbb{P}^\star = \int \mathbb{T}_x\, \mathbb{m}^\star(\mathrm{d}x),$$

where $\mathbb{m}^\star \in \mathcal{H}(\mu_s, S_s) \cap \mathcal{P}_n$ is the worst-case distribution achieving the bound $B(L_s, \mathcal{P}_n, \mu_s, S_s)$. From the previous discussion it must now follow that $\mathbb{P}^\star$ is supported on the rays $\{t x_k \;:\; t \in \mathrm{R}_+\}$ where $x_k \in \mathrm{R}^n$ are those points at which the optimal dual function in the dual characterization of $B(L_s, \mathcal{P}_n, \mu_s, S_s)$ kisses the transformed loss function L_s. The previous statement is visually illustrated in Figure 7.1.

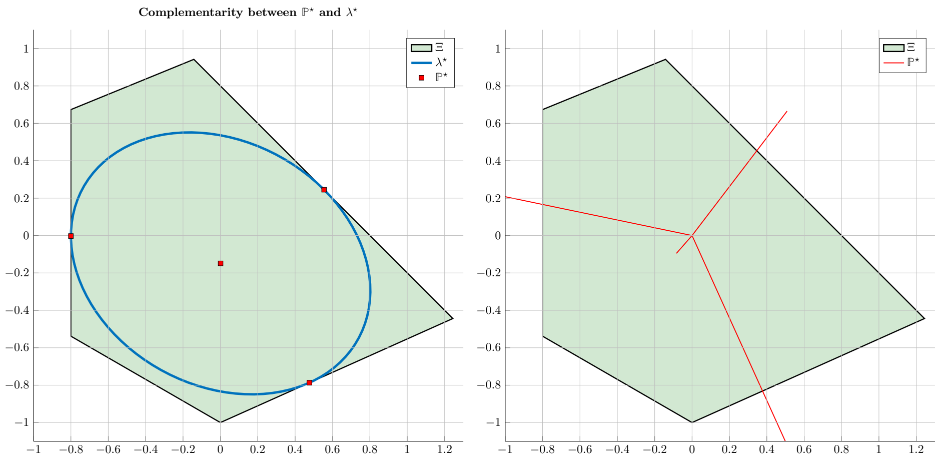
7.3 Structured uncertainty inequalities

We have thus far described expectation bounds $B(L, \mathcal{P}_n, \mu, S)$ over the standard probability simplex $\mathcal{P}_n$. The principal aim of this work, however, is to describe uncertainty quantification problems over more richly structured ambiguity sets $\mathcal{K} \subset \mathcal{P}_n$. In the current section, we will use the dual approach offered in Theorem 7.2 to tackle structured uncertainty problems as well.

The approach described here is similar to the one taken to deal with richly structured ambiguity sets in Section 6.3 of the previous chapter. That is, we intend to use the equivalence

$$B(L, \mathcal{K}, \mu, S) = B(L_s, \mathcal{P}_n, \mu_s, S_s)$$

to reduce a worst-case expectation problem over the structured ambiguity set $\mathcal{K}$ to its transformed equivalent over the standard probability simplex $\mathcal{P}_n$. The resulting unstructured worst-case expectation bound $B(L_s, \mathcal{P}_n, \mu_s, S_s)$ shall now however be approached via Theorem (7.2).



(a) Unstructured $\mathcal{K} = \mathcal{P}_n$ (b) Unimodal $\mathcal{K} = \mathcal{U}_n$

Figure 7.1: The relationship between worst-case probability distributions $\mathbb{P}^\star$ and dual optimal solutions $(Y^\star, y^\star, y_0^\star)$ for the generalized Chebyshev bound $G_\infty(\mu, S)$ discussed in Sections 6.2.2 and 7.2.1 is shown in Figure 7.1(a). The complementarity condition (7.14) states that $\mathbb{P}^\star$ must be supported on the points at which the optimal dual function $x^\top Y^\star x + 2{y^\star}^\top x + y_0^\star$ kisses the indicator function $\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}$. The figure shows the support of the worst-case probability distribution $\mathbb{P}^\star$ in red. The blue ellipsoid indicates the points at which the value of the dual function is one. The worst-case distribution $\mathbb{P}^\star$ for the generalized Gauss bound $G_n(\mu, S)$ is shown in 7.1(b) and is supported on the lines $[0, x_k]$ where x_k are those points at which the optimal dual function $x^\top Y^\star x + 2{y^\star}^\top x + y_0^\star$ kisses the loss function $L_s(x) = \int \mathbf{1}\{\mathrm{R}^n \setminus \Xi\}(y)\, \mathbb{u}_x^n(\mathrm{d}y)$.

The main results presented in this section, from a practitioners point of view, are summarized in the table below. We will focus mainly on indicator functions of polytopic sets Ξ which arise in worst-case probability inequalities and piecewise affine functions which arise when dealing with convex loss functions L. In doing so, we attempt to obtain counterparts to the uncertainty quantification problems discussed in Section 7.2.1 for more richly structured ambiguity sets $\mathcal{K}$.

Structure	Probability inequalities	Expectation & CVaR inequalities
Unimodal $\mathcal{U}_\alpha$	Corollary 7.2	Corollary 7.4
Monotone $\mathcal{M}_\gamma$	Corollary 7.3	Corollary 7.5

7.3.1 Probability inequalities

We address here the problem of bounding the probability of the event $\xi \notin \Xi$ where Ξ is an open convex polytope and $\mathbb{P} \in \mathcal{K}$ is a structured ambiguity set with known mean μ and second moment S. In this case we can use the standard identity between the probability of an event and the expectation of its indicator function to state

$$B(\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}, \mathcal{K}, \mu, S) = \begin{array}[t]{rl} \sup & \mathbb{P}(\xi \notin \Xi) \\ \text{s.t.} & \mathbb{P} \in \mathcal{H}(\mu, S), \\ & \mathbb{P} \in \mathcal{K}. \end{array}$$

We assume again in this section that the set $0 \in \Xi$ has a half-space representation in the form $\Xi := \left\{ x \in \mathrm{R}^n \ : \ a_i^\top x < b_i, \ \forall i \in I \right\}$.

We remark that when the ambiguity set $\mathcal{K}$ is taken to be the set of all α-unimodal distributions $\mathcal{U}_\alpha$ we recover the worst-case probability bound $G_\alpha(\mu, S)$ discussed in Section 6.3 of the preceding chapter. In this chapter we will generalize this result and show that the worst-case probability problem $B(\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}, \mathcal{K}, \mu, S)$ admits a tractable reformulation whenever $\mathcal{K}$ is a Choquet star simplex.

The indicator function $\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}$ can be represented as the point-wise maximum of the indicator functions associated with the half-spaces from which the set Ξ is composed, i.e.

$$L = \max_{i \in I} \ \mathbf{1}\left\{a_i^\top x \geq b_i\right\} = \mathbf{1}\{\mathrm{R}^n \setminus \Xi\}. \tag{7.15}$$

The next proposition shows how to transform, via the transformation (6.4), such an indicator function for radial extreme distributions $\mathbb{T}_x$ into a loss function L_s for use in Theorem 7.2:

Proposition 7.1. *If the set $\mathcal{K}$ admits a Choquet star representation with generating distribution $\mathbb{T}$, then*

$$L_s(x) = \int \max_{i \in I} \ \mathbf{1}\left\{a_i^\top x \geq b_i\right\}(y)\, \mathbb{T}_x(\mathrm{d}y) = \max_{i \in I} \ \mathbb{T}\left(\left[b_i / a_i^\top x, \infty\right)\right).$$

Proof. The proof follows from direct elementary manipulations and is thus omitted. □

We have, according to the reduction Theorem 7.1, that the worst-case probability problem over $\mathcal{K}$ can be reduced to an equivalent worst-case probability problem over the standard simplex $\mathcal{P}_n$

$$\begin{aligned}\sup_{\mathbb{P} \in \mathcal{K} \cap \mathcal{P}(\mu, S)} \mathbb{P}(\xi \notin \Xi) &= B\left(\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}, \mathcal{K}, \mu, S\right), \\ &= B\left(\max_{i \in I} \ \mathbb{T}\left(\left[b_i / a_i^\top x, \infty\right)\right), \mathcal{P}_n, \mu_s, S_s\right),\end{aligned}$$

where the final worst-case expectation bound is in the form required in Theorem 7.2 with $\ell_i(x) = \mathbb{T}\left(\left[b_i / a_i^\top x, \infty\right)\right)$. The univariate semi-infinite constraint (C_2) for these particular functions ℓ_i can be represented for many generators $\mathbb{T}$ as LMIs via sum-of-squares reformulations.

We will now make the result in Proposition 7.1 concrete for α-unimodal and γ-monotone distributions. Specifically, our method is as follows: Examples 3.2 and 3.3 provide us with the appropriate generating distributions $\mathbb{T}$ for α-unimodal or γ-monotone distributions. We then use these generating distributions $\mathbb{T}$ to transform $(L, \mu, S) \mapsto (L_s, \mu_s, S_s)$ via Theorem 7.1, where the mapping $L \mapsto L_s$ in particular is supplied in Proposition 7.1. The resulting worst-case probability bound is amendable to Theorem 7.2 for which we then identify the appropriate expression for the constraint (C_2) for our particular functions ℓ_i.

α-Unimodal distributions When the ambiguity set $\mathcal{K}$ is taken to be the set of all α-unimodal distributions $\mathcal{U}_\alpha$, we recover the worst-case probability bound $G_\alpha(\mu, S)$ discussed in Section 6.3 of the preceding chapter. In that case we can use Proposition 7.1 to provide an alternative tractable reformulation of the worst-case probability bound $G_\alpha(\mu, S)$ based on its dual characterization.

Corollary 7.2 (α-Unimodal probability inequalities)**.** *For any rational $0 \leq \alpha = \frac{v}{w}$, with $(v, w) \in \mathrm{N}$ and $0 \in \Xi$ we have the equality $B(\mathbf{1}\{\mathrm{R}^n \setminus \Xi\}, \mathcal{U}_\alpha, \mu, S) =$*

$$\begin{aligned}\inf \quad & \mathrm{Tr}\left\{\begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \cdot \begin{pmatrix} S_\alpha & \mu_\alpha \\ \mu_\alpha^\top & 1 \end{pmatrix}\right\} \\ \text{s.t.} \quad & (C_1), \\ & q^{2w+v} b_i^2 T_{1,i} + 2q^{w+v} b_i T_{2,i} + q^v (T_{3,i} - 1) + 1 \geq 0, \quad \forall q \geq 0\end{aligned}$$

where $S_\alpha \succ \mu_\alpha \mu_\alpha^\top$ for $S_\alpha = \frac{\alpha+2}{\alpha} S$ and $\mu_\alpha = \frac{\alpha+1}{\alpha} \mu$.

We remark here that Corollary 7.2 generalizes the results presented in Theorem 6.13 as it no longer matters that the unimodality parameter α satisfies $\alpha \geq 1$. However, where the result in Theorem 6.3 follows from a direct reformulation of the primal problem (6.2), the result in Corollary 7.2 hinges on the dual problem (7.4) to be strong. Strong duality calls for the additional Slater type condition $S_\alpha \succ \mu_\alpha \mu_\alpha^\top$ to hold, which explains the strict inequality in the corollaries stated hereafter. The proofs of the corollaries presented in this section are deferred to Appendix B.

γ-Monotone distributions Our approach is identical to that in previous paragraph, except that we now look to Example 3.3 to provide us with the appropriate generating distribution $\mathbb{T}$ for γ-monotone distributions.

Corollary 7.3 (γ-Monotone probability inequalities)**.** *For any $\gamma \in \mathrm{N}_0$ we have the equality $B(\mathbf{1}\left\{\mathrm{R}^n \setminus \Xi\right\}, \mathcal{M}_\gamma, \mu, S) =$*

$$
\begin{aligned}
\inf \quad & \operatorname{Tr}\left\{\begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \cdot \begin{pmatrix} S_\gamma & \mu_\gamma \\ \mu_\gamma^\top & 1 \end{pmatrix}\right\} \\
\text{s.t.} \quad & (C_1), \\
& T_{1,i} b_i^2 q^{n+\gamma+1} + 2 b_i T_{2,i} q^{n+\gamma} + \left(T_{3,i} - 1\right) q^{n+\gamma-1} + \\
& \qquad \frac{1}{B(n,\gamma)} \sum_{k=0}^{\gamma-1} \frac{(-1)^k}{n+k} \binom{\gamma-1}{k} q^{\gamma-k-1} \geq 0, \quad \forall q \geq 1
\end{aligned}
$$

where $S_\gamma \succ \mu_\gamma \mu_\gamma^\top$ for $S_\gamma = \frac{n+\gamma}{n} \frac{n+\gamma+1}{n+1} S$ and $\mu_\gamma = \frac{n+\gamma}{n} \mu$.

7.3.2 Expectation inequalities

The worst-case expectation bound $B(L, \mathcal{P}_n, \mu, S)$ over the standard simplex is well known to be tractable when the considered loss function L takes the form

$$
L(x) = \max_{i \in I} \; a_i^\top x - b_i \tag{7.16}
$$

and thus convex. The previous class of worst-case expectation bounds can indeed be directly dealt with using the approach taken in Example 7.3. Worst-case expectation problems $B(L, \mathcal{P}_n, \mu, S)$ for piece-wise affine loss functions find their application in a wide range of practical problem; see for instance Bertsimas and Popescu [17], Zymler et al. [142] or Smith [119] plus the many references therein. Because the set of all functions consisting of the point-wise maximum of affine functions coincides with the class of lower semicontinuous (lsc) convex functions, the following fact is of interest.

Fact 7.1. *If the set $\mathcal{K}$ admits a Choquet star representation with generating distribution $\mathbb{T}$ and L is convex, then*

$$
L_s(x) = \int_0^\infty L(tx) \; \mathbb{T}(\mathrm{d}t)
$$

is convex as well.

Proof. The statement can be proved almost immediately from the definition of convexity. For all $\theta \in [0, 1]$

$$\begin{aligned} L_s(\theta a + (1-\theta)b) &= \int_0^\infty L(t(\theta a + (1-\theta)b))\ \mathbb{T}(\mathrm{d}t) \\ &= \int_0^\infty L(\theta(ta) + (1-\theta)(tb))\ \mathbb{T}(\mathrm{d}t) \\ &\leq \int_0^\infty \theta L(ta) + (1-\theta)L(tb)\ \mathbb{T}(\mathrm{d}t) \\ &\leq \theta L_s(a) + (1-\theta)L_s(b) \end{aligned}$$

showing convexity of L_s. □

Despite the previous encouraging result, it is generally *not* the case that the function L_s can be represented as the maximum of a *finite* number of affine functions when L is in the form (7.16). Indeed, Fact 7.1 merely establishes that convexity is preserved but does not otherwise address the structure of L_s. This is problematic as the application of Theorem 7.2 to the transformed worst-case expectation problem $B(L_s, \mathcal{P}_n, \mu_s, S_s) = B(L, \mathcal{K}, \mu, S)$ demands the loss function L_s to be represented as a finite maximum. Hence, a generalization of Theorem 7.3 for general piece-wise affine functions with richly structured distribution $\mathcal{K} \subset P_n$ does not seem obvious.

Instead of considering convex piecewise linear loss functions L as those in (7.16), we focus our attention in what follows on loss functions in the form

$$L(x) = (d \circ \kappa_\Xi)(x), \tag{7.17}$$

where $d : \mathrm{R}_+ \to \mathrm{R}$ is a monotonically increasing function and $0 \in \Xi$ a convex set. The function κ_Ξ is the gauge function of the set $0 \in \Xi$ and increases with decreasing proximity to the set Ξ. Loss functions in the form (7.17) arise in the distributionally robust optimization problems of Zymler et al. [142, 143] and control problems by this author [131, 132] when bounding the expected violation of a constraint $\xi \in \Xi$ using

$$\mathbf{E}_{\mathbb{P}}[L(\xi)] \leq \alpha, \quad \forall \mathbb{P} \in \mathcal{C}$$

as the loss function L quantifies constraint violation severity and is increasing with decreasing proximity to the feasible set Ξ. In Chapter 8 constraints of the aforementioned distributionally robust type will be encountered again. The next proposition shows how to transform, via the transformation (6.4), the loss function (7.17) for radial extreme distributions $\mathbb{T}_x$ into a loss function L_s for use in Theorem 7.2:

Proposition 7.2. *If the set $\mathcal{K}$ admits a Choquet star representation with generating distribution $\mathbb{T}$ and L is in the form (7.17) with $0 \in \Xi = \{x \in \mathrm{R}^n \ :\ a_i^\top x < b_i,\, \forall i \in I\}$, then*

$$L_s(x) = (d_s \circ \kappa_\Xi)(x) = \max_{i \in I}\ d_s(a_i^\top x / b_i),$$

with $d_s(q) := \int_0^\infty d(tq)\ \mathbb{T}(\mathrm{d}t)$.

Proof. We have the following chain of equalities proving the claim

$$\begin{aligned} L_s(x) &= \int_0^\infty L(tx)\ \mathbb{T}(\mathrm{d}t) = \int_0^\infty d\left(\kappa_\Xi(tx)\right)\ \mathbb{T}(\mathrm{d}t) \\ &= \int_0^\infty d\left(t \cdot \kappa_\Xi(x)\right)\ \mathbb{T}(\mathrm{d}t), \end{aligned}$$

where the last equality follows from the positive homogeneity of κ_Ξ. □

We have, according to the reduction Theorem 7.1, that the worst-case expectation problem over $\mathcal{K}$ can be reduced to an equivalent worst-case probability problem over the standard simplex $\mathcal{P}_n$, i.e.

$$B(d \circ \kappa_\Xi, \mathcal{K}, \mu, S) = B\left(\max_{i\in I}\ d_s(a_i^\top x/b_i), \mathcal{P}_n, \mu_s, S_s\right),$$

where the second worst-case expectation bound is in the form required in Theorem 7.2 with $\ell_i(a_i^\top x) = d_s(a_i^\top x/b_i)$. The univariate semi-infinite constraint (C_2) for these particular functions ℓ_i can once again be represented for many generators $\mathbb{T}$ as LMIs via sum-of-squares reformulations. In doing so we thus obtain a tractable reformulation of the worst-case expectation problem $B(L, \mathcal{K}, \mu, S)$ for a limited class of piece-wise affine functions and by the reduction (7.9) for the corresponding CVaR problem as well.

In the remainder of this section we discuss specific ambiguity sets $\mathcal{K}$ that admit Choquet star representations, with a focus on unimodal and monotone distributions. To illustrate the power of Proposition 7.2 we assume that $d(t) = (t-1)_+$ so that the loss function L in (7.17) is piece-wise affine. We do remark however that the results stated hereafter apply equally well to many other choices for the function d. The method here follows the approach taken in Section 7.3.1 closely. Examples 3.2 and 3.3 provide us with the appropriate generating distributions $\mathbb{T}$ for α-unimodal or γ-monotone distributions. We then use this generating distribution $\mathbb{T}$ to transform $(L, \mu, S) \mapsto (L_s, \mu_s, S_s)$ via Theorem 7.1, where the mapping $L \mapsto L_s$ in particular is now supplied in Proposition 7.2. The resulting worst-case expectation bound is amendable to Theorem 7.2 for which we then identify the appropriate expression for the constraint (C_2) for our particular functions ℓ_i.

α-Unimodal distributions

Corollary 7.4 (α-Unimodal expectation inequalities)**.** *For any rational $0 \le \alpha = \frac{v}{w} \in \mathbb{Q}$, with v, $w \in \mathrm{N}$, we have the equality $B(\max\{0, \kappa_\Xi(x) - 1\}, \mathcal{U}_\alpha, \mu, S) =$*

$$\begin{aligned}
\inf \quad & \mathrm{Tr}\left\{\begin{pmatrix} Y & y \\ y^\top & y_0\end{pmatrix}\cdot\begin{pmatrix} S_\alpha & \mu_\alpha \\ \mu_\alpha^\top & 1\end{pmatrix}\right\} \\
\text{s.t.} \quad & (C_1), \\
& q^{2w+v}\, b_i^2 T_{1,i} + q^{w+v}\left(2b_i T_{2,i} - \frac{\alpha}{\alpha+1}\right) + q^v\,(1 + T_{3,i}) - \frac{1}{\alpha+1} \ge 0, \quad \forall q \ge 1
\end{aligned}$$

where $S_\alpha \succ \mu_\alpha \mu_\alpha^\top$ for $S_\alpha = \frac{\alpha+2}{\alpha} S$ and $\mu_\alpha = \frac{\alpha+1}{\alpha}\mu$.

γ-Monotone distributions

Corollary 7.5 (γ-Monotone expectation inequalities)**.** *For any $\gamma \in \mathrm{N}_0$ we have the equality $B(\max\{0, \kappa_\Xi(x) - 1\}, \mathcal{M}_\gamma, \mu, S) =$*

$$\begin{aligned}
\inf \quad & \mathrm{Tr}\left\{\begin{pmatrix} Y & y \\ y^\top & y_0\end{pmatrix}\cdot\begin{pmatrix} S_\gamma & \mu_\gamma \\ \mu_\gamma^\top & 1\end{pmatrix}\right\} \\
\text{s.t.} \quad & (C_1), \\
& T_{1,i} b_i^2 q^{n+\gamma+1} + \left(2b_i T_{2,i} - \frac{n}{n+\gamma}\right) q^{n+\gamma} + (T_{3,i} + 1)\, q^{n+\gamma-1} - \\
& \qquad \frac{1}{B(n,\gamma)} \sum_{k=0}^{\gamma-1} \frac{(-1)^k}{(n+k)(n+k+1)} \binom{\gamma-1}{k} q^{\gamma-k-1} \ge 0, \quad \forall q \ge 1
\end{aligned}$$

where $S_\gamma \succ \mu_\gamma \mu_\gamma^\top$ for $S_\gamma = \frac{n+\gamma}{n}\frac{n+\gamma+1}{n+1} S$ and $\mu_\gamma = \frac{n+\gamma}{n}\mu$.

As mentioned in the beginning of this section, the polynomial inequalities appearing in Corollaries 7.2 to 7.5 admit exact SDP representations based on sum-of-squares representations as stated in Theorem 4.2. Standard software tools, such as `YALMIP` of Löfberg [81], are available which implement this transformation automatically. We did not state the resulting SDP constraints explicitly as they offer no further insight and would only clutter the statement of previous corollaries further.

7.4 Numerical examples

We illustrate the optimal inequalities presented in this paper by bounding the value of European stock portfolios inspired by Bertsimas and Popescu [17] and by computing worst-case bounds when aggregating random variables with known marginal information as done in Embrechts et al. [47]. The resulting SDP problems are implemented in `Matlab` using the interface `YALMIP` and solved numerically using `SDPT3`.

7.4.1 Optimal pricing of stock portfolios

In this example we are interested in finding an upper bound on the price of a European stock option with random pay-off

$$\Phi(\xi) := \max\{0,\ a^\top \xi - k\} = k\,(\kappa_\Xi(\xi) - 1)^+$$

for $\Xi = \left\{x \in \mathrm{R}^n \ :\ a^\top x \leq k\right\}$ similar as in Bertsimas and Popescu [17]. This option allows its holder to buy a portfolio $a \in \mathrm{R}^n$ of stocks at a price $k \in \mathrm{R}_+$ at maturity. The payoff Φ is hence positive if the uncertain value $\xi \in \mathrm{R}^n$ of the stocks at maturity in the portfolio $a \in \mathrm{R}^n$ exceeds the negotiated price $k \in \mathrm{R}_+$. If the price of portfolio of stocks $a^\top \xi$ in the market at maturity is less then k, then the holder will not exercise his right to buy the stock portfolio at price k.

When we denote with $\mathbb{P}^\star$ the distribution of ξ, then for the issuer of the option it is of interest to know

$$p := \sup_{\mathbb{P} \in \mathcal{C}}\ \mathbf{E}_{\mathbb{P}}[\Phi(\xi)]$$

for $\mathcal{C}$ a set of probability distributions for which the option issuer is convinced that $\mathbb{P}^\star \in \mathcal{C}$. Indeed, the issuer would like to demand a price of the stock option buyer which exceeds p, as in this case he or she is convinced that on average a profit is made.

In the remainder of this section, we assume that our portfolio $\xi = (\xi_{\texttt{IBM}}, \xi_{\texttt{APPLE}})$ consists of $a = (1,1)^\top$ an equal part of `IBM` and `APPLE` stocks. The stock holder is convinced that the distribution of ξ satisfies

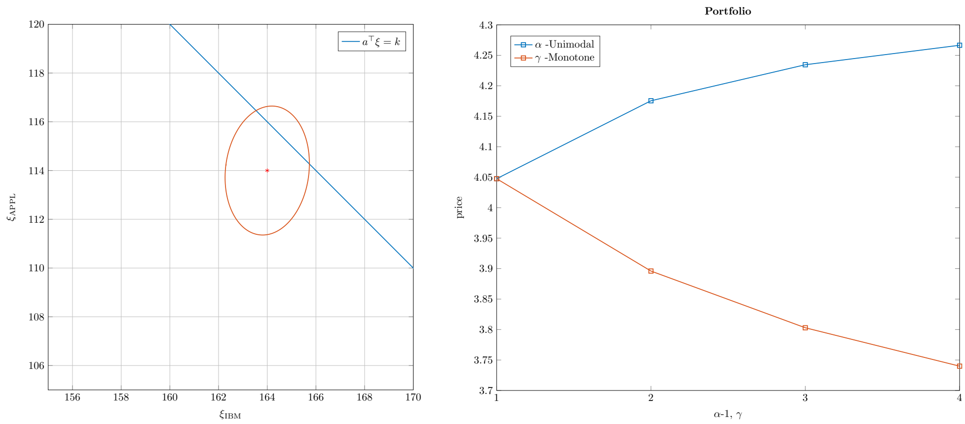
$$\mathbb{P}^\star \in \mathcal{C}\left(\begin{pmatrix}164\\114\end{pmatrix}, \begin{pmatrix}20 & 5\\ 5 & 60\end{pmatrix} + \begin{pmatrix}164\\114\end{pmatrix}\begin{pmatrix}164\\114\end{pmatrix}^\top\right)$$

for a strike price at maturity $k = 280$. This situation is sketched in Figure 7.2(a). The stock holder is also convinced that the distribution of ξ should be well-behaved and has a mode which coincides with its mean. In Figure 7.2(b), the optimal price p is given when the stock holder believes that either $\mathbb{P}^\star \in \mathcal{M}_\gamma$ or $\mathbb{P}^\star \in \mathcal{U}_\alpha$ in function of $\gamma \in \{1, \ldots, 5\}$ and $\alpha \in \{2, \ldots, 6\}$. As remarked before the bounds converge to either the bounds for arbitrary probability distributions when $\alpha \to \infty$ or completely monotone distributions in case $\gamma \to \infty$.

7.4.2 Factor models in insurance

As reported by Embrechts et al. [47], insurance companies most commonly model the size of claims ξ_i incurred as a result of different types of insurance policies separately from another. The claims ξ_i factor the total claim

$$S_n := \textstyle\sum_{i=1}^n \xi_i$$



(a) (b)

Figure 7.2: Optimal pricing of a portfolio containing an equal amount of `IBM` and `APPLE` stocks. Figure 7.2(a) indicates the distribution of $(\xi_{\texttt{IBM}}, \xi_{\texttt{APPLE}})$ visually. The red half line indicates realizations beyond which a profit is made.

as a sum of n separate claims ξ_i without a specified dependence structure. The problem of quantifying a certain statistic of $L(S_n)$ for a given loss function L based on (partial) marginal information of the distributions of the factors ξ_i is denoted by Rüschendorf [111] as a Fréchet problem.

We consider a portfolio containing four types of insurance policies, i.e. car, life, fire and medical insurances. We will assume that only information on the means $\mu_i := \mathbf{E}_{\mathbb{P}}[\xi_i]$ and second moments $s_i^2 := \mathbf{E}_{\mathbb{P}}\left[\xi_i^2\right]$ of the size of the corresponding insurance claims is given. Suppose we are interested in large aggregate claims S_n occurring with probability at most $\epsilon = 5\%$, where that part of the claim S_n exceeding the threshold $k = 150.000$ CHF is covered by a reinsurer. In what follows we therefore consider the problem of quantifying the least upper bound on the conditional value at risk $\mathrm{CVaR}_\epsilon\left(L(S_n)\right)$, where

$$L(S_n) = \min\left(\max\left(S_n, 0\right), k\right),$$

using only the marginal means μ_i and second moments $s_i^2 = \sigma_i^2 + \mu_i^2$ as given in Table 7.1. Additionally, it is assumed that the joint probability distribution $\mathbb{P}$ of $(\xi_1, \dots, \xi_4)$ is star unimodal. We are hence interested in the worst-case CVaR problem

$$\begin{array}{rl} \sup_{\mathbb{P}\in\mathcal{K}} & \mathbb{P}\text{-}\mathrm{CVaR}_\epsilon\left(L(S_n)\right) \\ \text{s.t.} & \displaystyle\int x_i\,\mathbb{P}(\mathrm{d}x) = \mu_i, \quad \forall i \\ & \displaystyle\int x_i^2\,\mathbb{P}(\mathrm{d}x) = s_i^2, \quad \forall i. \end{array} \tag{7.18}$$

The worst-case CVaR problem can be reduced to a worst-case expectation problem as indicated in Section 5.4.2 using the golden search method for the outer minimization problem over $\beta \in [0, k]$. Note that as the off-diagonal entries of the second moment matrix S are not given, the dual formulation (7.1) of the resulting worst-case expectation bound must be adapted slightly by requiring that the off-diagonal elements of the dual variable Y are zero.

According to Theorem 7.1 the resulting worst-case expectation bounds over the set of all star unimodal distributions can be transformed to equivalent worst-case expectation bounds over the standard probability simplex. In Appendix B we show that the resulting transformed loss function L_s is in the form required by Theorem 7.2. The worst-case CVaR bound (7.18) thus admits a tractable reformulation.

CHF	Average μ_i	Standard deviation σ_i
Car insurance	15.000	2.000
Life insurance	7.000	1.000
Fire insurance	3.000	5.000
Medical insurance	20.000	2.000

Table 7.1: Marginal means and standard deviations of the size of the claims incurred by the four types of insurance policies in the portfolio.

The worst-case excepted aggregate claim above the 5th percentile, i.e. $\mathrm{CVaR}_\alpha\left(L(S_n)\right)$, was numerically determined to be 123.325 CHF in approximately 15 seconds using `Matlab` on a PC[1] operated by `Debian GNU/Linux 7 (wheezy)`.

[1]An `Intel(R) Core(TM) Xeon(R) CPU E5540 @ 2.53GHz` machine.

Part III

Robust optimization and control

8 Robust optimization with second-order moment information

Robust optimization is understood to be a methodology for optimization despite uncertainty. Consider for the moment the uncertain optimization problem

$$\begin{aligned} \inf \quad & f(u) \\ \text{s.t.} \quad & u \in C, \\ & x(u,\xi) \in X \end{aligned} \tag{8.1}$$

where the vector u in R^d is the decision variable. The uncertain outcome $x(u,\xi)$ valued in R^c depends both on the decision u and a random influence ξ valued in R^n. The function f will be referred to as the objective function and the final constraint in (8.1) is denoted as the uncertain constraint. The fact that the final constraint in the optimization problem (8.1) is uncertain signals that in many real-world problems constraints are imposed on outcomes $x(u,\xi)$ which depend on uncertain events ξ which can not be influenced by the decision maker.

In our approach the uncertain influence ξ is taken to be governed by a probability distribution that itself is only partially known or ambiguous. The distribution $\mathbb{P}$ of the uncertain data ξ is merely assumed to belong to an ambiguity set $\mathcal{C}$ which ideally should be composed of all distributions consistent with the information available on the uncertain data ξ. The decision maker must hence take a decision u which remains feasible, whatever the distribution of the exogenous influence ξ within the ambiguity set $\mathcal{C}$.

We consider two types of such distributionally robust constraints. In the first case, we require that the uncertain constraint $x(u,\xi) \in X$ holds with a given probability for all probability distributions in the ambiguity set $\mathcal{C}$. These constraints are commonly referred to as distributionally robust chance constraints. In the second case, we impose CVaR constraints to bound the expected violation of the uncertain constraint $x(u,\xi) \in X$ for all probability distributions within $\mathcal{C}$. Such constraints are referred to as distributionally robust CVaR constraints.

We will consider in the last chapter of the dissertation the dynamic counterpart of the static optimization problem (8.1) as well. In the dynamic setting the uncertain outcome $x(u,\xi)$ shall represent the state of a linear dynamical system that depends both on previous control inputs u and exogenous disturbances ξ.

8.1 Uncertain constraints

There are several common approaches on how one can go about putting the ambiguous requirement $x(u,\xi) \in X$ on a mathematically sound basis. At this point the uncertain constraint is intentionally not made mathematically rigorous yet. In this chapter we will consider two types of distributionally robust formulations for the uncertain constraint $x(u,\xi) \in X$. We are particularly interested in characterizing under which conditions either formulation is sensible and amendable to practical computation.

Chance constraints are a popular means of modeling constraints on uncertain outcomes that need only to hold with a certain probability. Formally, the requirement that the random outcome $x(u,\xi)$ should be contained in the constraint set X with high probability is expressed as

$$\mathbb{P}(x(u,\xi) \in X) \geq 1 - \epsilon, \tag{8.2}$$

where $\epsilon \in (0,1)$ is a prescribed safety parameter that controls the level of acceptable constraint violations. The probability distribution $\mathbb{P}$ denotes here the true but possibly unknown probability distribution of the random variable ξ.

Chance constraints are often more practical than the worst-case robust constraints discussed in Chapter 1, which can be viewed as degenerate chance constraints with $\epsilon = 0$ and which tend to encourage overly conservative decisions. More importantly, in many practical problems worst-case state constraints typically become infeasible in the presence of an unbounded (e.g. Gaussian) disturbance ξ.

In spite of their conceptual appeal, chance constraints have not yet found wide application for a variety of reasons. On the one hand, the feasibility of a chance constraint can only be checked if the distribution $\mathbb{P}$ of the random vector ξ is precisely known. In practice, however, almost invariably this distribution must be estimated from noisy data and is therefore itself subject to ambiguity. This is problematic because, as shown by Zymler et al. [142], even small changes in the distribution can have a dramatic impact on the geometry and size of the set of decision variables u feasible within the chance constraint (8.2). Moreover, incorporating chance constraints into otherwise tractable optimization problems typically results in a non-convex problem, and consequently to computational intractability.

Finally, chance constraints of the type (8.2) bound the probability of constraint violation but do not impose any restrictions on the degree of the violations encountered. However, severe constraint violations, i.e. scenarios in which the uncertain outcome $x(u,\xi)$ strays far outside of the constraint set X, are often much more harmful than mild violations in which the uncertain outcome $x(u,\xi)$ remains close to the boundary of X. Chance constraints hence fail to distinguish between these two situations and provide no mechanism to penalize severe constraint violations relative to mild ones.

In order to address these deficiencies, we first require some terminology and notation. We will assume throughout this chapter that the constraint set X is characterized by the intersection of zero sublevel sets of finitely many convex functions $\ell_i : \mathrm{R}^c \to \mathrm{R}$, so that

$$X := \{x \in \mathrm{R}^c \ : \ \ell_i(x) < 0, \ \forall i \in [1, \ldots, k]\}.$$

We will refer to the functions ℓ_i as loss functions. We refer to (8.2) as a single chance constraint if $k = 1$ and as a joint chance constraint if $k > 1$. Every joint chance constraint can easily be reduced to an individual chance constraint by reexpressing X as $\{x \in \mathrm{R}^c \ : \ L^\alpha(x) < 0\}$, where the aggregate loss function

$$L^\alpha(x) := \max_{i \in [1,\ldots,k]} \alpha_i \ell_i(x) \tag{8.3}$$

remains convex in x and depends on a set of strictly positive scaling parameters $\alpha \in \mathrm{R}^k_{++}$. Note that the particular choice of α has no impact on the zero sublevel set of the function L^α, and consequently no impact on the set X or the associated chance constraint (8.2). The reader may therefore regard α initially as a positive parameter that can be chosen arbitrarily. However, the flexibility to select α will be useful either to control the tightness of a tractable approximation of the chance constraint (8.2), or to penalize the degree of constraint violation of statistical outliers in (8.2) as discussed further in this chapter.

We assume that the parameter $\alpha \in \mathrm{R}^k_{++}$ is given, either as an attempt to approximate a chance constraint or as an indicator of the relative importance of the loss severity measures ℓ_i.

Assumption 8.1. *The aggregated convex loss function $L^\alpha : \mathrm{R}^c \to \mathrm{R}$ with $\alpha \in \mathrm{R}^k_{++}$ is given as*

$$L^\alpha(x) = \max_{i\in[1,\ldots,k]} \left[\alpha_i \left(a_i^\top x - b_i\right)\right],$$

where $a_i \in \mathrm{R}^c$, $b_i \in \mathrm{R}$.

Recall that the constraint set X corresponds to the zero sub level set of the loss function L^α, and consequently is assumed to be polytopic in this chapter.

Throughout the chapter we will exploit an interesting connection between chance constraints of the type (8.2) and the VaR risk measure defined in Section 3.4. We emphasize that the *value*-at-risk in our particular context typically relates to the degree of violation of some physical constraint, and is unrelated to the loss of economic currency as in the usual interpretation in economics. In the context of this dissertation, *violation*-at-risk might therefore be a more appropriate interpretation.

By construction, the VaR measure of an uncertain quantity coincides with the $(1-\epsilon)$-quantile of the probability distribution of that quantity. Moreover, the reader may easily verify that the chance constraint (8.2) can be reformulated as a constraint on the VaR at level ϵ of the aggregate loss function $L^\alpha(x)$, that is,

$$\mathbb{P}\text{-VaR}_\epsilon\left(L^\alpha(x(u,\xi))\right) \leq 0 \quad \Longleftrightarrow \quad \mathbb{P}(x(u,\xi) \in X) \geq 1-\epsilon. \tag{8.4}$$

A major deficiency of the VaR is its non-convexity. In fact, it is well known that the function $\mathbb{P}\text{-VaR}_\epsilon(L^\alpha(x(u,\xi)))$ is generally non-convex in x even for linear loss functions L^α. A commonly employed alternative, convex, risk measure closely related to the VaR is the CVaR.

The CVaR enjoys a number of practical advantages over VaR in this context, since it is monotone, homogeneous and convex with respect to the loss function L^α. In addition, it represents a conservative (upper) approximation to VaR, and consequently a conservative means of approximating chance constraints. Indeed, it is easily shown that

$$\mathbb{P}^\star\text{-CVaR}_\epsilon\left(L^\alpha(x(u,\xi))\right) \leq 0 \quad \Longrightarrow \quad \mathbb{P}^\star(x(u,\xi) \in X) \geq 1-\epsilon. \tag{8.5}$$

Note that for convex loss functions L^α the set of all random outcomes $x(u,\xi)$ satisfying the CVaR constraint in (8.5) is convex due to the convexity and monotonicity of the CVaR.

In economic theory, CVaR traditionally measures an economic loss, hence the function L^α is specified ab initio. In optimization practice however, one is typically given a constraint set X and is free to select any loss functions ℓ_i compatible with X, i.e. one can choose any ℓ_i satisfying $X = \{x \in \mathrm{R}^c \;:\; L^\alpha(x) < 0\}$. The choice of the positive weights α_i can then be used to indicate the relative importance of the individual loss functions ℓ_i, i.e. the level of significance that the decision maker attaches to the degree of violation of individual constraints in the event that they occur.

CVaR constraints address the principal shortcoming of chance constraints. Indeed, CVaR constraints impose a higher penalty on realizations of x that materialize far outside the constraint set X (i.e. with $L^\alpha(x) \gg 0$) and therefore penalize severe constraint violations more aggressively than mild ones. In contrast, chance constraints impose uniform penalties on all constraint violations irrespective of their degree of infeasibility.

Unfortunately, checking the feasibility of CVaR constraints still requires precise knowledge of the probability distribution $\mathbb{P}$ of the random variable ξ. In practice, only limited information about $\mathbb{P}$ may be available, such as the support or some descriptive distributions of the location and dispersion of random variables under $\mathbb{P}$. Abstractly, we can represent the limited available information about $\mathbb{P}$ by an ambiguity set $\mathcal{C}$ of probability distributions with the following properties: (i) It is known that $\mathbb{P} \in \mathcal{C}$, and (ii) $\mathcal{C}$ is the smallest set of probability distributions for which we can guarantee that $\mathbb{P} \in \mathcal{C}$.

In order to facilitate statements about computational tractability, we require some structural assumptions concerning the ambiguity set $\mathcal{C}$. As in the preceding chapters, we will assume that the distributional information regarding the random variable ξ is limited to second-order moment information. We henceforth assume that the ambiguity set

$$\mathcal{C} = \mathcal{H}(0, \Sigma) \cap \mathcal{K}$$

contains all zero mean probability distributions having a variance matrix $\Sigma \in \mathrm{S}^n_+$. For the sake of exposition, we assumed without loss of generality that the random variable ξ has zero mean $\mu = 0$. This can always be achieved by considering an appropriate coordinate translation. Furthermore, in what follows we will take the usual assumption that the set $\mathcal{K}$ is a Choquet star simplex with generating univariate distribution $\mathbb{T}$. The set $\mathcal{K}$ can again be used to model any further structural information regarding the distribution of ξ such as unimodality or monotonicity.

To immunize the chance constraint (8.2) against distributional ambiguity, we can require that it should hold for each probability distribution in the ambiguity set $\mathcal{C}$. The resulting distributionally robust chance constraint can be represented as

$$\mathbb{P}(x(u,\xi) \in X) \geq 1 - \epsilon \quad \forall \mathbb{P} \in \mathcal{C} \iff \inf_{\mathbb{P} \in \mathcal{C}} \mathbb{P}(x(u,\xi) \in X) \geq 1 - \epsilon. \tag{8.6}$$

Similarly, recalling that $X = \{x \in \mathrm{R}^c : L^\alpha(x) < 0\}$ for any $\alpha \in \mathrm{R}^k_{++}$, we can immunize the CVaR constraint on the left hand side of (8.5) against distributional ambiguity as well. The resulting distributionally robust CVaR constraint takes the form

$$\mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x(u,\xi))) \leq 0 \ \forall \mathbb{P} \in \mathcal{C} \iff \sup_{\mathbb{P} \in \mathcal{C}} \mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x(u,\xi))) \leq 0. \tag{8.7}$$

As in the classical setting without distributional ambiguity, it can be shown that (8.7) provides a conservative approximation for (8.6); see Chen et al. [37] or Zymler et al. [142]. In other words,

$$\sup_{\mathbb{P} \in \mathcal{C}} \mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x(u,\xi))) \leq 0 \implies \inf_{\mathbb{P} \in \mathcal{C}} \mathbb{P}(x(u,\xi) \in X) \geq 1 - \epsilon. \tag{8.8}$$

From the preceding two characterizations (8.6) and (8.7) the relationship between distributionally robust constraints and uncertainty quantification becomes very clear. Indeed, the problem of checking whether a fixed decision u is feasible in either formulation reduces to an uncertainty quantification problem of the type discussed in Chapters 6 and 7. This chapter will hence benefit greatly from the results made in the preceding part of the dissertation.

In the remainder of the chapter, we will assume that the uncertain outcome $x(u,\xi)$ depends affinely on both its arguments separately. Specifically, we will assume that $x(u,\xi)$ can be written in the canonical form

$$x(u,\xi) = A(u) + B(u)\xi \tag{8.9}$$

where $A(u) : \mathrm{R}^d \to \mathrm{R}^c$ and $B(u) : \mathrm{R}^d \to \mathrm{R}^{c \times n}$ are both affine functions of the decision variable. We remark that in a large class of decision problems in engineering and finance the uncertain outcome can be represented in the form suggested in (8.9).

As the zero mean random variable ξ can be interpreted as a deviation, the vector valued term $A(u)$ can be seen to represent the nominal outcome $x(u, 0)$ perturbed linearly in ξ in accordance with the matrix $B(u)$. Following the previous interpretation, both distributionally robust formulations (8.6) and (8.7) provide hence a robust counterpart to the nominal constraint

$$x(u, 0) \in X \iff A(u) \in X \tag{8.10}$$

in which the random variable ξ has been replaced by its expectation. It is easily argued that for any ambiguity set in the form $\mathcal{C} = \mathcal{H}(0, \Sigma) \cap \mathcal{K}$, both the corresponding distributionally robust chance constraint (8.6) and the distributionally robust CVaR constraint (8.7) have the desirable property of implying the nominal constraint (8.10).

8.2 Single uncertain constraints

As reflected in Assumption 9.2, we will focus in this chapter on uncertain constraints $x(u, \xi) \in X$ for polytopic sets X represented as a finite number of half-space constraints. For uncertain constraints, the watershed between what is tractable and what remains intractable seems to be roughly between single and joint uncertain constraints as also noticed by Calafiore and El Ghaoui [30] and Zymler et al. [142]. As we will discuss in this chapter, many single uncertain constraints indeed admit exact tractable reformulations while their joint counterparts often lead to intractability and hence need approximation.

In this section, we will pay thus special attention to optimization problems of the type (8.1) where the uncertain constraint is in the particular form

$$x(u, \xi) \in X = \{x \in \mathrm{R}^c \ : \ \ell_1(x) < 0\}\,. \tag{8.11}$$

As discussed in previous section, the constraint (8.11) is referred to as a single uncertain constraint as the constraint set X is defined through a single loss function ℓ_1. Moreover, we required in Assumption 9.2 that the constraint function is affine and represented as $\ell_1(x) = a_1^\top x - b_1$. This means that the constraint set X in the uncertain constraint (8.11) consists of a single half-space constraint.

In this section we will be interested in conditions under which the distributionally robust chance constrained formulation (8.6) and the distributionally robust CVaR constrained formulation (8.7) for single uncertain constraints of the type (8.11) admit a tractable representation. It will be argued that in case of a single uncertain constraint an exact tractable reformulation can be found in either case in terms of a tractable SOC constraint.

It will be advantageous to the exposition in the remainder of this section to state the uncertain constraint (8.11) explicitly in terms of the random variable ξ. We have the following chain of equivalences

$$x(u, \xi) \in X \iff \ell_1(x(u, \xi)) < 0 \iff c(u)^\top \xi < d(u), \tag{8.12}$$

where the functions $c(u) : \mathrm{R}^d \to \mathrm{R}^n$ and $d(u) : \mathrm{R}^d \to \mathrm{R}$ depend affinely on the decision variable u and can be determined explicitly as $c(u) := B(u)^\top a$ and $d(u) := b - a^\top A(u)$.

8.2.1 Single chance constraints

The uncertain constraint (8.12) can be reformulated as a distributionally robust chance constraint as indicated in equation (8.6) of the preceding section. In the particular case of the single uncertain constraint (8.12), the distributionally robust chance constrained formulation reduces to the following condition on the decision u:

$$\mathbb{P}\left(c(u)^\top \xi < d(u)\right) \geq 1 - \epsilon, \quad \forall \mathbb{P} \in \mathcal{C}. \tag{8.13}$$

This single linear chance constraint (8.13) has already been studied extensively in the literature for several distinct types of ambiguity sets $\mathcal{C}$. Calafiore and El Ghaoui [30] show that they admit an exact tractable reformulation in the terms of a SOC constraint when the ambiguity set $\mathcal{C} = \mathcal{H}(0, \Sigma) \cap \mathcal{K}$ consists of all zero mean probability distributions with variance matrix Σ, i.e. $\mathcal{K} = \mathcal{P}_n$. As discussed at length in the preceding part of the thesis, the corresponding chance constraint can be quite pessimistic as the ambiguity set $\mathcal{C}$ in that case contains many probability distributions which in practice may not be very relevant. Yu et al. [139] made significant progress by providing an exact SOC representation for the distributionally robust chance constraint (8.13) in case $\mathcal{K}$ consists of symmetric linear unimodal distributions.

In this section we will analyze the single uncertain constraint (8.13) when the ambiguity set $\mathcal{C}$ is richly structured. As discussed previously, we consider the distributionally robust chance constraint (8.13) when the ambiguity set $\mathcal{K}$ admits a Choquet star representation in terms of the generating distribution $\mathbb{T}$ in which case

$$\mathrm{ex}\,\mathcal{K} = \{\mathbb{T}_x \ : \ x \in \mathrm{R}^n\}.$$

We will prove that the single distributionally robust chance constraint (8.13) admits a tractable SOC representation for any ambiguity set $\mathcal{K}$ admitting a Choquet star representation. We will make this somewhat abstract result concrete by considering the particular case of α-unimodal and γ-monotone distributions.

Before we state our main result concerning single distributionally robust chance constraints, we introduce first a function $f_{\mathbb{T}}$ defined through the following univariate uncertainty quantification problem:

$$\begin{aligned} f_{\mathbb{T}}(k) := \sup \quad & \mathbb{Q}(z \geq k) \\ \text{s.t.} \quad & \mathbb{Q} \in \mathcal{H}(0,1) \cap \mathcal{K}. \end{aligned} \tag{8.14}$$

The function $f_{\mathbb{T}}$ hence represents the worst-case probability of the event that a standardized univariate random variable z exceeds the threshold k when z is distributed in $\mathcal{K}$. We remark that the uncertainty quantification problem (8.14) was discussed in both the Chapters 6 and 7. The function $f_{\mathbb{T}}$ plays an important role in the reformulation of the chance constraint (8.13) as it will represent the influence of the structure imposed on the ambiguity set $\mathcal{C}$ through the set $\mathcal{K}$.

Proposition 8.1. *The single distributionally robust chance constraint* (8.13) *with ambiguity set* $\mathcal{C} = \mathcal{H}(0, \Sigma) \cap \mathcal{K}$ *admits an exact tractable representation in the terms of a convex SOC constraint*

$$b - a^\top A(u) \geq f_{\mathbb{T}}^{-1}(\epsilon) \left\| \Sigma^{1/2} B(u)^\top a \right\|_2, \tag{8.15}$$

where the function $f_{\mathbb{T}}$ *is defined as in* (8.14).

Proof. Define the univariate random variable $z_R := c(u)^\top \xi$. Because of the projection Theorem 3.4 for Choquet star representable sets, we have that the ambiguity concerning the distribution of the random variable z_R is given as the ambiguity set

$$\mathcal{H}(0, \sigma_R^2) \cap \mathrm{mix}\,\{\mathbb{T}_x \ : \ x \in \mathrm{R}\}$$

where $\sigma_R = \left\| \Sigma^{1/2} c(u) \right\|_2$. The chance constraint (8.13) can now equivalently be written in terms of the random variable z_R. We have indeed equivalently that

$$\mathbb{Q}\left(\frac{z_R}{\sigma_R} \geq \frac{d(u)}{\sigma_R} \right) \leq \epsilon, \quad \forall \mathbb{Q} \in \mathcal{H}(0, \sigma_R^2) \cap \mathrm{mix}\,\{\mathbb{T}_x \ : \ x \in \mathrm{R}\}.$$

As the random variable z_R/σ_R has zero mean and unit variance, the previous constraint is equivalent to $f_{\mathbb{T}}(d(u)/\sigma_R) \leq \alpha$. The function $f_{\mathbb{T}}(k)$ is evidently monotonically decreasing in k and thus admits the unique generalized inverse $f_{\mathbb{T}}^{-1}(\epsilon) = \inf\{k \in \mathrm{R} \ : \ f(k) \leq \epsilon\}$. The proposition follows now immediately from elementary manipulations. ☐

In what remains of this section we will show how the abstract result stated in Proposition 8.1 can be made concrete in the particular case of α-unimodal and γ-monotone distributions. Nevertheless we begin by discussing the case in which no structure is specified, i.e. $\mathcal{K} = \mathcal{P}_n$, as can be found in for instance in the work of Calafiore and El Ghaoui [30]. In this case the corresponding function $f_{\mathbb{T}}$ defined though the uncertainty quantification problem (8.14) even admits a closed form expression.

Unstructured probability distributions As promised we start by showing how the results of Calafiore and El Ghaoui [30] on distributionally robust chance constraints with unstructured probability distributions $\mathcal{K} = \mathcal{P}_n$ can be seen as a simple corollary of Proposition 8.1. Under previously mentioned conditions

$$\begin{aligned} f_\delta(k) = \sup \quad & \mathbb{Q}\left[z \geq k\right] \\ \text{s.t.} \quad & \mathbb{Q} \in \mathcal{H}(0,1) \cap \mathcal{P}_n. \end{aligned}$$

The last uncertainty quantification problem can be seen as the unilateral counterpart of the classical Chebyshev inequality (1.6). Indeed, $f_\delta(k)$ is the probability that a standardized random variable exceeds a certain threshold k and was studied already by Cantelli [34]. The classical Cantelli inequality establishes that

$$f_\delta(k) = \begin{cases} \frac{1}{1+k^2} & \text{if } k \geq 0 \\ 1 & \text{otherwise} \end{cases}$$

admitting the inverse $f_\delta^{-1}(\epsilon) = \sqrt{(1-\epsilon)/\epsilon}$.

Unimodal and monotone distributions For general Choquet star representable sets $\mathcal{K}$, the uncertainty quantification problem (8.14) does not seem to admit closed form expressions. Nevertheless, explicit tractable reformulations were provided for the function

$$f_{\mathbb{T}}(k) = B(\mathbf{1}\left\{z \geq k\right\}, \mathcal{K}, 0, 1)$$

in the Chapters 6 and 7.

In particular when $\mathcal{K}$ consists of the set of all α-unimodal distributions $\mathcal{U}_\alpha$, the function $f_{\mathbb{T}}(k)$ reduces to the worst-case probability bound $G_\alpha(0,1)$ for the polytopic set $\Xi = \{x \in \mathrm{R} \, : \, x \geq k\}$ and admits the tractable reformulation provided in Theorem 6.4. Alternatively when working with γ-monotone distributions $\mathcal{K} = \mathcal{M}_\gamma$, the corresponding worst-case probability bound $B(\mathbf{1}\left\{z \geq k\right\}, \mathcal{M}_\gamma, 0, 1)$ can be readily solved as described in Corollary 7.3.

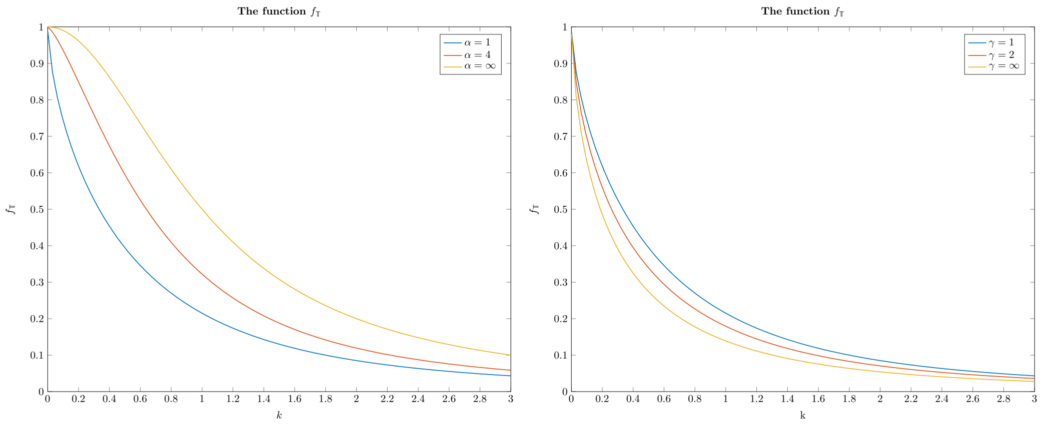
In either case, the function $f_{\mathbb{T}}$ can be evaluated by solving a small tractable convex optimization problem. For all practical purposes, these convex optimization problems provide a de facto closed form expression for the corresponding functions $f_{\mathbb{T}}$. The results are made concrete in case of unimodal and monotone distributions in Figure 8.1.

8.2.2 Single CVaR constraints

The uncertain constraint (8.12) can also be reformulated as a CVaR constraint instead as indicated in equation (8.7) of the preceding section. In the particular case of the single uncertain constraint (8.12) the distributionally robust CVaR constrained formulation in equation (8.7) reduces to following condition on the decision u:

$$\mathbb{P}\text{-CVaR}_\epsilon\left(\ell_1\left(x(u,\xi)\right)\right) \leq 0, \quad \forall \mathbb{P} \in \mathcal{C} \tag{8.16}$$

independent of the choice of $\alpha_1 \in \mathrm{R}_{++}$ as the CVaR is homogenous. Similar to the distributionally robust chance constrained formulation (8.13), the previous single CVaR constraint (8.16) has



(a) α-unimodal (b) γ-monotone

Figure 8.1: The function $f_{\mathbb{T}}$ for α-unimodal and γ-monotone distributions in R. The case $\alpha = \infty$ denotes the function f_δ corresponding no structural information, i.e. $\mathcal{K} = \mathcal{P}_n$. Similarly, the case $\alpha = \infty$ denotes the function $f_{\mathbb{T}}$ for completely monotone distributions in R.

been studied by Calafiore and El Ghaoui [30] albeit to a far lesser extend. Nevertheless, they are known to admit an exact tractable reformulation again in the form of a SOC constraint when the ambiguity set $\mathcal{C} = \mathcal{H}(0, \Sigma) \cap \mathcal{K}$ consists of all zero mean probability distributions with variance Σ, i.e. the ambiguity set $\mathcal{C}$ obtained with the particular choice $\mathcal{K} = \mathcal{P}_n$. For reasons similar to the distributionally robust chance constrained formulation, the corresponding distributionally robust CVaR reformulation can be quite pessimistic. Yu et al. [139] presents an exact SOC representation for the distributionally robust CVaR constraint (8.16) in case the set $\mathcal{K}$ consists of symmetric linear unimodal distributions.

We will set out to study the distributionally robust CVaR constraint (8.16) when the ambiguity set $\mathcal{C}$ is more richly structured. In particular, we will consider again the situation in which the ambiguity set $\mathcal{K}$ admits a Choquet star representation in terms of the generating distribution $\mathbb{T}$ in which case

$$\mathrm{ex}\,\mathcal{K} = \{\mathbb{T}_x \ : \ x \in \mathrm{R}^n\}\,.$$

We will first prove that the single distributionally robust CVaR constraint (8.16) admits a tractable SOC representation for any ambiguity set $\mathcal{K}$ admitting a Choquet star representation. We will make this abstract result concrete by considering the particular cases of α-unimodal and γ-monotone distributions.

Before we state the main result concerning distributionally robust CVaR constraints, we introduce first a function $g_{\mathbb{T}}$ defined through a univariate uncertainty quantification problem:

$$\begin{aligned} g_{\mathbb{T}}(\epsilon) := \sup \quad & \mathbb{Q}\text{-CVaR}_\epsilon(z) \\ \text{s.t.} \quad & \mathbb{Q} \in \mathcal{H}(0,1) \cap \mathcal{K}. \end{aligned} \tag{8.17}$$

It will be indicated shortly hereafter that the function $g_{\mathbb{T}}$ plays the same role in the reformulation of the distributionally robust CVaR constraint as the function $f_{\mathbb{T}}^{-1}$ played in Proposition 8.1.

The function $g_{\mathbb{T}}(\epsilon)$ represents the worst-case CVaR at level ϵ of a standardized univariate random variable z distributed in $\mathcal{K}$. The function $g_{\mathbb{T}}(\epsilon)$ is evidently monotonically decreasing in ϵ. We remark that previous uncertainty quantification problem (8.17) was discussed in Chapter 7. The function $g_{\mathbb{T}}$ plays an important role in the reformulation of the distributionally robust CVaR constraint (8.16) as it will represent the influence of the structure imposed on the ambiguity set $\mathcal{C}$ through the set $\mathcal{K}$.

Proposition 8.2. *The single distributionally robust CVaR constraint* (8.16) *with ambiguity set* $\mathcal{C} = \mathcal{H}(0, \Sigma) \cap \mathcal{K}$ *admits an exact representation in terms of a convex SOC constraint*

$$b - a^\top A(u) \geq g_{\mathbb{T}}(\epsilon) \left\| \Sigma^{1/2} B(u)^\top a \right\|_2, \tag{8.18}$$

where the function $g_{\mathbb{T}}$ *is defined as in* (8.17)*.*

Proof. Define the univariate random variable $z_R := c(u)^\top \xi$. Because of the projection Theorem 3.4 for Choquet star representable sets, we have that the ambiguity concerning the distribution of the random variable z_R is given as the ambiguity set

$$\mathcal{H}(0, \sigma_R^2) \cap \text{mix}\left\{\mathbb{T}_x \ : \ x \in \mathrm{R}\right\},$$

where $\sigma_R = \left\|\Sigma^{1/2} c(u)\right\|_2$. The CVaR constraint can equivalently be written with in terms of the random variable z_R. We have indeed equivalently that

$$\mathbb{Q}\text{-CVaR}_\epsilon \left(\frac{z_R}{\sigma_R} - \frac{d(u)}{\sigma_R} \right) \leq 0, \quad \forall \mathbb{Q} \in \mathcal{H}(0, \sigma_R^2) \cap \text{mix}\left\{\mathbb{T}_x \ : \ x \in \mathrm{R}\right\}$$

or

$$\mathbb{Q}\text{-CVaR}_\epsilon \left(\frac{z_R}{\sigma_R} \right) \leq \frac{d(u)}{\sigma_R}, \qquad \forall \mathbb{Q} \in \mathcal{H}(0, \sigma_R^2) \cap \text{mix}\left\{\mathbb{T}_x \ : \ x \in \mathrm{R}\right\}$$

using the homogeneity of the CVaR measure. As the random variable z_R/σ_R has zero mean and unit variance, the previous constraint is equivalent to $g_{\mathbb{T}}(\epsilon) \leq d(u)/\sigma_R$ from which the proposition follows immediately. □

In what remains of this section we will show how the result stated in Proposition 8.2 can be made concrete in the particular case of α-unimodal and γ-monotone distributions. Nevertheless, we start again by discussing the case in which no structure is specified, i.e. $\mathcal{K} = \mathcal{P}_n$ as found for instance in the works of Calafiore and El Ghaoui [30] or Zymler et al. [142]. In this case the corresponding function $g_{\mathbb{T}}$ defined though the uncertainty quantification problem (8.17) admits again a closed form expression.

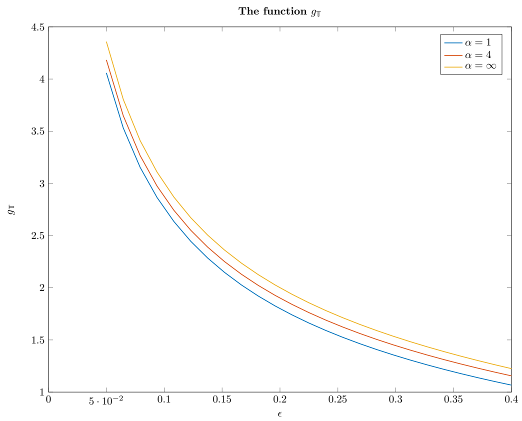
Unstructured probability distributions As promised we start by showing how the results of Calafiore and El Ghaoui [30] and Zymler et al. [142] on distributionally robust CVaR constraints with unstructured probability distributions $\mathcal{K} = \mathcal{P}_n$ can be seen as a simple corollary of Proposition 8.2. Under aforementioned conditions

$$\begin{aligned} g_\delta(\epsilon) = \sup \quad & \mathbb{Q}\text{-CVaR}_\epsilon(z) \\ \text{s.t.} \quad & \mathbb{Q} \in \mathcal{H}(0,1) \cap \mathcal{P}_n. \end{aligned}$$

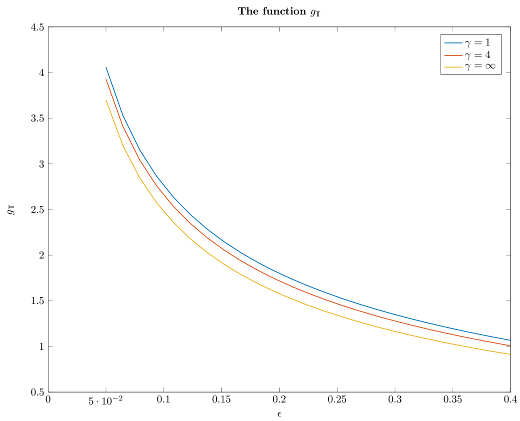
The previous worst-case CVaR problem is shown by Yu et al. [139] to in fact admit a closed form expression which is given as

$$g_\delta(\epsilon) = \sqrt{\frac{1-\epsilon}{\epsilon}}.$$

We remark here that in the unstructured case $\mathcal{K} = \mathcal{P}_n$, the distributionally robust chance constraint and CVaR constraint reformulations are equivalent. Indeed, we have that $g_\delta = f_\delta^{-1}$ and both chance and CVaR reformulations are seen to be equivalent by merit of Propositions 8.1 and 8.2. This remarkable observation was made before by Zymler et al. [142], where it was in fact shown to hold for any single uncertain constraint with a quadratically representable loss function ℓ_1. This equivalence between distributionally robust chance and CVaR constraints is



(a) α-unimodal



(b) γ-monotone

Figure 8.2: The function $g_{\mathbb{T}}$ for α-unimodal and γ-monotone distributions in R. The case $\alpha = \infty$ denotes the function g_δ corresponding no structural information, i.e. $\mathcal{K} = \mathcal{P}_n$. Similarly, the case $\alpha = \infty$ denotes the function $g_{\mathbb{T}}$ for completely monotone distributions in R.

however peculiar to the situation $\mathcal{K} = \mathcal{P}_n$ as shown by a counterexample in Yu et al. [139]. Instead, we have the inequality

$$f_{\mathbb{T}}^{-1}(\epsilon) \leq g_{\mathbb{T}}(\epsilon)$$

for any $\epsilon \in (0, 1)$ as a direct consequence of Propositions 8.1 and 8.2 in conjunction with the implication (8.8).

Unimodal and monotone distributions For general Choquet star representable sets $\mathcal{K}$, the uncertainty quantification problem (8.17) does unfortunately not seem to admit closed form expressions. As indicated in Section 5.4.2 the worst-case CVaR can be reduced to a worst-case expectation bound

$$g_{\mathbb{T}}(k) = \min_{\beta} \; \beta + \frac{1}{\epsilon} B((z-\beta)_+, \mathcal{K}, 0, 1).$$

Tractable reformulations of the inner problem $B((z-\beta)_+, \mathcal{K}, 0, 1)$ can be found in Chapter 7. As mentioned before, the outer minimization problem over β is convex and can be done using any univariate optimization method such as those found in Kiefer [68].

In particular when $\mathcal{K} = \mathcal{U}_\alpha$ consists of the set of all α-unimodal distributions, the worst-case bound $B((z-\beta)_+, \mathcal{U}_\alpha, 0, 1)$ admits the tractable reformulation provided in Corollary 7.4. Alternatively, when working with γ-monotone distributions $\mathcal{K} = \mathcal{M}_\gamma$ the corresponding worst-case bound $B((z-\beta)_+, \mathcal{M}_\gamma, 0, 1)$ can be readily solved as described in Corollary 7.5.

In either case, the function $g_{\mathbb{T}}$ can be evaluated by solving a small number of tractable convex optimization problems. For practical purposes, the convex optimization problems provide a de facto closed form solution for the corresponding functions $g_{\mathbb{T}}$. The results are made concrete in case of unimodal and monotone distributions in Figure 8.2.

8.3 Joint uncertain constraints

In the last part of the chapter we will return back to the more general problem of joint uncertain constraints. In the case of single uncertain constraints discussed in the previous section the difference between distributionally robust chance and CVaR constraints ran only skin deep. Indeed, both the single distributionally robust chance and CVaR constraints were shown to admit similarly structured SOC representations by virtue of Propositions 8.1 and 8.2, respectively. Joint uncertain constraints deepen the divide between the distributionally robust chance formulation and the distributionally robust CVaR formulation significantly. Where distributionally robust CVaR constraints are often computationally tractable, Zymler et al. [142] argued that their chance constraint counterpart is often not. In the present section we will outline when exact tractable reformulations are available, and when not, what type of approximation can be used instead.

In what remains we will consider an uncertain optimization problem of the type (8.1) with joint uncertain constraints of the particular form

$$x(u,\xi) \in X = \{x \in \mathrm{R}^c \ : \ L^\alpha(x) < 0\}\,. \tag{8.19}$$

As discussed in Section 8.1, the final constraint in the uncertain problem (8.19) is denoted as a joint chance constraint, as the set X is defined now through multiple loss functions ℓ_i defining the aggregated loss function $L^\alpha(x) := \max_{i\in[1,\ldots,k]} \alpha_i \ell_i(x)$. Assumption 9.2 requires each loss function $\ell_i(x) = a_i^\top x - b_i$ to be affine. The corresponding constraint set X is thus necessarily polytopic. Again we will be interested in the conditions under which the distributionally robust chance constrained formulation (8.6) and the distributionally robust CVaR constrained formulation (8.7) of the uncertain constraint (8.19) admit a tractable representation.

It will benefit the exposition in the remainder of this section to state the uncertain constraint of the optimization problem (8.19) explicitly in terms of the random variable ξ. We have the following chain of equivalences

$$x(u,\xi) \in X \iff L^\alpha(x(u,\xi)) < 0 \iff C(u)^\top \xi < D(u), \tag{8.20}$$

where the functions $C(u) : \mathrm{R}^d \to \mathrm{R}^{n\times k}$ and $D(u) : \mathrm{R}^d \to \mathrm{R}^k$ are matrix and vector valued, respectively, and depend affinely on the decision variable u. Both functions can be determined column-wise as $C_i(u) := B(u)^\top a_i$ and $D_i(u) := b_i - a_i^\top A(u)$ for $i \in [1,\ldots,k]$.

8.3.1 Joint chance constraints

The uncertain constraint (8.20) can be reformulated as a distributionally robust chance constraint as indicated in equation (8.6) of the preceding section. In the particular case of the joint uncertain constraint (8.12) for the polytopic constraint set X the chance constrained formulation in equation (8.6) reduces to the following condition on the decision u:

$$\mathbb{P}\left(C(u)^\top \xi < D(u)\right) \geq 1-\epsilon, \quad \forall \mathbb{P} \in \mathcal{C}. \tag{8.21}$$

This joint linear chance constraint (8.21) has already been studied extensively by Zymler et al. [142] in case the ambiguity set $\mathcal{C} = \mathcal{H}(0,\Sigma)\cap\mathcal{P}_n$ consists of all zero mean probability distributions with variance Σ. Unfortunately, distributionally robust chance constraints seem to be intractable in general. Indeed, there is no known tractable reformulation of the joint chance constraint (8.21), even in case the ambiguity set $\mathcal{K} = \mathcal{P}_n$ is left unstructured. Nevertheless, we will indicate in what follows how the results in the Chapters 6 and 7 can be brought to bear in the construction of tractable approximations of distributionally robust chance constraints.

The feasibility problem We start by remarking here that checking the feasibility of a fixed decision u in the joint chance constraint (8.21) requires the solution of an uncertainty quantification problem. Denote with $\Xi(u) := \left\{x \in \mathrm{R}^n \ : \ C(u)^\top x < D(u)\right\}$ the set of realizations of the random variable ξ such that the uncertain outcome $x(u,\xi)$ realizes within the constraint set X for a given decision u.

The joint chance constraint (8.21) can indeed equivalently be characterized in terms of an uncertainty quantification problem

$$B(L[u], \mathcal{K}, 0, \Sigma) = \sup \quad \mathbb{P}(\xi \notin \Xi(u)) \qquad \leq \ \epsilon, \qquad \text{s.t.} \quad \mathbb{P} \in \mathcal{H}(0, \Sigma) \cap \mathcal{K}$$

where the loss function is taken to be the indicator function $L[u](x) = \mathbf{1}\left\{\mathrm{R}^n \setminus \Xi(u)\right\}(x)$. For a fixed decision u the bound $B(L[u], \mathcal{K}, 0, \Sigma)$ reduces to an uncertainty quantification problem of the type discussed in Section 7.3.1. Checking the feasibility of a fixed decision can hence be done through solving a tractable convex optimization problem by merit of Proposition 7.1. The tractable reformulation in case the ambiguity set $\mathcal{K}$ consists of the set of all α-unimodal or γ-monotone distributions is given explicitly in Corollaries 7.2 and 7.3, respectively. We would like to remark that Proposition 7.1 is in fact only applicable when $0 \in \Xi(u)$. The previous assumption is not limiting here as it is equivalent to the requirement of nominal constraint satisfaction $x(u, 0) \in X$ and thus is fulfilled when the decision u is feasible.

We remark here that for the purpose of reformulating the joint chance constraint (8.21) the uncertainty quantification problem defining the bound $B(L[u], \mathcal{K}, 0, \Sigma)$ is best stated in its dual form. Indeed, the set of feasible decisions u in the joint chance constraint (8.21) can be characterized using the dual characterization (7.4) of the worst-case expectation problem $B(L[u], \mathcal{K}, 0, \Sigma)$ and is given as

$$(8.21) \iff \left\{ u \in \mathrm{R}^d \ : \ \begin{array}{l} \exists \begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \in \mathrm{S}_+^{n+1}, \quad \mathrm{Tr}\left\{Y\Sigma_s\right\} + y_0 \leq \epsilon, \\ x^\top Y x + 2x^\top y + y_0 - L_s[u](x) \geq 0, \quad \forall x \in \mathrm{R}^n \end{array} \right\} \tag{8.22}$$

for $L_s[u](x) := \int_0^\infty L[u](tx)\,\mathbb{T}(\mathrm{d}t)$ and $\Sigma_s \cdot \int_0^\infty t^2\,\mathbb{T}(\mathrm{d}t) = \Sigma \succ 0$. As pointed out previously, Proposition 7.1 guarantees that the feasibility of a given decision u in the set (8.22) reduces to an LMI. Although checking feasibility of a given decision is tractable there is unfortunately no known tractable reformulation of the set (8.22) or equivalently the distributionally robust chance constraint (8.21). This is the case even when no additional structure is imposed, i.e. $\mathcal{K} = \mathcal{P}_n$. In what remains of this section we discuss instead two popular approaches to derive tractable approximations of the distributionally robust chance constraint (8.21).

The Bonferroni approximation A very popular method to approximate the joint chance constraint (8.21) makes use of Bonferroni's inequality; see for instance Galambos [50]. In the context of the distributionally robust chance constraint (8.21) the Bonferroni inequality translates to

$$\mathbb{P}\left(\cup_{i\in[1,\ldots,k]} \ \left\{C_i(u)^\top \xi \geq D_i(u)\right\}\right) \ \leq \ \sum_{i=1}^k \mathbb{P}\left(C_i(u)^\top \xi \geq D_i(u)\right).$$

The power of the Bonferroni inequality lies in the fact that it limits the probability of the joint event $\{L^\alpha(x(u,\xi)) \geq 0\}$ in terms of the probability of k individual events $\{\ell_i(x(u,\xi)) \geq 0\}$. The distributionally robust joint chance constraint (8.21) can now indeed be approximated in terms of k distributionally robust single chance constraints each with a violation budget ϵ_i, i.e.

$$\forall \mathbb{P} \in \mathcal{C}: \quad \mathbb{P}\left(C_i(u)^\top \xi \leq D_i(u)\right) \leq 1 - \epsilon_i, \quad \sum_i \epsilon_i \leq \epsilon \implies (8.21)$$

Each distributionally robust single chance constraints can be represented as a tractable SOC constraint as indicated in Proposition 8.1.

A shortcoming of the Bonferroni approximation is that its quality depends critically on the choice of the violation budgets ϵ_i. Unfortunately, Nemirovski and Shapiro [88] show that the problem of determining the best choice for the budgets for a generic joint chance constraint of the type (8.21) is non-convex and believed to be intractable. The Bonferroni approximation is furthermore shown by Chen et al. [37] to be conservative even if the violation budget is allocated optimally. As a consequence, in most applications of the Bonferroni approximation to the distributionally robust chance constraint (8.21) the budget is divided equally amongst all individual constraints by taking $\epsilon_i = \epsilon/k$.

CVaR approximation Recent attempts by Chen et al. [37] and Zymler et al. [142] to improve the Bonferroni approximation are based on approximating the distributionally robust joint chance constraint (8.21) in terms of a distributionally robust CVaR constraint. The approximation is based on implication (8.8) which guarantees that a distributionally robust CVaR constraint implies its corresponding distributionally robust chance constraint.

As the constraint set $X = \{x \in \mathrm{R}^c \ : \ L^\alpha(x) < 0\}$ is independent of the choice of $\alpha \in \mathrm{R}^k_{++}$, the implication

$$\forall \mathbb{P} \in \mathcal{C} : \quad \mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x(u,\xi))) \leq 0 \implies (8.21)$$

holds uniformly in α. The parameter α can hence be chosen to make the CVaR approximation as tight as possible. However, the problem of finding the best values for α exactly seems to be computationally formidable. Nevertheless, in practice the quality of the CVaR approximation does not seem to depend as critically on the value of α as compared to the dependency of the Bonferroni approximation on the budgets ϵ_i.

Whether or not the CVaR approximation of distributionally robust chance constraints is sensible depends on whether distributionally robust CVaR constraints allow for a tractable reformulation. In what remains of this chapter we will indicate that is in fact the case for unstructured distributions $\mathcal{K} = \mathcal{P}_n$.

8.3.2 Joint CVaR constraints

The uncertain constraint (8.20) can also be reformulated as a distributionally robust CVaR constraint instead as indicated in equation (8.7) of the preceding section. In the particular case of the joint uncertain constraint (8.20) the distributionally robust CVaR constrained formulation in equation (8.7) reduces to the following condition on the decision u:

$$\mathbb{P}\text{-CVaR}_\epsilon \left(L^\alpha(x(u,\xi))\right) \leq 0, \quad \forall \mathbb{P} \in \mathcal{C}. \tag{8.23}$$

As discussed before, distributionally robust CVaR constraints alleviate some of the deficiencies suffered by chance constraints such as blindness towards extreme events in which the uncertain outcome $x(u,\xi)$ realizes far outside the constraint set X. We will indicate in this section that next to these more desirable theoretical properties, distributionally robust CVaR constraints have the additional benefit of being less computationally demanding.

Fact 8.1 (Distributionally robust CVaR constraints)**.** *The distributionally robust CVaR constraint* (8.23) *presents a convex constraint on the decision variable u, i.e. the set*

$$\left\{u \in \mathrm{R}^d \ : \ \mathbb{P}\text{-}CVaR_\epsilon \left(L^\alpha(x(u,\xi))\right) \leq 0, \quad \forall \mathbb{P} \in \mathcal{C}\right\}$$

is a convex set.

Proof. The fact can be proven directly from the definition of a convex set. Let the decisions u_1 and u_2 be feasible in the distributionally robust CVaR constraint (8.23), then for any decision $u = tu_1 + (1-\lambda)u_2$ with $t \in [0,1]$ we have

$$\begin{aligned}\mathbb{P}\text{-CVaR}_\epsilon \left(L^\alpha(x(u,\xi))\right) &= \mathbb{P}\text{-CVaR}_\epsilon \left(L^\alpha(tx(u_1,\xi) + (1-t)x(u_2,\xi))\right) \\ &\leq t\mathbb{P}\text{-CVaR}_\epsilon \left(L^\alpha(x(u_1,\xi))\right) + (1-t)\mathbb{P}\text{-CVaR}_\epsilon \left(L^\alpha(x(u_2,\xi))\right),\end{aligned}$$

using the affinity of $x(u,\xi)$ in u and the convexity of the function L^α and CVaR measure. When both u_1 and u_2 are feasible, the decision $tu_1 + (1-t)u_2$ for $t \in [0,1]$ is feasible as well. $\square$

We will assume for now that no structure in specified, i.e. $\mathcal{K} = \mathcal{P}_n$, and hence that the ambiguity set $C = \mathcal{H}(0, \Sigma)$ consists of all zero mean probability distributions with variance Σ. We will discuss the case of more richly structured ambiguity sets, i.e. $\mathcal{K} \subset \mathcal{P}_n$, in the last part of this section.

The feasibility problem We start our analysis again by remarking here that checking the feasibility of a fixed decision u in the joint distributionally robust CVaR constraint (8.23) requires the solution of an uncertainty quantification problem. Denote with $L[u](\xi) = L^\alpha(x(u,\xi))$ the function mapping the realizations of the random variable ξ to the severity of the uncertain outcome $x(u,\xi)$ as measured by the aggregated loss function L^α for a given decision u.

The distributionally robust CVaR constraint (8.23) can indeed equivalently be characterized in terms of an uncertainty quantification problem

$$\begin{aligned} B_{\mathrm{CVaR}}(L[u], \mathcal{P}_n, 0, \Sigma) = \sup \quad & \mathbb{P}\text{-CVaR}_\epsilon(L[u](\xi)) \quad \leq 0 \\ \text{s.t.} \quad & \mathbb{P} \in \mathcal{H}(0,\Sigma) \cap \mathcal{P}_n \end{aligned} \tag{8.24}$$

as discussed in Section 5.4.2. For a fixed and given decision u, the previous worst-case CVaR bound $B_{\mathrm{CVaR}}(L[u], \mathcal{P}_n, \mu, S)$ reduces to an uncertainty quantification problem of the type discussed in Section 7.2.1. Checking the feasibility of a fixed decision can hence be done through solving the tractable convex optimization problem given in Theorem 7.3.

We remark here that for the purpose of reformulating the distributionally robust CVaR constraint (8.23) the uncertainty quantification problem (8.24) is best stated in its dual form. Indeed, the set of feasible decisions u in the distributionally robust CVaR constraint (8.23) can be characterized using the dual characterization given in Theorem 7.3.

Theorem 8.1 (Distributionally robust CVaR contraints [142])**.** *The distributionally robust CVaR constraint* (8.23) *for ambiguity set $\mathcal{C} = \mathcal{H}(0,\Sigma) \cap \mathcal{P}_n$ and loss functions L^α satisfying Assumption 9.2 is equivalent to the LMI*

$$\left\{ u \in \mathrm{R}^d \;:\; \begin{array}{l} \exists \begin{pmatrix} Y & y \\ y^\top & y_0 \end{pmatrix} \in \mathrm{S}^{n+1}_+ \;\; \text{s.t.} \;\; \beta + \mathrm{Tr}\{Y\Sigma\} + y_0 \leq 0 \\ \begin{pmatrix} Y & y \\ y^\top & y_0 + \beta \end{pmatrix} \succeq \alpha_i \begin{pmatrix} 0 & \frac{1}{2}C_i(u) \\ \frac{1}{2}C_i(u)^\top & -D_i(u) \end{pmatrix}, \; \forall i \in [1,\ldots,k] \end{array} \right\} \tag{8.25}$$

if the variance matrix satisfies $\Sigma \succ 0$.

Proof. The proof is a simple corollary of equation (8.24) in combination with Theorem 7.3. $\square$

The constraint set (8.25) is an exact representation of the set of decisions u feasible in the distributionally robust CVaR constraint (8.23). As the functions $B(u)$ and $D(u)$ are both affine in the decision variable u, the feasible set of the distributionally robust CVaR constraint (8.23) is indeed represented as a tractable LMI condition. Theorem (8.25) is hence powerful as it allows for the exact and tractable reformulation of distributionally robust CVaR constraints.

Structured ambiguity sets When the ambiguity set $\mathcal{K}$ is more richly structured, the distributionally robust CVaR constraint (8.23) can be stated again in terms of an worst-case CVaR bound

$$\begin{aligned} B_{\mathrm{CVaR}}(L[u], \mathcal{K}, 0, \Sigma) = \sup \quad & \mathbb{P}\text{-CVaR}_\epsilon(L[u](\xi)) \quad \leq 0. \\ \text{s.t.} \quad & \mathbb{P} \in \mathcal{H}(0, \Sigma) \cap \mathcal{K} \end{aligned} \tag{8.26}$$

For a given decision u, the previous worst-case CVaR bound reduces to an uncertainty quantification problem of the type discussed in Section 7.3.2. Using the reduction put forward in Theorems 6.1 and 7.1, the worst-case CVaR bound in previous constraint is equivalent to $B_{\mathrm{CVaR}}(L_s[u], \mathcal{P}_n, 0, \Sigma_s)$ for the transformed arguments

$$L_s[u] = \int_0^\infty L[u](tx)\, \mathbb{T}(\mathrm{d}t) \quad \text{and} \quad \Sigma_s \cdot \int_0^\infty t^2\, \mathbb{T}(\mathrm{d}t) = \Sigma.$$

As shown in Fact 7.1, the function $L_s[u](x)$ is convex in x for any fixed u and vice-versa. Unfortunately, the loss function $L_s[u]$ is not in the form required by Theorem 7.2 and consequently the tractable representation (8.25) for the distributionally robust CVaR constraint (8.23) seems not to admit an immediate extension to the case of more richly structured sets $\mathcal{K}$.

8.4 Conclusions

We have introduced in this chapter distributionally robust chance and CVaR constraints. We discussed their properties in both the single and joint constraint case. For single uncertain constraints, the difference between distributionally robust chance and CVaR constraints runs only skin deep. Both the chance and the CVaR constrained formulation indeed present a convex robust constraint on the decision variable which can in fact be stated explicitly in terms of a tractable SOC constraint. In the joint case, only the CVaR constrained formulation is convex in general. Moreover, for unstructured distributions the CVaR constrained formulation reduces to an LMI in the decision variable. Although feasibility of a fixed decision in a distributionally robust chance constraint reduces to a tractable uncertainty quantification problem, distributionally robust chance constraints do not seem to present tractable constraints on the decision variable and hence must be approximated.

In Chapter 9 the distributionally robust chance and CVaR constraints discussed here will be considered in a dynamic setting. We will illustrate there that the proposed distributionally robust constraints constitute sensible design objectives in practice as well as in theory.

9 Robust optimal control with second-order moment information

We investigate the control of constrained stochastic linear systems when faced with limited information regarding the disturbance process, i.e. when only the first two moments of the disturbance distribution are known. We consider the two types of distributionally robust constraints already encountered in the static optimization setting in Chapter 8. In the first case, we require that the constraints hold with a given probability for all disturbance distributions sharing the known moments. These constraints were referred to as distributionally robust chance constraints. In the second case, we impose CVaR constraints to bound the expected constraint violation for all disturbance distributions consistent with the given moment information. Such constraints were referred to as distributionally robust CVaR constraints with second-order moment specifications.

We propose a method for designing linear controllers for systems with such constraints that is both computationally tractable and practically meaningful for both finite and infinite horizon problems. We prove in the infinite horizon case that our design procedure produces the globally optimal linear output feedback controller for distributionally robust CVaR and chance constrained problems. The proposed methods are illustrated for a wind blade control design case study for which distributionally robust constraints constitute sensible design objectives.

The current chapter will in fact present the dynamic counterpart of Chapter 8 on distributionally robust optimization. The robust optimization problem discussed in the previous chapter consisted of taking a single static decision u which satisfied the constraints despite uncertainty ξ. A robust control problem on the other hand considers taking several decisions u_t. The decisions u_t are taken as to satisfy the constraints at each time despite several uncertainties ξ_t. As the decisions u_t are not just static, but rather are taken dynamically over time, they may be adapted to the uncertainties ξ_t in accordance to a causal control law.

9.1 Finite horizon problems

We consider a discrete-time linear time-invariant (DLTI) system with n states x, m control inputs u, r outputs y, d exogenous inputs or disturbances ξ:

$$\begin{cases} x_{t+1} = Ax_t + Bu_t + C\xi_t \quad \text{and} \quad x_0 = x \\ \quad y_t = Dx_t + E\xi_t, \end{cases} \tag{S}$$

where all matrices are of appropriate dimension and the disturbances ξ_t model both process noise (via the term $C\xi_t$) and measurement noise (via $E\xi_t$).

When addressing chance- or CVaR- constrained control of the uncertain system S, we will require a more sophisticated measure space $(\Omega, \mathcal{F})$ than the measure spaces on finite dimensional vector spaces considered in the previous chapters. We will henceforward assume that the sample space Ω is sufficiently rich such that any (joint) distribution of all the random variables appearing in the system S on the Cartesian product of their individual range spaces is induced by a probability distribution on $(\Omega, \mathcal{F})$, and we will denote by $\mathcal{P}_\infty$ the set of all such probability distributions[1].

Our goal is to design a finite-horizon control law for the system S that minimizes an average quadratic cost, subject to an additional requirement that the state satisfies the constraint '$x_t \in X$' in either a chance- or CVaR-constrained sense. The control inputs u_t will be restricted to be $\mathcal{F}_t^y := \sigma(y_0, \ldots, y_t)$-measurable throughout. Simply put, the previous statement means that the adaptation of the control input u_t at any time t may only depend on the observed outputs $(y_0, \ldots, y_t)$ at that time and is enforced to guarantee causality. We remark that this is the essential difference between the static optimization problems discussed in Chapter 8 in which a decision needed to be taken independently of the disturbance, and the dynamic control problems discussed here in which feedback of the past measured disturbances to the input is essential.

We wish to achieve our control objectives despite some ambiguity on the disturbance distribution. Specifically, we assume only that the following information is available about the disturbance process:

Assumption 9.1 (Weak sense stationary disturbances)**.** *We assume that in the DLTI system S, the disturbance ξ_t is a weak sense stationary (w.s.s.) white noise process with variance matrix $\mathbf{E}_\mathbb{P}\left[(\xi_t - \mu) \cdot (\xi_t - \mu)^\top\right] = \Sigma$ and mean $\mathbf{E}_\mathbb{P}[\xi_t] = \mu$ for all $t \in \mathrm{N}_0$.*

The w.s.s. assumption appears frequently in signal processing literature such as Papoulis [96], but is far less common in the control literature. In effect, it assumes that only the autocorrelation $R_{\xi\xi}(t) := \mathbf{E}_\mathbb{P}\left[\xi_i \cdot \xi_{i-t}^\top\right]$ is known, with $R_{\xi\xi}(0) = \Sigma + \mu\mu^\top$ and $R_{\xi\xi}(t) = \mu\mu^\top$ otherwise. Furthermore, knowing the first two moments of a w.s.s. process is, by merit of the Wiener-Khintchine Theorem, equivalent to knowing its power spectrum. Estimating the spectral density of the disturbance ξ_t, for example from historical data, is significantly easier in practice than determining the complete marginal distribution of ξ_t with respect to $\mathbb{P}$. This particular type of estimation problem is also referred to as spectral density estimation in the signal processing community, see for instance Stoica and Moses [124].

The w.s.s. assumption implies that the only information available about the disturbance distribution is its autocorrelation function. Hence, the underlying probability distribution $\mathbb{P}^\star$ is only known to be an element of the ambiguity set

$$\mathcal{C}^\infty(\mu, \Sigma) := \left\{ \mathbb{P} \ : \ \mathbf{E}_\mathbb{P}\left[(\xi_i^\top, 1)^\top \cdot (\xi_j^\top, 1)^\top\right] = \begin{pmatrix} \Sigma\delta_{ij} + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{pmatrix}, \quad \forall i, j \in \mathrm{N}_0 \right\}.$$

The set $\mathcal{C}^\infty(\mu, \Sigma)$ contains all probability distributions consistent with the known moment information about the system disturbance. Notice that the probability distributions in $\mathcal{C}^\infty(\mu, \Sigma)$ are defined for an infinite horizon. This will permit us to work with the same ambiguity set $\mathcal{C}^\infty(\mu, \Sigma)$ despite varying horizons in the finite horizon setting. When choosing a control policy for the system S, we will require that it be distributionally robust with respect to the ambiguity set $\mathcal{C}^\infty(\mu, \Sigma)$, in either a chance constrained or CVaR sense, for the constraint '$x_t \in X$'. In order to achieve this control design objective, the notion of a distributionally robust constraint, introduced in Chapter 8, is now used to formulate our control problem.

[1] This means that we can think of Ω as the Cartesian product of all the random variables' range spaces, in which case $\mathcal{F}$ is identified with the Borel σ-algebra on Ω, while each random variable reduces to a coordinate projection.

Control constraints: We will consider distributionally robust constraints for the system S enforced over a finite time horizon of length N, i.e.

$$\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x_t)) \leq 0 \quad \forall t \in [0,\dots,N-1]. \tag{9.1}$$

We will refer to the parameter N as the horizon length of the finite optimal control problem studied. In the remainder of this chapter, we will be particularly interested in quadratically representable loss functions L^α as indicated in the following assumption.

Assumption 9.2. *An aggregated loss function $L^\alpha : \mathrm{R}^n \to \mathrm{R}$ with $\alpha \in \mathrm{R}^I_{++}$ for the distributionally robust CVaR constraints* (9.1) *is given as*

$$L^\alpha(x) = \max_{i\in\{1,\dots,k\}} [\alpha_i \ell_i(x)] = \max_{i\in\{1,\dots,k\}} \left[\alpha_i \left(x^\top E_i x + 2e_i^\top x + e_i^0\right)\right],$$

where $E_i \in \mathrm{S}^n_+$, $e_i \in \mathrm{R}_n$, $e_i^0 \in \mathrm{R}$.

We assume that the parameter $\alpha \in \mathrm{R}^n_{++}$ is known or given, either as an attempt to approximate the distributionally robust chance constraint

$$\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}(x_t \in X) \geq 1-\epsilon \quad \forall t \in [0,\dots,N-1],$$

or as an indicator of the relative importance of the loss severity measures ℓ_i as explained in Chapter 8. Recall that the constraint set X corresponds to the zero sub level set of the loss function L^α, and consequently is assumed to be a finite intersection of half-spaces and generalized ellipsoids by virtue of Assumption 9.2. Moreover as shown by Zymler et al. [142] and discussed in Section 8.2.2, the distributionally robust CVaR constraint (9.1) coincides with its chance constraint counterpart in case X is a single ellipsoid ($k = 1$). In the remainder of the chapter we will hence focus exclusively on the CVaR constraint formulation.

We remark here that distributionally robust constraint (9.1) presents a condition on the inputs $u := (u_0, u_1, \dots)$ as it can be restated as

$$\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x_t(u,\xi))) \leq 0 \quad \forall t \in [0,\dots,N-1].$$

with $\xi = (\xi_0, \xi_1, \dots)$ the collection of all disturbances. For the sake of exposition, we will not make the dependence of the states x_t explicit on the controls and disturbances. Nevertheless, the distribution $\mathbb{P}$ will always refer to the distribution of the disturbance ξ even if not made explicit. The distribution of the state x_t will be referred to as $\mathbb{Q}_t$ where its dependence on the control inputs is once again not made explicit.

For the system S we define a causal control policy $\pi_N := \{u_0, u_1, \dots, u_{N-1}\}$, such that the control input selected at each time $t \in [0,\dots,N-1]$ is a function mapping prior measurements to actions, i.e. u_t is $\mathcal{F}^y_t$-measurable, where we assume that the initial state $x_0 = x$ is known without any loss of generality[2]. Remark that the inputs $u_t(\xi)$ are thus all causal functions of the measured part of the disturbance ξ. This last dependence is however suppressed in this chapter to ease notation. We denote the set of all causal policies as Π_N. We wish to find, if it exists, a policy $\pi_N \in \Pi_N$ such that system S satisfies the CVaR constraints (9.1) over a finite horizon. We refer to such a policy as admissible with respect to the system S and the CVaR constraints (9.1).

Objective function: Our aim is to find a causal control policy $\pi_N \in \Pi_N$ that is admissible with respect to the CVaR constraints while minimizing a given objective function J_N. We will assume

[2] In the case that the initial state $x_0 = x$ is itself uncertain, one can always add an additional leading state $x_{-1} = 0$ and a state update equation $x_0 = Ax_{-1} + \xi_{-1}$, where ξ_{-1} equals x in distribution.

here that the objective function $J_N : \mathrm{R}^n \times \Pi_N \to \mathrm{R}_+$ is a discounted sum of quadratic stage costs, i.e. that it is in the form

$$J_N(x, \pi_N) := \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbf{E}_\mathbb{P}\left[\sum_{t=0}^{N-1} \beta^t \left[x_t^\top Q x_t + u_t^\top R u_t\right] + \beta^N x_N^\top Q_f x_N\right], \tag{9.2}$$

where we refer to $\beta \in [0, 1)$ as the discount factor of the control cost. It is assumed that the objective function J_N is convex, i.e. Q, $Q_f \in \mathrm{S}_+$ and $R \in \mathrm{S}_{++}$. We are therefore interested in the solution to the optimal control problem

$$\begin{array}{lll} \inf\limits_{\pi_N\in\Pi_N} & J_N(x, \pi_N) & \\ \text{s.t.} & x_{t+1} = Ax_t + Bu_t + C\xi_t,\ x_0 = x & \\ & \sup\limits_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x_t)) \leq 0, \quad \forall t \in [0, \ldots, N-1]. & \end{array} \tag{R_N}$$

There are however no known methods of solving the above problem in its full generality. The hardness of the problem can be attributed to two observations; (i) optimizing directly over arbitrary measurable policies π_N in Π_N where Π_N is infinite dimensional is out the question; and (ii) distributionally robust constraints such as (9.1), even for convex loss functions L, are hard to deal with directly when $x_t \in \mathcal{F}_t^y$ is a general non-linear function of the past measurements $(y_0, \ldots, y_t)$. Hence, in what follows we restrict attention to control policies that are affine in the past disturbances. Restricted policies of this type are well known in the operations research and control community, where they are commonly referred to either as *linear decision rules* such as in Ben-Tal et al. [6] or *affine feedback policies* by Goulart et al. [55]. Although such policies are typically suboptimal, recent research effort by Hadjiyiannis et al. [58], Van Parys et al. [129] and Iancu et al. [62] has focussed on providing sub-optimality bounds when applied to systems with worst-case constraints.

Denote by $x := (x_0^\top, \ldots, x_N^\top)^\top$, $u := (u_0^\top, \ldots, u_{N-1}^\top)^\top$ and $y := (y_0^\top, \ldots, y_{N-1}^\top)^\top$ the collection of states, inputs and measurements, respectively, over the given finite horizon. Similarly define a vector of disturbances as

$$\xi := (1, \xi_0^\top, \ldots, \xi_{N-1}^\top)^\top, \tag{9.3}$$

augmented with a leading one. This leading term is included for notational convenience so that any affine function of $(\xi_0, \ldots, \xi_{N-1})$ can be written as $X\xi$ for some matrix X with appropriate dimensions. Because of the w.s.s. condition on the disturbance process in Assumption 9.1, we have that $\mathbf{E}_\mathbb{P}\left[\xi \cdot \xi^\top\right] = M_w \in \mathrm{S}_{++}^{Nd+1}$ with

$$M_w := \left(1, \mu^\top, \ldots, \mu^\top\right)^\top \left(1, \mu^\top, \ldots, \mu^\top\right) + \operatorname{diag}(0, \mathbb{I}_N \otimes \Sigma).$$

The dynamics of the linear system S over the finite horizon N can then be written as

$$x = \mathbf{B}u + \mathbf{C}\xi,\ y = \mathbf{D}u + \mathbf{E}\xi, \tag{9.4}$$

for some matrices $(\mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{E})$ that can be derived from the system matrices and initial state $x_0 = x$; see Appendix B. Note in particular that the leading one in (9.3) means that the term $\mathbf{C}\xi$ is an affine function of both the disturbances and the initial state $x_0 = x$. Our approach will be to restrict u to be affine in the past disturbances, i.e. $u = U\xi$ for some causal feedback matrix $U \in \mathbf{N}$.

The set of causal policies $\mathbf{N}$ must ensure that the resulting feedback policy u_t is $\mathcal{F}_t^y$-measurable, i.e. that the feedback policy u_t depends only on the initial state x and observed outputs $[y_0, \ldots, y_t]$. This can be achieved by a reparametrization of the feedback policy $u = \tilde{U}\eta$ as an affine function of the *purified observations* $\eta = (\mathbf{DC} + \mathbf{E})\,\xi$ as discussed by Ben-Tal et al. [6, Section 14.4.2]. The causality set can then be defined as

$$\mathbf{N} := \left\{ U \in \mathrm{R}^{N_x \times N_w} \;:\; U = \begin{pmatrix} u_0 & 0 & 0 & 0 \\ u_1 & U_{1,0} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_{N-1} & U_{N-1,0} & \dots & U_{N-1,N-1} \end{pmatrix} (\mathbf{DC} + \mathbf{E}) \right\}$$

which ensures that u_t is $\mathcal{F}_t^y$-measurable. Assume we have such an affine policy $u = U\xi$, then the cost of this policy according to the cost function (9.2) is

$$\tilde{J}_N(x, U) := \mathrm{Tr}\left\{U^\top \left(J_u + \mathbf{B}J_x\mathbf{B}\right) U M_w + 2\mathbf{C}J_x\mathbf{B}UM_w + \mathbf{C}^\top J_x \mathbf{C} M_w\right\},$$

where $J_x := \mathrm{diag}\big(\mathrm{diag}\left(\beta^0, \dots, \beta^{N-1}\right) \otimes Q, \beta^N Q_f\big)$ and $J_u := \mathrm{diag}\left(\beta^0, \dots, \beta^{N-1}\right) \otimes R$. Note that $\tilde{J}_N(x, U)$ is convex quadratic in U since $\mathrm{diag}(Q, R) \in \mathrm{S}_+$. We are now ready to state the main result of this section, which shows that finding the best affine control policy for problem R_N can be reformulated as a tractable convex optimization problem.

Theorem 9.1 (CVaR constrained control)**.** *The best admissible affine control policy of problem R_N, i.e. a solution to the restricted problem*

$$\begin{aligned} \inf_{U \in \mathbf{N}} \quad & \tilde{J}_N(x, U) \\ \text{s.t.} \quad & x = \mathbf{B}u + \mathbf{C}\xi, \; u = U\xi \\ & \sup_{\mathbb{P} \in \mathcal{C}^\infty(\mu, \Sigma)} \mathbb{P}\text{-CVaR}_\epsilon \left(L^\alpha(x_t)\right) \le 0, \quad \forall t \in [0, \dots, N-1], \end{aligned} \tag{$\tilde{R}_N$}$$

where the loss function $L^\alpha : \mathrm{R}^n \to \mathrm{R}$ satisfies Assumption 9.2, can be found as a solution to the SDP

$$\begin{aligned} \inf \quad & \tilde{J}_N(x, U) \\ \text{s.t.} \quad & \left.\begin{aligned} & U \in \mathbf{N}, \; \beta_t \in \mathrm{R}, \; X_t \in \mathrm{S}_+^{Nd+2}, \; P_t^i \in \mathrm{S}_+^{Nd+1} \\ & \beta_t + \frac{1}{\epsilon}\mathrm{Tr}\left\{M_w X_t\right\} \le 0, \\ & X_t - \begin{pmatrix} \alpha_i P_t^i & \alpha_i(\mathbf{B}_t U + \mathbf{C}_t)^\top e_i \\ e_i^\top(\mathbf{B}_t U + \mathbf{C}_t)\alpha_i & \alpha_i e_i^0 - \beta_t \end{pmatrix} \succeq 0, \\ & \begin{pmatrix} P_t^i & (\mathbf{B}_t U + \mathbf{C}_t)^\top E_i^{\frac{1}{2}} \\ E_i^{\frac{1}{2}}(\mathbf{B}_t U + \mathbf{C}_t) & \mathbb{I}_n \end{pmatrix} \succeq 0, \end{aligned}\right\} \begin{aligned} & \forall t \in [0, \dots, N-1] \\ & \forall i \in [1, \dots, k] \end{aligned} \end{aligned} \tag{9.5}$$

where $\mathbf{B} =: (\mathbf{B}_0^\top, \dots, \mathbf{B}_{N-1}^\top)^\top$ *and* $\mathbf{C} =: (\mathbf{C}_0^\top, \dots, \mathbf{C}_{N-1}^\top)^\top$.

Proof. The proof follows by applying the tractability result provided in Theorem 7.3 to the constraints

$$\sup_{\mathbb{P} \in \mathcal{C}^\infty(\mu, \Sigma)} \mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(\overbrace{(\mathbf{B}_t U + \mathbf{C}_t)\,\xi}^{x_t})) \le 0.$$

Explicitly writing out the quadratic form in the preceding inequality as

$$\sup_{\mathbb{P} \in \mathcal{C}^\infty(\mu, \Sigma)} \mathbb{P}\text{-CVaR}_\epsilon \left(\max_i \alpha_i \xi^\top (\mathbf{B}_t U + \mathcal{C}_t)^\top E_i (\mathbf{B}_t U + \mathbf{C}_t)\xi + 2\alpha_i e_i^\top (\mathbf{B}_t U + \mathbf{C}_t)\xi + \alpha_i e_i^0\right) \le 0$$

yields a matrix inequality with quadratic terms in the variable U:

$$\exists \beta_t \in \mathrm{R}, \; X_t \in \mathrm{S}_+^{Nd+2} : \begin{cases} \beta_t + \frac{1}{\epsilon}\mathrm{Tr}\left\{M_w X_t\right\} \le 0 \\ X_t \succeq \begin{pmatrix} \alpha_i(\mathbf{B}_t U + \mathbf{C}_t)^\top E_i (\mathbf{B}_t U + \mathbf{C}_t) & (\mathbf{B}_t U + \mathbf{C}_t)^\top e_i \alpha_i \\ \alpha_i e_i^\top (\mathbf{B}_t U + \mathbf{C}_t) & e_i^0 \alpha_i - \beta_t \end{pmatrix}, \; \forall i \in [1, \dots, k]. \end{cases}$$

The final result claimed in the theorem is then found by applying classical Schur complements, and rewriting the quadratic matrix inequality as two LMIs using the additional variables $P_t^i \in \mathbb{S}_+^{Nd+1}$. □

As discussed before, the distributionally robust constraint (9.1) is sometimes used to approximate a corresponding chance constraint for a judicious choice of α. The equivalence between chance constraints and CVaR constraints when X is a simple ellipsoid or $k = 1$ is proven in Zymler et al. [142]. This result enables us to formulate the following corollary to Theorem 9.1.

Corollary 9.1 (Chance constrained control)**.** *The best admissible affine control policy of the restricted problem*

$$\begin{aligned} \inf_{U \in \mathbf{N}} \quad & \tilde{J}_N(x, U) \\ \text{s.t.} \quad & x = \mathbf{B}u + \mathbf{C}\xi, \;\; u = U\xi \\ & \inf_{\mathbb{P} \in \mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\left\{x_t \in X\right\} \geq 1 - \epsilon, \;\; \forall t \in [0, \ldots, N-1], \end{aligned}$$

where the constraint set $X = \left\{x \;:\; x^\top E_1 x + 2e_1^\top x + e_1^0 < 0\right\}$ is a single ellipsoid, can be found as a solution of the SDP (9.5) *with $k = 1$ and $\alpha_1 = 1$.*

9.2 Infinite horizon problems

Infinite horizon control problems lend themselves to applications in which transient behaviour is of lesser importance, but in which we are interested in steady state behaviour. In Section 9.3 we present a numerical example of such a problem in the context of wind turbine blade control. The problem setting is similar to the one presented in Section 9.1, in that we again consider the DLTI system S where the disturbance input process ξ_t satisfies Assumption 9.1. In addition, we assume that the disturbance ξ_t has zero mean $\mu_{w_t} = \mu = 0$, and a zero initial condition $x_0 = 0$ reflects our indifference towards transient behaviour.

Hence in this infinite horizon setting, we consider the following optimal control problem

$$\begin{aligned} \inf_{\pi \in \Pi_\infty} \quad & \sup_{\mathbb{P} \in \mathcal{C}^\infty(\mu,\Sigma)} \limsup_{N \to \infty} \frac{1}{N} \mathbf{E}_{\mathbb{P}} \left[\sum_{t=0}^{N-1} \left[x_t^\top Q x_t + u_t^\top R u_t\right] \right] \\ \text{s.t.} \quad & x_{t+1} = A x_t + B u_t + C\xi_t, \\ & \sup_{\mathbb{P} \in \mathcal{C}^\infty(\mu,\Sigma)} \limsup_{t \to \infty} \mathbb{P}\text{-CVaR}_\epsilon \left(L^\alpha\left(x_t\right)\right) \leq 0, \end{aligned} \tag{R_∞}$$

where the loss function $L^\alpha(x)$ satisfies again Assumption 9.2. We assume throughout this section that the pairs $(Q^{\frac{1}{2}}, A)$ and (C, A) are observable and that the pair (A, B) is stabilizable, which is sufficient to guarantee the existence of linear time-invariant exponentially stabilizing control policies. The feedback policies π are restricted to Π_∞, where Π_∞ is the set of all linear time-invariant and causal ($\mathcal{F}_t^y$-measurable) feedback policies. We restricted attention to linear control strategies for the same reasons mentioned in Section 9.1. It is also well known that such a restriction causes no loss of optimality when the distributionally robust constraint in R_∞ is disregarded; see Kwakernaak and Sivan [71]. Indeed, the classical linear-quadratic-Gaussian (LQG) controller is optimal for the unconstrained version of problem R_∞.

The cost function in R_∞ is the infinite horizon limit of the stage cost function in (9.2) for system S, with no discounting or terminal cost. By omitting the discounting factor, the cost of a control law π becomes independent of the initial condition x_0 reflecting an indifference towards the cost of transient behavior. The design goal in this case reduces to minimizing the average stage cost, so that the objective function becomes

$$\begin{aligned} J_\infty(\pi) &:= \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \limsup_{N\to\infty} \frac{1}{N}\mathbf{E}_{\mathbb{P}}\left[\sum_{t=0}^{N-1}\left[x_t^\top Q x_t + u_t^\top R u_t\right]\right], \\ &= \limsup_{N\to\infty} \frac{1}{N}\sum_{t=0}^{N-1}\mathbf{E}_{\mathbb{P}}\left[x_t^\top Q x_t + u_t^\top R u_t\right] \quad \forall \mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma), \end{aligned}$$

where the equality follows from the fact that the expectation of a quadratic cost is independent of $\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)$ for linear control policies.

The robust constraint in R_∞ can be seen as a distributionally robust version of the nominal requirement

$$\limsup_{t\to\infty} \mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x_t)) \le 0. \tag{9.6}$$

This constraint expresses the design requirement that $\mathbb{P}\text{-CVaR}_\epsilon(L^\alpha(x_t)) \le 0$ when t tends to infinity. If a steady state distribution $\mathbb{Q}_\infty$ exists, the constraint (9.6) can be read as a constraint on the steady state distribution of $\{x_t\}$, i.e.

$$\mathbb{Q}_\infty\text{-CVaR}_\epsilon(L^\alpha(x)) \le 0,$$

where the distributions $\mathbb{Q}_t$ of x_t converge to the steady state distribution $\mathbb{Q}_\infty$ of the random variable x for t tending to infinity. However, Assumption 9.1 is not sufficient to guarantee that x_t converges in distribution to any steady state distribution $\mathbb{Q}_\infty$, hence we cannot treat (9.6) as a steady state constraint in general. Nevertheless, we observe that, although $\{x_t\}$ need not converge in distribution, its first two moments are known to converge whenever π is a strictly stabilizing linear control law:

Theorem 9.2 (Steady state behavior [71, Theorem 6.23])**.** *Let the discrete-time stochastic process x_t be the solution of the stochastic difference equation $x_{t+1} = \bar{A}x_t + \bar{C}\xi_t$, where $\bar{A}\in\mathrm{R}^{n\times n}$, $\bar{C}\in\mathrm{R}^{n\times d}$ and ξ_t has zero mean and satisfies Assumption 9.1. Define the variance matrix*

$$C_{xx}(t) := \mathbf{E}_{\mathbb{P}}\left[[x_t - \mathbf{E}_{\mathbb{P}}[x_t]]\cdot[x_t - \mathbf{E}_{\mathbb{P}}[x_t]]^\top\right].$$

If $\bar{A}$ is asymptotically stable then the asymptotic variance matrix $P_\infty := \lim_{t\to\infty} C_{xx}(t)$ exists and is the unique solution of the discrete Lyapunov equation $P_\infty = \lim_{t\to\infty} C_{xx}(t) = \bar{A}P_\infty\bar{A}^\top + \bar{C}\bar{C}^\top$.

Despite the possible lack of convergence in distribution of x_t, Theorem 9.2 will enable us to represent the distributionally robust constraint of problem R_∞

$$\limsup_{t\to\infty} \mathbb{P}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_t\right)\right) \le 0, \quad \forall\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma), \tag{9.7}$$

as a tractable constraint on the linear control law π, provided that we can identify a dynamic counterpart to the tractable reformulation (7.11).

Note that a direct application of reformulation (7.11) to the constraint (9.7) is problematic. If one assumes momentarily that $\{x_t\}$ converges, then (9.7) could be reformulated as a distributionally robust constraint in the form

$$\mathbb{Q}_\infty\text{-CVaR}_\epsilon\left(L^\alpha\left(x\right)\right) \le 0, \quad \forall\mathbb{Q}_\infty\in\mathcal{Q}_\infty, \tag{9.8}$$

where $\mathcal{Q}_\infty := \left\{\mathbb{Q}_\infty \ : \ \int x\,\mathbb{Q}_\infty(\mathrm{d}x) = 0,\ \int xx^\top\,\mathbb{Q}_\infty(\mathrm{d}x) = P_\infty\right\}$ an ambiguity set of distributions of the steady state. One could then apply the reformulation (7.11) to produce a LMI representation of the constraint (9.8). However, not every probability distribution in $\mathcal{Q}_\infty$ is necessarily a steady state distribution obtainable as a limit distribution of $\{x_t\}$. Hence, even if a steady state distribution exists, a replacement of the infinite horizon constraint (9.7) with (9.8) is seemingly conservative.

However, in the finite-horizon case we have the following result:

Lemma 9.1. *Suppose that $x_0 = 0$ and each x_t is a linear function of $(\xi_0, \ldots, \xi_{t-1})$ resulting from some linear control policy $\pi \in \Pi_\infty$. Define $\mathcal{Q}_t := \left\{\mathbb{Q}_t \ : \ \int x\, \mathbb{Q}_t(\mathrm{d}x) = 0,\ \int xx^\top\, \mathbb{Q}_t(\mathrm{d}x) = C_{xx}(t)\right\}$ for each $t \in \mathrm{N}$. Then*

$$\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\text{-}CVaR_\epsilon\left(L^\alpha\left(x_t\right)\right) = \sup_{\mathbb{Q}_t\in\mathcal{Q}_t} \mathbb{Q}_t\text{-}CVaR_\epsilon\left(L^\alpha\left(x\right)\right).$$

Proof. To prove the claim, according to definition (3.11) of the CVaR, it suffices to show that $\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbf{E}_\mathbb{P}[g(x_t)] = \sup_{\mathbb{Q}_t\in\mathcal{Q}_t} \mathbf{E}_{\mathbb{Q}_t}[g(x)]$ holds for any measurable function $g : \mathrm{R}^n \to \mathrm{R}$. Because x_t is linear in the disturbances $x_t = R\cdot(\xi_0, \ldots, \xi_{t-1})$ for some $R \in \mathrm{R}^{n\times td}$ as $x_0 = 0$, we can write

$$\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbf{E}_\mathbb{P}[g\left(R\cdot(\xi_0, \ldots, \xi_{t-1})\right)] = \sup_{\mathbb{Q}_t\in\mathcal{Q}_t} \mathbf{E}_{\mathbb{Q}_t}[g(x)]$$

The last equality can easily be shown using the projection Theorem 3.4 as done in for instance by Yu et al. [139]. □

If one implements a stabilizing linear control policy $\pi \in \Pi_\infty$ such that $P_\infty = \lim_{t\to\infty} C_{xx}(t,t)$ can be shown to exist by application of Theorem 9.2, then it follows in the limit from Lemma 9.1 that

$$\limsup_{t\to\infty} \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_t\right)\right) = \sup_{\mathbb{Q}_\infty\in\mathcal{Q}_\infty} \mathbb{Q}_\infty\text{-CVaR}_\epsilon\left(L^\alpha\left(x\right)\right) \tag{9.9}$$

since the worst-case CVaR is continuous in its moment information. However the former worst-case CVaR bound in (9.9), although tractable as indicated in Theorem 7.3, could potentially lead to a conservative reformulation of the constraint of interest (9.7) since

$$\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \limsup_{t\to\infty} \mathbb{P}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_t\right)\right) \le \limsup_{t\to\infty} \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_t\right)\right), \tag{9.10}$$

and consequently

$$\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \limsup_{t\to\infty} \mathbb{P}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_t\right)\right) \le \sup_{\mathbb{Q}_\infty\in\mathcal{Q}_\infty} \mathbb{Q}_\infty\text{-CVaR}_\epsilon\left(L^\alpha\left(x\right)\right).$$

This conservatism follows from the possibility that for each time t the distributions attaining the worst-case bound of the right-hand side of previous inequality may depend on t. In other words, the worst-case bound of the right-hand side is not obviously obtainable as the limit when t tends to infinity for some fixed distribution $\mathbb{P} \in \mathcal{C}^\infty(\mu, \Sigma)$, similar as the situation discussed for condition (9.8). Fortunately, we can show that this is in fact not the case and that no conservatism is incurred.

Lemma 9.2. *Let the discrete-time stochastic process x_t be the solution of the stochastic difference equation $x_{t+1} = \bar{A}x_t + \bar{C}\xi_t$, where $\bar{A} \in \mathrm{R}^{n\times n}$, $\bar{C} \in \mathrm{R}^{n\times d}$ and ξ_t has zero mean and satisfies Assumption 9.1. If $\bar{A}$ is asymptotically stable then*

$$\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \limsup_{t\to\infty} \mathbb{P}\text{-}CVaR_\epsilon\left(L^\alpha\left(x_t\right)\right) = \sup_{\mathbb{Q}_\infty\in\mathcal{Q}_\infty} \mathbb{Q}_\infty\text{-}CVaR_\epsilon\left(L^\alpha\left(x\right)\right).$$

Proof. See Appendix B. □

The equivalence (7.12) in view of Lemma 9.2 now provides a probabilistic interpretation to what otherwise could be considered *ad hoc* variance constraints

$$\lim_{t\to\infty} \mathrm{Tr}\left\{E^{\frac{1}{2}}\mathbf{E}\left[x_t x_t^\top\right]E^{\frac{1}{2}}\right\} \le \epsilon$$

as for instance discussed in Zhu et al. [141]. Indeed, constraining the variance to be bounded using a trace norm can now be read as a distributionally robust probabilistic constraint on the state satisfying a centered ellipsoidal state constraint $X = \left\{x \ : \ x^\top Ex \le 1\right\}$. In general we have the following counterpart to the tractable reformulation (7.11).

Theorem 9.3 (Tractability of worst-case CVaR for linear systems). *Let the discrete-time stochastic process x_t be the solution of the stochastic difference equation $x_{t+1} = \bar{A}x_t + \bar{C}\xi_t$, where $\bar{A} \in \mathrm{R}^{n\times n}$, $\bar{C} \in \mathrm{R}^{n\times d}$ and ξ_t has zero mean and satisfies Assumption 9.1. If $\bar{A}$ is asymptotically stable then*

$$\forall \mathbb{P} \in \mathcal{C}^{\infty}(\mu, \Sigma) : \limsup_{t\to\infty} \mathbb{P}\text{-}CVaR_{\epsilon}\left(L^{\alpha}\left(x_t\right)\right) \leq 0 \iff \begin{array}{ll} \inf\limits_{\beta, X} & \beta + \dfrac{1}{\epsilon}\mathrm{Tr}\left\{\mathrm{diag}(P_{\infty}, 1) \cdot X\right\} \leq 0 \\ \text{s.t.} & X \in \mathrm{S}_{+}^{n+1},\ \beta \in \mathrm{R},\ \forall i \in [1, \ldots, k] : \\ & X - \begin{pmatrix} \alpha_i E_i & \alpha_i e_i \\ \alpha_i e_i^{\top} & \alpha_i e_i^0 - \beta \end{pmatrix} \succeq 0, \end{array}$$

where $P_{\infty} = \bar{A}P_{\infty}\bar{A}^{\top} + \bar{C}\bar{C}^{\top}$ the stationary variance of the state.

Proof. The theorem is a direct consequence of Lemma 9.2 combined with the reformulation (7.11). □

We remark that the condition in Theorem 9.3 offers an exact condition for constraints of the type (9.7) to hold under the disturbance Assumption 9.1. When a tighter condition is required, one must either resort to nonlinear control laws π or assume additional information regarding the disturbance process. We next show that the equivalence (7.12) in case of a single centered ellipsoid also implies that the optimal linear feedback law for problem R_{∞} has an order which equals the number of states n of system S, and is the combination of a Kalman filter and a static feedback gain.

Theorem 9.4 (Optimal linear feedback law). *The optimal linear feedback law $\pi^{\star}$ of problem R_{∞} in case of $L^{\alpha}(x) = x^{\top}E_1 x + e_1^0$ consists of a linear estimator-controller pair (S, K) and hence is of the form*

$$\pi^{\star} : \begin{cases} \hat{x}_{t+1} = A\hat{x}_t + Bu_t + S\left(y_{t+1} - C\left(Ax_t + Bu_t\right)\right) \\ \quad u_t = K\hat{x}_t, \end{cases} \tag{9.11}$$

with $S := YD^{\top}\left(DYD^{\top} + EE^{\top}\right)^{-1}$. The matrix Y is the unique positive definite solution of the discrete algebraic Riccati equation

$$Y = A\left(Y - YD^{\top}\left(DYD^{\top} + EE^{\top}\right)^{-1}DY\right)A^{\top} + CC^{\top},$$

which can be solved efficiently as done by Arnold and Laub [1]. The static feedback matrix is given by $K = Z^{\star}(P^{\star})^{-1}$, where $P^{\star} \in \mathrm{S}_{++}^{n}$ and $Z^{\star} \in \mathbb{R}^{m\times n}$ can be found as the optimal solution of the SDP

$$\begin{array}{ll} \inf & \mathrm{Tr}\, Q\left(\Sigma + P\right) + \mathrm{Tr}\, RX \\ \text{s.t.} & P \in \mathrm{S}_{+}^{n},\ Z \in \mathrm{R}^{m\times n},\ X \in \mathrm{S}_{+}^{m} \\ & \begin{pmatrix} X & Z \\ Z^{\top} & P \end{pmatrix} \succeq 0,\ e_1^0 + \frac{1}{\epsilon}\mathrm{Tr}\left\{E_1\left(\Sigma + P\right)\right\} \leq 0 \\ & \begin{pmatrix} P - APA^{\top} - BZA^{\top} - AZ^{\top}B^{\top} - \Gamma & BZ \\ Z^{\top}B^{\top} & P \end{pmatrix} \succeq 0, \end{array} \tag{9.12}$$

where $\Gamma := YD^{\top}\left(DYD^{\top} + EE^{\top}\right)^{-1}DY$ and $\Sigma = Y - \Gamma$. Since (9.11) can be decomposed into a Kalman estimator S and state feedback controller K, problem R_{∞} satisfies a separation or certainty equivalence principle.

Proof. See Appendix B. □

The Kalman filter in Theorem 9.4 depends only on the process and measurement noise characteristics and is independent of the distributionally robust constraint (9.7) and cost function J_{∞}. Finding the optimal static feedback gain K requires only the solution of the tractable convex problem (9.12).

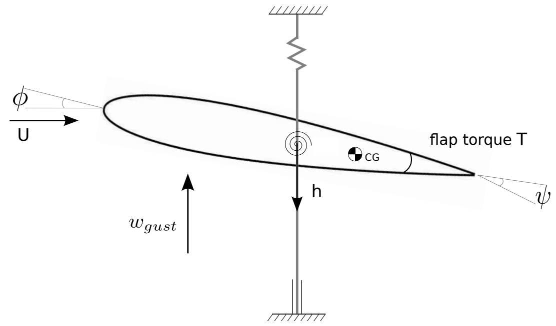


Figure 9.1: The geometry of the 2-DOF structural model. The overall model is linear continuous time invariant and has a modest size of 13 states, one endogenous and exogenous input T and ξ_{gust}, respectively.

9.3 Wind turbine blade control design problem

To illustrate the method introduced in the preceding section, we consider a wind turbine control problem similar to the one discussed in Ng et al. [91]. As the size of wind turbines is increased for larger energy capture, they are subject to greater risks of fatigue failure and extreme loading events. Therefore, most large wind turbines today are equipped with pitch control for speed regulation, which can also be used for load alleviation.

However, these pitch actuators are slow and limited by the inertia of the blades. Hence, as in Ng et al. [91], we assume that the blades are equipped with an actively controlled flap. The control objective is to minimize actuation energy while keeping some measure of blade loading within specified bounds. The disturbance acting on the turbine blades is mostly due to atmospheric turbulence, for which little more than the frequency spectrum is known; see Campbell [31]. According to the standard military design reference [86], atmospheric turbulence is typically treated as a Gaussian stochastic process defined by a standardized velocity spectrum. We follow this standard atmospheric turbulence model, modulo the normality assumption which is not well supported in reality. Hence, this is a natural setting in which the ideas developed in this paper are of practical interest.

An aerofoil section with flap can be modeled using a simple two degree of freedom (2-DOF) plunge-pitch aerofoil, restrained by a pair of springs as shown in Figure 9.1. The two dimensional aerofoil represents a cross section of one of the flexible wind turbine blades. For small elastic deformations and under the assumption of potential flow, we can use the classical methods provided by Theodorsen [126] to describe the behavior of our simple 2-DOF plunge-pitch aerofoil with a simple linear model. The modeling technique used here is by no means the only one possible, but results in a modest size system of only ten states. An alternative technique using classical vortex-panel methods by Katz and Plotkin [66] to get higher fidelity, but still linear, models is presented in Ng et al. [91]. We note that the methods described in this paper are not limited by the modest size of our control model, as indicated in further work of this author [132] where a high fidelity model is considered.

Since the disturbance modeling is important to our approach, we discuss it in slightly more detail in the next subsection.

9.3.1 Disturbance model

The majority of the disturbance acting upon the wind turbine blades is a direct result of atmospheric turbulence. Most commonly, atmospheric turbulence is represented as the convolution of (Gaussian) white noise through a linear time-invariant (LTI) shaping filter, referred to by the military reference [86] and Campbell [31] as a *von Kármán* filter. Hence

$$\xi_{\text{gust}} := H(n_1),$$

where n_1 is Gaussian white noise and H the *von Kármán* filter, which we choose to be a proper stable rational filter as in done in Campbell [31] with state space representation

$$\left[\begin{array}{ccc|c} -7.701 & -7.008 & -1.404 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 1.447 & 7.022 & 1.533 & 0 \end{array}\right]. \tag{H}$$

It is clear that the Gaussian assumption made on ξ_{gust} is unlikely to be fulfilled in practice, hence we assume only that n_1 is a scalar white w.s.s. noise process, i.e. $\mathbf{E}_{\mathbb{P}}\left[n_1^2(t)\right] = 1$ and thus not necessarily Gaussian. Hence, in practice we need only estimate the power spectrum of the atmospheric turbulence ξ_{gust}, e.g. from historical data.

The overall system of the wind turbine blade model with additional flap and disturbance filter is a linear continuous time invariant system with 13 states, 10 states for the 2-DOF airfoil model and 3 states for the turbulence model. The overall model has one endogenous input T and one exogenous input n_1. We assume that the states ϕ and h representing the pitch and plunge, see Figure 9.1, are measured with negligible measurement noise, i.e.

$$y = \begin{pmatrix} \phi \\ h \end{pmatrix} + \delta n_2,$$

where n_2 is a zero mean white noise signal with unit variance matrix, uncorrelated with n_1. To fit in the framework provided in the paper, we discretize the continuous time model using the zero order hold method at sampling frequency $f_s = 100$ Hz which captures most of the salient system dynamics for the model parameters we have selected.

9.3.2 Numerical results

A natural control design criterion in this setting is to ensure that the vector $(\dot{\alpha}, \dot{h})$ is kept small in order to bound the fatigue stress, usually caused by high variance dynamic loads. In addition we would like extreme static loading events to be rare, corresponding to the requirement that the deformation vector (α, h) remains close to zero. We express these two design criteria respectively as

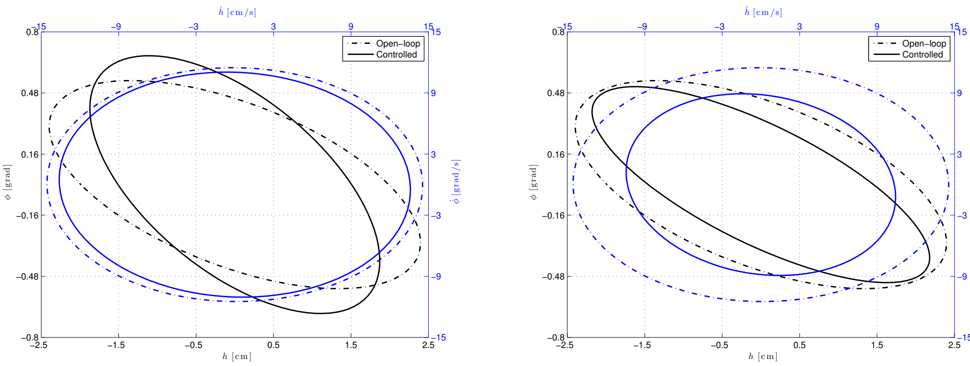
$$\inf_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \quad \limsup_{t\to\infty} \; \mathbb{P}\left\{(\dot{\phi}(t), \dot{h}(t)) \in B(55)\right\} \geq 1-\epsilon, \tag{9.13}$$

$$\inf_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \quad \limsup_{t\to\infty} \; \mathbb{P}\left\{(\phi(t), h(t)) \in B(6)\right\} \geq 1-\epsilon, \tag{9.14}$$

where $\epsilon = 0.1$, and $B(r)$ denotes a ball of radius r centered at the origin. The natural control objective in this setting is to minimize the expected actuation power usage. We express this by considering the cost function

$$J(\pi) = \limsup_{t\to\infty} \quad \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbf{E}_{\mathbb{P}}\left[\dot{\psi}^2(t)\right],$$

which must be minimized subject to the fatigue and loading constraints (9.13) and (9.14) respectively. Using the method described in Section 9.2, the optimal linear time invariant controller can



(a) (b)

Figure 9.2: Figure 9.2(a) shows the variance of the vectors (ϕ, h) and $(\dot{\phi}, \dot{h})$ when uncontrolled and with the optimal controller according to Section 9.2, as the sets $\left\{x \in \mathrm{R}^2 \ : \ x^\top \Sigma^{-1} x - 1 \leq 0\right\}$ with Σ the respective variance matrix. Similarly, Figure 9.2(b) shows the variance of the vectors (ϕ, h) and $(\dot{\phi}, \dot{h})$ when uncontrolled, and with the standard LQR controller $K_{\mathrm{LQR}}(0.1)$.

be computed efficiently. Although it should be noted that in Theorem 9.4 only one probability constraint is considered, the generalization to the case of finitely many constraints of type (9.7) is straightforward and omitted here. The difference between the variance of the vectors (ϕ, h) and $(\dot{\phi}, \dot{h})$, when uncontrolled or controlled with the synthesized controller $K^\star$, is visualized in Figure 9.2(a).

We compare this controller to the standard H_2-optimal controller found by tuning the cost function

$$\limsup_{t\to\infty} \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbf{E}_{\mathbb{P}}\left[\gamma\dot{\psi}^2(t) + \phi^2(t) + \dot{\phi}^2(t) + h^2(t) + \dot{h}^2(t)\right],$$

which weighs the actuation energy versus the size of the states $(\phi, \dot{\phi}, h, \dot{h})$, according to the tuning factor γ. A naïve method of designing a controller is to tune γ such that the closed loop system satisfies the fatigue (9.13) and loading (9.14) constraints.

We compare in Table 9.3.2 the cost of the optimal controller $K^\star$ and three naïvely tuned controllers $K_{\mathrm{LQR}}(\gamma_i)$. First it is noted that when uncontrolled, the control cost is zero. However, since $\epsilon = 0.1$ both design specifications (9.13) and (9.14) are violated. The optimal controller $K^\star$ has satisfied (9.13) and (9.14) exactly with no conservatism and relatively low cost. The LQR controller $K_{\mathrm{LQR}}(0.43)$ has the same cost as $K^\star$ but does not satisfy the constraints. The other LQR controllers either violate one of the constraints or have a massive cost compared to $K^\star$. The difference between the variance of the vectors (ϕ, h) and $(\dot{\phi}, \dot{h})$, when uncontrolled or controlled with the controller $K_{\mathrm{LQR}}(0.1)$, is visualized in Figure 9.2(b).

It can be seen from this example that the methodology of Section 9.2 provides an easy procedure to design controllers that handle constraints of the type (9.13) and (9.14). Again we point out that, by dropping the Gaussian assumption on the stochastic process (n_1, n_2), an assumption which in reality can not be justified anyway, the distributionally robust constraint formulation both makes practical sense and leads to a computationally tractable formulation.

Control	J	$(\phi, h) \notin B(6)$	$(\dot{\phi}, \dot{h}) \notin B(55)$
Uncontrolled	0	0.16	0.12
$K^\star$	82	0.10	0.10
$K_{\mathrm{LQR}}(0.43)$	82	0.16	0.09
$K_{\mathrm{LQR}}(0.1)$	425	0.15	0.07
$K_{\mathrm{LQR}}(3.2\mathrm{e}{-3})$	3730	0.10	0.05

Table 9.1: Numerical results for the wind turbine blade control problem. The third and fourth column show the worst-case probability that $(\phi, h) \notin B(6)$ and $(\dot{\phi}, \dot{h}) \notin B(55)$, respectively.

10 Discussion and outlook

10.1 Contributions of this dissertation

This dissertation presents novel contributions related to the three principal objectives put forward in Chapter 1. Each of these objectives was dealt with in its corresponding part of the dissertation. The specific contributions, organized by part, are as follows:

10.1.1 Convexity and probability

The first part of the dissertation attempts to argue that many concepts in probability theory enjoy an underlying convex structure very similar to convexity in R^n. Indeed, the sets of distributions defined through moment conditions, a unimodality or monotonicity property are all shown to be convex sets. This means in particular that many results found in this dissertation concerning probability theory and uncertainty quantification problems have a direct counterpart in either convex analysis or convex optimization in R^n. The first part of this dissertation states few truly novel results. Having said that, the perspective offered on uncertainty quantification problems as convex optimization problems over convex sets of probability distributions has, so we think, significant merit.

The offered perspective aspires to persuade the reader that essentially the same mathematical tools used to reformulate standard worst-case robust constraints apply equally well to distributionally robust formulations too. The crucial difference being that for the former constraints we exploit the convexity of the set of all possible realizations of an uncertain disturbance realizing in R^n and in the later constraints the convexity of the set of viable distributions of a random variable on R^n. The striking similarity between vectors in R^n and probability distributions on R^n is made explicit by the intentional analogy between Chapters 2 and 3 dealing with convex analysis and Chapters 4 and 5 on optimization over convex sets of vectors and probability distributions respectively.

Almost all results in this thesis draw from three fundamental ideas introduced in the first part of the dissertation:

1. Choquet representations of convex sets of distributions.

2. The fundamental theorem of linear programming.
3. Conic duality between measures and functions on R^n.

The three aforementioned results take an almost identical form in case of sets of vectors in R^n. Indeed, these three fundamental ideas are often best understood intuitively through the geometric interpretation offered by their finite dimensional counterparts.

10.1.2 Uncertainty quantification

In the second part of the dissertation we generalized the classical 19th century probability bounds discussed in Chapter 1 to events in arbitrary dimensions. This was done by considering worst-case expectation bounds for sets of structured distributions sharing a known mean vector and second moment matrix. Emphasis was put on the structured sets of all α-unimodal and γ-monotone distributions, although many of the result apply to more general structural properties equally well.

The central result in this part of the dissertation is the observation that an uncertainty quantification problem over a Choquet star simplex can be reduced to an equivalent transformed uncertainty quantification problem over the standard probability simplex. The previous reduction proves extremely beneficial to the exposition of the dissertation as only unstructured uncertainty quantification problems over the standard probability simplex need to be considered initially. These resulting unstructured uncertainty quantification problems can be analyzed further from either a primal or, equivalently, their dual optimization perspective.

In Chapter 6, via the primal optimization perspective, we revealed for the first time that the generalized Chebyshev bound reported by Vandenberghe et al. [133] admits a counterpart for unimodal distributions as well. In doing so we provide a much anticipated extension of the univariate classical Gauss bound to events in arbitrary dimensions. Using the notion of α-unimodality, the Gauss bound (for $\alpha = n$) and Chebyshev bound (letting $\alpha \to \infty$) are furthermore embedded as two extreme elements in a hierarchy of worst-case α-unimodal probability bounds all of which admit a tractable representation in terms of a semi-definite program.

The dual optimization perspective taken in Chapter 7 allowed for the generalization of many other worst-case probability and conditional value-at-risk bounds found in the literature to more richly structured sets of distributions. The dual perspective yields a semi-infinite convex optimization reformulation which is shown to be amendable to exact and tractable sum-of-squares reformulations. In case of α-unimodal and γ-monotone distributions, we explicitly provide for both corresponding structured worst-case expectation bounds an exact reformulation in terms of a semi-definite program.

10.1.3 Distributionally robust constraints

The results of the first two parts are put to good use in the context of distributionally robust chance and conditional value-at-risk constraints in the last part of the dissertation. The application and tractability of distributionally robust constraints with second-order moment information is discussed in the static optimization context of Chapter 8 and for dynamic optimal control problems in Chapter 9.

In the static optimization context of Chapter 8 we observe an important dichotomy between single and joint distributionally robust constraints in terms of their computational tractability. Single linear distributionally robust constraints are shown to admit a tractable reformulation in terms of a second-order cone constraint whatever the structural assumptions made regarding the corresponding ambiguity set. The more general case of joint polytopic distributionally robust constraints proved more challenging but can nevertheless be analyzed using the uncertainty quantification framework put forward in the second part of the dissertation.

In Chapter 9 we advanced the use of distributionally robust conditional value-at-risk constraints as mathematically sound and practically sensible design objectives for optimal control problems. The best linear control policy can be characterized using a tractable semi-definite program in both a finite and infinite horizon context. Furthermore, it is shown that the optimal linear controller separates still into a Kalman filter only influenced by the disturbance characteristics and state feedback gain characterized as a tractable semi-definite program parametric in only the control constraints and objective.

10.2 Directions for future research

In what remains of this chapter, we suggest two directions of research showing great potential along which the results in this dissertation could be extended even further.

10.2.1 Higher-order moment information

The theoretical results presented in the first part of the dissertation concerning the uncertainty quantification problem (5.1) are applicable to general ambiguity sets $\mathcal{C}$ defined through a finite number of moment conditions of the type $\int g_i(x)\,\mathbb{P}(\mathrm{d}x) = m_i$ and possibly further additional structure represented through the set $\mathcal{K}$. Nevertheless, all practical results stated subsequently deal only with quadratic moment functions g_i corresponding to second-order moment information, most often taking shape in a given mean vector or second moment matrix.

Recall that, ideally, the ambiguity set $\mathcal{C}$ should be taken as the smallest set such that the unknown distribution $\mathbb{P}$ of the disturbance ξ of interest is guaranteed to be an element of $\mathcal{C}$. In that case, the uncertainty quantification problem $\sup_{\mathbb{P}\in\mathcal{C}} \mathbf{E}_{\mathbb{P}}[L(\xi)]$ yields the best bound on the quantity of interest $\mathbf{E}_{\mathbb{P}}[L(\xi)]$ for a random variable ξ distributed within $\mathcal{C}$. When only second-moment information is considered, the related ambiguity set may be too large and the corresponding worst-case expectation bound overly pessimistic.

This pessimism innate to worst-case expectation bounds obtained using merely second-order moment information, is alleviated in this dissertation by requiring that the distributions in the ambiguity set $\mathcal{C}$ enjoy additional structure, e.g. unimodality or monotonicity. Higher-order moments such as skewness or Kurtosis could alternatively be taken into account as well by considering higher-order polynomial moment functions g_i thereby reducing the inherent pessimism of second-order probability and expectation bounds even further.

Unfortunately, there are negative results given in Bertsimas and Popescu [18] on the tractability of uncertainty quantification problems including higher-order moment information. Already unstructured uncertainty quantification problems are NP-hard when non-quadratic polynomial moment functions g_i are considered. It would be of interest however to know whether practically relevant restricted subclasses of uncertainty quantification problems with higher-order moment information can nevertheless be approximated systematically. One may think here of for instance the Fréchet type problems discussed in Rüschendorf [111] in which many higher order marginal moments are assumed known, but no or a very limited dependence structure is specified. The application of approximate sum-of-squares techniques similar to the approach taken by Popescu [103] to uncertainty quantification problems could present a good starting point when pursuing this direction of research.

10.2.2 Data-driven uncertainty quantification

An important motivation for the use of any distributionally robust approach is the observation that in practice distributions are never observed directly, but rather need to be estimated from noisy historical data and are thus necessarily incompletely characterized and ambiguous. Relating this to the approach taken in this dissertation, it is indeed the case that estimating the mean and second moment of a random variable is more sensible in practice then aiming to estimate its full distribution. This previous observation is most pronounced in case the random variable is multidimensional as estimating probability distributions in high dimensions is prone to the infamous curse of dimensionality reported by Donoho [45].

To what confidence the mean and second moment can be estimated, or for that matter to what extent structural properties such as unimodality or monotonicity can be detected, is not discussed in this dissertation. The work of Delage and Ye [43] presents valuable ideas on how one can go about constructing a confidence set M containing the mean and second moment $(\mu, S) \in M$ with high confidence starting from independent and identically distributed samples from the unknown distribution. When the confidence set M admits a semidefinite representation, the corresponding uncertainty quantification problems with moment ambiguity can readily be solved using the approach outlined in Section 6.4.

The previous two step approach in which a certain number of a priori fixed moments, e.g. mean and second moment, is estimated first and then a related uncertainty quantification problem is considered afterwards begs the question whether or not a more direct method could be developed. Recently the works of Bertsimas et al. [16] and Esfahani and Kuhn [48] showed great signs of promise in constructing ambiguity sets $\mathcal{C}$ for which $\mathbb{P} \in \mathcal{C}$ with high confidence directly from data without making much structural assumptions on the ambiguity set $\mathcal{C}$. It would be of interest to study to which extent a priori structural information such as unimodality or monotonicity of the distribution $\mathbb{P}$ could be brought to bear in the aforementioned data-driven context as well.

A Mathematical Preliminaries

A.1 The measure spaces $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$

We collect here the necessary material on measure spaces relevant to this dissertation in order to make its exposition somewhat self contained. As entire books, indeed entire bookshelves, are devoted to this topic our compilation will necessary be minimal. We will focus all attention to probability measures on R^n. The presented results can be found in almost any standard reference on measure theory; see for instance the work of Billingsley [22].

A pair $(\Omega, \mathcal{F})$, consisting of an non-empty set Ω and a σ-algebra $\mathcal{F}$ on Ω, is denoted a measure space. A set $\mathcal{F}$ is a σ-algebra on Ω whenever (i) the set $\mathcal{F}$ contains the universe Ω, (ii) for any $A \in \mathcal{F}$ then also its complement $\Omega \setminus A \in \mathcal{F}$, and lastly (iii) the set $\mathcal{F}$ is closed under countable unions, i.e. for all countable collection of sets $A_i \in \mathcal{F}$ then also its union $\cup_i A_i \in \mathcal{F}$.

Definition A.1 (Measurable functions)**.** *A function $\xi : \Omega_1 \to \Omega_2$ between two measure spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ is said to be measurable if*

$$\xi^{-1}(A) := \{x \in \Omega_1 \ : \ \xi(x) \in A\} \in \mathcal{F}_1 \quad \forall A \in \mathcal{F}_2.$$

A function $\mathbb{m} : \mathcal{F} \to \mathrm{R}$ is a denoted a (finite signed) measure if $\mathbb{m}(\emptyset) = 0$ and $\mathbb{m}$ is countable additive, i.e. for all countable collections A_i of pairwise disjoint sets in $\mathcal{F}$ we have $\mathbb{m}(\cup_i A_i) = \sum_i \mathbb{m}(A_i)$. If a measure $\mathbb{P} : \mathcal{F} \to \mathrm{R}_+$ has the property $\mathbb{P}(\Omega) = 1$ then it is called a probability distribution. Taken together, a measure space $(\Omega, \mathcal{F})$ and a probability distribution $\mathbb{P}$ compose a probability space in which case Ω is usually referred to as the universe and $\mathcal{F}$ as the event space. In the context of probability spaces, measurable functions are denoted as random variables. We can relate to a random variable a probability distribution on its range space.

Definition A.2 (Distribution)**.** *The probability distribution $\mathbb{P}_2 : \mathcal{F}_2 \to [0, 1]$ is denoted the distribution of the random variable ξ between the probability space $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and measure space $(\Omega_2, \mathcal{F}_2)$ if*

$$\mathbb{P}_2(A) = \mathbb{P}_1(\xi^{-1}(A)), \quad \forall A \in \mathcal{F}_2.$$

We use the shorthand $\xi \sim \mathbb{P}$ to denote that the random variable has distribution $\mathbb{P}$.

Definition A.3 (Borel σ-algebra)**.** *For a set S we denote with $\mathcal{B}(S)$ the smallest σ-algebra containing all open subset of S.*

We remark here that $\mathcal{B}(S)$ can alternatively, and equivalently, be defined as the intersection of all σ-algebras on S containing all its open subsets. It can be remarked that $\mathcal{B}(S)$ is only well

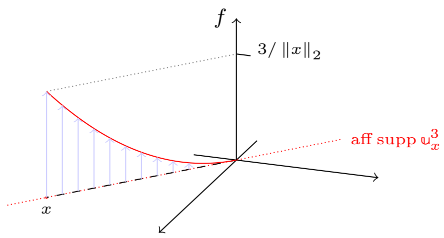


Figure A.1: The radial probability measure $\mathbb{u}_x^3$ is not absolutely continuous and hence admits no density for the Lebesgue measure on R^n. However, the probability measure $\mathbb{u}_x^3$ is absolutely continuous with respect to the Lebesgue measure on the linear subspace $\operatorname{aff}\operatorname{supp}\mathbb{u}_x^3$ generated by its support.

defined when S is endowed with a topology, i.e. its open and closed subsets are appropriately defined.

In this work we are only confronted with the measure spaces $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$ where $\mathcal{B}(\mathrm{R}^n)$ is the Borel σ-algebra on R^n. Throughout the dissertation we assume that R^n is endowed with the standard topology. That is, a set S in R^n is closed if for any sequence $x_i \in S$, $i \in \mathrm{N}$, we have that $\lim_{i\to\infty} x_i \in S$. In what follows, we restrict attention to the measure space $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$ although most of the stated results hold in more general measure spaces.

If $\mathbb{P}_1 \geq 0$ and $\mathbb{P}_2 \geq 0$ are two probability distributions on the same measure space $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$, then $\mathbb{P}_1$ is said to be absolutely continuous with respect to $\mathbb{P}_2$ if $\mathbb{P}_1(A) = 0$ for every set $A \in \mathcal{B}(\mathrm{R}^n)$ for which $\mathbb{P}_2(A) = 0$. The Radon-Nykodym theorem guarantees that if $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_2$ and both are σ-finite, then $\mathbb{P}_1$ admits a density function f with respect to $\mathbb{P}_2$, i.e.

$$\mathbb{P}_1(A) = \int_A f(\omega)\ \mathbb{P}_2(\mathrm{d}\omega), \quad \forall A \in \mathcal{B}(\mathrm{R}^n), \tag{A.1}$$

where $f : \mathrm{R}^n \to \mathrm{R}_+$ is a measurable function. The probability distribution $\mathbb{P}_1$ is said to be absolutely continuous if it is absolutely continuous with respect to the (σ-finite) Lebesgue measure on R^n. We say that f is a density function of $\mathbb{P}_1$ if equation (A.1) holds where $\mathbb{P}_2$ is taken to be the Lebesgue measure, i.e.

$$\mathbb{P}_1(A) = \int_A f(\omega)\ \mathrm{d}\omega, \quad \forall A \in \mathcal{B}(\mathrm{R}^n). \tag{A.2}$$

We remark here that the density function of a continuous measure is usually not unique. The values of f on a finite (or even countably infinite) set of points could be changed to other nonnegative values and equation (A.2) would still hold. We say that $\mathbb{P}_1$ admits a continuous density function if (A.2) holds for some positive continuous function f.

Example A.1 (Radial measures)**.** *The radial probability measures $\mathbb{u}_x^\alpha$ and $\mathbb{m}_x^\gamma$ are not absolutely continuous and hence admit no density for the Lebesgue measure on R^n. However, both probability measures are absolutely continuous with respect to the Lebesgue measure on the linear subspace generated by their support. Indeed, the corresponding density functions f_α and f_γ for $\mathbb{u}_x^\alpha$ and $\mathbb{m}_x^\gamma$ are*

$$f_\alpha(\omega) = \mathbf{1}\left\{0 \leq \omega \leq \|x\|_2\right\} \cdot \alpha \cdot \frac{\omega^{\alpha-1}}{\|x\|_2^\alpha}$$

and

$$f_\gamma(\omega) = \mathbf{1}\left\{0 \leq \omega \leq \|x\|_2\right\} \cdot B(n,\gamma)^{-1} \cdot \frac{\omega^{n-1}}{\|x\|_2^n} \cdot \left(1 - \frac{\omega}{\|x\|_2}\right)^{\gamma-1}$$

respectively. See also Figure A.1 in which the situation for $\mathbb{u}_x^3$ *is depicted.*

A.2 The vector spaces $\mathcal{E}_n$ and $\mathcal{E}_n^\star$

Many of the results stated in Chapters 3 and 5 are expressed in the language of topological vector spaces (over R). We thus deem it fitting to glean the most relevant results concerning topological vector spaces of measures on the measure space $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$ discussed in Section A.1 and put them to paper. The interested reader is referred to the work of Schaefer and Wolff [113] for a more complete and general treatment of the matter.

A vector space over the field R is a set of elements V, denoted as vectors, together with two operations called vector addition and scalar multiplication that satisfy certain axioms. The space $\mathcal{E}_n$ of all (finite signed) measures on the measure space $(\mathrm{R}^n, \mathcal{B}(\mathrm{R}^n))$ is a vector space for the vector addition and scalar multiplication defined, respectively, through

$$(\mathbb{m}_1 + \mathbb{m}_2)(A) := \mathbb{m}_1(A) + \mathbb{m}_2(A) \quad (\alpha\mathbb{m}_1)(A) := \alpha\mathbb{m}_1(A), \quad \forall A \in \mathcal{B}(\mathrm{R}^n)$$

for $\mathbb{m}_1$, $\mathbb{m}_2$ elements of $\mathcal{E}_n$ and $\alpha \in \mathrm{R}$.

In contrast to the intuition stemming from vectors in R^n, there is in general vector spaces no notion of nearness or distance. The mixing hull of a set $\mathcal{S}$ in $\mathcal{E}_n$ as defined in (3.3) does however require $\mathcal{E}_n$ to be endowed a topology. Throughout the dissertation we assume that a set $\mathcal{S}$ in $\mathcal{E}_n$ is closed if for any sequence $\mathbb{m}_i \in \mathcal{S}$, $i \in \mathrm{N}$, we have that $\lim_{i\to\infty} \mathbb{m}_i \in \mathcal{S}$ in the weak sense.

Definition A.4 (Weak convergence)**.** *A sequence* $\{\mathbb{m}_i\}$ *of elements in* $\mathcal{E}_n$ *is said to converge weakly to an element* $\mathbb{m}$ *in* $\mathcal{E}_n$ *if*

$$\int g(x)\,\mathbb{m}_i(\mathrm{d}x) \to \int g(x)\,\mathbb{m}(\mathrm{d}x)$$

for every bounded and continuous measurable function $g : \mathrm{R}^n \to \mathrm{R}$.

Two standard references on weak convergence are the books by Billingsley [21] and Parthasarathy [97]. An important result on weak convergence is the following result due to Prokhorov [106].

Theorem A.1 (Prokhorov)**.** *A subset* $\mathcal{S}$ *of* $\mathcal{E}_n$ *has compact closure under weak convergence if, and only if, for every* $\epsilon > 0$ *there is a compact set* $B \subset \mathrm{R}^n$ *such that*

$$\mathbb{m}(B) \geq 1 - \epsilon$$

for all $\mathbb{m} \in \mathcal{S}$.

We say that two vector spaces V and $V^\star$ are paired, if there is defined a bilinear form $\langle\cdot,\cdot\rangle : V^\star \times V \to \mathrm{R}$. In other words, for any $v^\star \in V$ and $v \in V$, we have that $\langle v^\star, \cdot\rangle$ and $\langle \cdot, v\rangle$ are linear functionals on the spaces V and $V^\star$, respectively. In this dissertation the duality presented in the infinite dimensional setting is based on the pairing of vector spaces.

The topological vector space $\mathcal{E}_n$ endowed with the weak topology can be paired with the space $\mathcal{E}_n^\star$ of all measurable functions $f : \mathrm{R}^n \to \mathrm{R}$ using a bilinear product defined here as

$$\langle f, \mathbb{m}\rangle := \int f(x)\ \mathbb{m}(\mathrm{d}x).$$

Please note that we do not need to equip the dual space $\mathcal{E}_n^\star$ with a topology in this dissertation.

A.3 Equality constrained quadratic programs

We will state here a relevant result concerning equality constrained QPs used throughout the rest of this paper. Assume we define a function $Q : \mathrm{R}^d \to \mathrm{R}$ as follows

$$Q(b) := \min_{x\in\mathrm{R}^n} \quad x^\top Gx + 2x^\top c + y \qquad \text{s.t.} \quad Ax = b,$$

with $A \in \mathrm{R}^{d\times n}$ having full row rank and G positive semidefinite. It is assumed that the function $x^\top Gx + 2x^\top c$ in bounded from below such that $Q(b) > \infty$. We can now represent the quadratic function Q using a dual representation as indicated in the following theorem.

Theorem A.2 (Parametric representation of Q)**.** *The function Q is lower bounded by*

$$Q(b) \geq b^\top T_1 b + 2b^\top T_2 + T_3 \tag{A.3}$$

for all $T_1 \in \mathrm{S}^d$, $T_2 \in \mathrm{R}^d$ and $T_3 \in \mathrm{R}$ such that there exist $\Lambda_1 \in \mathrm{R}^{d\times d}$, $\Lambda_2 \in \mathrm{R}^d$ with

$$\begin{pmatrix} \Lambda_1 + \Lambda_1^\top - T_1 & \Lambda_2 - T_2 & -\Lambda_1^\top A \\ \Lambda_2^\top - T_2^\top & y - T_3 & c^\top - \Lambda_2^\top A \\ -A^\top \Lambda_1 & c - A^\top \Lambda_2 & G \end{pmatrix} \succeq 0. \tag{A.4}$$

Moreover, inequality (A.3) *is tight uniformly in $b \in \mathrm{R}^d$ for some T_1, T_2 and T_3 satisfying condition* (A.4)*.*

Proof. The Lagrangian of the optimization problem defining $Q(b)$ is given as

$$\mathcal{L}(x,\lambda) := x^\top Gx + 2x^\top \left(c + A^\top \lambda\right) - 2\lambda^\top b + y.$$

As $x^\top Gx + 2x^\top c$ is bounded from below on R^n, we have that for all $b \in \mathrm{R}^d$ there exists a minimizer $x^\star$ such that $Q(b) = (x^\star)^\top Gx^\star + 2(x^\star)^\top c + y$ and $Ax^\star = b$. From the first order optimality conditions for convex QPs which can be found in Nocedal and Wright [92, Lemma 16.1], we have that $\min_x \max_\lambda \ \mathcal{L}(x,\lambda) = \mathcal{L}(x^\star,\lambda^\star) = \max_\lambda \min_x \ \mathcal{L}(x,\lambda)$ where the saddle point $(x^\star,\lambda^\star)$ is any solution of the linear system

$$\begin{pmatrix} G & A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} x^\star \\ \lambda^\star \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}. \tag{A.5}$$

The quadratic optimization problem $\max_x \mathcal{L}(x,\lambda^\star)$ admits a maximizer if, and only if, $(c+A^\top \lambda^\star)$ is in the range of G. It must thus hold that

$$\left(\mathbb{I}_d - GG^\dagger\right)\left(c + A^\top \lambda^\star\right) = 0. \tag{A.6}$$

Hence when dualizing the problem defining $Q(b)$, we get its dual representation

$$Q(b) = \max_\lambda \ -\left(c + A^\top \lambda\right)^\top G^\dagger \left(c + A^\top \lambda\right) - 2\lambda^\top b + y.$$

From equation (A.5) it follows that $\lambda^\star$ is any solution of the linear equation $b + AG^\dagger A^\top \lambda^\star + AG^\dagger c = 0$. Therefore there exists an affine $\lambda^\star(b) = -\Lambda_1^\star b - \Lambda_2^\star$ with $\Lambda_1^\star \in \mathrm{R}^{d\times d}$ and $\Lambda_2^\star \in \mathrm{R}^d$ such that

$$Q(b) = -\left(c - A^\top \Lambda_1^\star b - A^\top \Lambda_2^\star\right)^\top G^\dagger \left(c - A^\top \Lambda_1^\star b - A^\top \Lambda_2^\star\right) + 2b^\top {\Lambda_1^\star}^\top b + 2{\Lambda_2^\star}^\top b + y. \tag{A.7}$$

From equation (A.6) it follows that for all $b \in \mathrm{R}^d$ it holds that $\left(\mathbb{I}_d - GG^\dagger\right)\left(c - A^\top \Lambda_1^\star b - A^\top \Lambda_2^\star\right) = 0$. We must hence also have that

$$\left(\mathbb{I}_d - GG^\dagger\right)\left(-A^\top \Lambda_1^\star,\ c - A^\top \Lambda_2^\star\right) = 0 \tag{A.8}$$

The dual reprenstation of $Q(b)$ guarantees that for all $\lambda(b) = -\Lambda_1 b - \Lambda_2$ with $\Lambda_1 \in \mathrm{R}^{d\times d}$ and $\Lambda_2 \in \mathrm{R}^d$

$$Q(b) \geq -\left(c - A^\top \Lambda_1 b - A^\top \Lambda_2\right)^\top G^\dagger \left(c - A^\top \Lambda_1 b - A^\top \Lambda_2\right) + 2b^\top \Lambda_1^\top b + 2\Lambda_2^\top b$$

Lower bounding the right hand side of the previous inequality with $b^\top T_1 b + 2T_2^\top b + T_3$ yields $Q(b) \geq b^\top T_1 b + 2T_2^\top b + T_3$ if for all b in R^d it holds that

$$\begin{pmatrix} b \\ 1 \end{pmatrix}^\top \left[\begin{pmatrix} \Lambda_1 + \Lambda_1^\top - T_1 & \Lambda_2 - T_2 \\ \Lambda_2^\top - T_2^\top & y - T_3 \end{pmatrix} - \begin{pmatrix} -\Lambda_1^\top A \\ c^\top - \Lambda_2^\top A \end{pmatrix} G^\dagger \begin{pmatrix} -A^\top \Lambda_1 & c - A^\top \Lambda_2 \end{pmatrix}\right] \begin{pmatrix} b \\ 1 \end{pmatrix} \geq 0$$

and

$$\left(\mathbb{I}_d - GG^\dagger\right)\left(-A^\top \Lambda_1,\, c - A^\top \Lambda_2\right) = 0.$$

Following Gallier [51, Thm 4.3], we obtain the first part of the theorem after taking a Schur complement

$$\exists \Lambda_1, \Lambda_2 : \quad \begin{pmatrix} \Lambda_1 + \Lambda_1^\top - T_1 & \Lambda_2 - T_2 & -\Lambda_1^\top A \\ \Lambda_2^\top - T_2^\top & y - T_3 & c^\top - \Lambda_2^\top A \\ -A^\top \Lambda_1 & c - A^\top \Lambda_2 & G \end{pmatrix} \succeq 0 \implies Q(b) \geq b^\top T_1 b + 2T_2^\top b + T_3.$$

As $Q(b)$ is a quadratic function there exist $T_1^\star$, $T_2^\star$ and $T_3^\star$ such that $Q(b) = b^\top T_1^\star b + 2{T_2^\star}^\top b + T_3^\star$. The equations (A.7) and (A.8) guarantee [51, Thm 4.3] that

$$\begin{pmatrix} \Lambda_1^\star + {\Lambda_1^\star}^\top - T_1^\star & \Lambda_2^\star - T_2^\star & -{\Lambda_1^\star}^\top A \\ {\Lambda_2^\star}^\top - {T_2^\star}^\top & y - T_3^\star & c^\top - {\Lambda_2^\star}^\top A \\ -A^\top \Lambda_1^\star & c - A^\top \Lambda_2^\star & G \end{pmatrix} \succeq 0$$

completing the proof. □

B Proofs

Corollary 7.2:

From Example 3.2, we have that the generating distribution $\mathbb{T}$ for the set of α-unimodal distributions $\mathcal{U}_\alpha$ satisfies

$$\mathbb{T}([0,t]) = \alpha \int_0^t \lambda^{\alpha-1}\,\mathrm{d}\lambda, \quad \forall t \in [0,1].$$

The moment transformations from Theorem 6.1 become

$$\mu_\alpha := \left[\int_0^\infty \lambda\,\mathbb{T}(\mathrm{d}\lambda)\right]^{-1}\mu = \left[\alpha \int_0^1 \lambda^\alpha(\mathrm{d}\lambda)\right]^{-1}\mu = \frac{\alpha+1}{\alpha}\mu$$

$$S_\alpha := \left[\int_0^\infty \lambda^2\,\mathbb{T}(\mathrm{d}\lambda)\right]^{-1}S = \left[\alpha \int_0^1 \lambda^{\alpha+1}(\mathrm{d}\lambda)\right]^{-1}S = \frac{\alpha+2}{\alpha}S.$$

From Proposition 7.1, the transformed loss function L_s required in Theorem 6.1 can be found as

$$\begin{aligned} L_s(x) &= \max_{i\in I} \mathbb{T}\left(\left[b_i/a_i^\top x, \infty\right)\right) \\ &= \max_{i\in I} f_i(a_i^\top x), \end{aligned}$$

where

$$f_i(q) = \begin{cases} \alpha \displaystyle\int_{b_i/q}^1 \lambda^{\alpha-1}\,\mathrm{d}\lambda, & q \ge b_i, \\ 0 & \text{otherwise.} \end{cases}$$

In order to apply Theorem 7.2, we now need only reformulate the semi-infinite constraint (C_2), i.e. the constraint

$$T_{3,i} + 2qT_{2,i} + q^2T_{1,i} \ge f_i(q) \quad \forall q \in \mathrm{R}, \quad \forall i \in I.$$

Because $0 \in \Xi$ and hence $b_i > 0$, we have equivalently, for each $i \in I$, and for all $q \in \mathrm{R}^+$

$$T_{3,i} + 2qT_{2,i} + q^2T_{1,i} \ge \begin{cases} 1 - (b_i/q)^\alpha & q \ge b_i, \\ 0 & \text{otherwise.} \end{cases}$$

which can be seen to reduce to

$$T_{3,i} + 2qT_{2,i} + q^2T_{1,i} \ge 1 - \frac{b_i^\alpha}{q^\alpha}, \quad \forall q \ge 0.$$

Defining a new scalar variable $\tilde{q}$ and applying the variable substitution $\tilde{q}^{\,w} = q$, this can be rewritten as

$$\tilde{q}^{2\,w+v}T_{1,i} + 2\tilde{q}^{w+v}T_{2,i} + \tilde{q}^v(T_{3,i}-1) + b_i^\alpha \ge 0, \quad \forall \tilde{q} \ge 0$$

after multiplying both sides with $\tilde{q}^v > 0$. The final result is obtained after the substitution $b_i^{1/w}\bar{q} = \tilde{q}$.

Corollary 7.3:

We follow the same approach as the proof of Corollary 7.2, but this time use the generating distribution $\mathbb{T}$ for γ-monotone distributions from Example 3.3, i.e.

$$\mathbb{T}([0,t]) = \frac{1}{B(n,\gamma)} \cdot \int_0^t \lambda^{n-1} \cdot (1-\lambda)^{\gamma-1}\, \mathrm{d}\lambda, \quad \forall t \in [0,1].$$

In this case the moment transformations from Theorem 6.1 become

$$\mu_\gamma := \left[\int_0^\infty \lambda\, \mathbb{T}(\mathrm{d}\lambda)\right]^{-1} \mu \;= \left[\frac{1}{B(n,\gamma)} \int_0^1 \lambda^{n}(1-\lambda)^{\gamma-1}(\mathrm{d}\lambda)\right]^{-1} \mu \;= \frac{n+\gamma}{n}\mu$$

$$S_\gamma := \left[\int_0^\infty \lambda^2\, \mathbb{T}(\mathrm{d}\lambda)\right]^{-1} S = \left[\frac{1}{B(n,\gamma)} \int_0^1 \lambda^{n+1}(1-\lambda)^{\gamma-1}(\mathrm{d}\lambda)\right]^{-1} S = \frac{n+\gamma}{n}\frac{n+\gamma+1}{n+1} S.$$

From Proposition 7.1, the transformed loss function L_s required in Theorem 6.1 become

$$\begin{aligned} L_s(y) &= \max_{i\in I}\; \mathbb{T}\left(\left[b_i/a_i^\top y, \infty\right)\right) \\ &= \max_{i\in I}\; f_i(a_i^\top y). \end{aligned}$$

where

$$f_i(q) = \begin{cases} \dfrac{1}{B(n,\gamma)} \displaystyle\int_{b_i/q}^1 \lambda^{n-1}(1-\lambda)^{\gamma-1}\, \mathrm{d}\lambda, & q \ge b_i, \\ 0 & \text{otherwise.} \end{cases}$$

For $q \ge b_i$, we can use a binomial expansion to evaluate this integral[1], obtaining

$$\begin{aligned} B(n,\gamma) f_i(q) &= B(n,\gamma) - \int_0^{b_i/q} \lambda^{n-1} \cdot (1-\lambda)^{\gamma-1}\, \mathrm{d}\lambda \\ &= B(n,\gamma) - \sum_{k=0}^{\gamma-1} \int_0^{b_i/q} (-1)^k \binom{\gamma-1}{k} \lambda^{n+k-1}\, \mathrm{d}\lambda \\ &= B(n,\gamma) - b_i^n \sum_{k=0}^{\gamma-1} \frac{(-b_i)^k}{n+k} \binom{\gamma-1}{k} \frac{1}{q^{n+k}} \end{aligned}$$

In order to apply Theorem 7.2, we now need only reformulate the semi-infinite constraint (C_2). We obtain, for each $i \in I$, the constraint

$$T_{3,i} + 2qT_{2,i} + q^2 T_{1,i} \ge 1 - \frac{b_i^n}{B(n,\gamma)} \sum_{k=0}^{\gamma-1} \frac{(-b_i)^k}{n+k} \binom{\gamma-1}{k} \frac{1}{q^{n+k}},\ \forall q \ge b_i.$$

recalling that $0 \in \Xi$ and hence $b_i > 0$. We multiply both sides by $q^{n+\gamma-1} > 0$ to produce, for each $i \in I$ the constraint

$$T_{1,i} q^{n+\gamma+1} + 2T_{2,i} q^{n+\gamma} + (T_{3,i} - 1)\, q^{n+\gamma-1} + \frac{b_i^n}{B(n,\gamma)} \sum_{k=0}^{\gamma-1} \frac{(-b_i)^k}{n+k} \binom{\gamma-1}{k} q^{\gamma-k-1} \ge 0, \quad \forall q \ge b_i.$$

The final result is obtained after the substitution $b_i \bar{q} = q$.

[1] Note that the integral amounts to $1 - B(n,\gamma)^{-1} \int_0^{b_i/q} \lambda^{n-1}(1-\lambda)^{\gamma-1}\, \mathrm{d}\lambda = 1 - I_{b_i/q}(n,\gamma)$, where $I_{b_i/q}(n,\gamma)$ is the so-called *regularized incomplete beta function*, i.e. the cumulative distribution function for the beta distribution with shape parameters (n,γ).

Corollary 7.4:

The method of proof follows that of Corollary 7.2, except that we now apply Proposition 7.2 to generate the transformed loss function L_s.

In this case the loss function L is equivalent to $L = d \circ \kappa_\Xi$ with $d(t) = \max\{0, t-1\}$. Recalling from Example 3.2 the generating distribution $\mathbb{T}$ for α-unimodal distributions, we set

$$\begin{aligned} d_s(t) &= \int_0^\infty d(\lambda t)\, \mathbb{T}(\mathrm{d}\lambda) \\ &= \alpha \int_0^1 \max\{0, (\lambda t - 1)\} \lambda^{\alpha-1} \mathrm{d}\lambda, \end{aligned}$$

which is zero for any $t \leq 1$. For $t \geq 1$, we can evaluate the integral to get

$$\begin{aligned} \forall t \geq 1 \ : \ d_s(t) &= \alpha \int_{1/t}^1 (t\lambda^\alpha - \lambda^{\alpha-1}) \mathrm{d}\lambda \\ &= \frac{\alpha}{\alpha+1} t - 1 + \frac{1}{\alpha+1} \left(\frac{1}{t}\right)^\alpha \end{aligned}$$

and then set $L_s(x) = \max_{i \in I} f_i(a_i^\top x)$ where each $f_i(q) := d_s(q/b_i)$.

We can now apply Theorem 7.2 by reformulating the constraint (C_2) for this choice of f_i for each $i \in I$, resulting in the constraint

$$T_{3,i} + 2qT_{2,i} + q^2 T_{1,i} \geq \frac{\alpha}{\alpha+1} \frac{q}{b_i} - 1 + \frac{1}{\alpha+1} \frac{b_i^\alpha}{q^\alpha} \qquad \forall q \geq b_i$$

because $0 \in \Xi$ and hence $b_i > 0$. We define a new scalar variable $\tilde{q}$ and apply the variable substitution $\tilde{q}^{\,w} = q$, resulting in the constraint

$$\tilde{q}^{2w+v} T_{1,i} + \tilde{q}^{w+v} \left(2T_{2,i} - \frac{\alpha}{(\alpha+1) b_i} \right) + \tilde{q}^v \left(1 + T_{3,i}\right) - \frac{b_i^\alpha}{\alpha+1} \geq 0, \quad \forall \tilde{q} \geq b_i^{1/w}$$

after multiplying both sides by $\tilde{q}^v > 0$. The final result is obtained after the substitution $b_i^{1/w} \bar{q} = \tilde{q}$.

Corollary 7.5:

The method of proof parallels that of Corollary 7.4, but this time using the generating distribution $\mathbb{T}$ for γ-monotone distributions from Example 3.3. In this case we set

$$d_s(t) = \frac{1}{B(n,\gamma)} \int_0^1 \max\{0, (\lambda t - 1)\} \lambda^{n-1} (1-\lambda)^{\gamma-1} \mathrm{d}\lambda,$$

which is zero for any $t \leq 1$. For any $t \geq 1$, using a binomial expansion we can evaluate the integral to get

$$\begin{aligned} \forall t \geq 1 \ : \ B(n,\gamma) d_s(t) &= t \int_{1/t}^1 \lambda^n (1-\lambda)^{\gamma-1}\ \mathrm{d}\lambda - \int_{1/t}^1 \lambda^{n-1} (1-\lambda)^{\gamma-1}\ \mathrm{d}\lambda \\ &= tB(n+1,\gamma) - B(n,\gamma) + \int_0^{1/t} \lambda^{n-1} (1-\lambda)^{\gamma-1} \mathrm{d}\lambda - t \int_0^{1/t} \lambda^n (1-\lambda)^{\gamma-1} \mathrm{d}\lambda \\ &= tB(n+1,\gamma) - B(n,\gamma) + \sum_{k=0}^{\gamma-1} \left[(-1)^k \binom{\gamma-1}{k} \int_0^{1/t} \left(\lambda^{n-1} - t\lambda^n\right) \lambda^k\ \mathrm{d}\lambda \right] \\ &= tB(n+1,\gamma) - B(n,\gamma) + \sum_{k=0}^{\gamma-1} \frac{(-1)^k}{(n+k)(n+k+1)} \binom{\gamma-1}{k} \left(\frac{1}{t}\right)^{n+k} \end{aligned}$$

and then set $L_s(x) = \max_{i\in I} f_i(a_i^\top x)$ where each $f_i(q) := d_s(q/b_i)$. In order to apply Theorem 7.2, we now need only reformulate the semi-infinite constraint (C_2). We obtain, for each $i \in I$, the constraint

$$T_{3,i} + 2qT_{2,i} + q^2 T_{1,i} \geq \frac{B(n+1,\gamma)}{b_i B(n,\gamma)} q - 1 + \frac{b_i^n}{B(n,\gamma)} \sum_{k=0}^{\gamma-1} \frac{(-b_i)^k}{(n+k)(n+k+1)} \binom{\gamma-1}{k} \frac{1}{q^{n+k}} \quad \forall q \geq b_i$$

because $0 \in \Xi$ and hence $b_i > 0$. We multiply both sides by $q^{n+\gamma-1} > 0$ to produce the constraint

$$T_{1,i} q^{n+\gamma+1} + \left(2T_{2,i} - \frac{B(n+1,\gamma)}{b_i B(n,\gamma)}\right) q^{n+\gamma} + (T_{3,i}+1)\, q^{n+\gamma-1} - \frac{b_i^n}{B(n,\gamma)} \sum_{k=0}^{\gamma-1} \frac{(-b_i)^k}{(n+k)(n+k+1)} \binom{\gamma-1}{k} q^{\gamma-k-1} \geq 0, \quad \forall q \geq b_i.$$

The final result is obtained after the substitution $b_i \bar{q} = \tilde{q}$.

Factor models in insurance

As mentioned in Section 5.4.2, any worst-case CVaR problem can be reduced to a related worst-case expectation problem. We are therefore interested in loss functions of the form $L(S_n) = \min(\max(S_n, 0), k) - \beta$ for $0 \leq \beta \leq k$. We have that the loss function $L(S_n)$ can be written as the gauge function $L(S_n) = d \circ \kappa_\Xi(S_n)$ for

$$\Xi = \left\{ x \in \mathrm{R}^n \ : \ \sum_{i=1}^n x_i \geq 1 \right\}$$

and

$$d = \begin{cases} 0 & \text{if } t \leq \beta, \\ t - \beta & \text{if } \beta \leq t < k, \\ k - \beta & \text{if } t \geq k. \end{cases}$$

Recalling from Example 3.2 the generating distribution $\mathbb{T}$ for α-unimodal distributions, we set $d_s(t) = \int_0^\infty d(\lambda t)\, \mathbb{T}(\mathrm{d}\lambda)$ which is zero for any $t \leq \beta$. For $\beta \leq t < k$, we can evaluate the integral to get

$$\begin{aligned} \beta \leq \forall t < k \ : \ d_s(t) &= \alpha \int_{\beta/t}^1 (\lambda t - \beta) \lambda^{\alpha-1} \mathrm{d}\lambda \\ &= \frac{\alpha}{\alpha+1} t - \beta + \frac{\beta^{\alpha+1}}{\alpha+1} \frac{1}{t^\alpha}. \end{aligned}$$

Similarly for $t \geq k$, we get

$$\begin{aligned} \forall t \geq k \ : \ d_s(t) &= \alpha \int_{\beta/t}^{k/t} (\lambda t - \beta) \lambda^{\alpha-1} \mathrm{d}\lambda + \alpha \int_{k/t}^1 (k - \beta) \lambda^{\alpha-1} \mathrm{d}\lambda \\ &= k - \beta - \frac{k^{\alpha+1} - \beta^{\alpha+1}}{\alpha+1} \frac{1}{t^\alpha} \end{aligned}$$

and then set $L_s(x) = d_s(\sum_{i=1}^n x_i)$. In order to apply Theorem 7.2, we now need only reformulate the semi-infinite constraint (C_2). This can be done using methods analogous to the method described in the proof of Corollary 7.4, but is omitted here for the sake of brevity. We get finally

$$\left\{\begin{aligned} & T_{1,i}\beta^2 q^{2w+v} + q^{w+v}\beta\left(2T_{2,i} - \frac{\alpha}{\alpha+1}\right) + q^v\left(T_{3,i}+\beta\right) - \frac{\beta}{\alpha+1} \geq 0, \quad 1 \leq \forall q < \left(\frac{k}{\beta}\right)^{1/w} \\ & T_{1,i}k^2 q^{2w+v} + 2kq^{w+v}T_{2,i} + q^v\left(T_{3,i}+\beta-k\right) + k\frac{1-(\beta/k)^{\alpha+1}}{\alpha+1} \geq 0, \quad \forall q \geq 1 \end{aligned}\right\} (C_2)$$

Proof of Lemma 9.2

Recalling (9.9), it is sufficient to prove that the inequality (9.10) actually holds with equality. In other words, it suffices to show that

$$\limsup_{t\to\infty} \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_t\right)\right) \leq \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \limsup_{t\to\infty} \mathbb{P}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_t\right)\right). \tag{B.1}$$

Choose any $\delta > 0$ and $N' \in \mathrm{N}$. From the definitions of the limit superior and the supremum appearing in the left-hand side of (B.1), there exists some time instance $N \geq N' > 0$ and some probability measure $\tilde{\mathbb{P}} \in \mathcal{C}^\infty(\mu,\Sigma)$ such that the left-hand side of (B.1) is bounded by

$$\limsup_{t\to\infty} \sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)} \mathbb{P}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_t\right)\right) \leq \tilde{\mathbb{P}}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_N\right)\right) + \delta. \tag{B.2}$$

Consider now the subsequence $\{x_{kN}\}_{k=0}^\infty$, i.e. the subsequence obtained by taking every N^{th} element of the sequence $\{x_t\}$ beginning from x_0. The elements of this subsequence are related by

$$x_{(k+1)N} = \bar{A}^N x_{kN} + \left[\bar{A}^{N-1}\bar{C},\ldots,\bar{A}\bar{C},\bar{C}\right]\Delta\xi_k, \quad k \in \mathrm{N}$$

where each $\Delta\xi_k := (\xi_{kN};\ldots;\xi_{(k+1)N-1})$ is a collection of disturbances N steps long.

Consider the distribution of $\Delta\xi_0$ under the measure $\tilde{\mathbb{P}}$, i.e. the marginal distribution of the first N elements of the disturbance sequence $\{\xi_t\}$ under $\tilde{\mathbb{P}}$. Construct a probability measure $\mathbb{P}' \in \mathcal{C}^\infty(\mu,\Sigma)$ such that the subsequences $\Delta\xi_k$ are independent and identically distributed (i.i.d.) and such that $\Delta\xi_0$ has the same distribution under both $\mathbb{P}'$ and $\tilde{\mathbb{P}}$. The marginal distribution of the state x_N will then likewise be the same under both $\mathbb{P}'$ and $\tilde{\mathbb{P}}$. Consequently, we must also have

$$\tilde{\mathbb{P}}\text{-CVaR}_\epsilon\left(L^\alpha\left(x_N\right)\right) = \mathbb{P}'\text{-CVaR}_\epsilon\left(L^\alpha\left(x_N\right)\right)$$

as the CVaR is a law invariant risk measure; see Chapter 3.

The loss function L^α is convex and thus satisfies $L^\alpha(x+y) \geq L^\alpha(x) + \partial L^\alpha(x)^\top y$ with ∂L^α a subgradient of L^α. As the CVaR measure is monotone, we obtain the inequality

$$\begin{aligned} \mathbb{P}'\text{-CVaR}_\epsilon\left(L^\alpha\left(x_{(k+1)N}\right)\right) &= \mathbb{P}'\text{-CVaR}_\epsilon\left(L^\alpha\left(\bar{A}^N x_{kN} + \left[\bar{A}^{N-1}\bar{C},\ldots,\bar{A}\bar{C},\bar{C}\right]\Delta\xi_k\right)\right) \\ &\geq \mathbb{P}'\text{-CVaR}_\epsilon\Big(L^\alpha\left(\left[\bar{A}^{N-1}\bar{C},\ldots,\bar{A}\bar{C},\bar{C}\right]\Delta\xi_k\right) + \partial L^\alpha\left(\left[\bar{A}^{N-1}\bar{C},\ldots,\bar{A}\bar{C},\bar{C}\right]\Delta\xi_k\right)^\top \bar{A}^N x_{kN}\Big). \end{aligned}$$

From the definition of the CVaR given in (3.11) and the inequality $(a+b)^+ \geq (a)^+ - |b|$ we can then conclude

$$\begin{aligned} \mathbb{P}'\text{-CVaR}_\epsilon\left(L^\alpha\left(x_{(k+1)N}\right)\right) \geq\ & \mathbb{P}'\text{-CVaR}_\epsilon\left(L^\alpha\left(\left[\bar{A}^{N-1}\bar{C},\ldots,\bar{A}\bar{C},\bar{C}\right]\Delta\xi_k\right)\right) \\ & - \frac{1}{\epsilon}\mathbf{E}_{\mathbb{P}'}\Big[\left|\partial L^\alpha\left(\left[\bar{A}^{N-1}\bar{C},\ldots,\bar{A}\bar{C},\bar{C}\right]\Delta\xi_k\right)^\top \bar{A}^N x_{kN}\right|\Big]. \end{aligned}$$

For every $\delta > 0$ there exists p such that $|x| \leq \epsilon\delta + px^2$. Since the disturbance subsequences $\Delta\xi_k$ are assumed i.i.d. with $x_0 = 0$, the random variable x_N has the same distribution as

$\left[\bar{A}^{N-1}\bar{C},\ldots,\bar{C}\right]\Delta\xi_k$ for all k. Using the fact that x_{kN} is independent from $\Delta\xi_k$ we then obtain

$$\begin{aligned}
&\mathbf{E}_{\mathbb{P}'}\left[\left|\partial L^{\alpha}\left(\left[\bar{A}^{N-1}\bar{C},\ldots,\bar{A}\bar{C},\bar{C}\right]\Delta\xi_k\right)^{\top}(\bar{A}^N)^{\top}x_{kN}\right|\right]\\
&\quad\le p\mathbf{E}_{\mathbb{P}'}\left[\left(\partial L^{\alpha}(x_N)^{\top}\bar{A}^N x_{kN}\right)^{\top}\cdot\left(\partial L^{\alpha}(x_N)^{\top}\bar{A}^N x_{kN}\right)\right]+\epsilon\delta\\
&\quad\le p\mathrm{Tr}\left\{\mathbf{E}_{\mathbb{P}'}\left[\partial L^{\alpha}(x_N)\partial L^{\alpha}(x_N)^{\top}\right]\bar{A}^N\mathbf{E}_{\mathbb{P}'}\left[x_{kN}x_{kN}^{\top}\right](\bar{A}^N)^{\top}\right\}+\epsilon\delta
\end{aligned}$$

using additionally that $C_{xx}(kN)\preceq P_\infty$ for all k. Since N' could be chosen arbitrarily large, we now assume that $N\ge N'>0$ is large enough that

$$p\mathrm{Tr}\left\{\mathbf{E}_{\mathbb{P}'}\left[\partial L^{\alpha}(x_N)\partial L^{\alpha}(x_N)^{\top}\right]\bar{A}^N P_\infty(\bar{A}^N)^{\top}\right\}\le\epsilon\delta,$$

which is always possible when $\bar{A}$ is asymptotically stable and because

$$\mathbf{E}_{\mathbb{P}'}\left[\partial L^{\alpha}(x_N)\partial L^{\alpha}(x_N)^{\top}\right]\preceq E\ \mathbf{E}_{\mathbb{P}'}\left[x_N\cdot x_N^{\top}\right]E^{\top}+e\preceq\infty$$

for any $E\succeq\alpha_i E_i$ and $e\succeq\alpha_i e_i\cdot e_i^{\top}$ for all $i\in[1,\ldots,k]$ is bounded from above uniformly in N. It then follows that

$$\mathbb{P}'\text{-CVaR}_\epsilon\left(L_\alpha\left(x_{(k+1)N}\right)\right)+2\delta\ge\mathbb{P}'\text{-CVaR}_\epsilon\left(L_\alpha\left(x_N\right)\right)=\tilde{\mathbb{P}}\text{-CVaR}_\epsilon\left(L_\alpha\left(x_N\right)\right),\quad\forall k.\qquad\text{(B.3)}$$

Combining the preceding inequalities (B.2) and (B.3), we obtain the inequality

$$\limsup_{t\to\infty}\sup_{\mathbb{P}\in\mathcal{C}^\infty(\mu,\Sigma)}\mathbb{P}\text{-CVaR}_\epsilon\left(L^{\alpha}\left(x_t\right)\right)\le\limsup_{t\to\infty}\mathbb{P}'\text{-CVaR}_\epsilon\left(L^{\alpha}\left(x_t\right)\right)+3\delta.$$

Since $\delta>0$ could be chosen to be arbitrarily small, (B.1) immediately follows and the proof is complete.

Proof of Theorem 9.4

We have according to Lemma 9.2 and Corollary 7.1 the equivalences

$$\begin{aligned}
\sup_{\mathbb{P}\in\mathcal{C}_\infty(\mu,\Sigma)}\limsup_{t\to\infty}\mathbb{P}\text{-CVaR}_\epsilon\left(L_1\left(x_t\right)\right)\le0 &\iff\sup_{\mathbb{Q}_\infty\in\mathcal{Q}_\infty}\mathbb{Q}_\infty\text{-CVaR}_\epsilon\left(L_1\left(x\right)\right)\le0.\\
&\iff\limsup_{t\to\infty}\ \mathrm{Tr}\left\{E_1^{1/2}\mathbf{E}_{\mathbb{P}}\left[x_t x_t^{\top}\right]E_1^{1/2}\right\}\le-e_1^0\epsilon.
\end{aligned}$$

when the closed loop system is stable. Notice that this without loss of generality as an unstable system would yield an unbounded cost J_∞ and hence can be discarded.

The following inequalities

$$\liminf_{t\to\infty}\ \mathbf{E}_{\mathbb{P}}\left[x_t^{\top}Qx_t+u_t^{\top}Ru_t\right]\le J_\infty(\pi)\le\limsup_{t\to\infty}\ \mathbf{E}_{\mathbb{P}}\left[x_t^{\top}Qx_t+u_t^{\top}Ru_t\right],\quad\forall\mathbb{P}\in\mathcal{C}^\infty(\mu,S)$$

follow immediately from the definition of limit inferior and limit superior, respectively. The objective function now can be written in the form of a standard H_2-problem,

$$J_{\mathrm{lqr}}=\lim_{t\to\infty}\mathrm{Tr}\left\{Q^{\frac12}\mathbf{E}_{\mathbb{P}}\left[x_t x_t^{\top}\right]Q^{\frac12}+R^{\frac12}\mathbf{E}_{\mathbb{P}}\left[u_t u_t^{\top}\right]R^{\frac12}\right\},$$

using the fact that the expectation operator is linear and the trace identity $\mathrm{Tr}\{AB\}=\mathrm{Tr}\{BA\}$ and $\mathbf{E}_{\mathbb{P}}\left[x_t x_t^{\top}\right]$ converges for $t\to\infty$. Hence, when restricted to linear control strategies, problem R_∞ reduces to

$$\begin{array}{ll}
\inf_\pi & \lim_{t\to\infty}\ \mathbf{E}_{\mathbb{P}}\left[x_t^{\top}Qx_t+u_t^{\top}Ru_t\right]\\
\text{s.t.} & x_{t+1}=Ax_t+Bu_t+C\xi_t,\\
 & \lim_{t\to\infty}\ \mathrm{Tr}\left\{E_1^{\frac12}\mathbf{E}_{\mathbb{P}}\left[x_t x_t^{\top}\right]E_1^{\frac12}\right\}\le-e_1^0\epsilon.
\end{array}$$

However, the last problem is an instance of a standard multi-criterion H_2-problem, see Boyd et al. [26, Section 12.2.1]. The fact that the optimal control law is of the form (9.11) is a result of the fact that it solves an H_2-problem with a different cost measure, i.e. there exists an unconstrained H_2-problem with state and input penalty matrices $\tilde{Q}$, $\tilde{R}$ for which the solution satisfies the omitted trace constraint as shown by Boyd et al. [26, Section 6.5.1]. The fact that K can be found as the solution to an SDP can be found as well in Boyd et al. [26], and essentially follows from standard LMI manipulations.

Definition of system matrices

Define the matrices $\mathbf{B} \in \mathrm{R}^{N_x \times N_u}$, $\mathbf{C} \in \mathrm{R}^{N_x \times N_w}$, $\mathbf{D} \in \mathrm{R}^{N_y \times N_u}$ and $\mathbf{E} \in \mathrm{R}^{N_y \times N_w}$ as follows

$$\mathbf{B} := \begin{pmatrix} 0 & & & & \\ B & 0 & & & \\ AB & B & 0 & & \\ \vdots & & \ddots & \ddots & \\ \vdots & & & B & 0 \\ A^{N-1}B & A^{N-2}B & \dots & AB & B \end{pmatrix} =: \begin{pmatrix} \mathbf{B}_0 \\ \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \vdots \\ \mathbf{B}_N \end{pmatrix}, \quad \mathbf{D} := \begin{pmatrix} 0 & & & & \\ D & 0 & & & \\ & D & 0 & & \\ & & \ddots & \ddots & \\ & & & D & 0 \end{pmatrix} =: \begin{pmatrix} \mathbf{D}_0 \\ \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_{N-1} \end{pmatrix},$$

$$\mathbf{C} := \begin{pmatrix} x_0 & & & & \\ Ax_0 & C & & & \\ A^2x_0 & AC & C & & \\ \vdots & & \ddots & \ddots & \\ A^N x_0 & A^{N-1}C & \dots & AC & C \end{pmatrix} =: \begin{pmatrix} \mathbf{C}_0 \\ \mathbf{C}_1 \\ \mathbf{C}_2 \\ \vdots \\ \mathcal{C}_N \end{pmatrix}, \quad \mathbf{E} := \begin{pmatrix} 1 & & & \\ & E & & \\ & & E & \\ & & & \ddots \\ & & & & E \end{pmatrix} =: \begin{pmatrix} \mathbf{E}_0 \\ \mathbf{E}_1 \\ \mathbf{E}_2 \\ \vdots \\ \mathbf{E}_{N-1} \end{pmatrix},$$

where x_0 is the initial state of system S, and $N_x := (N+1)n$, $N_u := Nm$, $N_w := Nd + 1$ and $N_y = rN$.

Bibliography

[1] W.F. Arnold III and A.J. Laub. Generalized eigenproblem algorithms and software for algebraic Riccati equations. *Proceedings of the IEEE*, 72(12):1746–1754, 1984.

[2] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

[3] B.R. Barmish and C.M. Lagoa. The uniform distribution: A rigorous justification for its use in robustness analysis. *Mathematics of Control, Signals, and Systems*, 10(3):203–222, 1997.

[4] O.E. Barndorff-Nielsen, M. Maejima, and K.-I. Sato. Some classes of multivariate infinitely divisible distributions admitting stochastic integral representations. *Bernoulli*, 12(1):1–33, 2006.

[5] A. Barvinok. *A Course in Convexity.* American Mathematical Society, 2002.

[6] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization.* Princeton University Press, 2009.

[7] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

[8] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications.* SIAM, 2001.

[9] A. Ben-Tal and A. Nemirovski. Robust optimization–methodology and applications. *Mathematical Programming*, 92(3):453–480, 2002.

[10] A. Ben-Tal and A. Nemirovski. Selected topics in robust convex optimization. *Mathematical Programming*, 112(1):125–158, 2008.

[11] P.O. Berge. A note on a form of Tchebycheff's theorem for two variables. *Biometrika*, 29(3-4):405–406, 1938.

[12] S.N. Bernstein. Sur les fonctions absolument monotones. *Acta Mathematica*, 52(1):1–66, 1929.

[13] E.M.J. Bertin, R. Theodorescu, and I. Cuculescu. *Unimodality of Probability Measures.* Mathematics and its applications. Springer, 1997.

[14] D. Bertsekas and I.B. Rhodes. Sufficiently informative functions and the minimax feedback control of uncertain dynamic systems. *IEEE Transactions on Automatic Control*, 18(2):117–124, 1973.

[15] D. Bertsimas, D.B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.

[16] D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *arXiv preprint arXiv:1401.0212*, 2013.

[17] D. Bertsimas and I. Popescu. On the relation between option and stock prices: A convex optimization approach. *Operations Research*, 50(2):358–374, 2002.

[18] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

[19] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[20] I.-J. Bienaymé. Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *Comptes Rendus de l'Académie des Sciences*, 37:159–184, 1853.

[21] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 1968.

[22] P. Billingsley. *Probability and Measure*. John Wiley & Sons, 1995.

[23] Z.W. Birnbaum, J. Raymond, and H.S. Zuckerman. A generalization of Tshebyshev's inequality to two dimensions. *The Annals of Mathematical Statistics*, 18(1):70–79, 1947.

[24] E. Bishop and K. De Leeuw. The representations of linear functionals by measures on sets of extreme points. In *Annales de l'institut Fourier*, volume 9, pages 305–331. Institut Fourier, 1959.

[25] F. Blanchini. Set invariance in control. *Automatica*, 35(11):1747–1767, 1999.

[26] S.P. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[27] S.P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[28] G.C. Calafiore. Random convex programs. *SIAM Journal on Optimization*, 20(6):3427–3464, 2010.

[29] G.C. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

[30] G.C. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.

[31] C.W. Campbell. Monte Carlo turbulence simulation using rational approximations to von Kármán spectra. *AIAA Journal*, 24(1):62–66, 1986.

[32] M.C. Campi and S. Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

[33] M. Cannon, B. Kouvaritakis, and X. Wu. Probabilistic constrained MPC for multiplicative and additive stochastic uncertainty. *IEEE Transactions on Automatic Control*, 54(7):1626–1632, 2009.

[34] F.P. Cantelli. *Intorno ad un teorema fondamentale della teoria del rischio*. Tip. degli operai, 1910.

[35] A. Charnes, W.W. Cooper, and G.H. Symonds. Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Science*, 4(3):235–263, 1958.

[36] P. Chebyshev. Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées*, 12(2):177–184, 1867.

[37] W. Chen, M. Sim, J. Sun, and C.-P. Teo. From CVaR to uncertainty set: Implications in joint chance-constrained optimization. *Operations Research*, 58(2):470–485, 2010.

[38] J. Cheng, E. Delage, and A. Lisser. Distributionally robust stochastic knapsack problem. Technical report, HEC Montréal, 2013.

[39] J. Danielsson, P. Embrechts, C. Goodhart, C. Keating, F. Muennich, O. Renault, and H.S. Shin. An academic response to Basel II, 2001.

[40] G.B. Dantzig. Linear programming under uncertainty. *Management Science*, 1(3-4):197–206, 1955.

[41] G.B. Dantzig. Reminiscences about the origins of linear programming. *Operations Research Letters*, 1(2):43–48, 1982.

[42] D.P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

[43] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[44] S.W. Dharmadhikari and K. Joag-Dev. *Unimodality, Convexity, and Applications*, volume 27 of *Probability and Mathematical Statistics*. Academic Press, 1988.

[45] D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, pages 1–32, 2000.

[46] L. El Ghaoui, F. Oustry, and H. Lebret. Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, 9(1):33–52, 1998.

[47] P. Embrechts, G. Puccetti, and L. Rüschendorf. Model uncertainty and VaR aggregation. *Journal of Banking & Finance*, 37(8):2750–2764, 2013.

[48] P.M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

[49] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. In *INFOCOM'97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies.*, volume 3, pages 1096–1104. IEEE, 1997.

[50] J. Galambos. Bonferroni inequalities. *The Annals of Probability*, 5(4):577–581, 1977.

[51] J. Gallier. The Schur complement and symmetric positive semidefinite (and definite) matrices. Technical report, Penn Engineering, 2010.

[52] F.R. Gantmacher and M.G. Krein. Oscillation matrices and kernals and small vibrations of dynamical systems. *Gostekhizdat, Moscow*, 1950.

[53] C.E. Garcia, D.M. Prett, and M. Morari. Model predictive control: Theory and practice – a survey. *Automatica*, 25(3):335–348, 1989.

[54] C.F. Gauss. Theoria combinationis observationum erroribus minimis obnoxiae, pars prior. *Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores*, 33:321–327, 1821.

[55] P.J. Goulart, E.C. Kerrigan, and J.M. Maciejowski. Optimization over state feedback policies for robust control with constraints. *Automatica*, 42(4):523–533, 2006.

[56] M. Grant, S. Boyd, and Y. Ye. *Disciplined Convex Programming.* Springer, 2006.

[57] B.D. Grundy. Option prices and the underlying asset's return distribution. *The Journal of Finance*, 46(3):1045–1069, 1991.

[58] M.J. Hadjiyiannis, P.J. Goulart, and D. Kuhn. An efficient method to estimate the suboptimality of affine controllers. *IEEE Transactions on Automatic Control*, 56(12):2841–2853, 2011.

[59] S. Han, M. Tao, U. Topcu, H. Owhadi, and R.M. Murray. Convex optimal uncertainty quantification. *arXiv preprint arXiv:1311.7130*, 2013.

[60] C. M. Harris and N.D. Singpurwalla. Life distributions derived from stochastic hazard functions. *IEEE Transactions on Reliability*, 17(2):70–79, 1968.

[61] R. Hettich and K.O. Kortanek. Semi-infinite programming: Theory, methods, and applications. *SIAM Review*, 35(3):380–429, 1993.

[62] D.A. Iancu, M. Sharma, and M. Sviridenko. Supermodularity and affine policies in dynamic robust optimization. *Operations Research*, 61(4):941–958, 2013.

[63] K. Isii. On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, 1962.

[64] N.P. Jewell. Mixtures of exponential distributions. *The Annals of Statistics*, pages 479–484, 1982.

[65] S. Karlin and W.J. Studden. *Tchebycheff systems: With applications in analysis and statistics*, volume 376. Interscience Publishers, 1966.

[66] J. Katz and A. Plotkin. *Low-Speed Aerodynamics*. Number 13 in Cambridge aerospace series. Cambridge University Press, 2001.

[67] E.C. Kerrigan. *Robust constraint satisfaction: Invariant sets and predictive control*. PhD thesis, University of Cambridge, 2001.

[68] J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American Mathematical Society*, 4(3):502–506, 1953.

[69] M. Krein and D. Milman. On extreme points of regular convex sets. *Studia Mathematica*, 9(1):133–138, 1940.

[70] M.G. Krein and A.A. Nudelman. *The Markov Moment Problem and Extremal Problems: Ideas and Problems of P.L. Čebyšev and A.A. Markov and their Further Development*. American Mathematical Society, 1977.

[71] H. Kwakernaak and R. Sivan. *Linear Optimal Control Systems*. Wiley-Interscience, 1972.

[72] C.M. Lagoa and B.R. Barmish. Distributionally robust monte carlo simulation: A tutorial survey. In *Proceedings of the 15th IFAC world congress*, pages 358–358, 2002.

[73] C.M. Lagoa, P.S. Shcherbakov, and B.R. Barmish. Probabilistic enhancement of classical robustness margins: The unirectangularity concept. *Systems & Control Letters*, 35(1):31–43, 1998.

[74] G.R.G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M.I. Jordan. A robust minimax approach to classification. *The Journal of Machine Learning Research*, 3:555–582, 2003.

[75] H.J. Landau. *Moments in mathematics*. American Mathematical Society, 1987.

[76] J.B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[77] P. Lévy. Extensions d'un théorème de D. Dugué et M. Girault. *Probability Theory and Related Fields*, 1(2):159–173, 1962.

[78] A.W. Lo. Semi-parametric upper bounds for option prices and expected payoffs. *Journal of Financial Economics*, 19(2):373–387, 1987.

[79] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284(1):193–228, 1998.

[80] J. Löfberg. Yalmip: A toolbox for modeling and optimization in matlab. In *Proceedings of the CACSD Conference*, pages 284–289, 2004.

[81] J. Löfberg. Pre- and post-processing sum-of-squares programs in practice. *IEEE Transactions on Automatic Control*, 54(5):1007–1011, 2009.

[82] A. Markov. *On Certain Applications of Algebraic Continued Fractions.* PhD thesis, St. Petersburg, 1884.

[83] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[84] D.Q. Mayne, M.M. Seron, and S.V. Raković. Robust model predictive control of constrained linear systems with bounded disturbances. *Automatica*, 41(2):219–224, 2005.

[85] L.M. Meaux, J.W. Seaman Jr, and T.L. Boullion. Calculation of multivariate Chebyshev-type inequalities. *Computers & Mathematics with Applications*, 20(12):55–60, 1990.

[86] U.S. Military. *U.S. Military Handbook MIL-HDBK-1797*, 1997.

[87] B.L. Miller and H.M. Wagner. Chance constrained programming with joint constraints. *Operations Research*, 13(6):930–945, 1965.

[88] A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.

[89] Y. Nesterov. Squared functional systems and optimization problems. In *High Performance Optimization*, pages 405–440. Springer, 2000.

[90] Y. Nesterov and A. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *Studies in Applied and Numerical Mathematics.* SIAM, 1994.

[91] B.F. Ng, R. Palacios, J.M.R. Graham, and E.C. Kerrigan. Robust control synthesis for gust load alleviation from large aeroelastic models with relaxation of spatial discretisation. *Proceedings of the EWEA*, 2012.

[92] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer New York, 2006.

[93] F. Oldewurtel, A. Parisio, C.N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy and Buildings*, 45:15–27, 2012.

[94] R.A. Olshen and L.J. Savage. A Generalized Unimodality. Technical Report 143, Stanford University, 1969.

[95] H. Owhadi, C. Scovel, T.J. Sullivan, M. McKerns, and M. Ortiz. Optimal uncertainty quantification. *SIAM Review*, 55(2):271–345, 2013.

[96] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 2002.

[97] K.R. Parthasarathy. *Probability Measures on Metric Spaces.* American Mathematical Society, 1967.

[98] K. Pearson. On generalised Tchebycheff theorems in the mathematical theory of statistics. *Biometrika*, 12(3/4):284–296, 1919.

[99] D. Pestana and S. Mendonça. Higher-order monotone functions and probability theory. In *Generalized Convexity and Generalized Monotonicity*, volume 502 of *Lecture Notes in Economics and Mathematical Systems*, pages 317–331. Springer, 2001.

[100] G.C. Pflug. Some remarks on the value-at-risk and the conditional value-at-risk. In *Probabilistic Constrained Optimization*, pages 272–281. Springer, 2000.

[101] R.R. Phelps. *Lectures on Choquet's Theorem*, volume 1757 of *Lecture Notes in Mathematics*. Springer, 2001.

[102] I. Pólik and T. Terlaky. A survey of the S-lemma. *SIAM Review*, 49(3):371–418, 2007.

[103] I. Popescu. A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research*, 30(3):632–657, 2005.

[104] W.B. Powell. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons, 2007.

[105] A. Prekopa. On probabilistic constrained programming. In *Proceedings of the Princeton Symposium on Mathematical Programming*, pages 113–138. Princeton University Press, 1970.

[106] Y.V. Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

[107] R.T. Rockafellar. Extension of Fenchel's duality theorem for convex functions. *Duke Mathematical Journal*, 33(1):81–89, 1966.

[108] R.T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

[109] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1998.

[110] W.W. Rogosinski. Moments of non-negative mass. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 245(1240):1–27, 1958.

[111] L. Rüschendorf. Fréchet-bounds and their applications. In *Advances in Probability Distributions with Given Marginals*, volume 67 of *Mathematics and its Applications*, pages 151–187. Springer, 1991.

[112] I.R. Savage. Probability inequalities of the Tchebycheff type. *Journal of Research of the National Bureau of Standards-B. Mathematics and Mathematical Physics*, 65:211–222, 1961.

[113] H.H. Schaefer and M.P. Wolff. *Locally Convex Topological Vector Spaces*. Springer, 1999.

[114] A.T. Schwarm and M. Nikolaou. Chance-constrained model predictive control. *AIChE Journal*, 45(8):1743–1752, 1999.

[115] T. Sellke. Generalized Gauss–Chebyshev inequalities for unimodal distributions. *Metrika*, 43(1):107–121, 1996.

[116] A. Shapiro. On duality theory of conic linear problems. *Nonconvex Optimization and its Applications*, 57:135–155, 2001.

[117] A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542, 2002.

[118] M. Slater. Lagrange multipliers revisited. In *Traces and Emergence of Nonlinear Programming*, pages 293–306. Springer, 2014.

[119] J.E. Smith. Generalized Chebychev inequalities: Theory and applications in decision analysis. *Operations Research*, 43(5):807–825, 1995.

[120] A.L. Soyster. Technical note–Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21(5):1154–1157, 1973.

[121] B. Stellato. Data-driven chance constrained optimization. Master thesis, ETH Zürich, 2014.

[122] T.J. Stieltjes. Quelques recherches sur la théorie des quadratures dites mécaniques. 3(1):409–426, 1884.

[123] S.M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465–474, 1981.

[124] P. Stoica and R.L. Moses. *Introduction to Spectral Analysis.* Prentice hall, 1997.

[125] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[126] T. Theodorsen. *General Theory of Aerodynamic Instability and the Mechanism of Flutter.* NACA, 1935.

[127] R.H. Tütüncü, K.C. Toh, and M.J. Todd. Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95(2):189–217, 2003.

[128] B.P.G. Van Parys, P.J. Goulart, and D. Kuhn. Generalized Gauss inequalities via semidefinite programming. *Mathematical Programming*, pages 1–32, 2015.

[129] B.P.G. Van Parys, P.J. Goulart, and M. Morari. Infinite horizon performance bounds for uncertain constrained systems. *IEEE Transactions on Automatic Control*, 58(11):2803–2817, Nov 2013.

[130] B.P.G. Van Parys, P.J. Goulart, and M. Morari. Distributionally robust expectation inequalities for structured distributions. *Optimization Online*, 2015.

[131] B.P.G. Van Parys, D. Kuhn, P.J. Goulart, and M. Morari. Distributionally robust control of constrained stochastic systems. *IEEE Transactions on Automatic Control*, 2013.

[132] B.P.G. Van Parys, B.F. Ng, P.J. Goulart, and R. Palacios. Optimal control for load alleviation in wind turbines. In *32nd ASME Wind Energy Symposium.* AIAA, 2014.

[133] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized Chebyshev bounds via semidefinite programming. *SIAM Review*, 49(1):52–64, 2007.

[134] S.A. Vorobyov, H. Chen, and A.B. Gershman. On the relationship between robust minimum variance beamformers with probabilistic and worst-case distortionless response constraints. *IEEE Transactions on Signal Processing*, 56(11):5719–5724, 2008.

[135] H.S. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal on Control*, 6(1):131–147, 1968.

[136] H. Xu, C. Caramanis, and S. Mannor. Optimization under probabilistic envelope constraints. *Operations Research*, 60(3):682–699, 2012.

[137] Y. Yamada and J.A. Primbs. Value-at-risk estimation for dynamic hedging. *International Journal of Theoretical and Applied Finance*, 5(4):333–354, 2002.

[138] Y. Ye. *Interior Point Algorithms: Theory and Analysis*, volume 44 of *Wiley Series in Discrete Mathematics and Optimization*. John Wiley & Sons, 2011.

[139] Y.-L. Yu, Y. Li, D. Schuurmans, and C. Szepesvári. A general projection property for distribution families. In *Advances in Neural Information Processing Systems*, pages 2232–2240, 2009.

[140] C. Zalinescu. *Convex Analysis in General Vector Spaces*. World Scientific, 2002.

[141] G. Zhu, M.A. Rotea, and R. Skelton. A convergent algorithm for the output covariance constraint control problem. *SIAM Journal on Control and Optimization*, 35(1):341–361, 1997.

[142] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198, 2013.

[143] S. Zymler, D. Kuhn, and B. Rustem. Worst-case value-at-risk of non-linear portfolios. *Management Science*, 59(1):172–188, 2013.

Curriculum Vitae

Timeline

Sept. 2011 – Aug. 2015	Scientific Assistent IfA, ETH Zurich, Switzerland
Summer 2010	Research Intern IWR, University of Heidelberg, Germany
Sept. 2009 – July 2011	ME in Engineering Mathematics University of Leuven, Belgium
Summer 2009	Research Intern COSIC, University of Leuven, Belgium
Sept. 2006 – July 2009	BE in Electrical Engineering University of Leuven, Belgium

Scholarships & prizes

November 2014	Finalist INFORMS George Nicholson student paper competition with the article "*Generalized Gauss inequalities via semidefinite programming*"
Februari 2013	Marie Curie Scholarship, PITN-GA-2010-264940