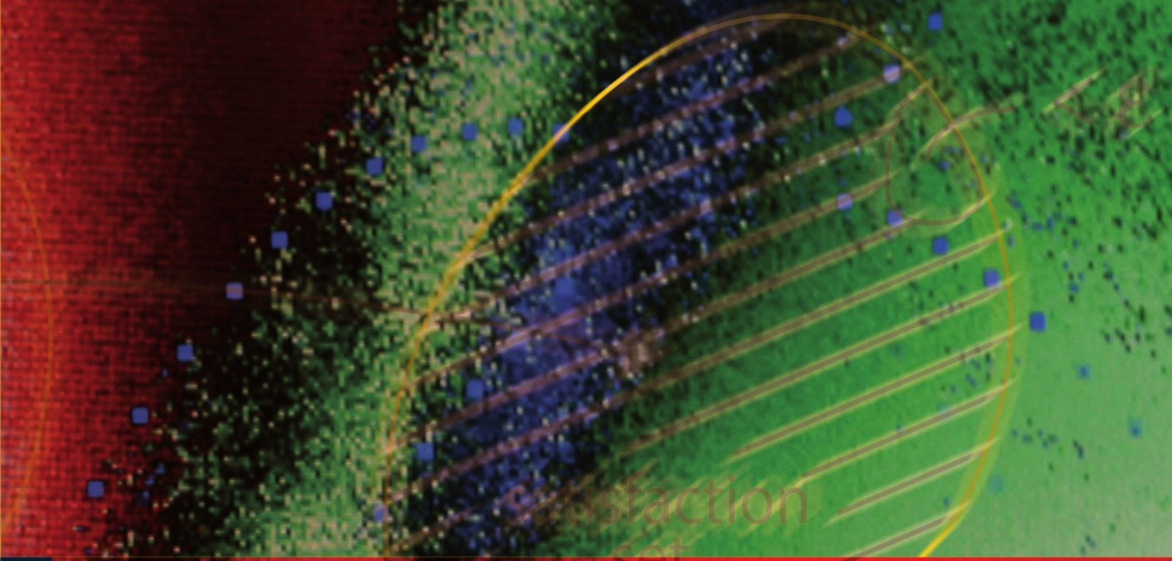


Probabilistic and Randomized Methods for Design under Uncertainty

Giuseppe Calafiore
Fabrizio Dabbene
Editors

Springer

Probabilistic and Randomized Methods for Design under Uncertainty

Giuseppe Calafiore and Fabrizio Dabbene (Eds.)

Probabilistic and Randomized Methods for Design under Uncertainty

With 21 Figures

Springer

Giuseppe Calafiore, PhD
Dipartimento di Automatica e Informatica
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino
Italy

Fabrizio Dabbene, PhD
IEIIT-CNR
Politecnico di Torino
Corso Duca degli Abruzzi, 24
10129 Torino
Italy

British Library Cataloguing in Publication Data
Probabilistic and randomized methods for design under
uncertainty
1. Engineering design - Mathematical models 2. Mathematical
optimization 3. Stochastic systems 4. Sampling (Statistics)
5. Uncertainty (Information theory)
I. Calafiore, Giuseppe, 1969- II. Dabbene, Fabrizio, 1968-
620'.0042'015196
ISBN-10: 184628094X

Library of Congress Control Number: 2005935660

ISBN-10: 1-84628-094-X e-ISBN 1-84628-095-8 Printed on acid-free paper
ISBN-13: 978-1-84628-094-8

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Printed in Germany

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springer.com

To our old and new families

Preface

Solum certum nihil esse certi
Plinius, *Naturalis Historia*

A central issue in many engineering design endeavors is the presence of *uncertainty* in the problem description. Different application fields employ different characterizations of the uncertainty (for instance, deterministic unknown-but-bounded description *vs* stochastic description) and correspondingly adopt different techniques to devise 'designs' that are in some way insensitive, or *robust*, with respect to uncertainty.

The classes of problems considered in this book refer to uncertain control systems, as well as to generic decision or optimization problems in which the data are not exactly known.

In the area of robust control, the approach relying on a purely deterministic unknown-but-bounded description of the uncertainty spawned researches that yielded significant results in the last thirty years. However, this deterministic, or worst-case, approach also showed some inherent limitations that can be resumed in the fundamental tradeoff between computational complexity and 'conservatism'.

From a more philosophical perspective, a worst-case design, even if it could come at a cheap computational cost, may not be a desirable design in practice, since it contemplates *all* possible uncertainty scenarios, including those that are extremely unlikely to happen. On the other hand, one might think that a worst-case approach is necessary in situations where even rare events may lead to disastrous consequences. It should nevertheless be noticed that a design based on an unknown-but-bounded description of the uncertainty may lead to a 'false' belief of safety, since it provides no guarantees for uncertainty outcomes that, for some unforeseen reason, happen to fall outside the *a-priori* assumed uncertainty set. In this respect, an alternative approach based on a stochastic description of the uncertainty makes it clear right from the outset that, in reality, no statement can be given with absolute certainty.

In the field of optimization, the stochastic approach is indeed a classical one, dating back to the late fifties with the work of Dantzig [96] on linear programming under uncertainty and Charnes and Cooper [80] on chance-constrained optimization. In the robust control area, instead, the stochastic approach still enjoys limited attention, since this area has been dominated lately by the worst-case deterministic viewpoint; early exceptions include [192, 195, 342] and later [207, 290, 347, 402]. From an historical point of view, we notice that the two areas of optimization and control followed two different routes: the optimization area has always been dominated by the stochastic paradigm, and only recently the works of El Ghaoui and Lebret [121] and Ben-Tal and Nemirovski [35] brought the worst-case approach into this area. Conversely, since the early eighties robust control has been mainly based on the worst-case paradigm; see the seminal works of Zames [407] on $\mathcal{H}_\infty$ control, of Kharitonov [183] on parametric uncertainty, and the structured singular value theory of Doyle [114] and Safonov [316]. Lately, however, the probabilistic paradigm gained new interest among control researchers, see, *e.g.*, [24, 74, 252, 359, 381] and the many references therein.

Book Scope and Structure

This book brings together leading researchers from both the optimization and control areas, with the intent of highlighting the interactions between the two fields, and with a focus on randomized and probabilistic techniques for solving design problems in the presence of stochastic uncertainty.

The book is divided into three parts. The first part presents three contributions dealing with the general theory and solution methodologies for probability-constrained and stochastic optimization problems and a contribution on the theory of risk measures. The second part of the book contains five chapters devoted to explicit robust design methods based on uncertainty randomization and sampling. The first chapter of the third part of the book presents a novel statistical learning theory framework for system identification, whereas the other six chapters in this part focus on applications of randomized methods for analysis and design of robust control systems.

Acknowledgments

We sincerely thank all the outstanding researchers and friends that made this project possible with their contributions. The financial support of the Italian Ministry of University and Research through an FIRB grant is gratefully acknowledged.

This book goal will be attained if it will stimulate further exchange and dialogue among scientists from the control and the optimization areas, and if it will foster further interest and attract new researchers to these fields.

G. Calafiore
F. Dabbene

Torino, November 2005

Contents

Part I Chance-Constrained and Stochastic Optimization

Part II Robust Optimization and Random Sampling

List of Contributors

Chaouki T. Abdallah
Dept. Electrical & Computer Eng.
University of New Mexico
Albuquerque - New Mexico,
87131-0001
chaouki@ece.unm.edu

Jorge Aravena
Dept. Electrical & Computer Eng.
Louisiana State University
Baton Rouge - Louisiana, 70803
aravena@ece.lsu.edu

Giuseppe Calafiore
Dip. Automatica e Informatica
Politecnico di Torino
Torino - Italy, 10129
giuseppe.calafiore@polito.it

Marco C. Campi
Dip. Elettronica per l'Automazione
Università di Brescia
Brescia - Italy, 25123
campi@ing.unibs.it

Xinjia Chen
Dept. Electrical & Computer Eng.
Louisiana State University
Baton Rouge - Louisiana, 70803
chan@ece.lsu.edu

Fabrizio Dabbene
IEEIT-CNR
Politecnico di Torino
Torino - Italy, 10129
fabrizio.dabbene@polito.it

Darinka Dentcheva
Dept. Mathematical Sciences
Stevens Institute of Technology
Castle Point on Hudson
Hoboken - New Jersey, 07030
ddentche@stevens-tech.edu

Vivek F. Farias
Dept. Electrical Engineering
Stanford University
vff@stanford.edu

Yasumasa Fujisaki
Dept. Computer & Systems
Engineering
Kobe University
Nada, Kobe - Japan, 657-8501
fujisaki@cs.kobe-u.ac.jp

Stacy D. Hill
Applied Physics Laboratory
The Johns Hopkins University
11100 Johns Hopkins Road
Laurel - Maryland, 20723-6099
stacy.hill@jhuapl.edu

Peter F. Hokayem
Coordinated Science Laboratory
University of Illinois
1308 W. Main Street
Urbana - Illinois, 61801
hal@uiuc.edu

Stoyan Kanev
DCSC, TU-Delft
Mekelweg 2
CD Delft - the Netherlands, 2628
s.kanev@dcsc.tudelft.nl

Rajeeva L. Karandikar
Indian Statistical Institute
S.J.S. Sansawal Marg
New Delhi - India, 110 016
rlk@isid.ac.in

Yasuaki Kozawa
Grad. School Science & Technology
Kobe University
Nada, Kobe - Japan, 657-8501

Constantino M. Lagoa
Dept. Electrical Engineering
The Pennsylvania State University
University Park - Penn., 16802
lagoa@engr.psu.edu

Xiang Li
Dept. Electrical Engineering
The Pennsylvania State University
University Park - Penn., 16802
xiangli@psu.edu

Silvia Mastellone
Coordinated Science Laboratory
University of Illinois
1308 W. Main Street
Urbana - Illinois 61801
smastel2@uiuc.edu

Maria Cecilia Mazzaro
Dept. Electrical Engineering
The Pennsylvania State University
University Park - Penn., 16802
cmazzaro@gandalf.ee.psu.edu

Sean P. Meyn
Dept. Electrical & Computer Eng.
Coordinated Science Laboratory
University of Illinois
Urbana-Champaign - Illinois, 61801
meyn@uiuc.edu

Arkadi Nemirovski
Technion
Israel Institute of Technology
Haifa - Israel, 32000
nemirovs@ie.technion.ac.il

Yasuaki Oishi
Dept. Mathematical Informatics
Grad. School Inf. Science & Tech.
The University of Tokyo
Tokyo - Japan, 113-8656
oishi@mist.i.u-tokyo.ac.jp

Andrzej Ruszczyński
Rutgers University
Piscataway - New Jersey, 08854
rusz@rutcor.rutgers.edu

Alexander Shapiro
Georgia Institute of Technology
Atlanta - Georgia, 30332-0205
ashapiro@isye.gatech.edu

James C. Spall
Applied Physics Laboratory
The Johns Hopkins University
11100 Johns Hopkins Road
Laurel - Maryland, 20723-6099
james.spall@jhuapl.edu

David R. Stark
Applied Physics Laboratory
The Johns Hopkins University
11100 Johns Hopkins Road
Laurel - Maryland, 20723-6099
david.stark@jhuapl.edu

Robert F. Stengel
Mechanical & Aerospace Engineering
Princeton University
Princeton - New Jersey, 08544
stengel@princeton.edu

Mario Sznaier
Dept. Electrical Engineering
The Pennsylvania State University
University Park - Penn., 16802
msznaier@frodo.ee.psu.edu

Vladislav B. Tadić
Dept. Aut. Control & Sys. Eng.
University of Sheffield
Sheffield - United Kingdom, S1 3JD
v.tadic@sheffield.ac.uk

Roberto Tempo
IEEIT-CNR
Politecnico di Torino
Torino - Italy, 10129
tempo@polito.it

Benjamin Van Roy
Dept. Electrical Engineering
Stanford University
bvr@stanford.edu

Michel Verhaegen
DCSC, TU-Delft
Mekelweg 2,
CD Delft - the Netherlands, 2628
m.verhaegen@dcsc.tudelft.nl

Mathukumalli Vidyasagar
Tata Consultancy Services
No. 1, Software Units Layout
Hyderabad - India, 500 081
sagar@atc.tcs.co.in

Qian Wang
Mechanical Engineering
The Pennsylvania State University
University Park - Penn., 16802
quw6@psu.edu

Kemin Zhou
Dept. Electrical & Computer Eng.
Louisiana State University
Baton Rouge - Louisiana, 70803
kemin@ece.lsu.edu

Part I

Chance-Constrained and Stochastic Optimization

1

Scenario Approximations of Chance Constraints

Arkadi Nemirovski[1] and Alexander Shapiro[2]

[1] Technion – Israel Institute of Technology, Haifa 32000, Israel,
nemirovs@ie.technion.ac.il

[2] Georgia Institute of Technology, Atlanta, Georgia 30332-0205, USA,
ashapiro@isye.gatech.edu

Summary. We consider an optimization problem of minimization of a linear function subject to the chance constraint $\mathbb{P}\{G(x,\xi)\in C\}\geq 1-\varepsilon$, where C is a convex set, $G(x,\xi)$ is bi-affine mapping and ξ is a vector of random perturbations with known distribution. When C is multi-dimensional and ε is small, like 10^{-6} or 10^{-10}, this problem is, generically, a problem of minimizing under a nonconvex and difficult to compute constraint and as such is computationally intractable. We investigate the potential of conceptually simple *scenario approximation* of the chance constraint. That is, approximation of the form $G(x,\eta^j)\in C$, $j=1,...,N$, where $\{\eta^j\}_{j=1}^N$ is a sample drawn from a properly chosen trial distribution. The emphasis is on the situation where the solution to the approximation should, with probability at least $1-\delta$, be feasible for the problem of interest, while the sample size N should be polynomial in the size of this problem and in $\ln(1/\varepsilon)$, $\ln(1/\delta)$.

1.1 Introduction

Consider the following optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) \;\text{ subject to }\; G(x,\xi)\in C, \tag{1.1}$$

where $C\subset\mathbb{R}^m$ is a closed convex set and $f(x)$ is a real valued function. We assume that the constraint mapping $G:\mathbb{R}^n\times\mathbb{R}^d\to\mathbb{R}^m$ depends on uncertain parameters represented by vector ξ which can vary in a set $\Xi\subset\mathbb{R}^d$. Of course, for a fixed $\xi\in\Xi$, the constraint $G(x,\xi)\in C$ means existence of $z\in C$ such that $G(x,\xi)=z$. In particular, suppose that the set C is given in the form

$$C \doteq \big\{z : z = Wy - w,\ y\in\mathbb{R}^\ell,\ w\in\mathbb{R}^m_+\big\}, \tag{1.2}$$

where W is a given matrix. Then the constraint $G(x,\xi)\in C$ means that the system $Wy\geq G(x,\xi)$ has a feasible solution $y=y(\xi)$. Given x and ξ, we refer to the problem of finding $y\in\mathbb{R}^\ell$ satisfying $Wy\geq G(x,\xi)$ as the second stage feasibility problem.

We didn't specify yet for what values of the uncertain parameters the corresponding constraints should be satisfied. One way of dealing with this is to require the constraints to hold for *every* possible realization $\xi \in \varXi$. If we view ξ as a random vector with a (known) probability distribution having support[3] $\varXi$, this requires the second stage feasibility problem to be solvable (feasible) with probability one. In many situations this may be too conservative, and a more realistic requirement is to ensure feasibility of the second stage problem with probability close to one, say at least with probability $1-\varepsilon$. When ε is really small, like $\varepsilon = 10^{-6}$ or $\varepsilon = 10^{-12}$, for all practical purposes confidence $1-\varepsilon$ is as good as confidence 1. At the same time, it is well known that passing from $\varepsilon = 0$ to a positive ε, even as small as 10^{-12}, may improve significantly the optimal value in the corresponding two-stage problem.

The chance constraints version of problem (1.1) involves constraints of the form

$$\mathbb{P}\big\{G(x,\xi) \in C\big\} \geq 1-\varepsilon. \tag{1.3}$$

Chance constrained problems were studied extensively in the stochastic programming literature (see, *e.g.*, [281] and references therein). We call $\varepsilon > 0$ the *confidence parameter* of chance constraint (1.3), and every x satisfying (1.3) as an $(1-\varepsilon)$-*confident* solution to (1.1). Our goal is to describe the set X_ε of $(1-\varepsilon)$-confident solutions in a 'computationally meaningful' way allowing for subsequent optimization of a given objective over this set. Unless stated otherwise we assume that the constraint mapping is linear in ξ and has the form

$$G(x,\xi) \doteq A_0(x) + \sigma \sum_{i=1}^{d} \xi_i A_i(x), \tag{1.4}$$

where $\sigma \geq 0$ is a coefficient, representing the perturbation level of the problem, and $A_i : \mathbb{R}^n \to \mathbb{R}^m$, $i = 0, ..., d$, are given affine mappings. Of course, the coefficient σ can be absorbed into the perturbation vector ξ. However, in the sequel we use techniques which involve change of the perturbation level of the data. Sometimes we use notation $G_\sigma(x,\xi)$ for the right hand side of (1.4) in order to emphasize its dependence on the perturbation level of the problem.

Example 1. Suppose that we want to design a communication network with p terminal nodes and n arcs. The topology of the network is given, and all we need to specify is vector x of capacities of the arcs; $c^T x$ is the cost of the network to be minimized. The load d in the would-be network (that is, the amounts of data d_{rs}, $r, s = 1, ..., p$, to be transmitted from terminal node r to terminal node s per unit time) is uncertain and is modelled as $d_{rs} = d^*_{rs} + \xi_{rs}$, where d^* is the nominal demand and $\xi = \{\xi_{rs}\}$ is a vector of random perturbations which is supposed to vary in a given set $\varXi$. The network can carry load d if the associated multicommodity flow problem (to assign arcs γ with flows $y^\gamma_{rs} \geq 0$ – amounts of data with origin at r and destination at s

[3] The support of the probability distribution of random vector ξ is the smallest closed set $\varXi \subset \mathbb{R}^d$ such that the probability of the event $\{\xi \in \varXi\}$ is equal to one.

passing through γ – obeying the standard flow conservation constraints with 'boundary conditions' d and the capacity bounds $\sum_{r,s} y^{\gamma}_{rs} \leq x_{\gamma}$) is solvable. This requirement can be formulated as existence of vector y such that $Wy \geq G(x,d)$, where W is a matrix and $G(x,d)$ is an affine function of x and the load d, associated with the considered network. When the design specifications require 'absolute reliability' of the network, *i.e.*, it should be capable to carry every realization of random load, the network design problem can be modelled as problem (1.1) with the requirement that the corresponding constraints $G(x,\xi) \in C$ should be satisfied for every $\xi \in \varXi$. This, however, can lead to a decision which is too conservative for practical purposes.

As an illustration, consider the simplest case of the network design problem, where p 'customer nodes' are linked by arcs of infinite capacity with a central node ('server') c, which, in turn is linked by an arc (with capacity x to be specified) with 'ground node' g, and all data to be transmitted are those from the customer nodes to the ground one; in fact, we are speaking about p jobs sharing a common server with performance x. Suppose that the loads d_r created by jobs r, $r = 1, ..., p$, are independent random variables with, say, uniform distributions in the respective segments $[d^*_r(1-\sigma), d^*_r(1+\sigma)]$, where $\sigma \in (0,1)$ is a given parameter. Then the 'absolutely reliable' optimal solution clearly is

$$x^* = \sum_{r=1}^{p} d^*_r(1+\sigma).$$

At the same time, it can be shown[4] that for $\tau \geq 0$,

$$\mathbb{P}\left\{\sum_r d_r > \sum_r d^*_r + \tau\sigma\sqrt{\sum_r (d^*_r)^2}\right\} \leq e^{-\tau^2/2}.$$

It follows that whenever $\varepsilon \in (0,1)$ and for $D \doteq \sum_{r=1}^{p} d^*_r$, the solution

$$x(\varepsilon) = D + \sigma\sqrt{2\ln(1/\varepsilon)}\sqrt{\sum_r (d^*_r)^2}$$

is $(1-\varepsilon)$-confident. The cost of this solution is by the factor

$$\kappa = \frac{1+\sigma}{1+\sigma\sqrt{2\ln(1/\varepsilon)}\left(\sum_r (d^*_r)^2\right)^{1/2} D^{-1}}$$

less than the cost of the absolutely reliable solution. For example, with $\varepsilon = 10^{-9}$, $p = 1000$ and all d^*_r, $r = 1, ..., p$, equal to each other, we get κ as large as 1.66; reducing ε to 10^{-12}, we still get $\kappa = 1.62$.

[4] This follows from the following inequality due to Hoeffding: if $X_1, ..., X_n$ are independent random variables such that $a_i \leq X_i \leq b_i$, $i = 1, ..., n$, then for $t \geq 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right\} \leq \exp\left\{\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\}.$$

The difference between the absolutely reliable and $(1-\varepsilon)$-confident solutions will be even more dramatic if we assume that d_r are normally distributed independent random variables. Then the corresponding random vector d is supported on the whole space and hence the demand cannot be satisfied with probability one for any value of x, while for any $\varepsilon > 0$, there exists a finite $(1-\varepsilon)$-confident solution.

It is important to point out that 'computationally meaningful' *precise* description of the solution set X_ε of (1.3) seems to be intractable, except for few simple particular cases. Indeed, clearly a necessary condition for existence of a 'computationally meaningful' description of the set X_ε is the possibility to solve efficiently the associated problem for a fixed first stage decision vector: 'given x, check whether $x \in X_\varepsilon$'. To the best of our knowledge, the only generic case where the function

$$\phi(x) \doteq \mathbb{P}\big\{G(x,\xi) \in C\big\}$$

can be efficiently computed analytically is the case where ξ has a normal distribution and C is a segment in $\mathbb{R}$, which is pretty restrictive. Computing $\phi(x)$ is known to be NP-hard already in the situation as simple as the one where ξ is uniformly distributed in a box and C is a polytope.

Of course, there is always a possibility to evaluate $\phi(x)$ by Monte Carlo simulation, provided that C is computationally tractable which basically means that we can check efficiently whether a given point belongs to C. Straightforward simulation, however, requires sample sizes of order ε^{-1} and becomes therefore impractical for ε like 10^{-8} or 10^{-12}. We are not aware of generic cases where this difficulty[5] can be avoided.

Aside from difficulties with efficient computation of $\phi(x)$, there is another severe problem: the set X_ε typically is nonconvex. The only generic exception we know of is again the case of randomly perturbed linear constraint, where C is a segment, with ξ having a normal distribution. Nonconvexity of X_ε makes our ultimate goal (to optimize efficiently over X_ε) highly problematic.

In view of the outlined difficulties, we pass from the announced goal to its relaxed version, where we are looking for 'tractable approximations' of chance constraint (1.3). Specifically, we are looking for *sufficient* conditions for the validity of (1.3), conditions which should be both efficiently verifiable and define a convex set in the space of design variables. The corresponding rationale is clear; we want to stay at the safe side, this is why we are looking for *sufficient* conditions for the validity of (1.3), and we want to be able to optimize efficiently objectives (at least simple ones) under these conditions.

[5] It should be stressed that the difficulties with Monte Carlo estimation of $\mathbb{P}\{\xi \in Q_x\}$, where $Q_x \doteq \{\xi : G(x,\xi) \notin C\}$, come from nonconvexity of Q_x rather than from the fact that we are interested in rare events. Indeed, at least for uniformly distributed ξ, advanced Monte Carlo techniques allow for polynomial time estimation of the quantity $\mathbb{P}\{\xi \in Q\}$ with every fixed relative accuracy, provided that Q is convex, [115,179].

This is why the conditions should be efficiently verifiable and define convex feasible sets.

There are two major avenues for building tractable approximations of chance constraints. The first is to consider one by one interesting generic randomly perturbed constraints (linear, conic quadratic, semidefinite, *etc.*) and to look for specific tractable approximations of their chance counterparts. This approach is easy to implement for linear constraints with $m = 1$ and $C \doteq \mathbb{R}_+$. Then the constraint $G(x, \xi) \in C$ is equivalent to $a^T x + \xi^T A(x) \leq b$, with $A(x)$ being affine in x. Assuming that we know an upper bound V on the covariance matrix of ξ, so that $\mathbb{E}\{(h^T \xi)^2\} \leq h^T V h$ for every vector h, a natural 'safe version' of the random constraint in question is

$$a^T x + \gamma \sqrt{A^T(x) V A(x)} \leq b, \tag{1.5}$$

where $\gamma = \gamma(\varepsilon)$ is a 'safety parameter' which should satisfy the condition

$$\mathbb{P}\big\{\xi : h^T \xi > \gamma \sqrt{h^T V h}\big\} \leq \varepsilon \ \text{ for any } h \in \mathbb{R}^d.$$

An appropriate value of γ can be derived from the standard results on probabilities of large deviations for scalar random variables. For example, for the case when ξ has 'light[6] tail', it suffices to take $\gamma(\varepsilon) = 2\sqrt{1 + \ln(\varepsilon^{-1})}$.

Results of the outlined type can be obtained for randomly perturbed conic quadratic[7] constraints $\|Ax - b\| \leq \tau$ (here $C \doteq \{(y, t) : t \geq \|y\|\}$ is the Lorentz cone), as well as for randomly perturbed semidefinite constraints (C is the semidefinite cone in the space of matrices), see [240]. However, the outlined approach has severe limitations: it hardly could handle the case when C possesses complicated geometry. For example, using 'safe version' (1.5) of a single randomly perturbed linear inequality, one can easily build an approximation of the chance constraint corresponding to the case when C is a polyhedral set given by a list of linear inequalities. At the same time, it seems hopeless to implement the approach in question in the case of a simple two-stage stochastic program, where we need a safe version of the constraint $G(x, \xi) \in C$ with the set C given in the form (1.2). Here the set C, although polyhedral, is *not* given by an explicit list of linear inequalities (such a list can be exponentially long), which makes the aforementioned tools completely inapplicable.

The second avenue of building tractable approximations of chance constraints is the *scenario approach* based on Monte Carlo simulation. Specifically, given the probability distribution $\mathbf{P}$ of random data vector ξ and level of perturbations σ, we choose somehow a 'trial' distribution $\mathbf{F}$ (which does not need to be the same as $\mathbf{P}$). Consequently, we generate a sample $\eta^1, ..., \eta^N$ of N realizations, called *scenarios*, of ξ drawn from the distribution $\mathbf{F}$, and treat the system of constraints

[6] Specifically, $\mathbb{E}\left[\exp\left\{\frac{(h^T\xi)^2}{4h^T V h}\right\}\right] \leq \exp\{1\}$ for every $h \in \mathbb{R}^d$, as in the case where ξ has normal distribution with zero mean and covariance matrix V.

[7] Unless stated otherwise, $\|z\| \doteq (z^T z)^{1/2}$ denotes the Euclidean norm.

$$G(x,\eta^j) \in C,\ j=1,...,N, \tag{1.6}$$

as an approximation of chance constraint (1.3). This is the approach we investigate in this chapter.

The rationale behind this scenario based approach is as follows. First of all, (1.6) is of the same level of 'computational tractability' as the *unperturbed* constraint, so that (1.6) is computationally tractable, provided that C is so and that the number of scenarios N is reasonable. Thus, all we should understand is what can be achieved with a reasonable N. For the time being, let us forget about optimization with respect to x, fix $x=\bar{x}$ and let us ask ourselves what are the relations between the predicates '$\bar{x}$ satisfies (1.3)' and '$\bar{x}$ satisfies (1.6)'. Recall that the random sample $\{\eta^j\}_{j=1}^N$ is drawn from the trial distribution $\mathbf{F}$. We assume in the remainder of this section the following.

> *The trial distribution* $\mathbf{F}$ *is the distribution of* $s\xi$*, where* $s \geq 1$ *is fixed and* $\mathbf{P}$ *is the probability distribution of random vector* ξ*.*

Because of (1.4) we have that $G_\sigma(\bar{x},\xi) \in C$ if and only if $\xi \in Q_{\bar{x},\sigma}$, where

$$Q_{\bar{x},\sigma} \doteq \left\{ z \in \mathbb{R}^d : \sigma \sum_{i=1}^d z_i A_i(\bar{x}) \in C - A_0(\bar{x}) \right\}. \tag{1.7}$$

Note that the set $Q_{\bar{x},\sigma}$ is closed and convex along with C, and for any $s>0$,

$$s^{-1} Q_{\bar{x},\sigma} = \{\xi : G_{s\sigma}(\bar{x},\xi) \in C\}.$$

Now, in 'good cases' $\mathbf{P}$ possesses the following 'concentration' property.

(!) For every closed convex set $Q \subset \mathbb{R}^d$ with $\mathbf{P}(Q)$ not too small, *e.g.*, $\mathbf{P}(Q) \geq 0.9$, the mass $\mathbf{P}(sQ)$ of s-fold enlargement of Q rapidly approaches 1 as s grows. That is, if Q is closed and convex and $\mathbf{P}(Q) \geq 0.9$, then there exists $\kappa > 0$ such that for $s \geq 1$ it holds that

$$\mathbf{P}\big(\{\xi \notin sQ\}\big) \leq e^{-\kappa s^2}. \tag{1.8}$$

(we shall see that, for example, in the case of normal distribution, estimate (1.8) holds true with $\kappa = 0.82$).

Assuming that the above property (!) holds, and given small $\varepsilon > 0$, let us set[8]

$$s \doteq \sqrt{\kappa^{-1}\ln(\varepsilon^{-1})} \ \text{ and } \ N \doteq \lceil \ln(\delta)/\ln(0.9) \rceil, \tag{1.9}$$

where $\delta > 0$ is a small reliability parameter, say, $\delta = \varepsilon$. Now, if $\bar{x}$ satisfies the constraint

$$\mathbf{P}\big(\{\xi : G_{s\sigma}(\bar{x},\xi) \notin C\}\big) \leq \varepsilon,$$

[8] Notation $\lceil a \rceil$ stands for the smallest integer which is greater than or equal to $a \in \mathbb{R}$.

that is, $\bar{x}$ satisfies the strengthened version of (1.3) obtained by replacing the original level of perturbations σ with $s\sigma$, then the probability to get a sample $\{\eta^j\}_{j=1}^N$ such that $\bar{x}$ does *not* satisfy (1.6) is at most

$$\sum_{j=1}^{N} \mathbb{P}\big(\{G(\bar{x},\eta^j) \notin C\}\big) \le \varepsilon N = O(1)\varepsilon \ln(\delta^{-1}),$$

where the constant $O(1)$ is slightly bigger than $[\ln(0.9^{-1})]^{-1} = 9.5$. For $\delta = \varepsilon$, say, this probability is nearly of order ε.

Let $Q \doteq s^{-1}Q_{\bar{x},\sigma}$, and hence

$$\mathbf{P}(\{\xi \in Q\}) = \mathbf{P}(\{s\xi \in Q_{\bar{x},\sigma}\}) = \mathbb{P}\big(\{G(\bar{x},\eta^j) \in C\}\big).$$

Consequently, if $\mathbf{P}(\{\xi \in Q\}) < 0.9$, then the probability p of getting a sample $\{\eta^j\}_{j=1}^N$ for which $\bar{x}$ satisfies (1.6), is the probability to get N successes in N Bernoulli trials with success probability for a single experiment less than 0.9. That is, $p \le 0.9^N$, and by (1.9) we obtain $p \le \delta$. For small $\delta = \varepsilon$, such an event is highly unlikely. And if $\mathbf{P}(\{\xi \in Q\}) \ge 0.9$, then by using (1.8) and because of (1.9) we have

$$\mathbf{P}(\{\xi \notin Q_{\bar{x},\sigma}\}) = \mathbf{P}(\{\xi \notin sQ\}) \le e^{-\kappa s^2} = \varepsilon.$$

That is, $\bar{x}$ satisfies the chance constraint (1.3).

We can summarize the above discussion as follows.

(!!) If $\bar{x}$ satisfies the chance constraint (1.3) with a moderately increased level of perturbations (by factor $s = \sqrt{O(\ln(\varepsilon^{-1}))}$), then it is highly unlikely that $\bar{x}$ does not satisfy (1.6) (probability of that event is less than $O(1)\varepsilon \ln(\varepsilon^{-1})$).
If $\bar{x}$ does satisfy (1.6), then it is highly unlikely that $\bar{x}$ is infeasible for (1.3) at the original level of perturbations (probability of that event is then less than $\delta = \varepsilon$). Note that the sample size which ensures this conclusion is just of order $O(1)\ln(\varepsilon^{-1})$.

The approach we follow is closely related the *importance sampling* method, where one samples from a properly chosen artificial distribution rather than from the actual one in order to make the rare event in question ‘more frequent’. The difference with the traditional importance sampling scheme is that the latter is aimed at estimating the expected value of a given functional and uses change of the probability measure in order to reduce the variance of the estimator. In contrast to this, we do not try to estimate the quantity of interest (which in our context is $\mathbb{P}\{\xi \notin Q\}$, where Q is a given convex set) because of evident hopeless of the estimation problem. Indeed, we are interested in multidimensional case and dimension independent constructions and results, while the traditional importance sampling is heavily affected by the ‘curse of dimensionality’. For example, the distributions of two proportional to each other with coefficient 2 normally distributed vectors ξ and η of dimension 200 are 'nearly singular' with respect to each other: one can find two *nonintersecting* sets U, V in $\mathbb{R}^{200}$ such that $\mathbb{P}\{\xi \notin U\} = \mathbb{P}\{\eta \notin V\} < 1.2 \times 10^{-11}$. Given this fact, it seems ridiculous to *estimate* a quantity related to one of these distributions via a sample drawn from the other one. What could be done (and what we intend to do) is to use the sample of realizations of the larger random vector η to make conclusions of the type 'if all elements of a random sample of size $N = 10,000$ of η belong to a given convex set Q, then, up to chance of 'bad sampling' as small as 10^{-6}, the probability for the smaller vector ξ to take value outside Q is at most 4.6×10^{-9}'. Another difference between what we are doing and the usual results on importance sampling is in the fact that in our context the convexity of Q is crucial for the statements (and the proofs), while in the traditional importance sampling it plays no significant role.

Scenario approach is widely used in Stochastic Optimization. We may refer to [332], and references therein, for a discussion of the Monte Carlo sampling approach to solving two-stage stochastic programming problems of the generic form

$$\min_{x \in X} \mathbb{E}\left[F(x,\xi)\right], \tag{1.10}$$

where $F(x,\xi)$ is the optimal value of the corresponding second stage problem. That theory presents moderate upper bounds on the number of scenarios required to solve the problem within a given accuracy and confidence. However, all results of this type known to us postulate from the very beginning that $F(x,\xi)$ is finite valued with probability one, *i.e.*, that the problem has a relatively complete recourse.

As far as problems with chance constraints of the form (1.3) are concerned, seemingly the only possibility to convert such a problem into one with simple (relatively complete) recourse is to penalize violations of constraints. That is, to approximate the problem of minimization of $f(x) \doteq c^T x$ subject to $Ax \geq b$ and chance constraint (1.3), by the problem

$$\min_{x} c^T x + \gamma \mathbb{E}\left[\inf_{y,t}\left\{t \geq 0 : Wy + t \geq G(x,\xi)\right\}\right] \text{ s.t. } Ax \geq b, \tag{1.11}$$

where $\gamma > 0$ is a penalty parameter. The difficulty, however, is that in order to solve (1.10) within a fixed absolute accuracy in terms of the objective, the number of scenarios N should be of order of the maximal, over $x \in X$, variance of $F(x,\xi)$. For problem (1.11), that means N should be of order of γ^2; in turn, the penalty parameter γ should be inverse proportional to the required confidence parameter ε, and we arrive at the same difficulty as in the case of straightforward Monte Carlo simulation: the necessity to work with prohibitively large samples of scenarios when high level of confidence is required.

To the best of our knowledge, the most recent and advanced results on chance versions of randomly perturbed convex programs are those of Calafiore and Campi [70, 71]. These elegant and general results state that whatever are the distributions of random perturbations (perhaps entering nonlinearly into the objective and the constraints) affecting a convex program with n decision variables, $O(1)n\varepsilon^{-1}\ln(1/\delta)$-scenario sample is sufficient to solve the problem within confidence $1-\varepsilon$ with reliability $1-\delta$ (that is, with probability of bad sampling at most δ), see also Chapter 5 of this book. Here again everything is fine except for the fact that the sample size is proportional to ε^{-1}, which makes the approach impractical when high level of confidence is required.

The rest of the chapter is organized as follows. In Section 1.2, we develop our techniques as applied to the *analysis* problem, as in the motivating discussion above. Note that validity of our scheme for the analysis problem does not yield automatically its validity for the synthesis one, where one optimizes a given objective over the feasible set[9] of (1.6). Applications of the methodology in the synthesis context form the subject of Section 1.3. Technical proofs are relegated to Appendix.

We use the following notation: $\mathbb{E}_{\mathbf{P}}\{\cdot\}$ stands for the expectation with respect to a probability distribution $\mathbf{P}$ on $\mathbb{R}^n$ (we skip index $\mathbf{P}$, when the distribution is clear from the context). By default, all probability distributions are Borel ones with finite first moments. For $\lambda \in \mathbb{R}$, a distribution $\mathbf{P}$ on $\mathbb{R}^n$ and $\xi \sim \mathbf{P}$, we denote by $\mathbf{P}^{(\lambda)}$ the distribution of random vector $\lambda\xi$. Finally, in the sequel, 'symmetric' for sets and distributions always means 'symmetric with respect to the origin'. Unless stated otherwise all considered norms on $\mathbb{R}^d$ are Euclidean norms.

1.2 The Analysis Problem

In this section, the assumption that the mappings $A_i(\cdot)$, $i = 1, ..., d$, are affine plays no role and is discarded. Recall that the Analysis version of (1.3) is to check, given $\bar{x}$, σ, $\varepsilon > 0$, and (perhaps, partial) information on the distribution $\mathbf{P}$ of ξ, whether $\mathbf{P}\big(\{\xi : G_\sigma(\bar{x}, \xi) \notin C\}\big) \le \varepsilon$. Consider the set $Q \doteq Q_{\bar{x},\sigma}$, where $Q_{\bar{x},\sigma}$ is defined in (1.7). Recall that Q is closed and convex. The Analysis problem can be formulated as to check whether the relation

$$\mathbf{P}\big(\{\xi : \xi \notin Q\}\big) \le \varepsilon \tag{1.12}$$

holds true. The scenario approach, presented in Section 1.1, results in the following generic test:

[9] Indeed, our motivating discussion implies only that every *fixed* point $\bar{x}$ which does not satisfy (1.3) is highly unlikely to be feasible for (1.6) – the probability of the corresponding 'pathological' sample $\{\eta^j\}$ is as small as δ. This, however, does not exclude the possibility that a point x which depends on the sample, *e.g.*, the point which optimizes a given objective over the feasible set of (1.6) – is not that unlikely to violate (1.3).

(T) *Given confidence parameter $\varepsilon \in (0,1)$, reliability parameter $\delta \in (0,1)$, and information on (zero mean) distribution $\mathbf{P}$ on $\mathbb{R}^d$, we act as follows:*
(i) We specify a trial distribution $\mathbf{F}$ on $\mathbb{R}^d$ along with integers $N > 0$ (sample size) and $K \geq 0$ (acceptance level), where $K < N$.
(ii) We generate a sample $\{\eta^j\}_{j=1}^N$, drawn from trial distribution $\mathbf{F}$, and check whether at most K of the N sample elements violate the condition[10]

$$\eta^j \in Q.$$

If it is the case, we claim that (1.12) *is satisfied ('acceptance conclusion'), otherwise we make no conclusion at all.*

We are about to analyze this test, with emphasis on the following major questions:

A. How to specify the 'parameters' of the test, that is, trial distribution $\mathbf{F}$, sample size N and acceptance level K, in a way which ensures the validity of the acceptance with reliability at least $1-\delta$, so that the probability of false acceptance (*i.e.*, generating a sample $\{\eta^j\}$ which results in the acceptance conclusion in the case when (1.12) is false) is less than δ.

B. What is the 'resolution' of the test (for specified parameters)? Here 'resolution' is defined as a factor $r = r(\varepsilon, \delta) \geq 1$ such that whenever $\mathbf{P}(\{\xi \in Q_{\bar{x},r\sigma}\}) \leq \varepsilon$ (that is, $\bar{x}$ satisfies (1.3) with the level of perturbations increased by factor r), the probability of *not* getting the acceptance conclusion is at most δ.

1.2.1 Majorization

In Section 1.1 we focused on scenario approach in the case when the scenario perturbations are multiples of the 'true' ones. In fact we can avoid this restriction; all we need is the assumption that the trial distribution *majorizes* the actual distribution of perturbations in the following sense.

Definition 1. Let $\mathbf{F}$, $\mathbf{P}$ be probability distributions on $\mathbb{R}^d$. It is said that $\mathbf{F}$ majorizes[11] $\mathbf{P}$ (written $\mathbf{F} \succeq \mathbf{P}$) if for every convex lower semicontinuous function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ one has $\mathbb{E}_{\mathbf{F}}[f] \geq \mathbb{E}_{\mathbf{P}}[f]$, provided that these expectations are well defined.

It is well known that the above majorization is a partial order (see, *e.g.*, [237]). Some other basic properties of majorization are summarized in the following proposition.

[10] Recall that $\eta^j \in Q$ is equivalent to $G(\bar{x}, \eta^j) \in C$.

[11] In the literature on stochastic orderings the relation '$\succeq$' is called the *convex order*, [237].

Proposition 1. *The following statements hold.*
(i) $\mathbf{F} \succeq \mathbf{P}$ *if and only if* $\mathbb{E}_{\mathbf{F}}[f] \geq \mathbb{E}_{\mathbf{P}}[f]$ *for every convex function f with linear growth (that is, a real valued convex function f such that $|f(x)| \leq O(\|x\|)$ as $\|x\| \to \infty$).*
(ii) *The distribution of the sum $\xi + \eta$ of two independent random vectors $\xi, \eta \in \mathbb{R}^d$ majorizes the distribution of ξ, provided that $\mathbb{E}[\eta] = 0$.*
(iii) *If $\mathbf{F} \succeq \mathbf{P}$ and $\mathbf{F}' \succeq \mathbf{P}'$, then $\lambda\mathbf{F} + (1-\lambda)\mathbf{F}' \succeq \lambda\mathbf{P} + (1-\lambda)\mathbf{P}'$ whenever $\lambda \in [0,1]$.*
(iv) *Let $\mathbf{F} \succeq \mathbf{P}$ be distributions on $\mathbb{R}^p \times \mathbb{R}^q$, and $\tilde{\mathbf{F}}, \tilde{\mathbf{P}}$ be the associated marginal distributions on $\mathbb{R}^p$. Then $\tilde{\mathbf{F}} \succeq \tilde{\mathbf{P}}$.*
(v) *If $\mathbf{F}, \mathbf{P}$ are distributions on $\mathbb{R}^p$ and $\mathbf{F}', \mathbf{P}'$ are distributions on $\mathbb{R}^q$, then the distribution $\mathbf{F} \times \mathbf{F}'$ on $\mathbb{R}^{p+q}$ majorizes the distribution $\mathbf{P} \times \mathbf{P}'$ if and only if both $\mathbf{F} \succeq \mathbf{P}$ and $\mathbf{F}' \succeq \mathbf{P}'$.*
(vi) *Let ξ, η be random vectors in $\mathbb{R}^d$ and A be an $m \times d$ matrix. If the distribution of ξ majorizes the one of η, then the distribution of $A\xi$ majorizes the one of $A\eta$.*
(vii) *For symmetric distributions $\mathbf{F}, \mathbf{P}$, it holds that $\mathbf{F} \succeq \mathbf{P}$ if and only if $\mathbb{E}_{\mathbf{F}}\{f\} \geq \mathbb{E}_{\mathbf{P}}\{f\}$ for all even convex functions f with linear growth such that $f(0) = 0$.*
(viii) *For $\alpha \geq 1$ and symmetrically distributed random vector ξ, the distribution of $\alpha\xi$ majorizes the distribution of ξ.*

Proof. (i) This is evident, since every lower semicontinous convex function on $\mathbb{R}^d$ is pointwise limit of a nondecreasing sequence of finite convex functions with linear growth.

(ii) For a real valued convex f we have

$$\mathbb{E}_{\xi+\eta}[f(\eta+\xi)] = \mathbb{E}_{\xi}\{\mathbb{E}_{\eta}[f(\eta+\xi)]\} \geq \mathbb{E}_{\xi}\{f(\mathbb{E}_{\eta}[\eta+\xi])\} = \mathbb{E}_{\xi}[f(\xi)],$$

where the inequality follows by the Jensen inequality.

(iii), (iv), (vi) and (vii) are evident, and (viii) is readily given by (vii).

(v) Assuming $\mathbf{F} \succeq \mathbf{P}$ and $\mathbf{F}' \succeq \mathbf{P}'$, for a convex function $f(u,u')$ with linear growth ($u \in \mathbb{R}^p$, $u' \in \mathbb{R}^q$) we have

$$\int f(u,u')\mathbf{F}(du)\mathbf{F}'(du') = \int \left\{ \int f(u,u')\mathbf{F}'(du') \right\} \mathbf{F}(du) \geq$$
$$\int \left\{ \int f(u,u')\mathbb{P}'(du') \right\} \mathbf{F}(du) \geq \int \left\{ \int f(u,u')\mathbf{P}'(du') \right\} \mathbf{P}(du) =$$
$$\int f(u,u')\mathbf{P}(du)\mathbf{P}'(du')$$

(we have used the fact that $\int f(u,u')\mathbf{P}'(du')$ is a convex function of u with linear growth). We see that $\mathbf{F} \times \mathbf{F}' \succeq \mathbf{P} \times \mathbf{P}'$. The inverse implication $\mathbf{F} \times \mathbf{F}' \succeq \mathbf{P} \times \mathbf{P}' \Rightarrow \{\mathbf{F} \succeq \mathbf{P} \;\&\; \mathbf{F}' \succeq \mathbf{P}'\}$ is readily given by (iv). □

Let us also make the following simple observation:

Proposition 2. *Let* $\mathbf{F}$ *and* $\mathbf{P}$ *be symmetric distributions on* $\mathbb{R}$ *such that* $\mathbf{P}$ *is supported on* $[-a,a]$ *and the first absolute moment of* $\mathbf{F}$ *is* $\geq a$. *Then* $\mathbf{F} \succeq \mathbf{P}$. *In particular, we have:*

(i) *The distribution of the random variable taking values* ± 1 *with probabilities* $1/2$ *majorizes every symmetric distribution supported in* $[-1,1]$*;*

(ii) *The normal distribution* $\mathcal{N}(0,\pi/2)$ *majorizes every symmetric distribution supported in* $[-1,1]$.

Proof. Given symmetric probability distribution $\mathbf{P}$ supported on $[-a,a]$ and a symmetric distribution $\mathbf{F}$ with the first absolute moment $\geq a$, we should prove that for every convex function f with linear growth on $\mathbb{R}$ it holds that $\mathbb{E}_{\mathbf{P}}[f] \leq \mathbb{E}_{\mathbf{F}}[f]$. Replacing $f(x)$ with $(f(x)+f(-x))/2+c$, which does not affect the quantities to be compared, we reduce the situation to the one where f is even convex function with $f(0)=0$. The left hand side of the inequality to be proven is linear in $\mathbf{P}$, thus, it suffices to prove the inequality for a weakly dense subset of the set of extreme points in the space of symmetric probability distributions on $[-a,a]$, *e.g.*, for distributions assigning masses $1/2$ to points $\pm\alpha$ with $\alpha \in (0,a]$. Thus, we should prove that if f is nondecreasing finite convex function on the ray $\mathbb{R}_+ \doteq \{x : x \geq 0\}$ such that $f(0)=0$ and $\alpha \in (0,a]$, then $f(\alpha) \leq 2\int_0^\infty f(x)\mathbf{F}(dx)$. When proving this fact, we can assume without loss of generality that $\mathbf{F}$ possesses continuous density $p(x)$. Since f is convex, nondecreasing and non-negative on $\mathbb{R}_+$ and $f(0)=0$, for $x \geq 0$ we have $f(x) \geq g(x) \doteq \max[0, f(\alpha)+f'(\alpha)(x-\alpha)]$, and $g(0)=0$, so that $g(x)=c(x-\beta)_+$ for certain $c \geq 0$ and $\beta \in [0,\alpha]$. Replacing f with g, we do not affect the left hand side of the inequality to be proven and can only decrease the right hand side of it. Thus, it suffices to consider the case when $f(x)=(x-\beta)_+$ for certain $\beta \in [0,\alpha]$. The difference

$$h(\beta) = f(\alpha) - 2\int\limits_0^\infty f(x)\mathbf{F}(dx) = \alpha - \beta - 2\int\limits_\beta^\infty (x-\beta)p(x)dx,$$

which we should prove is nonpositive for $\beta \in [0,\alpha]$, is nonincreasing in β. Indeed, $h'(\beta) = -1 + 2\int_\beta^\infty p(x)dx \leq 0$. Consequently,

$$h(\beta) \leq h(0) = \alpha - 2\int\limits_0^\infty xp(x)dx = \alpha - \int |x|\mathbf{F}(dx) \leq 0$$

due to $\alpha \leq a \leq \int |x|\mathbf{F}(dx)$. □

Corollary 1. *Let* $\mathbf{P}$ *be a probability distribution on d-dimensional unit*[12] *cube* $\{z \in \mathbb{R}^d : \|z\|_\infty \leq 1\}$ *which is 'sign-symmetric', that is, if* $\xi \sim \mathbf{P}$ *and* E *is a diagonal matrix with diagonal entries* ± 1, *then* $E\xi \sim \mathbf{P}$. *Let, further,* $\mathbf{U}$ *be the uniform distributions on the vertices of the unit cube, and*[13] *let* $\mathbf{F} \sim \mathcal{N}(0, \frac{\pi}{2} I_d)$. *Then* $\mathbf{P} \preceq \mathbf{U} \preceq \mathbf{F}$.

[12] The norm $\|z\|_\infty$ is the max-norm, *i.e.*, $\|z\|_\infty \doteq \max\{|z_1|,...,|z_d|\}$.

[13] By $\mathcal{N}(\mu,\Sigma)$ we denote normal distribution with mean μ and covariance matrix Σ, and by I_d we denote the $d \times d$ unit matrix.

Proof. Without loss of generality we can assume that $\mathbf{P}$ has density. The restriction of $\mathbf{P}$ on the non-negative orthant is a weak limit of convex combinations of masses $\mathbf{P}(\mathbb{R}^d_+) = 2^{-d}$ sitting at points from the intersection of the unit cube and $\mathbb{R}^d_+$, Consequently, $\mathbf{P}$ itself is a weak limit of uniform distributions on the vertices of boxes of the form $\{x : |x_i| \le a_i \le 1,\ i = 1, ..., d\}$, that is, limit of direct products $\mathbf{U}_a$ of uniform distributions sitting at the points $\pm a_i$. By Proposition 1(iii), in order to prove that $\mathbf{P} \preceq \mathbf{U}$ it suffices to verify that $\mathbf{U} \succeq \mathbf{U}_a$ for all a with $0 \le a_i \le 1$. By Proposition 1(v), to prove the latter fact it suffices to verify that the uniform distribution on $\{-1; 1\}$ majorizes uniform distribution on $\{-a; a\}$ for every $a \in [0, 1]$, which indeed is the case by Proposition 2. To prove that $\mathbf{F} \succeq \mathbf{U}$, by Proposition 1(v) it suffices to verify that the $\mathcal{N}(0, \frac{\pi}{2})$-distribution on the axis majorizes the uniform distribution on $\{-1; 1\}$, which again is stated by Proposition 2. □

Another observation of the same type as in Proposition 2 is as follows.

Proposition 3. *The uniform distribution on $[-a, a]$ majorizes every symmetric unimodal distribution $\mathbf{P}$ on the segment (that is, distribution with density which is nonincreasing function of $|x|$ and vanishes for $|x| > a$) and is majorized by normal distribution $\mathcal{N}(0, \sigma^2)$ with $\sigma = \frac{\sqrt{2\pi}}{4} \approx 0.6267$.*

Proof. The first statement is evident. To prove the second statement is the same as to prove that the uniform distribution on $[-a, a]$ with $a = 4/\sqrt{2\pi}$ is majorized by the standard normal distribution $\mathcal{N}(0, 1)$. To this end, same as in the proof of Proposition 2, it suffices to verify that

$$\int_0^a a^{-1} f(x) dx \le \frac{2}{\sqrt{2\pi}} \int_0^\infty f(x) \exp\{-x^2/2\} dx$$

for every real valued nondecreasing convex function $f(x)$ on $[0, \infty]$ such that $f(0) = 0$. Functions of this type clearly can be approximated by linear combinations, with non-negative coefficients, of functions of the form $(x - \beta)_+$, with $\beta \ge 0$. Thus, it suffices to prove the inequality in question for $f(x) = (x - \beta)_+$, which is straightforward. □

1.2.2 Concentration

Let us consider the following 'concentration' property.

Definition 2. Let $\bar\theta \in [\frac{1}{2}, 1)$ and $\psi(\theta, \gamma)$ be a function of $\theta \in (\bar\theta, 1]$ and $\gamma \ge 1$ which is convex, nondecreasing and nonconstant as a function of $\gamma \in [1, \infty)$. We say that a probability distribution $\mathbf{F}$ on $\mathbb{R}^d$ possesses $(\bar\theta, \psi)$-concentration property (notation: $\mathbf{F} \in \mathcal{C}(\bar\theta, \psi)$), if for every closed convex set $Q \subset \mathbb{R}^d$ one has

$$\mathbf{F}(Q) \ge \theta > \bar\theta \text{ and } \gamma \ge 1 \quad \Rightarrow \quad \mathbf{F}(\{x \notin \gamma Q\}) \le \exp\{-\psi(\theta, \gamma)\}.$$

If the above implication is valid under additional assumption that Q is symmetric, we say that $\mathbf{F}$ possesses symmetric $(\bar\theta,\psi)$-concentration property (notation: $\mathbf{F}\in\mathcal{SC}(\bar\theta,\psi)$).

Distributions with such concentration properties admit a certain calculus summarized in the following proposition.

Proposition 4. *The following statements hold.*
(i) *A symmetric distribution which possesses a symmetric concentration property possesses concentration property as well: if* $\mathbf{F}\in\mathcal{SC}(\bar\theta,\psi)$ *is symmetric, then* $\mathbf{F}\in\mathcal{C}(\widehat\theta,\widehat\psi)$ *with* $\widehat\theta\doteq(1+\bar\theta)/2$ *and* $\widehat\psi(\theta,\gamma)\doteq\psi(2\theta-1,\gamma)$.
(ii) *Let* $\xi\sim\mathbf{F}$ *be a random vector in* $\mathbb{R}^d$, A *be an* $m\times d$ *matrix and* $\mathbf{F}^{(A)}$ *be the distribution of* $A\xi$. *Then* $\mathbf{F}\in\mathcal{C}(\bar\theta,\psi)$ *implies that* $\mathbf{F}^{(A)}\in\mathcal{C}(\bar\theta,\psi)$.
(iii) *Let* $\mathbf{F}\in\mathcal{C}(\bar\theta,\psi)$ *be a distribution on* $\mathbb{R}^p\times\mathbb{R}^q$, *and* $\tilde{\mathbf{F}}$ *be the associated marginal distribution on* $\mathbb{R}^p$. *Then* $\tilde{\mathbf{F}}\in\mathcal{C}(\bar\theta,\psi)$.
(iv) *Let* ξ^i, $i=1,...,p$, *be independent random vectors in* $\mathbb{R}^d$ *with symmetric distributions* $\mathbf{F}_1,...,\mathbf{F}_p$, *such that* $\mathbf{F}_i\in\mathcal{C}(\bar\theta,\psi)$, $i=1,...,p$. *Then the distribution* $\mathbf{F}$ *of* $\eta=\xi^1+...+\xi^p$ *belongs to* $\mathcal{C}(\widehat\theta,\widehat\psi)$ *with* $\widehat\theta\doteq2\bar\theta-1$ *and* $\widehat\psi(\theta,\cdot)$ *given by the convex hull*[14] *of the function*

$$\varphi(\gamma)\doteq\begin{cases}\ln\left(\frac{1}{1-\theta}\right), & 1\le\gamma<p,\\ \max\left\{\ln\left(\frac{1}{1-\theta}\right),\psi(2\theta-1,\gamma/p)-\ln p\right\}, & \gamma\ge p,\end{cases}$$

where $\gamma\in[1,\infty)$ *and* $\theta>\widehat\theta$.
(v) *Let* $\mathbf{F}_i\in\mathcal{C}(\bar\theta,\psi)$ *be distributions on* $\mathbb{R}^{m_i}$, $i=1,...,p$, *and assume that all* $\mathbf{F}_i$ *are symmetric. Then* $\mathbf{F}_1\times...\times\mathbf{F}_p\in\mathcal{C}(\widehat\theta,\widehat\psi)$ *with* $\widehat\theta$ *and* $\widehat\psi$ *exactly as in* (iv).

Moreover, statements (ii) – (v) *remain valid if the class* $\mathcal{C}(\bar\theta,\psi)$ *in the premises and in the conclusions is replaced with* $\mathcal{SC}(\bar\theta,\psi)$.

Proof. (i) Let $\mathbf{F}$ satisfy the premise of (i), and let Q be a closed convex set such that $\mathbf{F}(Q)\ge\theta>\widehat\theta$. By symmetry of $\mathbf{F}$, we have $\mathbf{F}(Q\cap(-Q))\ge2\theta-1>\bar\theta$, and hence

$$\mathbf{F}(\{\xi\not\in\gamma Q\})\le\mathbf{F}(\{\xi\not\in\gamma(Q\cap(-Q))\})\le\exp\{-\psi(2\theta-1,\gamma)\}.$$

The statements (ii) and (iii) are evident.

(iv) Let Q be a closed convex set such that $\theta\doteq\mathbf{F}(Q)>\widehat\theta$. We claim that then

$$\mathbf{F}_i(Q)\ge2\theta-1>\bar\theta,\ \ i=1,...,p. \tag{1.13}$$

Indeed, let us fix i, and let ζ be the sum of all ξ^j except for ξ^i, so that $\eta=\zeta+\xi^i$, ζ and ξ^i are independent and ζ is symmetrically distributed. Observe that conditional, given the value u of ξ^i, probability for ζ to be

[14] The convex hull of a function φ is the largest convex function majorized by φ.

outside Q is at least 1/2, provided that $u \notin Q$. Indeed, when $u \notin Q$, there exists a closed half-space Π_u containing u which does not intersect Q (recall that Q is closed and convex); since ζ is symmetrically distributed, $u+\zeta \in \Pi_u$ with probability at least 1/2, as claimed. From our observation it follows that if $\xi^i \notin Q$ with probability s, then $\eta \notin Q$ with probability at least $s/2$; the latter probability is at most $1-\theta$, whence $s \le 2-2\theta$, and (1.13) follows.

Assuming $\gamma \ge p$ and $\mathbb{P}\{\xi^1 + ... + \xi^p \in Q\} \ge \theta > \widehat{\theta}$, we have

$$\mathbb{P}\{\xi^1 + ... + \xi^p \notin \gamma Q\} \le \sum_{i=1}^{p} \mathbb{P}\{\xi^i \notin (\gamma/p)Q\} \le p\exp\{-\psi(2\theta-1,\gamma/p)\},$$

where the concluding inequality is given by (1.13) and the inclusions $\mathbf{F}_i \in \mathcal{C}(\bar{\theta},\psi)$. Now, the distribution of η is symmetric, so that $\mathbf{F}(\{\eta \in Q\}) > \widehat{\theta} \ge 1/2$ implies that Q intersect $-Q$, that is, that $0 \in Q$. Due to the latter inclusion, for $\gamma \ge 1$ one has $\mathbf{F}(\{\eta \in \gamma Q\}) \ge \mathbf{F}(\{\eta \in Q\}) \ge \theta$. Thus,

$$\mathbf{F}(\{\eta \notin \gamma Q\}) \le \begin{cases} 1-\theta, & 1 \le \gamma \le p, \\ p\exp\{-\psi(2\theta-1,\gamma/p\}, & \gamma \ge p, \end{cases}$$

and (iv) follows.

(v) Let $\xi^i \sim \mathbf{F}_i$ be independent, $i=1,...,p$, and let $\bar{\mathbf{F}}_i$ be the distribution of the $(m_1+...+m_p)$-dimensional random vector

$$\zeta^i = (0_{m_1+...+m_{i-1}}, \xi^i, 0_{m_{i+1}+...+m_p}).$$

Clearly, $\bar{\mathbf{F}}_i \in \mathcal{C}(\bar{\theta},\psi)$ due to similar inclusion for $\mathbf{F}_i$. It remains to note that $\sum\limits_i \zeta^i \sim \mathbf{F}_1 \times ... \times \mathbf{F}_p$ and to use (iv). □

We intend now to present a number of concrete distributions possessing the concentration property.

Example 2 (Normal distribution). Consider the cumulative distribution function $\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} \exp\{-z^2/2\}dz$ of the standard normal distribution and let[15] $\phi(\theta) \doteq \Phi^{-1}(\theta)$ for $\theta \in (0,1)$.

Theorem 1. *Let B be a closed convex set in $\mathbb{R}^d$. Then the following holds.* (i) *If $\eta \sim N(0,I_d)$ and $\mathbb{P}\{\eta \in B\} \ge \theta > \frac{1}{2}$, then for $\alpha \in (0,1)$:*

$$\mathbb{P}\{\alpha\eta \in B\} \ge 1 - \exp\left\{-\frac{\phi^2(\theta)}{2\alpha^2}\right\}.$$

(ii) *If $\zeta \sim N(0,\Sigma)$ and $\mathbb{P}\{\zeta \notin B\} \equiv 1-\theta < \frac{1}{2}$, then for $\gamma \ge 1$:*

$$\mathbb{P}\{\zeta \notin \gamma B\} \le \min\left\{1-\theta, \exp\left(-\tfrac{1}{2}\phi^2(\theta)\gamma^2\right)\right\}.$$

In other words, a zero mean normal distribution on $\mathbb{R}^d$ belongs to $\mathcal{C}(\frac{1}{2},\psi)$ with

$$\psi(\theta,\gamma) \doteq \max\left\{\ln[(1-\theta)^{-1}], \tfrac{1}{2}\phi^2(\theta)\gamma^2\right\}.$$

[15] The inverse function $\phi(\theta) \doteq \Phi^{-1}(\theta)$ is defined by the equation $\Phi(\phi(\theta)) = \theta$.

Proof. Our proof of this result is based on the following result due to Borell [52]:

(!!!) For $\eta \sim \mathcal{N}(0, I_d)$, every $\gamma > 0$, $\varepsilon \geq 0$ and every closed set $X \subset \mathbb{R}^d$ such that $\mathbb{P}\{\eta \in X\} \geq \gamma$, one has

$$\mathbb{P}\left\{\operatorname{dist}(\eta, X) > \varepsilon\right\} \leq 1 - \Phi(\phi(\gamma) + \varepsilon), \tag{1.14}$$

where $\operatorname{dist}(a, X) \doteq \inf_{x \in X} \|a - x\|$.

Now let η, ζ be independent $\mathcal{N}(0, I_d)$ random vectors, and let

$$p(\alpha) = \mathbb{P}\{\alpha\eta \notin B\}.$$

We have that $\alpha\eta + \sqrt{1-\alpha^2}\,\zeta \sim N(0, I_d)$, and hence

$$\mathbb{P}\{\operatorname{dist}(\alpha\eta + \sqrt{1-\alpha^2}\zeta, B) > t\} \leq 1 - \Phi(\phi(\theta) + t)$$

by (1.14). On the other hand, let $\alpha\eta \notin B$, and let $e = e(\eta)$ be a vector such that $\|e\| = 1$ and $e^T[\alpha\eta] > \max\limits_{x \in B} e^T x$. If ζ is such that $\sqrt{1-\alpha^2}e^T\zeta > t$, then $\operatorname{dist}(\alpha\eta + \sqrt{1-\alpha^2}\zeta, B) > t$, and hence if $\alpha\eta \notin B$, then

$$\mathbb{P}\left\{\zeta : \operatorname{dist}(\alpha\eta + \sqrt{1-\alpha^2}\zeta, B) > t\right\} \geq 1 - \Phi(t/\sqrt{1-\alpha^2}).$$

Whence for all $t \geq 0$ such that $\delta(t) \doteq \phi(\theta) + t - t/\sqrt{1-\alpha^2} \geq 0$ one has

$$\begin{aligned} p(\alpha)[1 - \Phi(t/\sqrt{1-\alpha^2})] &\leq \mathbb{P}\left\{\operatorname{dist}(\alpha\eta + \sqrt{1-\alpha^2}\zeta, B) > t\right\} \\ &\leq 1 - \Phi(\phi(\theta) + t). \end{aligned}$$

It follows that

$$\begin{aligned} p(\alpha) \leq \tfrac{1-\Phi(\phi(\theta)+t)}{1-\Phi(t/\sqrt{1-\alpha^2})} &= \frac{\int\limits_{t/\sqrt{1-\alpha^2}}^{\infty} \exp\{-(s+\delta(t))^2/2\}ds}{\int\limits_{t/\sqrt{1-\alpha^2}}^{\infty} \exp\{-s^2/2\}ds} \\ = \frac{\int\limits_{t/\sqrt{1-\alpha^2}}^{\infty} \exp\{-s^2/2 - s\delta(t) - \delta^2(t)/2\}ds}{\int\limits_{t/\sqrt{1-\alpha^2}}^{\infty} \exp\{-s^2/2\}ds} &\leq \exp\{-t\delta(t)/\sqrt{1-\alpha^2} - \delta^2(t)/2\}. \end{aligned}$$

Setting in the resulting inequality $t = \frac{\phi(\theta)(1-\alpha^2)}{\alpha^2}$, we get

$$p(\alpha) \leq \exp\left\{-\frac{\phi^2(\theta)}{2\alpha^2}\right\}. \qquad \square$$

Example 3 (Uniform distribution on the vertices of a cube). We start with the following known fact (which is the Talagrand Inequality in its extended form given in [172]).

Theorem 2. *Let $(E_t, \|\cdot\|_{E_t})$ be finite-dimensional normed spaces, $t = 1, ..., d$, F be the direct product of $E_1, ..., E_d$ equipped with the norm $\|(x^1, ..., x^d)\|_F \doteq \sqrt{\sum\limits_{t=1}^{d} \|x^t\|^2_{E_t}}$, $\mathbf{F}_t$ be Borel probability distributions on the unit balls of E_t and $\mathbf{F}$ be the product of these distributions. Given a closed convex set $A \subset F$, let $\operatorname{dist}(x, A) = \min_{y\in A} \|x - y\|_F$. Then*

$$\mathbb{E}_{\mathbf{F}}\left[\exp\left\{\tfrac{1}{16}\operatorname{dist}^2(x, A)\right\}\right] \le \tfrac{1}{\mathbf{F}(A)}.$$

This result immediately implies the following.

Theorem 3. *Let $\mathbf{P}$ be the uniform distribution on the vertices of the unit cube $\left\{x \in \mathbb{R}^d : \|x\|_\infty \le 1\right\}$. Then $\mathbf{P} \in \mathcal{SC}(\bar\theta, \psi)$ with the parameters given by*

$$\begin{aligned}
&\bar\theta = \tfrac{1+\exp\{-\pi^2/8\}}{2} \approx 0.6456,\\
&\rho(\theta) = \sup_{\omega\in(0,\pi/2]} \left\{\omega^{-1} \arccos\left(\tfrac{1+\exp\{-\omega^2/2\}-\theta}{\theta}\right) : 1 + \exp\{-\omega^2/2\} < 2\theta\right\},\\
&\psi(\theta, \gamma) = \max\left\{\ln\tfrac{1}{1-\theta}, \ln\tfrac{\theta}{1-\theta^2} + \tfrac{\rho^2(\theta)(\gamma-1)^2}{16}\right\}.
\end{aligned} \tag{1.15}$$

In order to prove this result we need the following lemma.

Lemma 1. *Let ξ_j be independent random variables taking values ±1 with probabilities $1/2$ and let $\zeta \doteq \sum\limits_{j=1}^{d} a_j\xi_j$ with $\|a\| = 1$. Then for every $\rho \in [0, 1]$ and every $\omega \in [0, \pi/2]$ one has*

$$\mathbb{P}\left\{|\zeta| \le \rho\right\}\cos(\rho\omega) - \mathbb{P}\left\{|\zeta| > \rho\right\} \le \cos^d(\tfrac{\omega}{\sqrt{d}}) \le \exp\{-\omega^2/2\}.$$

In particular, if

$$\theta \doteq \mathbb{P}\left\{|\zeta| \le \rho\right\} > \bar\theta \doteq \frac{1 + \exp\{-\pi^2/8\}}{2},$$

then $\rho \ge \rho(\theta)$, where $\rho(\theta)$ is defined in (1.15).

Proof. For $\omega \in [0, \pi/2]$ we have

$$\mathbb{E}\{\exp\{i\zeta\omega\}\} = \prod_j \mathbb{E}\{\exp\{ia_j\xi_j\omega\}\} = \prod_j \cos(a_j\omega).$$

Observe that the function $f(s) = \ln\cos(\sqrt{s})$ is concave on $[0, (\pi/2)^2]$. Indeed, $f'(s) = -\tan(\sqrt{s})\frac{1}{2\sqrt{s}}$ and

$$f'(s) = -\tfrac{1}{\cos^2(\sqrt{s})}\tfrac{1}{4s} + \tan(\sqrt{s})\tfrac{1}{4s\sqrt{s}} = -\tfrac{1}{4s^2\cos^2(s)}\left[\sqrt{s} - \sin(\sqrt{s})\cos(\sqrt{s})\right] \le 0.$$

Consequently, for $0 \le \omega \le \pi/2$ we have

$$\sum_j \ln(\cos(a_j\omega)) = \sum_j f(a_j^2\omega^2) \leq \max_{\substack{0\leq s_j\leq(\pi/2)^2\\ \sum_j s_j=\omega^2}} \sum_j f(s_j) = df(\omega^2/d) \leq \exp\{-\omega^2/2\},$$

and we see that

$$0 \leq \omega \leq \tfrac{\pi}{2} \Rightarrow \mathbb{E}\{\exp\{\imath\zeta\omega\}\} \leq \cos^d(\tfrac{\omega}{\sqrt{d}}) \leq \exp\{-\omega^2/2\}.$$

On the other hand, ζ is symmetrically distributed, and therefore for $0 \leq \rho \leq 1$ and $\omega \in [0,\pi/2]$ we have, setting $\mu \doteq \mathbb{P}\{|\zeta| \leq \rho\}$:

$$\mathbb{E}\left\{\exp\{i\omega\zeta\}\right\} \geq \mu\cos(\rho\omega) - (1-\mu),$$

and we arrive at the announced result. □

Proof of Theorem 3. Let Q be a symmetric closed convex set in $\mathbb{R}^d$ such that

$$\mathbb{P}\{\xi \in Q\} \geq \theta > \bar{\theta}.$$

We claim that then Q contains the centered at the origin Euclidean ball of the radius $\rho(\theta)$. Indeed, otherwise Q would be contained in the strip $\Pi = \{x : |a^Tx| \leq c\}$ with $c < \rho(\theta)$ and $\|a\| = 1$. Setting $\zeta = a^T\xi$, we get

$$\mathbb{P}\{|\zeta| \leq c\} = \mathbb{P}\{\xi \in \Pi\} \geq \mathbb{P}\{\xi \in Q\} \geq \theta,$$

whence by Lemma 1, $c \geq \rho(\theta)$, which is a contradiction.

For $s \geq 1$ from $x \notin sQ$ it follows that the set $x+(s-1)Q$ does not intersect Q; since this set contains the $\|\cdot\|$-ball centered at x of the radius $(s-1)\rho(\theta)$, the Euclidean distance $d_Q(x) \doteq \mathrm{dist}(x,Q)$, from x to Q, is at least $(s-1)\rho(\theta)$. At the same time, by Talagrand Inequality we have

$$\mathbb{E}\left[\exp\left\{\tfrac{d_Q^2(\xi)}{16}\right\}\right] \leq \tfrac{1}{\mathbb{P}\{\xi\in Q\}} \leq \tfrac{1}{\theta}.$$

On the other hand, when $\gamma \geq 1$ we have, by the above arguments,

$$\mathbb{E}\left[\exp\left\{\tfrac{d_Q^2(\xi)}{16}\right\}\right] \geq \mathbb{P}\left\{\xi \in Q\right\} + \exp\left\{\tfrac{(\gamma-1)^2\rho^2(\theta)}{16}\right\}\mathbb{P}\left\{\xi \notin \gamma Q\right\},$$

whence if $\gamma \geq 1$, then

$$\mathbb{P}\left\{\xi \notin \gamma Q\right\} \leq \tfrac{1-\theta^2}{\theta}\exp\left\{-\tfrac{(\gamma-1)^2\rho^2(\theta)}{16}\right\},$$

and of course

$$\gamma \geq 1 \Rightarrow \mathbb{P}\left\{\xi \notin \gamma Q\right\} \leq 1-\theta,$$

and the result follows. □

Example 4 (Uniform distribution on the cube). This example is similar to the previous one.

Theorem 4. *Let* $\mathbf{P}$ *be the uniform distribution on the unit cube* $\{x \in \mathbb{R}^d : \|x\|_\infty \leq 1\}$. *Then* $\mathbf{P} \in \mathcal{SC}(\bar{\theta}, \psi)$ *with the parameters given by*

$$\begin{aligned}
&\bar{\theta} = \tfrac{1+\exp\{-\pi^2/24\}}{2} \approx 0.8314,\\
&\rho(\theta) = \sup_{\omega \in (0,\pi/2]} \left\{\omega^{-1} \arccos\left(\tfrac{1+\exp\{-\omega^2/6\}-\theta}{\theta}\right) : 1+\exp\{-\omega^2/6\} < 2\theta\right\},\\
&\psi(\theta,\gamma) = \max\left\{\ln\left(\tfrac{1}{1-\theta}\right), \ln\left(\tfrac{\theta}{1-\theta^2}\right) + \tfrac{\rho^2(\theta)(\gamma-1)^2}{16}\right\}.
\end{aligned} \tag{1.16}$$

We have the following analog of Lemma 1.

Lemma 2. *Let* ξ_j *be independent random variables uniformly distributed in* $[-1,1]$ *and* $\zeta = \sum\limits_{j=1}^{d} a_j \xi_j$ *with* $\|a\| = 1$. *Then for every* $\rho \in [0,1]$ *and every* $\omega \in [0,\pi/2]$ *one has*

$$\mathbb{P}\{|\zeta| \leq \rho\} \cos(\rho\omega) - \mathbb{P}\{|\zeta| > \rho\} \leq \left(\frac{\sin(\omega d^{-1/2})}{\omega d^{-1/2}}\right)^d \leq \exp\left\{-\omega^2/6\right\}. \tag{1.17}$$

In particular, if

$$\theta \doteq \mathbb{P}\{|\zeta| \leq \rho\} > \bar{\theta} \doteq \frac{1+\exp\{-\pi^2/24\}}{2},$$

then $\rho \geq \rho(\theta)$, *where* $\rho(\theta)$ *is defined in* (1.16).

Proof. For $\omega \in [0,\pi/2]$ we have

$$\mathbb{E}\{\exp\{i\zeta\omega\}\} = \prod_j \mathbb{E}\{\exp\{ia_j\xi_j\omega\}\} = \prod_j \tfrac{\sin(a_j\omega)}{a_j\omega}.$$

Observe that the function $f(s) = \ln(\sin(\sqrt{s})) - \frac{1}{2}\ln s$ is concave on $[0,(\pi/2)^2]$. Indeed, $f'(s) = \cot(\sqrt{s})\frac{1}{2\sqrt{s}} - \frac{1}{2s}$ and

$$f'(s) = -\tfrac{1}{\sin^2(\sqrt{s})}\tfrac{1}{4s} - \cot(\sqrt{s})\tfrac{1}{4s\sqrt{s}} + \tfrac{1}{2s^2} = \tfrac{h(\sqrt{s})}{4s^2\sin^2(\sqrt{s})},$$

where

$$h(r) = 2\sin^2(r) - r\sin(r)\cos(r) - r^2 = 1 - \cos(2r) - (r/2)\sin(2r) - r^2.$$

We have $h(0) = 0$, $h'(r) = (3/2)\sin(2r) - r\cos(2r) - 2r$, so that $h'(0) = 0$, $h'(r) = 2\cos(2r) + 2r\sin(2r) - 2$, so that $h'(0) = 0$, and finally $h''(r) = -2\sin(2r) + 4r\cos(2r)$, so that $h''(0) = 0$ and $h''(r) \leq 0$, $0 \leq r \leq \pi/2$, due to $\tan(u) \geq u$ for $0 \leq u < \pi/2$. It follows that $h(\cdot) \leq 0$ on $[0,\pi/2]$, as claimed.

From log-concavity of f on $[0,(\pi/2)^2]$, same as in the proof of Lemma 1, we get the first inequality in (1.17); the second is straightforward. □

The remaining steps in the proof of Theorem 4 are completely similar to those for Example 3.

Remark 1. Examples 3 and 4 admit natural extensions. Specifically, let ξ be a random vector in $\mathbb{R}^d$ with independent symmetrically distributed on $[-1,1]$ coordinates ξ_i, and let the distributions $\mathbf{P}_i$ of ξ_i be 'not too concentrated at the origin', *e.g.*, be such that: (i) $\mathbb{E}\{\xi_i^2\} \geq \alpha^2 > 0$, $i = 1,...,d$, or (ii) $\mathbf{P}_i$ possesses density which is bounded by $1/\alpha$, $i = 1,...,d$. Let $\mathbf{P}$ be the distribution of ξ. Then $\xi \in \mathcal{C}(\bar{\theta}, a(\theta) + b(\theta)\gamma^2)$ with $\bar{\theta}$ and $a(\cdot)$, $b(\cdot) > 0$ depending solely on α. The proof is completely similar to those in Examples 2 and 3.

Remark 2. We have proven that the uniform distributions on the vertices of the unit cube $\{\|x\|_\infty \leq 1\}$ and on the entire cube possess symmetric concentration property. In fact they possess as well the general concentration property with slightly 'spoiled' $\bar{\theta}$, $\psi(\cdot,\cdot)$ due to Proposition 4(i).

1.2.3 Main Result

Proposition 5. *Let* $\mathbf{F}$, $\mathbf{P}$ *be probability distributions on* $\mathbb{R}^d$ *such that* $\mathbf{P} \preceq \mathbf{F}$, $\mathbf{F}$ *is symmetric and* $\mathbf{F} \in \mathcal{C}(\bar{\theta}, \psi)$. *Let, further,* Q *be a closed convex set in* $\mathbb{R}^d$ *such that*

$$\mathbf{F}(Q) \geq \theta > \bar{\theta}$$

and let $p_Q(x)$ *be the Minkowski function*[16] *of* Q. *Then for every convex continuous and nondecreasing function* $\Psi : \mathbb{R}_+ \to \mathbb{R}$ *one has*

$$\mathbb{E}_{\mathbf{P}}\left[\Psi(p_Q(\xi))\right] \leq \left(\theta + e^{-\psi(\theta,1)}\right)\Psi(1) + \int\limits_1^\infty \Psi'(\gamma)e^{-\psi(\theta,\gamma)}d\gamma. \tag{1.18}$$

If the assumption $\mathbf{F} \in \mathcal{C}(\bar{\theta}, \psi)$ *is weakened to* $\mathbf{F} \in \mathcal{SC}(\bar{\theta}, \psi)$, *then the conclusion remains valid under the additional assumption that* Q *is symmetric.*

Proof. Let $f(x) \doteq \Psi(p_Q(x))$, so that f is a convex lower semicontinuous function on $\mathbb{R}^d$, and let

$$P(\gamma) \doteq \mathbf{F}(\{x \notin \gamma Q\}) = \mathbf{F}(\{x : p_Q(x) > \gamma\}),$$

so that

$$\gamma \geq 1 \Rightarrow P(\gamma) \leq S(\gamma) \doteq \exp\{-\psi(\theta,\gamma)\}.$$

We have that $\mathbb{E}_{\mathbf{P}}\{f\} \leq \mathbb{E}_{\mathbf{F}}\{f\}$, since $\mathbf{P} \preceq \mathbf{F}$, and

$$\begin{aligned}\mathbb{E}_{\mathbf{F}}\{f\} &\leq \Psi(1)\mathbf{F}(Q) - \int\limits_1^\infty \Psi(\gamma)dP(\gamma) \leq \theta\Psi(1) + \Psi(1)P(1) + \int\limits_1^\infty \Psi'(\gamma)P(\gamma)d\gamma \\ &\leq (\theta + S(1))\Psi(1) + \int\limits_1^\infty \Psi'(\gamma)S(\gamma)d\gamma,\end{aligned}$$

as claimed. □

[16] Minkowski function is defined as $p_Q(x) \doteq \inf\{t : t^{-1}x \in Q,\ t > 0\}$. Under our premise, $0 \in Q$ due to symmetry of $\mathbf{F}$ and $\mathbf{F}(Q) > \bar{\theta} > 1/2$. Consequently, $p_Q(\cdot)$ is a lower semicontinous convex function with values in $\mathbb{R} \cup \{+\infty\}$.

Theorem 5. *Let* $\mathbf{F}$*,* $\mathbf{P}$ *be probability distributions on* $\mathbb{R}^d$ *such that* $\mathbf{P} \preceq \mathbf{F}$*,* $\mathbf{F}$ *is symmetric and* $\mathbf{F} \in \mathcal{C}(\bar{\theta}, \psi)$*. Let, further,* Q *be a closed convex set in* $\mathbb{R}^d$ *such that*

$$\mathbf{F}(Q) \geq \theta > \bar{\theta}$$

and let $p_Q(x)$ *be the Minkowski function of* Q*. Then for every* $s > 1$ *one has*

$$\mathbf{P}(\{x : x \notin sQ\}) \leq \mathrm{Err}(s, \theta) \doteq \inf_{1 \leq \beta < s} \tfrac{1}{s-\beta} \int\limits_{\beta}^{\infty} \exp\{-\psi(\theta, \gamma)\} d\gamma. \tag{1.19}$$

In particular, if $\psi(\theta, \gamma) \geq a(\theta) + b(\theta)\gamma^2/2$ *with* $b(\theta) > 0$*, then*

$$\mathrm{Err}(s, \theta) \leq \frac{4 \exp\{-a(\theta) - b(\theta)(s+1)^2/8\}}{b(\theta)(s^2 - 1)}. \tag{1.20}$$

If the assumption $\mathbf{F} \in \mathcal{C}(\bar{\theta}, \psi)$ *is weakened to* $\mathbf{F} \in \mathcal{SC}(\bar{\theta}, \psi)$*, then the conclusion remains valid under the additional assumption that* Q *is symmetric.*

Proof. Let $\beta \in [1, s)$, and let $\Psi(\gamma) = \frac{(\gamma - \beta)_+}{s - \beta}$. Applying (1.18), we get

$$\begin{aligned} \mathbf{P}(\{x : x \notin sQ\}) &= \mathbf{P}(\{x : x \notin sQ\})\Psi(s) \leq \mathbb{E}_{\mathbf{P}}\{\Psi \circ p_Q\} \\ &\leq \tfrac{1}{s-\beta} \int\limits_{\beta}^{\infty} \exp\{-\psi(\theta, \gamma)\} d\gamma. \end{aligned}$$

Since this relation holds true for every $\beta \in [1, s)$, (1.19) follows.

Now let θ be such that $\psi(\theta, \gamma) \geq a + b\gamma^2/2$ for all $\gamma \geq 1$, where $b > 0$. Then (1.19) implies that

$$\begin{aligned} &\mathbf{P}(\{x : x \notin sQ\}) \leq \left[\tfrac{1}{s-\beta} \int\limits_{\beta}^{\infty} \exp\{-a - b\gamma^2/2\} d\gamma \right] \Bigg|_{\beta = \frac{1+s}{2}} \\ &= \tfrac{2\exp\{-a\}}{s-1} \int\limits_{\frac{1+s}{2}}^{\infty} \exp\{-b\gamma^2/2\} d\gamma \leq \tfrac{2\exp\{-a\}}{s-1} \int\limits_{\frac{1+s}{2}}^{\infty} \tfrac{\gamma}{\frac{1+s}{2}} \exp\{-b\gamma^2/2\} d\gamma \\ &= \tfrac{4\exp\{-a - b(s+1)^2/8\}}{b(s^2-1)}. \end{aligned}$$

□

1.2.4 Putting Blocks Together

Now we are ready to address the questions A and B posed at the beginning of this Section.

Setup for the test

Theorem 5 suggests the following *Basic Setup* for test (T):
Input. Closed convex set $Q \subset \mathbb{R}^d$, zero mean distribution $\mathbf{P}$ on $\mathbb{R}^d$, confidence

parameter $\varepsilon \in (0,1)$, reliability parameter $\delta \in (0,1)$. The goal is to justify the hypothesis

$$\mathbf{P}(\{\xi \notin Q\}) \leq \varepsilon.$$

Choosing 'pre-trial' distribution. We choose a *symmetric* 'pre-trial' distribution $\bar{\mathbf{F}}$ on $\mathbb{R}^d$ in such a way that

I(1) $\bar{\mathbf{F}} \succeq \mathbf{P}$;
I(2) $\bar{\mathbf{F}}$ possesses the concentration property: $\bar{\mathbf{F}} \in \mathcal{C}(\bar{\theta}, \psi)$ with known $\bar{\theta}$ and ψ.

After $\bar{\mathbf{F}}$ is chosen, we compute the associated 'error function' (*cf.* (1.19))

$$\mathrm{Err}(s,\theta) = \inf_{1 \leq \beta < s} \frac{1}{s-\beta} \int\limits_{\beta}^{\infty} \exp\{-\psi(\theta,\gamma)\} d\gamma. \tag{1.21}$$

Choosing trial distribution, sample size and acceptance level. We choose somehow design parameters $\theta \in (\bar{\theta}, 1)$ and $s > 1$ ('amplification') such that

$$\mathrm{Err}(s,\theta) \leq \varepsilon$$

and specify the trial distribution $\mathbf{F}$ as $\bar{\mathbf{F}}^{(s)}$. We further specify sample size N and acceptance level K in such a way that the probability to get at least $N-K$ successes in N Bernoulli experiments with probability θ of success in a single experiment is at most δ:

$$\sum_{r=0}^{K} \binom{N}{r} \theta^{N-r} (1-\theta)^r \leq \delta. \tag{1.22}$$

For example, one can set

$$K \doteq 0, \;\; N \doteq \left\lceil \frac{\ln(\delta)}{\ln(\theta)} \right\rceil. \tag{1.23}$$

Theorem 6. *With the outlined setup, the probability of false acceptance for the resulting test* (T) *is* $\leq \delta$.

Proof. Let $B = s^{-1}Q$. Assume first that $\bar{\mathbf{F}}(B) \geq \theta$. Applying (1.19), we get

$$\mathbf{P}(\{\xi \notin Q\}) = \mathbf{P}(\{\xi \notin sB\}) \leq \mathrm{Err}(s,\theta) \leq \varepsilon,$$

that is, in the case in question false acceptance is impossible. Now consider the case of $\bar{\mathbf{F}}(B) < \theta$, or, which is the same, $\mathbf{F}(Q) < \theta$. In this case, by (1.22), the probability to make acceptance conclusion is at most δ. □

Remark 3. The outlined reasoning demonstrates that when Q is symmetric, Theorem 6 remains valid when the requirement $\bar{\mathbf{F}} \in \mathcal{C}(\bar{\theta}, \psi)$ is weakened to $\bar{\mathbf{F}} \in \mathcal{SC}(\bar{\theta}, \psi)$. The same is true for Theorem 8 below.

Resolution

Let us try to understand how conservative is our test. The answer is easy when the trial distribution coincides with the actual one.

Theorem 7. *Let* $\mathbf{P}$ *be symmetric and possess the concentration property:* $\mathbf{P} \in \mathcal{C}(\bar\theta, \psi)$*, so that the choice* $\bar{\mathbf{F}} = \mathbf{P}$ *satisfies* I(1) *and* I(2) *(from the Basic Setup), and let* N, K, θ, s *be the parameters given by the Basic Setup for this choice of pre-trial distribution. Let* $\theta_* \doteq \mathbf{P}(s^{-1}Q)$*.*

Then the probability for (T) *not to make the acceptance conclusion is at most*

$$\delta_* = 1 - \sum_{r=0}^{K} \theta_*^{N-r}(1-\theta_*)^r.$$

When Q *is symmetric, the conclusion remains valid when the assumption* $\mathbf{P} \in \mathcal{C}(\bar\theta, \psi)$ *is weakened to* $\mathbf{P} \in \mathcal{SC}(\bar\theta, \psi)$*.*

Proof. The statement is essentially a tautology: since $\mathbf{F} = \bar{\mathbf{F}}^{(s)} = \mathbf{P}^{(s)}$, we have $\mathbf{F}(Q) = \mathbf{P}(s^{-1}Q) = \theta_*$, and the probability for (T) not to make the acceptance conclusion is exactly δ_*. □

In terms of Question B, Theorem 7 states that the resolution of (T) is not worse than s, provided that

$$1 - \sum_{r=0}^{K} (1-\varepsilon)^{N-r}\varepsilon^r \le \delta. \tag{1.24}$$

When the setup parameters N, K are chosen according to (1.23), that is, $K = 0$, $N = \left\lceil \frac{\ln(\delta)}{\ln(\theta)} \right\rceil$, condition (1.24) becomes $1 - (1-\varepsilon)^N \le \delta$, which is for sure true when $2\varepsilon \ln(1/\delta) \le \delta \ln(1/\theta)$.

Situation with resolution in the case when the trial distribution is not a scaling $\mathbf{P}^{(s)}$ of the actual one is much more complicated, and its detailed investigation goes beyond the scope of this chapter. Here we restrict ourselves to demonstration of a phenomenon which can occur in the general case. Let $\mathbf{P}$ be the uniform distribution on the vertices of the unit d-dimensional cube Q, and $\mathbf{F}$ be normal distribution $\mathcal{N}(0, \frac{\pi}{2} I_d)$, so that $\mathbf{F} \succeq \mathbf{P}$ by Proposition 2. We have $\mathbf{P}(Q) = 1$, while a typical realization of $\mathbf{F}$ is outside the box $\frac{1}{2}\pi\kappa\sqrt{2\ln d}\, Q$, $\kappa < 1$ with probability tending to 1 as $d \to \infty$, provided that $\kappa < 1$. It follows that in the situation in question the resolution of (T) is dimension-dependent and deteriorates, although pretty slow, as dimension grows.

Homogenization

We next present a slight modification of test (T) – the *homogenized* analysis test (HT) which is better suited for many applications.

Input. Closed convex set $Q \subset \mathbb{R}^d$, zero mean distribution $\mathbf{P}$ on $\mathbb{R}^d$, scale parameter $\bar{\sigma} > 0$, reliability parameter $\delta \in (0,1)$. The goal is to get upper bounds for the probabilities

$$\mathbf{P}(\{s^{-1}\bar{\sigma}\xi \notin Q\}), \text{ for } s > 1.$$

Setup.

Trial distribution. We choose a symmetric distribution $\mathbf{F}$ on $\mathbb{R}^d$ such that $\mathbf{F} \succeq \mathbf{P}$ and $\mathbf{F} \in \mathcal{C}(\bar{\theta}, \psi)$ with known $\bar{\theta}$ and ψ, and compute the corresponding function $\mathrm{Err}(\cdot,\cdot)$ according to (1.21).

Sample size and acceptance level. We choose somehow $\theta \in (\bar{\theta}, 1)$, sample size N and acceptance level K satisfying (1.22).

Execution. We generate N-element sample $\{\eta^j\}_{j=1}^N$ from the trial distribution and check whether

$$\mathrm{card}(\{j \leq N : \bar{\sigma}\eta^j \notin Q\}) \leq K.$$

If it is the case, we say that (HT) is successful, and claim that

$$\mathbf{P}(\{s^{-1}\bar{\sigma}\xi \notin Q\}) \leq \mathrm{Err}(s, \theta), \text{ for all } s > 1, \tag{1.25}$$

otherwise we say that (HT) is unsuccessful.

The analogy of Theorem 6 for (HT) is as follows.

Theorem 8. *With the outlined setup, bounds* (1.25)*, if yielded by* (HT)*, are valid with reliability at least* $1 - \delta$*. Equivalently: in the case when not all of the bounds are valid, the probability for* (HT) *to be successful is at most* δ*.*

Indeed, in the case when $\mathbf{F}(\{\eta : \bar{\sigma}^{-1}\eta \in Q\}) \geq \theta$, bounds (1.25) are valid by (1.19), and in the case when $\mathbf{F}(\{\eta : \bar{\sigma}^{-1}\eta \in Q\}) < \theta$, the probability of successful termination is $\leq \delta$ by (1.22).

The difference between (T) and (HT) is clear. The goal of (T) is to justify the hypothesis that $\xi \sim \mathbf{P}$ takes its value outside a given convex set Q with probability at most ε. The goal of (HT) is to bound from above the probability for ξ to take value outside of set $s\bar{\sigma}^{-1}Q$ as a function of $s > 1$. This second goal is slightly easier than the first one, in the sense that now a *single* sample allows to build bounds for the indicated probabilities simultaneously for all $s > 1$.

1.2.5 Numerical Illustration

Here we illustrate our constructions, by a numerical example.

The situation

We consider a discrete time linear dynamical system

$$z(t+1) = Az(t), \quad A = \frac{1}{203}\begin{bmatrix} 39 & 69 & 41 & -11 & 69 & 84 \\ 56 & -38 & -92 & 82 & 28 & 57 \\ -85 & -40 & -98 & -41 & 72 & -78 \\ 61 & 86 & -83 & -43 & -31 & 38 \\ -5 & -96 & 51 & -96 & 66 & -77 \\ 54 & 2 & 21 & 27 & 34 & 57 \end{bmatrix} \qquad (S)$$

Recall that a necessary and sufficient stability condition ‘all trajectories converge to 0 as $t \to \infty$’ for a system of the form (S) is the existence of a *Lyapunov stability certificate* – a matrix $X \succ 0$ and $\gamma \in [0, 1)$ satisfying the relation

$$\left[\begin{array}{c|c} \gamma^2 X & A^T X \\ \hline XA & X \end{array}\right] \succeq 0. \qquad (1.26)$$

System (S) is stable; as the corresponding certificate, one can take

$$X = \bar{X} = \begin{bmatrix} 1954 & 199 & 170 & 136 & 35 & 191 \\ 199 & 1861 & -30 & -136 & 222 & 137 \\ 170 & -30 & 1656 & 17 & -370 & -35 \\ 136 & -136 & 17 & 1779 & 296 & 112 \\ 35 & 222 & -370 & 296 & 1416 & 25 \\ 191 & 137 & -35 & 112 & 25 & 2179 \end{bmatrix},$$

and $\gamma = \bar{\gamma} = 0.95$. The question we are interested in is: assume that entries in A are subject to random perturbations

$$A_{ij} \mapsto A_{ij}(1 + \sigma\xi_{ij}), \qquad (1.27)$$

where ξ_{ij} are independent random perturbations uniformly distributed on $[-1, 1]$. How large could be the level of perturbations σ in order for $(\bar{X}, \gamma = 0.9999)$ to remain the Lyapunov stability certificate for the perturbed matrix with probability at least $1 - \varepsilon$, with ε like 10^{-8} or 10^{-12}?

For fixed X and γ, (1.26) is a Linear Matrix Inequality in A, so that the question we have posed can be reformulated as the question of how large could be σ under the restriction that

$$\mathbf{P}(\sigma^{-1}Q) \geq 1 - \varepsilon,$$

where $\mathbf{P}$ is the distribution of random 6×6 matrix with independent entries uniformly distributed in $[-1, 1]$ and Q is the closed convex set[17]

$$Q = \left\{ \xi \in \mathbb{R}^{6\times 6} : \left[\begin{array}{c|c} 0 & -[A\cdot\xi]^T\bar{X} \\ \hline -\bar{X}[A\cdot\xi] & 0 \end{array}\right] \preceq \left[\begin{array}{c|c} \gamma^2\bar{X} & A^T\bar{X} \\ \hline \bar{X}A & \bar{X} \end{array}\right] \right\}.$$

In order to answer this question, we use the (HT) test and act as follows.

(a) As the trial distribution $\mathbf{F}$, we use the zero mean normal distribution with covariance matrix $\frac{\pi}{8}I_{36}$ which, by Proposition 3, majorizes the uniform distribution $\mathbf{P}$.

At first glance, the choice of normal distribution in the role of $\mathbf{F}$ seems strange – the actual distribution itself possesses the concentration property, so that it would be natural to choose $\bar{\mathbf{F}} = \mathbf{P}$. Unfortunately, function ψ for the uniform distribution (see Theorem 4 and Remark 2), although of the same type as its normal-distribution counterpart (see Theorem 1), leads to more conservative estimates because of worse constant factors; this explains our choice of the trial distribution.

(b) We run a 'pilot' 1000-element simulation in order to get a rough safe guess $\bar{\sigma}$ of what is the level of perturbations in question. Specifically, we generate a 1000-element sample drawn from $\mathbf{F}$, for every element η of the sample compute the largest σ such that $\eta \in \sigma^{-1}Q$, and then take the minimum, over all elements of the sample, of the resulting quantities, thus obtaining the largest level of perturbations which is compatible with our sample. This level is slightly larger than 0.064, and we set $\bar{\sigma} = 0.064$.

(c) Finally, we run test (HT) itself. First, we specify the sample size N as 1000 and the acceptance level K as 0. Then we compute the largest θ satisfying (1.23) with reliability parameter $\delta = 10^{-6}$, that is, $\theta = \exp\{\frac{\ln(\delta)}{N}\} = 10^{-0.006} \approx 0.9863$. Second, we build 1000-element sample, drawn from $\mathbf{F}$, and check whether all elements η of the sample satisfy the inclusion $\bar{\sigma}\eta \in Q$, which indeed is the case. According to Theorem 8, the latter fact allows to claim, with reliability at least $1 - \delta$ (that is, with chances to make a wrong claim at most $\delta = 10^{-6}$), that for every $s > 1$ one has

$$\mathbf{P}(s^{-1}\bar{\sigma}\xi \notin Q) \leq \mathrm{Err}(s, \theta) = \mathrm{Err}(s, 0.9863)$$

with $\mathrm{Err}(\cdot, \cdot)$ given by (1.21) (where ψ is as in Theorem 1). In other words, up to probability of bad sampling as small as 10^{-6}, we can be sure that for every $s > 1$, at the level of perturbations $s^{-1}\bar{\sigma} = 0.064s^{-1}$ the probability for

[17] By $A \cdot B$ we denote the componentwise product of two matrices, *i.e.*, $[A\cdot B]_{ij} = A_{ij}B_{ij}$. This is called Hadamard product by some authors. The notation '$\preceq$' stands for the standard partial order in the space $\mathbf{S}^m$ of symmetric $m \times m$ matrices: $A \succeq B$ ($A \succ B$) if and only if $A - B$ is positive semidefinite (positive definite). Thus, '$\succeq$' ('$\preceq$') stand for two different relations, namely majorization as defined in Definition 1, and the partial order induced by the semidefinite cone. What indeed '$\succeq'$ means, will be clear from the context.

$(\bar{X}, 0.9999)$ to remain Lyapunov stability certificate for the perturbed matrix is at least $1 - \mathrm{Err}(s, \theta)$. From the data in Table 1.1 we see that moderate reduction in level of perturbations ρ ensures dramatic decrease in the probability ε of 'large deviations,' *cf.* (1.20).

A natural question is how conservative are our bounds? The experiment says that as far as the levels of perturbations are concerned, the bounds are accurate up to moderate constant factor. Indeed, according to our table, perturbation level $\sigma = 0.0128$ corresponds to confidence as high as $1 - \varepsilon$ with $\varepsilon = 5.9 \times 10^{-15}$; simulation demonstrates that ten times larger perturbations result in confidence as low as $1 - \varepsilon$ with $\varepsilon = 1.6 \times 10^{-2}$.

Table 1.1. $p(\sigma)$: probability of a perturbation (1.27) for which $(\bar{X}, 0.9999)$ fails to be a Lyapunov stability certificate

σ	0.0580	0.0456	0.0355	0.0290	0.0246	0.0228	0.0206
$p(\sigma) \le$	0.3560	0.0890	0.0331	0.0039	2.9e-4	6.9e-5	6.3e-6
σ	0.0188	0.0177	0.0168	0.0156	0.0148	0.0136	0.0128
$p(\sigma) \le$	4.6e-7	6.9e-9	9.4e-9	3.8e-10	4.0e-11	3.0e-13	5.9e-15

1.3 The Synthesis Problem

We now address the problem of optimizing under chance constraints

$$\min_{x \in X} c^T x \text{ subject to } \mathbb{P}\{G_\sigma(x, \xi) \in C\} \ge 1 - \varepsilon,$$

with $G_\sigma(x, \xi)$ defined in (1.4) and $\xi \sim \mathbf{P}$. We assume that $C \subset \mathbb{R}^m$ is a closed convex set and X is a compact convex set. As about the distribution $\mathbf{P}$ of perturbations, we assume in the sequel that it is symmetric. In this case, our chance constraint is essentially the same as the symmeterized constraint

$$\mathbb{P}\{G_\sigma(x, \xi) \in C \text{ and } G_\sigma(x, -\xi) \in C\} \ge 1 - \varepsilon.$$

Indeed, the validity of the symmeterized constraint implies the validity of the original one, and the validity of the original constraint, with ε replaced by $\varepsilon/2$, implies the validity of the symmeterized one. In our context of really small ε the difference between confidence $1 - \varepsilon$ and confidence $1 - \varepsilon/2$ plays no role, and by reasons to be explained later we prefer to switch from the original form of the chance constraint to its symmeterized form. Thus, from now on our problem of interest is

$$\min_{x \in X} c^T x \text{ subject to } \mathbb{P}\{G_\sigma(x, \pm\xi) \in C\} \ge 1 - \varepsilon. \tag{1.28}$$

We denote by $\mathrm{Opt}(\sigma, \varepsilon)$ the optimal value of the above problem (1.28).

Finally, we assume that the corresponding 'scenario counterpart' problems of the form

$$\min_{x \in X} c^T x \text{ subject to } G_\sigma(x, \pm\eta^j) \in C,\ j = 1, ..., N,$$

can be processed efficiently, which definitely is the case when the set C is computationally tractable (recall that the mappings $A_i(\cdot)$ are affine).

As it was mentioned in the Introduction section, we focus on the case when problem of interest (1.28), as it is, is too difficult for numerical processing. Our goal is to use scenario counterpart of (1.28) with randomly chosen scenarios η^j, $j = 1, ..., N$, in order to get a suboptimal solution $\widehat{x}$ to the problem of interest, in a way which ensures that:

1) [*Reliability*] The resulting solution, if any, should be feasible for (1.28) with reliability at least $1 - \delta$: the probability to generate a 'bad' scenario sample – such that $\widehat{x}$ is well defined and is *not* feasible for (1.28) – should be $\leq \delta$ for a given $\delta \in (0, 1)$;

2) [*Polynomiality*] The sample size N should be 'moderate' – polynomial in the sizes of the data describing (1.28) and in $\ln(\varepsilon^{-1})$, $\ln(\delta^{-1})$.

Under these *sine qua non* requirements, we are interested in tight scenario approximations. In our context, it is natural to quantify tightness as follows (*cf.* the definition of resolution):

> *A scenario-based approximation scheme is tight within factor* $\kappa = \kappa(\varepsilon, \delta) \geq 1$, *if whenever* (1.28) *possesses a solution* $\bar{x}$ *which remains feasible after the uncertainty level is increased by factor* κ*, the scheme, with probability at least* $1 - \delta$*, is productive (*$\widehat{x}$ *is well-defined) and ensures that* $c^T\widehat{x} \leq c^T\bar{x}$.

Informally speaking, a reliable κ-tight scenario approximation with probability at least $1 - 2\delta$ is 'in-between' the problem of interest (1.28) and similar problem with κ times larger uncertainty level: up to probability of bad sampling $\leq 2\delta$, the scheme yields an approximate solution which is feasible for the problem of interest and results in the value of the objective not worse than $\mathrm{Opt}(\kappa\sigma, \varepsilon)$.

We are about to present several approximation schemes aimed at achieving the outlined goals.

1.3.1 Naive Approximation

The conceptually simplest way to build a scenario-based approximation scheme for (1.28) is to apply the Analysis test (T) as developed in Section 1.2, with setup as stated in Section 1.2.4. It is convenient to make two conventions as follows:

– From now on, we allow for the pre-trial distribution $\bar{\mathbf{F}}$ to possess the symmetric concentration property. By Remark 3, this extension of the family of trial distributions we can use[18] keeps intact the conclusion of Theorem 6, provided that the Analysis test is applied to a closed convex and symmetric set Q, which will always be the case in the sequel.

– The parameters N, K of the test are those given by (1.23), that is, $K = 0$ and $N = \left\lceil \frac{\ln(\delta)}{\ln(\theta)} \right\rceil$.

Observe that setup of (T) – the pre-trial distribution $\bar{\mathbf{F}}$ and the quantities θ, s, N as defined in Section 1.2.4 – depends solely on the distribution $\mathbf{P}$ of perturbations and required reliability and confidence parameters δ, ε and is completely independent of the (symmetric) convex set Q the test is applied to. It follows, in particular, that a single setup fits all sets from the family

$$Q_{x,\sigma} \doteq \left\{ \xi \in \mathbb{R}^d : G_\sigma(x, \pm\xi) \in C \right\}, \quad x \in X,\, \sigma > 0.$$

Note that all sets from this family are convex, closed and symmetric.

A straightforward approximation scheme for (1.28) based on the Analysis test as applied to the sets $Q_{x,\sigma}$ would be as follows.

Naive approximation scheme: *With setup parameters $\bar{\mathbf{F}}, \theta, s, N$ as described above, we build a sample $\{\eta^j\}_{j=1}^N$ from distribution $\mathbf{F} = \bar{\mathbf{F}}^{(s)}$ and approximate problem* (1.28) *by its scenario counterpart*

$$\min_{x \in X} c^T x \;\text{ subject to } G_\sigma(x, \pm\eta^j) \in C,\; j = 1, ..., N. \tag{1.29}$$

If problem (1.29) *is feasible and therefore solvable (X was assumed to be compact), we take, as $\widehat{x}$, an optimal solution to the problem, otherwise $\widehat{x}$ is undefined (the sample is non-productive).*

By Theorem 6 and Remark 3, every *fixed in advance* point $\bar{x}$ which happens to be feasible for (1.29), with reliability at least $1 - \delta$ is feasible for (1.28). Moreover, in view of Theorem 7 and subsequent discussion, our approximation scheme is tight within factor s, provided that $\bar{\mathbf{F}} = \mathbf{P}$ and

$$2\varepsilon \ln(1/\delta) \leq \delta \ln(1/\theta). \tag{1.30}$$

Unfortunately, these good news about the naive scheme cannot overweight is crucial drawback: *we have no reasons to believe that the scheme satisfies the crucial for us Reliability requirement.* Indeed, the resulting approximate solution $\widehat{x}$ depends on the sample, which makes Theorem 6 inapplicable to $\widehat{x}$.

> The outlined severe drawback of the naive approximation scheme is not just a theoretical possibility. Indeed, assume that $X \doteq \{x \in \mathbb{R}^d : \|x\| \leq 100d^{1/2}\}$, vector c in (1.28) has unit length and the chance constraint in question is $\mathbb{P}\{-1 \leq \xi^T x \leq 1\} \geq 1-\varepsilon$, where $\xi \sim \mathbf{P} = \mathcal{N}(0, I_d)$. Note that all our constructions and bounds are not explicitly affected by the dimension of

[18] The desire to allow for this extension is the reason for requiring the symmetry of $\mathbf{P}$ and passing to the symmeterized form of the chance constraint.

ξ or by the size of X. In particular, when applied to the normal distribution $\mathbf{P} = \bar{\mathbf{F}}$ and given ε and δ, they yield sample size N which is independent of $d = \dim \xi$. For large d, therefore, we will get $2N < d$. In this situation, as it is immediately seen, with probability approaching 1 as $d \to \infty$ there will exist a unit vector x (depending on sample $\{\eta^j\}$) orthogonal to all elements of the sample and such that $e^T x \le -0.1d^{-1/2}$. For such an x, the vector $100d^{1/2}x$ will clearly be feasible for (1.29), whence the optimal value in this problem is ≤ -10. But then every optimal solution to (1.29), in particular, $\widehat{x}$, is of norm at least 10. Thus, the typical absolute values of $\xi^T\widehat{x} \sim \mathcal{N}(0, \|\widehat{x}\|^2)$ are significantly larger than 1, and $\widehat{x}$, with probability approaching 1 as d grows, will be very far from satisfying the chance constraint...

There is an easy way to cure, to some extent, the naive scheme. Specifically, when $\widehat{x}$ is well defined, we generate a new N-element sample from the trial distribution and subject $\widehat{x}$ to our Analysis test. In the case of acceptance conclusion, we treat $\widehat{x}$ as the approximate solution to (1.28) yielded by the modified approximation scheme, otherwise no approximate solution is yielded. This modification makes the naive scheme $(1-\delta)$-reliable, however, at the price of losing tightness. Specifically, let $\bar{\mathbf{F}} = \mathbf{P}$ and (1.30) hold true. In this case, as we have seen, the naive scheme is tight within factor s, while there are no reasons for the modified scheme to share this property.

Numerical illustration

To illustrate the modified naive scheme, consider dynamical system (S) from Section 1.2.5 and pose the following question: what is the largest level of perturbations $\bar{\sigma}$ for which all, up to probability $\varepsilon << 1$, perturbations of A admit a common Lyapunov stability certificate (X, γ) with $\gamma = 0.9999$ and the condition number of X not exceeding 10^5? Mathematically speaking, we are interested to solve the optimization problem

$$\begin{gathered}\max_{\sigma, X} \sigma \;\; \text{subject to } I \preceq X \preceq \alpha I \;\; \text{and} \\ \mathbb{P}\left\{\xi : \pm\sigma\left[\begin{array}{c|c} 0 & (A\cdot\xi)^T X \\ \hline X(A\cdot\xi) & 0\end{array}\right] \preceq \left[\begin{array}{c|c}\gamma^2 X & A^T X \\ \hline XA & X\end{array}\right]\right\} \ge 1-\varepsilon,\end{gathered} \tag{1.31}$$

where $\gamma = 0.9999$, $\alpha = 10^5$ and ξ is a 6×6 random matrix with independent entries uniformly distributed in $[-1, 1]$, and $A \cdot \xi$ denotes the Hadamard (*i.e.*, componentwise) product of matrices A and ξ.

Note that this problem is not exactly in the form of (1.28) – in the latter setting, the level of perturbations σ is fixed, and in (1.31) it becomes the variable to be optimized. Of course, we could apply bisection in σ in order to reduce (1.31) to a small series of feasibility problems of the form (1.28), but on a closest inspection these troubles are completely redundant. Indeed, when applying our methodology to the feasibility problem with a given σ, we were supposed to draw a sample of perturbations $\{s\eta^j\}_{j=1}^N$, with η^j being drawn from pre-trial distribution $\bar{\mathbf{F}}$, with amplification s determined by θ, σ and ε, and then check whether the resulting scenario counterpart of our feasibility problem, that is, the program

$$\begin{aligned}&\text{Find X such that } I \preceq X \preceq \alpha I \text{ and}\\ &\pm s\sigma\left[\begin{array}{c|c} 0 & (A\cdot\eta^j)^T X\\ \hline X(A\cdot\eta^j) & 0\end{array}\right] \preceq \left[\begin{array}{c|c}\gamma^2 X & A^T X\\ \hline XA & X\end{array}\right],\ j=1,...,N,\end{aligned}$$

is or is not feasible. But the answer to this question, given $\{\eta^j\}$, depends solely on the product of $s\sigma$, so that in fact the outlined bisection is equivalent to solving a single problem

$$\begin{aligned}&\max_{\sigma,X} \sigma \text{ subject to } I \preceq X \preceq \alpha I \text{ and}\\ &\pm \sigma\left[\begin{array}{c|c} 0 & (A\cdot\eta^j)^T X\\ \hline X(A\cdot\eta^j) & 0\end{array}\right] \preceq \left[\begin{array}{c|c}\gamma^2 X & A^T X\\ \hline XA & X\end{array}\right],\ j=1,...,N,\end{aligned} \tag{1.32}$$

with η^j drawn from the pre-trial distribution. The latter problem is quasiconvex and therefore can be efficiently solved. After its solution σ_*, X_* is found, we can apply Analysis test to check whether indeed $(X_*, \gamma = 0.9999)$ remains, with probability at least $1 - \varepsilon$, a Lyapunov stability certificate for random perturbations of A at the perturbation level σ_*.

In our experiment, we followed the outlined approach, with the only difference that at the concluding step we used the homogenized Analysis test rather than the basic one. Specifically, we acted as follows:

(a) As in Section 1.2.5, we chose $\mathcal{N}(0, \frac{\pi}{8} I_{36})$ as our pre-trial distribution $\bar{\mathbf{F}}$ and set the sample size N to 1000, which is the size given by (1.23) for $\delta = 10^{-6}$ and $\theta = 0.9863$.

(b) We drew $N = 1000$-element sample from $\bar{\mathbf{F}}$ and solved resulting problem (1.32), thus getting $\sigma_* \approx 0.0909$ and certain X_*.

(c) Our concluding step was to bound from below, for small values of ε, the perturbation levels for which $(X = X_*, \gamma = 0.9999)$ is, with probability $\geq 1-\varepsilon$, a stability certificate for a perturbation of A. This task is completely similar to the one considered in Section 1.2.5, and we acted exactly as explained there. The numerical results are presented in Table 1.2. Comparing the data in Tables 1.1 and 1.1, we see that optimization in X results, for every value of ε, in 'safe' perturbation levels twice as large as those before optimization. To feel the difference, note that at the perturbation level 0.0290 Table 1.1 guarantees preserving (certificate for) stability with confidence as poor as $1 - 0.0039$; Table 1.2 states that even at bit larger perturbation level 0.0297, stability is preserved with confidence as high as $1 - 4 \cdot 10^{-11}$, reliability of both claims being at least 0.999999.

1.3.2 Iterative Approximation

As we have seen, the naive approximation scheme has severe drawbacks: without modification, the scheme possesses certain tightness properties, but can

Table 1.2. $p(\sigma)$: probability of a perturbation (1.27) for which $(X_*, 0.9999)$ fails to be a Lyapunov stability certificate

σ	0.116	0.0912	0.0709	0.0580	0.0491	0.0456	0.0412
$p(\sigma) \leq$	0.3560	0.0890	0.0331	0.0039	2.9e-4	6.9e-5	6.3e-6
σ	0.0412	0.0375	0.0355	0.0336	0.0297	0.0272	0.0255
$p(\sigma) \leq$	4.6e-7	6.9e-9	9.4e-9	3.8e-10	4.0e-11	3.0e-13	5.9e-15

be unreliable; modification recovers reliability, but 'kills' tightness. We are about to present an *iterative* approximation scheme which is reliable *and* has reasonable tightness properties. In the sequel, we sketch the scheme, skipping straightforward and boring details.

Preliminaries: polynomial time black-box convex optimization

Consider a situation as follows. We are given:

(a) A convex compact set $X \subset \mathbb{R}^n$ with non-empty interior, which is contained in the centered at the origin Euclidean ball of a known radius R and is equipped with *Separation Oracle* $\mathcal{S}_Q$ – a routine which, given an input point $x \in \mathbb{R}^n$, reports whether $x \in X$, and if it is not the case, returns a *separator* – a linear inequality which is satisfied everywhere on X and is violated at x.

(b) A linear objective $c^T x$ to be minimized.

(c) Access to a 'wizard' working as follows. The wizard has in its disposal a once for ever fixed set $\mathcal{L}$ of linear inequalities with n variables; when invoked, it picks an inequality from this set and returns it to us. For the time being, we make absolutely no assumptions on how this inequality is chosen: wizard's choice can be randomized, can depend on past choices, *etc.*

(d) Positive parameters r ('feasibility margin') and ω (desired accuracy).

In Convex Programming, there are methods (*e.g.*, the Ellipsoid algorithm) capable to optimize (what precisely, it will become clear in a moment) in the outlined environment, specifically, as follows. The method generates, one after another, a predetermined number M of *search points* $x_t \in \mathbb{R}^n$, $t = 1, ..., M$. At step $t \geq 1$, the method already has in its disposal point x_{t-1} ($x_0 = 0$) and builds a vector e_t and the next search point x_t, namely, as follows:

- [generating e_t] We call the Separation oracle, x_{t-1} being the input. If the oracle reports that $x_{t-1} \notin X$, we call x_{t-1} *non-productive* and specify e_t as the gradient of the separator returned by the oracle. If $x_{t-1} \in X$, we make a predetermined number N of calls to the wizard and add the N linear inequalities returned by the wizard at step t to the collection of inequalities returned at the previous steps, thus getting a list of Nt linear inequalities. We then check whether x_{t-1} satisfies all these Nt inequalities. If there is an inequality in the list which is violated at x_{t-1}, we qualify x_{t-1} as non-productive and specify e_t as the gradient of the violated inequality. Finally, if $x_{t-1} \in X$ satisfies all inequalities returned so far by the wizard, we qualify x_{t-1} as productive and set $e_t = c$.

• [generating x_t] Given x_{t-1}, e_t and information coming from the previous steps (for the Ellipsoid method, the latter is summarized in a single $n \times n$ matrix B_{t-1}), we build x_t. How x_t is built, it depends on the method in question; the only issue which matters in our context is that the arithmetic cost of generating x_t should be polynomial in n (for the Ellipsoid method, the cost of building x_t *and* updating $B_{t-1} \mapsto B_t$ is just $O(1)n^2$ operations).

After all M search points are built, we treat the best (with the smallest value of $c^T x$) of the *productive* search points as the resulting approximate solution $\widehat{x}$; if no productive search points were generated, the result is undefined.

Now, upon termination, we have in our disposal a list $\mathcal{I}$ of NM linear inequalities $\ell(x) \leq 0$ which came from the wizard; these inequalities define the convex compact set

$$X^{\mathcal{I}} = \{x \in X : \ell(x) \leq 0,\ \ell \in \mathcal{I}\}.$$

The convex optimization algorithms we are speaking about ensure the following property:

> (P): *With properly chosen and polynomial in n and $\ln\left(\frac{nR}{r} \cdot \frac{R\|c\|}{\omega}\right)$ number of steps $M = M(n, R, r, \omega)$ (for the Ellipsoid method, $M = 2n^2 \ln\left(\frac{nR^2\|c\|}{r\omega} + 2\right)$), the following is true: whenever the set $X^{\mathcal{I}}$ contains Euclidean ball of radius r, $\widehat{x}$ is well defined and*
>
> $$c^T \widehat{x} \leq \min_{x \in X^{\mathcal{I}}} c^T x + \omega.$$

Now we can finally explain what is the optimization problem we were solving: this is the problem $\min\limits_{x \in X^{\mathcal{I}}} c^T x$ *defined in course of the solution process*[19].

Iterative approximation scheme

We are ready to present an *iterative approximation scheme* for solving (1.28). Assume that the domain X of (1.28) is contained in the centered at the origin ball of known radius R and that both X and C are equipped with

[19] In standard applications, this situation, of course, is not that strange: the problem we are solving is known in advance and is $\min\limits_x \left\{c^T x : x \in X, g(x) \leq 0\right\}$, where g is a convex function. The set $\mathcal{L}$ of linear inequalities is comprised of inequalities of the form $\ell_y(x) \equiv g(y) + (x-y)^T g'(y) \leq 0$, $y \in \mathbb{R}^n$, and the inequality returned by the wizard invoked at point x_{t-1} is $\ell_{x_{t-1}}(x) \leq 0$. In this case, $\widehat{x}$, if defined, is a feasible solution to the problem of interest, and if the feasible set of the latter problem contains a ball of radius r, then $\widehat{x}$ is well-defined and is an ω-optimal solution to the problem of interest, provided that the number of steps M is as in (P).

Separation Oracles. Given required confidence and reliability parameters ε, δ, let us choose trial distribution $\mathbf{F}$, θ, s and sample size N exactly in the same fashion as for naive scheme. Besides this, let us choose an optimization algorithm possessing property (P); for the sake of definiteness, let it be the Ellipsoid method. Finally, let us choose a small positive r and specify the number M of steps of the method according to (P), that is,

$$M = O(1)n^2 \ln\left(\frac{nR}{r} \cdot \frac{R\|c\|}{\omega} + 2\right),$$

where ω is the accuracy within which we want to solve (1.28). Now let us run the Ellipsoid method, mimicking the wizard as follows:

> The linear inequalities returned by the wizard at step t are uniquely defined by the search point x_{t-1} and a realization η^τ of a random vector $\eta \sim \mathbf{F}$; here τ counts the calls to the wizard, *and* $\eta^1, \eta^2, \ldots$ *are independent of each other*. Given x_{t-1} and η^τ, the wizard computes the points $y_\pm = G_\sigma(x_{t-1}, \pm\eta^\tau)$ and calls the Separation Oracle for C to check whether both these points belong to C. If it is the case, the wizard returns a trivial – identically true – inequality $\ell(x) \equiv 0^T x \le 0$. If at least one of the points, say, y_+, does *not* belong to C, the wizard acts as follows. Let $e(u) \le 0$ be the linear inequality returned by the Separation oracle; this inequality holds true for $u \in C$ and is violated at y_+. The wizard converts this inequality into the linear inequality
>
> $$\ell(x) \equiv e\left(A_0(x) + \sigma \sum_{i=1}^{d} \eta_i^\tau A_i(x)\right) \le 0$$
>
> and this is the inequality the wizard returns. Since $A_i(\cdot)$ are affine, this indeed is a linear inequality in variables x, and since $e(y_+) > 0$, this inequality is violated at x_{t-1}.

We have specified the wizard and thus a (randomized) optimization process; a realization of this process and the corresponding result $\widehat{x}$, if any, are uniquely defined by a realization of MN-element sample with independent elements drawn from the trial distribution. The resulting approximation scheme for (1.28) is successful if and only if $\widehat{x}$ is well defined, and in this case $\widehat{x}$ is the resulting approximate solution to (1.28).

Let us investigate the properties of our new approximation scheme. Our first observation is that the scheme is reliable.

Theorem 9. *The reliability of the iterative approximation is at least* $1 - M\delta$*, that is, the probability to generate a sample such that* $\widehat{x}$ *is well defined and is* not *feasible for* (1.28) *is at most* $M\delta$*.*

Proof. If $\widehat{x}$ is well defined, it is one of the productive points x_{t-1}, $1 \le t \le M$. Observe that for a given t the probability of 'bad sampling' at step t, that is, probability of the event E_t that x_{t-1} is declared productive and at the same time $\mathbf{F}(\{\eta : G_\sigma(x_{t-1}, \pm\xi) \in C\}) < \theta$, is at most δ. Indeed, by wizard's construction, the conditional, given what happened before step t, probability of this event is at most the probability to get N successes in N independent Bernoulli experiments 'check whether $G_\sigma(x_{t-1}, \pm\zeta^p) \in C$' with $\zeta^p \sim \mathbf{F}$, $p = 1, ..., N$, where the probability of success in a single experiment is $< \theta$; by (1.23), this probability is at most δ. Since the conditional, given the past, probability of E_t is $\leq \delta$, so is the unconditional probability of E_t, whence the probability of the event $E = E_1 \cup ... \cup E_M$ is at most $M\delta$. If the event E does not take place and $\widehat{x}$ is well-defined, then $\widehat{x}$ satisfies the requirement $\mathbf{F}(\{\eta : G_\sigma(\widehat{x}, \pm\eta) \in C\}) \geq \theta$, whence, by properties of our analysis test, $\mathbf{P}(\{\xi : G_\sigma(\widehat{x}, \pm\xi) \in C\}) \geq 1 - \varepsilon$. By construction, $\widehat{x}$, if well defined, belongs to X. Thus, $\widehat{x}$ indeed is feasible for (1.28) 'modulo event E of probability $\leq M\delta$'. □

Our next observation is that *when $\mathbf{P} = \bar{\mathbf{F}}$, the iterative scheme is nearly tight up to factor s*. The precise statement is as follows.

Theorem 10. *Let $\bar{\mathbf{F}} = \mathbf{P}$, and let there exist an Euclidean ball $U_r \subset X$ of radius nr such that all points $x \in U_r$ are feasible for* (1.28)*, the uncertainty level being increased by factor s:*

$$\mathbf{P}(\{\xi : G_{s\sigma}(x, \pm\xi) \in C\}) \geq 1 - \varepsilon, \quad \text{for all } x \in U_r.$$

Then, with reliability at least $1 - (n+2)MN\varepsilon$, $\widehat{x}$ is well defined and satisfies the relation

$$c^T\widehat{x} \leq \mathrm{Opt}(s\sigma, \varepsilon) + \omega, \tag{1.33}$$

where s is the amplification parameter of the scheme. In other words, the probability to generate a sample $\eta^1, ..., \eta^{MN}$ such that $\widehat{x}$ is undefined or is well defined but fails to satisfy (1.33) *is at most $(n+2)MN\varepsilon$.*

Proof. Let $\kappa > 0$, and let $\bar{x}_\kappa \in X$ be such that

$$c^T\bar{x}_\kappa \leq \mathrm{Opt}(s\sigma, \varepsilon) + \kappa \ \text{ and } \ \mathbf{P}(\{\xi : G_{s\sigma}(\bar{x}_\kappa, \pm\xi) \in C\}) \geq 1 - \varepsilon.$$

Now, let Δ be a perfect simplex with vertices $z_0, ..., z_n$ on the boundary of U_r; since the radius of U_r is nr, Δ contains a ball V of radius r. We now claim that up to probability of bad sampling $p = (n+2)MN\varepsilon$, the $n+2$ points $z_0, ..., z_n, \bar{x}_\kappa$ belong to $X^{\mathcal{I}}$. Indeed, let $z \in X$ be a fixed point satisfying the chance constraint

$$\mathbf{P}(\{\xi : G_{s\sigma}(z, \pm\xi) \in C\}) \geq 1 - \varepsilon$$

(as it is the case for $z_0, ..., z_n, \bar{x}_\kappa$). Due to $z \in X$ and the construction of our wizard, the event $z \notin X^{\mathcal{I}}$ takes place if and only if the underlying sample $\eta^1, ..., \eta^{MN}$ of MN independent realizations of random vector $\eta \sim \mathbf{F} = \mathbf{P}^{(s)}$ contains an element η^t such that either $e\,(G_\sigma(z, \eta^t)) > 0$ or $e\,(G_\sigma(z, -\eta^t)) > 0$, or both, where e is an affine function (depending on the sample) such that

$e(y) \leq 0$ for all $y \in C$. Thus, at least one of the two points $G_\sigma(z, \pm\eta^t)$ fails to belong to C. It follows that the event $z \notin X^{\mathcal{I}}$ is contained in the union, over $t = 1, ..., MN$, of the *complements* to the events $F_t = \{\eta : G_\sigma(z, \pm\eta^t) \in C\}$. Due to $\mathbf{F} = \mathbf{P}^{(s)}$, the $\mathbf{F}$-probability of F_t is nothing but the $\mathbf{P}$-probability of the event $\{\xi : G_{s\sigma}(z, \pm\xi) \in C\}$, that is, $\mathbf{F}(F_t) \geq 1 - \varepsilon$. It follows that the probability of the event $z \notin X^{\mathcal{I}}$ is at most $MN\varepsilon$. Applying this result to every one of the points $z_0, ..., z_n, \bar{x}_\kappa$, we conclude that the probability for at least one of these points to be outside of $X^{\mathcal{I}}$ is at most $(n+2)MN\varepsilon$, as claimed.

We are nearly done. Indeed, let E be the event

$$\{\eta^1, ..., \eta^{MN} : z_0, ..., z_n, \bar{x}_\kappa \in X^{\mathcal{I}}\}.$$

As we just have seen, the probability of this event is at least $1-(n+2)MN\varepsilon$. Since $X^{\mathcal{I}}$ is convex, in the case of E the set $X^{\mathcal{I}}$ contains the entire simplex Δ with the vertices $z_0, ..., z_n$ and thus contains the ball V_r of radius r. Invoking (P), we see that in this case $\widehat{x}$ is well defined and

$$c^T\widehat{x} \leq \omega + \min_{x \in X^{\mathcal{I}}} c^T x \leq \omega + c^T \bar{x}_\kappa \leq \omega + \kappa + \mathrm{Opt}(s\sigma, \varepsilon),$$

where the second inequality is given by the fact that in the case of E we have $\bar{x}_\kappa \in X^{\mathcal{I}}$. Thus, the probability of the event '$\widehat{x}$ is well defined and satisfies $c^T\widehat{x} \leq \mathrm{Opt}(s\sigma, \varepsilon)+\omega+\kappa$' is at least the one of E, that is, it is $\geq 1-(n+2)MN\varepsilon$. Since $\kappa > 0$ is arbitrary, (1.33) follows. □

Discussion

With the Ellipsoid method as the working horse, the number M of steps in the iterative approximation scheme is about $2n^2 \ln\left(\frac{nR^2\|c\|}{r\omega}\right)$. It follows that the *un*reliability level guaranteed by Theorem 9 does not exceed $2n^2 \ln\left(\frac{nR^2\|c\|}{r\omega}\right)\delta$; in order to make this unreliability at most a given $\chi << 1$, it suffices to take $\delta = \frac{1}{2}\chi n^{-2} \ln^{-1}\left(\frac{nR^2\|c\|}{r\omega}\right)$. Since relation (1.23) requires 'per step' sample size $N = \left\lceil \frac{\ln(\delta)}{\ln(\theta)} \right\rceil$, with our δ the total sample size MN is polynomial in $\frac{n}{1-\theta}$ and in *logarithms* of all remaining parameters (R, r, ω, χ). Thus, our approximation scheme is polynomial, which are good news. Further, with the outlined setup the *un*reliability level $\bar{\chi} = (n+2)MN\varepsilon$ indicated in Theorem 10 is linear in ε and polynomial in $\frac{n}{1-\theta}$ and logarithms of the remaining parameters, which again are good news. A not so good news is that the scheme requires an *ad hoc* choice of r. This, however, seems not that disastrous, since the only element of the construction which is affected by this choice (and affected just logarithmically) is the number of steps M. In reality, we can choose M as large as is allowed by side considerations like restrictions on execution time, thus making r as small as possible under these restrictions (or, equivalently, arriving at approximation as tight as possible, since the less is r, the more likely becomes the premise in Theorem 10).

As far as practicality of the iterative approximation scheme is concerned, the factor of primary importance is the design dimension n, since the reliability characteristics and the computational complexity of the scheme are much more sensitive to n than to parameters like $R, r, \omega, ...$ Let us look at this phenomenon in more details. With $(nR/r) \cdot (R\|c\|/\omega)$ bounded from above by 10^{12} (which seems to be sufficient for real life applications), we have $M = 55n^2$. Bounding the total number of scenarios MN by 10^6 and setting the reliability parameter χ to 10^{-6}, we get $N = 10^6 M^{-1} = 1.82 \cdot 10^4 \cdot n^{-2}$ and $\delta = M^{-1}\chi = 1.82 \cdot 10^{-8} \cdot n^{-2}$. Via (1.23), the resulting N and δ correspond to

$$\theta = \theta(n) \doteq \exp\{-\ln(1/\delta)/N\} = \exp\left\{-n^2 \frac{17.8 - 2\ln(n)}{1.82 \cdot 10^4}\right\}.$$

Let the pre-trial distribution be normal. Then $\theta(n)$ should be $> \bar{\theta} = 0.5$, which is the case for $n \leq 34$ only. For $n \leq 34$ and $\theta = \theta(n)$, the associated confidence parameter $\varepsilon = \mathrm{Err}(s, \theta(n))$ depends solely on the amplification parameter s; the tradeoff between s and ε is presented in Table 1.3. As we see, the required amplification level rapidly grows (*i.e.*, tightness rapidly deteriorates) as n grows. This is exactly what should be expected, given that the per step number of scenarios N under our assumptions is inverse proportional to n^2. The influence of n can be moderated by replacing our working horse, the Ellipsoid method, with more advanced convex optimization algorithms; this issue, however, goes beyond the scope of this contribution.

Table 1.3. Tradeoff between amplification s and confidence parameter ε for iterative approximation scheme (total sample size 10^6, normal trial distribution)

n	2	6	10	14	18	22	26	30
$\theta(n)$	0.9964	0.9723	0.9301	0.8738	0.8074	0.7432	0.6576	0.5805
N	4450	506	182	93	57	38	27	21
ε	s							
1.0e-3	1.17	1.88	2.54	3.39	4.63	6.68	10.80	23.07
1.0e-4	1.50	2.19	2.93	3.88	5.25	7.51	12.02	25.38
1.0e-5	1.70	2.46	3.27	4.31	5.80	8.26	13.14	27.49
1.0e-6	1.88	2.71	3.58	4.70	6.31	8.95	14.16	29.45
1.0e-7	2.05	2.93	3.86	5.06	6.78	9.58	15.12	31.30
1.0e-8	2.20	3.14	4.13	5.40	7.21	10.18	16.02	33.03
1.0e-9	2.34	3.33	4.38	5.71	7.63	10.75	16.87	34.67
1.0e-10	2.47	3.52	4.61	6.02	8.02	11.28	17.68	36.25
1.0e-11	2.60	3.69	4.84	6.30	8.39	11.79	18.45	37.76
1.0e-12	2.72	3.86	5.05	6.57	8.75	12.28	19.20	39.22
1.0e-13	2.83	4.02	5.26	6.84	9.09	12.75	19.91	40.63
1.0e-14	2.94	4.17	5.45	7.09	9.42	13.21	20.61	41.98

1.3.3 The Case of Chance Semidefinite Constraint

In this section, we focus on the case of 'relatively simple' geometry of C, specifically, assume that C can be represented as the intersection of the cone $\mathbf{S}^m_+$ of positive semidefinite symmetric $m \times m$ matrices and an affine plane, or, equivalently, that the randomly perturbed constraint in question is Linear Matrix Inequality (LMI)

$$x \in X \text{ and } A_\xi(x) \doteq \sum_{i=1}^{d} \xi_i A_i(x) \preceq A_0(x), \tag{1.34}$$

where $X \subset \mathbb{R}^n$ is the domain of our constraint (we assume the domain to be convex and compact), $A_i(x)$, $i = 0, ..., d$, are symmetric matrices affinely depending on $x \in \mathbb{R}^n$, $\xi_i \in \mathbb{R}$ are random perturbations. Without loss of generality we have set the level of perturbations σ to 1, so that σ is not present in (1.34) at all. Note that the family of cross-sections of the semidefinite cone is very rich, which allows to reformulate in the form of (1.34) a wide spectrum of systems of convex constraints, *e.g.*, (finite) systems of linear and conic quadratic inequalities. Besides this, LMI constraints arise naturally in many applications, especially in Control [59].

The question we address is as follows. Let $\mathbf{P}$ be the distribution of the perturbation $\xi = (\xi_1, ..., \xi_d)$, and let X_ε be the solution set of the chance constraint associated with (1.34):

$$X_\varepsilon = \{x \in X : \mathbf{P}(\{\xi : A_\xi(x) \preceq A_0(x)\}) \geq 1 - \varepsilon\}.$$

Now suppose that we choose somehow a symmetric pre-trial distribution $\bar{\mathbf{F}}$, draw an N-element sample $\eta[N] = \{\eta^j\}_{j=1}^N$ from the trial distribution $\mathbf{F} = \bar{\mathbf{F}}^{(s)}$ (s is the amplification level) and thus obtain the 'scenario approximation' of X_ε – the set

$$X(\eta[N]) = \left\{x \in X : A_{\eta^j}(x) \preceq A_0(x),\, j = 1, ..., N\right\}.$$

The question we are interested in is: *Under which circumstances the random scenario approximation $X(\eta[N])$ is, with reliability at least $1 - \delta$, a subset of X_ε, that is,*

$$\mathbb{P}\{\eta[N] : X(\eta[N]) \subset X_\varepsilon\} \geq 1 - \delta. \tag{1.35}$$

Note the difference between this question and the one addressed in Section 1.2. The results of Section 1.2, when translated into our present situation, explain under which circumstances, *given in advance* a point x and having observed that $x \in X(\eta[N])$, we may be pretty sure that $x \in X_\varepsilon$. Now we require much more: having observed $\eta[N]$ (and thus $X(\eta[N])$), we want to be pretty sure that *all* points from $X(\eta[N])$ belong to X_ε. Note that in the latter case every point of $X(\eta[N])$, *e.g.*, the one which minimizes a given objective $c^T x$ over $X(\eta[N])$, belongs to X_ε. In other words, in the case of (1.35), an approximation scheme where one minimizes $c^T x$ over $X(\eta[N])$ allows to find, with reliability $1-\delta$, *feasible* suboptimal solution to the problem $\min\limits_{x\in X_\varepsilon} c^T x$ of minimization under the chance constraint.

Preprocessing the situation

For the moment, let us restrict ourselves to the case where $\mathbf{P} = \mathbf{P}_1 \times ... \times \mathbf{P}_d$, where $\mathbf{P}_i$, $i = 1, ..., d$, is the distribution of ξ_i assumed to be symmetric. Note that if $a_i > 0$ are deterministic scalars, we can replace the perturbations ξ_i with $a_i\xi_i$, and mappings $A_i(x)$ with the mappings $a_i^{-1}A_i$ without affecting the feasible set of the chance constraint. In other words, we lose nothing when assuming that 'typical values' of ξ_i are at least of order of 1, specifically, that $\mathbf{P}_i(\{|\xi_i| \geq 1\}) \geq 0.2$, $i = 1, ..., d$. With this normalization, we immediately arrive at a rough *necessary* condition for the inclusion $x \in X_\varepsilon$, namely,

$$\pm A_i(x) \preceq A_0(x),\ i = 1, ..., d. \tag{1.36}$$

Indeed, let $x \in X_\varepsilon$ with $\varepsilon < 0.45$. Given $p \leq d$ and setting

$$A_\xi(x) \doteq \xi_p A_p(x) + S_\xi^p(x),$$

observe that ξ_p and $S_\xi^p(x)$ are independent and symmetrically distributed, which combines with $x \in X_\varepsilon$ to imply that

$$\mathbf{P}(\{\xi : \xi_p A_p(x) \pm S_\xi^p(x) \preceq A_0(x)\}) \geq 1 - 2\varepsilon.$$

By our normalization and due to the symmetry of $\mathbf{P}_p$, we have that $\mathbf{P}(\{\xi : \xi_p \geq 1\}) \geq 0.1$. It follows that

$$\mathbf{P}(\{\xi : \xi_p \geq 1\ \&\ \xi_p A_p(x) \pm S_\xi^p(x) \preceq A_0(x)\}) \geq 0.9 - 2\varepsilon > 0,$$

that is, the set $\{\xi : \xi_p \geq 1\ \&\ \xi_p A_p(x) \pm S_\xi^p(x) \preceq A_0(x)\}$ is non-empty, which is possible only when $t_+ A_p(x) \preceq A_0(x)$ for certain $t_+ \geq 1$. Similar reasoning proves that $-t_- A_p(x) \preceq A_0(x)$ for certain $t_- \geq 1$; due to these observations, $\pm A_i(x) \preceq A_0(x)$.

Note that (1.36) is a nice deterministic convex constraint, and it makes sense to include it into the definition of X; with this modification of X, we have $A_0(x) \succeq 0$ everywhere on X (since (1.36) implies $A_0(x) \succeq 0$). In order to simplify our subsequent analysis, let us strengthen the latter inequality to $A_0(x) \succ 0$ (which can be ensured by slight shrinkage of X to a point $\bar{x}$ such that $A_0(\bar{x}) \succ 0$, provided that such a point exists). Thus, from now on we make the following assumption:

A.I. *X is a closed and convex compact set such that relations* (1.36) *and $A_0(x) \succ 0$ take place everywhere on X.*

Now we formulate our assumptions on the actual and the pre-trial distributions. We discard temporary assumptions on $\mathbf{P}$ made at the beginning of this subsection (their only goal was to motivate **A.I**); what we actually need are similar in spirit assumptions on the pre-trial distribution. Here is what we assume from now on:

A.II. *The actual distribution* $\mathbf{P}$ *is with zero mean and is majorized by symmetric pre-trial distribution* $\bar{\mathbf{F}} \in \mathcal{C}(\bar{\theta}, \psi)$ *with known* $\bar{\theta}$, $\psi(\cdot,\cdot)$. *In addition,*

1) *For certain* $\widehat{\theta} \in (\bar{\theta}, 1)$ *and all* $\gamma \geq 1$ *one has*

$$\psi(\widehat{\theta}, \gamma) \geq a + b\gamma^2/2$$

with $b > 0$*;*

2) *For certain* c*, random vector* $\eta \sim \bar{\mathbf{F}}$ *satisfies the bound*

$$\mathbb{E}\{\|\eta\|^2\} \leq c^2 d.$$

The result we are about to establish (for the case of normal distributions, it was announced in [240]) is as follows.

Theorem 11. *Let* **A.I-II** *hold true. Given confidence and reliability parameters* $\varepsilon, \delta \in (0, 1/2)$*, let us set, for* $s > 1$*,*

$$\mathrm{Err}(s) = \inf_{1 \leq \beta < s} \tfrac{1}{s-\beta} \int\limits_{\beta}^{\infty} \exp\{-\psi(\widehat{\theta}, \gamma)\} d\gamma$$

(cf. (1.19)*) and specify the amplification parameter* s *in such a way that*

$$\mathrm{Err}(s) = \varepsilon;$$

note that

$$s \leq 2 + \sqrt{\tfrac{|a| + \ln\left(\frac{2}{b\varepsilon}\right)}{b}}$$

in view of (1.20)*.*

Let, further, the sample size N *be specified as*

$$N = \left\lceil \tfrac{\kappa}{1-\widehat{\theta}} \left(\ln(\delta^{-1}) + \kappa m^2 d \ln(Csd) \right) \right\rceil, \tag{1.37}$$

with appropriately chosen absolute constant κ *and constant* C *depending solely on* $\widehat{\theta}, a, b, c$*. Then, with sample* $\eta[N]$ *drawn from the trial distribution* $\mathbf{F} = \bar{\mathbf{F}}^{(s)}$*, one has*

$$\mathbb{P}\{X(\eta[N]) \subset X_\varepsilon\} \geq 1 - \delta.$$

For proof, see Appendix.

Note that when treating the parameters $\widehat{\theta}$, a, b, c involved into **A.I-II** as absolute constants (which is possible, *e.g.*, for the pre-trial distributions given by Examples 2–4, see Section 1.2.2), the sample size N as given by (1.37) is polynomial in the sizes m, d of the problem and in $\ln(1/\delta)$, $\ln(\ln(1/\varepsilon))$.

Tightness of the approximation scheme suggested by Theorem 11 admits the following evident quantification.

Proposition 6. *Let, in addition to Assumptions* **A.I-II***, the pre-trial distribution* $\bar{\mathbf{F}}$ *be identical to the actual distribution* $\mathbf{P}$*, and let* x *be a fixed in advance point of* X *which is feasible for the chance constraint with increased by factor* s *level of perturbations:*

$$\mathbf{P}\left(\{\xi : s\sum_{i=1}^{d}\xi_i A_i(x) \preceq A_0(x)\}\right) \geq 1-\varepsilon, \tag{1.38}$$

where s *is the amplification parameter specified in Theorem 11. Then* $x \in X(\eta[N])$*, the sample being drawn from the trial distribution as defined in Theorem 11, with probability at least* $1 - N\varepsilon$*, where* N *is given by* (1.37)*. In particular, optimizing a given objective* $c^T x$ *over* $X(\eta[N])$*, we, with reliability at least* $1 - \delta - N\varepsilon$*, get a point* $\widehat{x} \in X_\varepsilon$ *with the value of the objective not exceeding*

$$\min_x \left\{c^T x : x \in X \text{ satisfies (1.38)}\right\}$$

Note that the amplification factor s specified in Theorem 11 is $O(1)\sqrt{\ln(1/\varepsilon)}$, provided that we treat a, b as absolute constants; thus, under the premise of Proposition 6 the tightness of our approximation scheme is nearly independent of ε.

1.4 Concluding Remarks

In this chapter, our goal was to get reliable *inner* approximations of the feasible set of optimization problem (1.28) with chance constraint; we have seen that in good cases (*e.g.*, when the perturbations have normal or uniform distributions, and C is the semidefinite cone), the scenario approach allows to achieve this goal with polynomial in the sizes of the problem and logarithms of the reliability and confidence parameters number of scenarios and level of conservativeness as moderate as $O(1)\sqrt{\ln(1/\varepsilon)}$. A natural question is whether something similar can be done for *outer* approximation of the feasible set in question. The answer, in general, seems to be negative, as can be seen from the following example. Assume that the chance constraint is

$$\mathbb{P}\left\{x^T\xi \leq 1\right\} \geq 1-\varepsilon,$$

where $\xi \sim \mathcal{N}(0, I_n)$. The true feasible set X_ε of the chance constraint is the centered at the origin Euclidean ball E^ε of the radius $r = r(\varepsilon)$ given by $\frac{1}{\sqrt{2\pi}}\int\limits_r^\infty \exp\{-\gamma^2/2\}d\gamma = \varepsilon$, so that $r = (1+o(1))\sqrt{2\ln(1/\varepsilon)}$ as $\varepsilon \to +0$. At the same time, the radius of the largest centered at the origin ball U contained in the feasible set $\{x : x^T\xi^j \leq 1, j = 1, ..., N\}$ of the scenario counterpart, where ξ^j are drawn from $\mathcal{N}(0, \sigma^2 I_n)$, is, with probability approaching 1 as $n \to \infty$, as small as $\sigma^{-1}n^{-1/2}$ (since typical values of $\|\xi^j\|$ are as large as $\sigma\sqrt{n}$). Thus, unless σ we use goes to 0 as $O(n^{-1/2})$ as n grows (which would make no much sense), the scenario approximation of X_ε with high probability is much ‘thinner’ along certain *sample-depending* directions than X_ε itself.

1.5 Appendix: Proof of Theorem 11

Recall that d is the dimension of the perturbation vectors, m is the row size of the matrices $A_i(x)$. From now on, $O(1)'$s stand for appropriate positive *absolute* constants, and C_i are positive quantities depending solely on the quantities $\widehat{\theta}$, a, b, c involved into Assumption **A.II**.

Lemma 3. *Let $\eta \sim \mathbf{F}$. Then for $\rho \geq 0$,*

$$\mathbf{F}\left(\left\{\eta : \|\eta\| > \rho s\sqrt{d}\right\}\right) \leq 2\exp\left\{-C_1\rho^2\right\}. \tag{1.39}$$

Proof. By **A.II**.2) and Chebychev Inequality ,

$$\mathbf{F}\left(\{\eta : \|\eta\| \leq C_{1,1}s\sqrt{d}\}\right) \geq \widehat{\theta}$$

for appropriately chosen $C_{1,1}$. Due to the concentration property and **A.II.**1), it follows that whenever $\gamma \geq 1$, we have

$$\mathbf{F}\left(\{\eta : \|\eta\| \geq C_{1,1}s\sqrt{d}\gamma\}\right) \geq \exp\{-a - b\gamma^2/2\},$$

and (1.39) follows. □

Our next technical result is as follows.

Lemma 4. *Let $\mathcal{A} = \{(A_1, ..., A_d) : A_i \in \mathbf{S}^m, -I \preceq A_i \preceq I\}$. For $A = (A_1, ..., A_d) \in \mathcal{A}$, let*

$$B(A) = \left\{u \in \mathbb{R}^d : 0.9\sum_{i=1}^{d} u_iA_i \preceq I\right\}.$$

Further, let $\eta \sim \mathbf{F}$, N be a positive integer, let η^j, $j = 1, ..., N$, be independent realizations of η, and let $\mathbf{F}_N$ be the distribution of $\eta[N] = \{\eta^j\}_{j=1}^N$. Finally, let $\Delta \doteq \frac{1-\widehat{\theta}}{4}$ and

$$\Xi^N \doteq \left\{\eta[N] : \forall\left(A \in \mathcal{A} : \mathbf{F}(B(A)) < \widehat{\theta}\right) \exists t \leq N : \sum_{i=1}^{d} \eta_i^j A_i \not\preceq I\right\}. \tag{1.40}$$

Then

$$\mathbf{F}_N(\Xi^N) \geq 1 - \exp\{O(1)m^2d\ln(C_2sd) - O(1)(1-\widehat{\theta})N\} \tag{1.41}$$

with properly chosen C_2.

Proof. Let us equip the space of k-tuples of $m \times m$ symmetric matrices with the norm

$$\|(A_1, ..., A_d)\|_\infty = \max_i \|A_i\|,$$

where $\|A_i\|$ is the standard spectral norm of a symmetric matrix. Given $\omega > 0$, let $\mathcal{A}^\omega$ be a minimal ω-net in $\mathcal{A}$; by the standard reasons, we have

$$\mathrm{card}(\mathcal{A}^\omega) \leq \exp\{O(1)m^2 d\ln(2+\omega^{-1})\}. \tag{1.42}$$

Note that if $A, A' \in \mathcal{A}$, then

$$0.9\sum_{i=1}^d \eta_i A_i \preceq 0.9\sum_{i=1}^d \eta_i A_i' + 0.9\|\eta\|_1\|A'-A\|_\infty I,$$

whence

$$\left\{\eta : 0.9\sum_{i=1}^d \eta_i A_i \preceq I\right\} \supset \left\{\eta : \sum_{i=1}^d \eta_i A_i' \preceq 1.1I\right\} \bigcap \left\{\eta : 0.9\|\eta\|_1\|A'-A\|_\infty \leq 0.01\right\},$$

so that

$$\begin{array}{rl}\mathbf{F}(B(A)) \geq & \underbrace{\mathbf{F}(\{\eta : \sum_{i=1}^d \eta_i A_i' \preceq 1.1I\})}_{\phi(A')} - \mathbf{F}(\{\eta : 0.9\|\eta\|_1\|A'-A\|_\infty > 0.01\}) \\ \geq & \phi(A') - 2\exp\{-C_{2,1}\|A'-A\|_\infty^{-2}(ds)^{-2}\}\end{array} \tag{1.43}$$

for appropriately chosen $C_{2,1}$, where the concluding $\geq$ is given by (1.39) due to $\|\eta\|_1 \leq \sqrt{d}\|\eta\|$.

Now let

$$\mathcal{B}^\omega = \{A' \in \mathcal{A}^\omega : \mathbf{F}(\{\eta : \sum_{i=1}^d \eta_i A_i' \preceq 1.1I\}) \leq \widehat{\theta} + \varDelta\}$$

where $\varDelta$ is given by (1.40). According to (1.39), we can find $C_{2,2}$ such that

$$\mathbf{F}\left(\{\eta : \|\eta\| \geq C_{2,2}s\sqrt{d}\}\right) \leq \varDelta,$$

so that $A' \in \mathcal{B}^\omega$ implies

$$\mathbf{F}\left(\{\eta : \sum_{i=1}^d \eta_i A_i' \preceq 1.1I \text{ or } \|\eta\|_1 > C_{2,2}sd\}\right) \leq \widehat{\theta} + 2\varDelta = \frac{1+\widehat{\theta}}{2} < 1.$$

Setting

$$\varXi_\omega^N[A'] = \left\{\eta[N] : \forall (j \leq N) : \|\eta^j\| > C_{2,2}s\sqrt{d} \text{ or } \sum_{i=1}^d \eta_i^t A_i' \preceq 1.1I\right\},$$

we have by evident reasons

$$A' \in \mathcal{B}^\omega \Rightarrow \mathbf{F}_N(\varXi_\omega^N[A']) \leq \exp\{-O(1)(1-\widehat{\theta})N\},$$

whence

$$\begin{aligned}\mathbf{F}_N\left\{\cup_{A'\in\mathcal{B}^\omega}\Xi^N_\omega[A']\right\} &\leq \operatorname{card}(\mathcal{A}^\omega)\exp\{-O(1)(1-\widehat{\theta})N\}\\ &\leq \exp\{O(1)m^2d\ln(2+\omega^{-1})-O(1)(1-\widehat{\theta})N\}\end{aligned} \tag{1.44}$$

(we have used (1.42)). Now let us set $\omega = C_{2,3}(sd)^{-1}$ with $C_{2,3}$ chosen in such a way that $C_{2,2}\omega sd < 0.1$ and (1.43) implies that

$$A, A' \in \mathcal{A}, \|A'-A\|_\infty \leq \omega \Rightarrow \phi(A') \leq \mathbf{F}(B(A)) + \varDelta. \tag{1.45}$$

Let E be the complement of the set $\cup_{A'\in\mathcal{B}^\omega}\Xi^N_\omega[A']$; due to (1.44) and to our choice of ω, we have

$$\mathbf{F}_N(E) \geq 1 - \exp\{O(1)m^2d\ln(C_2sd) - O(1)(1-\widehat{\theta})N\}.$$

In view of this relation, in order to prove Lemma it suffices to verify that $E \subset \Xi^N$, that is,

$$\eta[N] \in E \Rightarrow \left[\forall\left(A\in\mathcal{A} : \mathbf{F}(B(A)) < \widehat{\theta}\right)\ \exists j \leq N : \sum_{i=1}^d \eta^j_i A_i \not\preceq I\right].$$

Indeed, given $A \in \mathcal{A}$ such that $\mathbf{F}(B(A)) < \widehat{\theta}$, let A' be the $\|\cdot\|_\infty$-closest to A point from $\mathcal{A}^\omega$, so that $\|A-A'\|_\infty \leq \omega$. By (1.45),

$$\phi(A') \doteq \mathbf{F}\left(\{\eta : \sum_{i=1}^d \eta_i A_i \preceq 1.1I\}\right) \leq \mathbf{F}(B(A)) + \varDelta \leq \widehat{\theta} + \varDelta,$$

whence $A' \in \mathcal{B}^\omega$. It follows that whenever $\eta[N] \in \Xi^N$, there exists $j \leq N$ such that

$$\|\eta^j\| \leq C_{2,2}s\sqrt{d} \quad\text{and}\quad \sum_{i=1}^d \eta^j_i A'_i \not\preceq 1.1I.$$

Since

$$\sum_{i=1}^d \eta^j_i A'_i \preceq \sum_{i=1}^d \eta^j_i A_i + \underbrace{\|\eta^j\|_1\|A-A'\|_\infty}_{\leq C_{2,2}sd\omega\leq 0.1} I \preceq \sum_{i=1}^d \eta^j_i A_i + 0.1I,$$

it follows that $\sum\limits_{i=1}^d \eta^j_i A_i \not\preceq I$, as claimed. □

We are ready to complete the proof of Theorem 11. Let Ξ^N be the set from Lemma 4. For $x \in X$, let

$$\begin{aligned} B_x &= \left\{u : 0.9\sum_{i=1}^d u_iA_i(x) \preceq A_0(x)\right\} = B(A_x),\\ A_x &= \left(A_0^{-1/2}(x)A_1(x)A_0^{-1/2}(x), ..., A_0^{-1/2}(x)A_d(x)A_0^{-1/2}(x)\right) \in \mathcal{A},\end{aligned}$$

where the concluding inclusion is given by Assumption **A.I**. We claim that

$$\forall\left(\eta[N]\in\Xi^N, x\in X(\eta[N])\right) : \mathbf{F}(B_x) \geq \widehat{\theta}. \tag{1.46}$$

Indeed, let $\eta[N] \in \Xi^N$ and $x \in X(\eta[N])$, so that $\eta^j \in B_x = B(A_x)$ for $j = 1, ..., N$. Assuming on contrary to (1.46), that $\mathbf{F}(B_x) < \widehat{\theta}$, or, which is the same due to $B_x = B(A_x)$, $\mathbf{F}(B(A_x)) < \widehat{\theta}$, we derive from (1.40) and the inclusion $\eta[N] \in \Xi^N$ that $\sum\limits_{i=1}^{d} \eta_i^j (A_x)_i \not\preceq I$ for certain $t \leq N$; but then $\eta^j \notin B_x$, which is a contradiction.

Now let $\eta[N] \in \Xi^N$ and $x \in X(\eta[N])$. Setting $Q_x = s^{-1}B_x$, by (1.46) we have

$$\widehat{\theta} \geq \mathbf{F}(B_x) \equiv \bar{\mathbf{F}}^{(s)}(B_x) \equiv \bar{\mathbf{F}}(\{\zeta : s\zeta \in B_x\}) = \bar{\mathbf{F}}(Q_x),$$

whence $\mathbf{P}(\{\xi \notin sQ_x \equiv B_x\}) \leq \mathrm{Err}(s) = \varepsilon$ by Theorem 5. Recalling definition of B_x, we conclude that

$$\eta[N] \in \Xi^N \Rightarrow X(\eta[N]) \subset X_\varepsilon.$$

Invoking (1.41), we see that with N as given by (1.37), the probability of generating a sample $\eta[N]$ with $X(\eta[N]) \not\subset X_\varepsilon$ is $\leq \delta$, provided that C is a properly chosen function of a, b, c and κ is a properly chosen absolute constant. □

2

Optimization Models with Probabilistic Constraints

Darinka Dentcheva

Stevens Institute of Technology, Department of Mathematical Sciences
Castle Point on Hudson, Hoboken, NJ 07030
ddentche@stevens-tech.edu

Summary. This chapter presents an overview of the theory and numerical techniques for optimization models involving one or more constraints on probability functions. We focus on recent developments involving nonlinear probabilistic models. The theoretical fundament includes the theory and examples of generalized concavity for functions and measures, and some specific properties of probability distributions, including discrete distributions. We analyze the structure and properties of the constraining probabilistic functions and of the probabilistically constrained sets. An important part of the analysis is the development of algebraic constraints equivalent to the probabilistic ones. Optimality and duality theory for such models is presented.

In the overview of numerical methods for solving probabilistic optimization problems the emphasis is put on recent numerical methods for nonlinear probabilistically constrained problems based on the optimality and duality theory presented here. The methods provide optimal solutions for convex problems. Otherwise, they solve certain relaxations of the problem and result in suboptimal solutions and upper and lower bounds for the optimal value. Special attention is paid to probabilistic constraints with discrete distributions.

Some numerical approaches via statistical approximations are discussed as well. Numerical techniques of bounding probability in higher dimensional spaces with satisfactory precision are mentioned briefly in the context of discrete distributions. Application of combinatorial techniques in this context is sketched.

2.1 Introduction

Deterministic optimization models are usually formulated as problems of minimizing or maximizing a certain objective functional $f(x)$ over x in a feasible set $\mathcal{D}$ described by a finite system of inequalities

$$g_j(x) \leq 0, \quad j \in J,$$

with some functionals g_j, $j \in J$.

When the objective functional or some of the constraint functionals depend not only on the decision vector x, but also on some random vector Z, the formulation of the optimization problem becomes unclear, and new precise definitions of the ‘objective’ and of the ‘feasible set’ are needed.

One way of dealing with that is to optimize the objective function and to require the satisfaction of the constraints *on average.* This leads to the following *stochastic optimization problem*:

$$\begin{aligned} &\min\ \mathbb{E}[f(x,Z)] \\ &\text{subject to } \mathbb{E}[g_j(x,Z)] \le 0, \quad j \in J. \end{aligned}$$

We have assumed for this formulation that the expected value functions are well defined. More importantly, it assumes that the average performance is representative for our decision problem. When some of the quantities $g_j(x,Z)$ have high variability a constraint on their expected value may not be satisfactory. When high uncertainty is involved another way to define the feasible set may be to impose constraints on probability functions, as in the following model:

$$\begin{aligned} &\min\ \mathbb{E}[f(x,Z)] \\ &\text{subject to } \mathbb{P}[g_j(x,Z) \le 0,\ j \in J] \ge p, \end{aligned} \tag{2.1}$$

where $p \in (0,1)$ is a modelling parameter expressing some fixed probability level. Constraints on probability are called *probabilistic* or *chance* constraints. The probability function can be formally understood as the expected value of the indicator function of the corresponding event. However, the discontinuity of the indicator function makes such problems qualitatively different than the expectation models.

In the following example probabilistic constraints arise in a natural way. Suppose we consider n investment opportunities, with random returns $R_1, \dots, R_n$ in the next year. We have certain initial capital K and our aim is to invest some of it in such a way that the expected value of our investment after a year is maximized, under the condition that the chance of losing no more than a given fixed amount $b > 0$ is at least p, where $p \in (0,1)$. Such a requirement is called the *Value at Risk* (VaR) constraint.

Let $x_1, \dots, x_n$ be the amounts invested in the n opportunities. Our investment changes in value after a year by $g(x,R) = \sum_{i=1}^{n} R_i x_i$. We can formulate the following stochastic optimization problem with probabilistic constraints:

$$
\begin{aligned}
&\max \sum_{i=1}^{n} \mathbb{E}[R_i] x_i \\
&\text{s.t. } \mathbb{P}\left[\sum_{i=1}^{n} R_i x_i \geq -b\right] \geq p \\
&\sum_{i=1}^{n} x_i \leq K \\
&x \geq 0.
\end{aligned}
\tag{2.2}
$$

The constraint

$$\mathbb{P}[g_j(x,Z) \leq 0,\ j \in J] \geq p$$

is called *joint probabilistic constraint*, while the constraints

$$\mathbb{P}[g_j(x,Z) \leq 0] \geq p_j,\ j \in J,\, p_j \in [0,1]$$

are called *individual probabilistic constraints*. Infinitely many individual probabilistic constraints appear naturally in the context of stochastic ordering constraints.

The notion of stochastic ordering or *stochastic dominance of first order* has been introduced in statistics in [203,212] and further applied and developed in economics [125,284]. It is defined as follows. For an integrable random variable X we consider its distribution function, $F_X(\eta) = \mathbb{P}[X \leq \eta]$, $\eta \in \mathbb{R}$. We say that a random variable X *dominates in the first order a random variable* Y if

$$F_X(\eta) \leq F_Y(\eta) \quad \text{for all} \quad \eta \in \mathbb{R}.$$

We denote this relation $X \succeq_{(1)} Y$. For two integrable random variables X and Y, we say that X *dominates* Y *in the second order* if

$$\int_{-\infty}^{\eta} F_X(\alpha)\, d\alpha \leq \int_{-\infty}^{\eta} F_Y(\alpha)\, d\alpha \quad \text{for all} \quad \eta \in \mathbb{R}.$$

We denote this relation $X \succeq_{(2)} Y$. The second order dominance has been introduced in [150]. A modern perspective on stochastic ordering is presented in [232, 237, 352].

Returning to our example, we can require that the net profit on our investment dominates certain benchmark outcome Y, which may be the return of our current portfolio or some acceptable index. Then the VaR constraint has to be satisfied at a continuum of points. Setting $\mathbb{P}\big[Y \leq \eta\big] = p_\eta$, model (2.2) becomes:

$$
\begin{aligned}
&\max \sum_{i=1}^{n} \mathbb{E}[R_i] x_i \\
&\text{s.t. } \mathbb{P}\big[\sum_{i=1}^{n} R_i x_i \le \eta\big] \le p_\eta \ \text{ for all } \eta \in \mathbb{R} \\
&\qquad \sum_{i=1}^{n} x_i \le K \\
&\qquad x \ge 0.
\end{aligned}
$$

Using the stochastic dominance notation we can formulate the model as follows:

$$
\begin{aligned}
&\max \sum_{i=1}^{n} \mathbb{E}[R_i] x_i \\
&\text{s.t. } \sum_{i=1}^{n} R_i x_i \succeq_{(1)} Y \\
&\qquad \sum_{i=1}^{n} x_i \le K \\
&\qquad x \ge 0.
\end{aligned}
$$

By changing the order of integration we can express the integrated distribution function as the expected shortfall: for each target value η we have

$$
F_X^{(2)}(\eta) = \int_{-\infty}^{\eta} F_X(\alpha)\, d\alpha = \mathbb{E}\big[(\eta - X)_+\big],
$$

where $(\eta - X)_+ = \max(\eta - X, 0)$. The integrated distribution function $F_X^{(2)}(\cdot)$ is continuous, convex, non-negative and nondecreasing. It is well defined for all random variables X with finite expected value. A second order dominance constraint can be formulated as follows:

$$
\sum_{i=1}^{n} R_i x_i \succeq_{(2)} Y \quad \Longleftrightarrow \quad \mathbb{E}\big[(\eta - \sum_{i=1}^{n} R_i x_i)_+\big] \le \mathbb{E}\big[(\eta - Y)_+\big] \ \text{ for all } \eta \in \mathbb{R}.
$$

We can formulate the above model replacing the first order dominance constraint with the following constraints:

$$
\mathbb{E}\big[(\eta - \sum_{i=1}^{n} R_i x_i)_+\big] \le \mathbb{E}\big[(\eta - Y)_+\big] \ \text{ for all } \eta \in \mathbb{R}.
$$

These second order dominance constraints can be viewed as a continuum of integrated chance constraints. In financial context it can be viewed as a continuum of *Conditional Value-at-Risk* (CVaR) constraints. For more information on this connection we refer to [110].

Models involving constraints on probability are introduced by Charnes *et al.* [81], Miller and Wagner [230], and Prékopa [275]. Problems with integrated chance constraints are considered in [152]. Models with stochastic dominance constraints are introduced and analyzed by Dentcheva and Ruszczyński in [107, 109, 111].

An essential contribution to the theory and solutions of problems with chance constraints was the theory of α-concave measures and functions. In [276, 277] the concept of logarithmic concave measures is introduced and studied. This notion was generalized to α-concave measures and functions in [51, 53, 63, 295], and further analyzed in [357], and [245]. Differentiability properties of probability functions are studied in [184, 185, 372, 373]. Statistical approximations of probabilistically constrained problems were analyzed by [178, 317]. For Monte Carlo approximations of chance constrained problems the reader is referred to [70, 71], see also Chapters 1 and 5 in this volume. Stability of models with probabilistic constraints is addresses in [103,155–157,303]. Nonlinear probabilistic problems are investigated in [104] where optimality conditions and duality results are established. Generalized concavity theory for discrete distributions and its consequences for probabilistic optimization is presented in [105, 106].

The formulation of the problem with probabilistic constraints is in harmony with the basic statistical principles used in testing statistical hypotheses and other statistical decisions. In engineering, reliability is frequently a central issue (*e.g.*, in telecommunication, transportation, hydrological network design and operation, engineering structure design, electronic manufacturing problems, *etc.*) and the problem with probabilistic constraints is very relevant. In finance, the concept of Value at Risk enjoys great popularity, [113, 300]. Integrated chance constraints represent a more general form of this concept. The concept of stochastic dominance is fundamental in economics and statistics (see [14, 107, 111, 132, 232]).

2.2 Structure and Properties of Probabilistically Constraint Sets

Fundamental questions to every optimization model concern the convexity of the feasible set, as well as continuity and differentiability of the constraint functions. The analysis of models with probability functions is based on specific properties of the underlying probability distributions. In particular, the *generalized concavity* theory plays a central role in probabilistic optimization. It facilitates the application of powerful tools of convex analysis.

2.2.1 Generalized Concavity of Functions and Measures

The generalized concavity discussed in this chapter is based on concavity of certain nonlinear transformation of the functions.

Definition 1. *A non-negative function $f(x)$ defined on a convex set $D \subset \mathbb{R}^s$ is said to be α-concave, where $\alpha \in [-\infty, +\infty]$, if for all $x, y \in D$ and all $\lambda \in [0,1]$ the following holds:*
If $\alpha = -\infty$ then

$$f(\lambda x + (1-\lambda)y) \geq \min(f(x), f(y));$$

If $\alpha = 0$ then

$$f(\lambda x + (1-\lambda)y) \geq f^{\lambda}(x) f^{1-\lambda}(y);$$

If $\alpha = \infty$ then

$$f(\lambda x + (1-\lambda)y) \geq \max(f(x), f(y));$$

For any other value of α

$$f(\lambda x + (1-\lambda)y) \geq [\lambda f^{\alpha}(x) + (1-\lambda) f^{\alpha}(y)]^{1/\alpha}.$$

Here we take the following conventions: $\ln 0 = -\infty$, $0(+\infty) = 0$, $0(-\infty) = 0$, $0^0 = 1$, $0^{-|\alpha|} = +\infty$, $\infty^{-|\alpha|} = 0$, $\infty^0 = 1$.

In the case of $\alpha = 0$ the function f is called *logarithmic concave*, and for $\alpha = 1$ it is simply *concave.*

If f is α-concave, then it is β-concave for all $\beta \leq \alpha$. Thus all α-concave functions are $(-\infty)$-concave, that is, quasi-concave.

Definition 2. *A probability measure $\mathbb{P}$ defined on the Borel subsets of a convex set $\Omega \subset \mathbb{R}^s$ is said to be α-concave if for any Borel measurable subsets A and B of Ω and for all λ we have the inequality*

$$\mathbb{P}(\lambda A + (1-\lambda)B) \geq \Big(\lambda [\mathbb{P}(A)]^{\alpha} + (1-\lambda)[\mathbb{P}(B)]^{\alpha}\Big)^{1/\alpha},$$

where $\lambda A + (1-\lambda)B = \{\lambda x + (1-\lambda)y : x \in A,\, y \in B\}$. All special cases of α and of one of the probabilities equal to 0 are treated as in Definition 1.

It is clear that if a random variable Z induces an α-concave probability measure on $\mathbb{R}^s$, then its distribution function $F_Z(x) = \mathbb{P}(Z \leq x)$ is an α-concave function.

As usual, concavity properties imply certain continuity. The following theorem is due to Borell [53].

Theorem 1. *If $\mathbb{P}$ is a quasi-concave measure on $\mathbb{R}^s$ and the dimension of its support is s, then $\mathbb{P}$ has a density (with respect to the Lebesgue measure).*

There is a relation between α-concavity properties of measures and their densities (see [63, 280, 295] and references therein).

Theorem 2. *Let Ω be an open convex subset of $\mathbb{R}^s$ and let m be the dimension of the smallest affine subspace L containing Ω. The probability measure $\mathbb{P}$ on Ω is γ-concave with $\gamma \in [-\infty, 1/m]$, if and only if its probability density function with respect to the Lebesgue measure on L is α-concave with*

$$\alpha = \begin{cases} \gamma/(1-m\gamma) & \textit{if } \gamma < 1/m, \\ +\infty & \textit{if } \gamma = 1/m. \end{cases}$$

Corollary 1. *Let an integrable non-negative function $f(x)$ be defined on a non-degenerated convex set $\Omega \subset \mathbb{R}^s$. If $f(x)$ is α-concave with $-1/s \leq \alpha \leq \infty$ and positive on the interior of Ω, then the measure $\mathbb{P}$ on Ω defined as*

$$\mathbb{P}(A) = \int_A f(x)\,dx, \quad A \subset \Omega,$$

is γ-concave with

$$\gamma = \begin{cases} \alpha/(1+s\alpha) & \textit{if } \alpha \neq -1/s, \\ -\infty & \textit{if } \alpha = -1/s. \end{cases}$$

The corollary states in particular that if a measure $\mathbb{P}$ on $\mathbb{R}^s$ has a density function $f(x)$ such that $f^{-1/s}$ is convex, then $\mathbb{P}$ is quasi-concave.

For the following two results we refer the reader to [281].

Theorem 3. *If the s-dimensional random vector Z has a log-concave probability distribution and A is a constant $m \times s$ matrix, then the m-dimensional random vector $Y = AZ$ has a log-concave probability distribution.*

Lemma 1. *If $\mathbb{P}$ is an α-concave probability distribution and $A \subset \mathbb{R}^s$ is a convex set, then the function $f(x) = \mathbb{P}(A + x)$ is α-concave.*

We extend the definition of generalized concavity to make it applicable to the case of discrete distributions. The first definition of discrete multivariate distributions is introduced in [29]. We adopt here the definition of [105] because it is more suitable to probabilistic optimization and it has essential consequences for optimality and duality theory of probabilistic optimization as it will become clear in Section 2.4.

Definition 3. *A distribution function F is called α-concave on the set $\mathcal{A} \subseteq \mathbb{R}^s$ with $\alpha \in [-\infty, \infty]$, if*

$$F(z) \geq \big(\lambda F(x)^\alpha + (1-\lambda)F(y)^\alpha\big)^{1/\alpha}$$

for all $z, x, y \in \mathcal{A}$ and $\lambda \in (0,1)$ such that $z \geq \lambda x + (1-\lambda)y$. The special cases $\alpha = 0$, and $\alpha = \pm\infty$ are treated the same way as in Definition 1.

Observe that if $\mathcal{A} = \mathbb{R}^s$ this definition coincides with the usual definition of α-concavity of a distribution function.

To illustrate the relation between Definition 1 and Definition 3 let us consider the case of integer random vectors which are roundups of continuously distributed random vectors. We denote the set of s-dimensional vectors with integer components by $\mathbb{Z}^s$.

Remark 1. If the distribution function of a random vector Z is α-concave on $\mathbb{R}^s$ then the distribution function of $Y = \lceil Z \rceil$ is α-concave on $\mathbb{Z}^s$.

The last property follows from the observation that at integer points both distribution functions coincide.

Furthermore for random vectors with independent components, we can relate the concavity of their marginal distributions to the concavity of the joint distribution.

Theorem 4. *Assume that $Z = (Z^1, \cdots, Z^L)$, where the s_l-dimensional subvectors Z_l, $l = 1, \cdots, L$, are independent $(\sum_{l=1}^{L} s_l = s)$. Furthermore, let the marginal distribution functions $F_l : \mathbb{R}^{s_l} \to [0,1]$ be α_l-concave on sets $\mathcal{A}_l \subset \mathbb{Z}^{s_l}$.*

1. *If $\alpha_l > 0$, $l = 1, \cdots, L$, then F_Z is α-concave on $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_L$ with $\alpha = (\sum_{l=1}^{L} \alpha_l^{-1})^{-1}$;*
2. *If $\alpha_l = 0$, $l = 1, \cdots, L$, then F_Z is log-concave on $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_L$.*

For an integer random variable, our definition of α-concavity is related to log-concavity of sequences.

Definition 4. *A sequence p_k, $k \in \mathbb{Z}$, is called log-concave, if*

$$p_k^2 \geq p_{k-1} p_{k+1} \quad \text{for all} \quad k \in \mathbb{Z}.$$

We have the following property (see [280, Theorem 4.7.2]):

Theorem 5. *Suppose that for an integer random variable Y the probabilities $p_k = \mathbb{P}\{Y = k\}$, $k \in \mathbb{Z}$ form a log-concave sequence. Then the distribution function of Y is α-concave on $\mathbb{Z}$ for every $\alpha \in [-\infty, 0]$.*

Another important property of α-concave measures is the existence of a so-called floating body for all probability levels $p \in [1/2, 1]$. Let us recall that the support function of a convex set $C \subset \mathbb{R}^s$ is defined as follows:

$$\sigma_C(h) = \sup\{\langle h, x \rangle : x \in C\}.$$

Definition 5. *A measure $\mathbb{P}$ on $\mathbb{R}^s$ has a floating body for a level $p > 0$ if there exists a convex set $C_p \subset \mathbb{R}^s$ such that, for all vectors $h \in \mathbb{R}^s$,*

$$\mathbb{P}\big(\{x \in \mathbb{R}^s : \langle h, x \rangle \geq \sigma_{C_p}(h)\}\big) = p.$$

The set C_p is called the floating body of $\mathbb{P}$ at level p.

All log-concave measures have a floating body, [225].

Theorem 6. *Any log-concave probability measure has a floating body C_p for all levels $p \in [1/2, 1]$.*

2.2.2 Examples of α-Concave Measures

1. The density of *the non-degenerate multivariate normal distribution* on $\mathbb{R}^s$:

$$f(x) = \frac{1}{\sqrt{(2\pi)^s \det \Sigma}} \exp\Big(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \Big),$$

where Σ is a positive definite matrix of dimension $s \times s$ and $\mu \in \mathbb{R}^s$. Since the function $\ln f(x)$ is concave (that is, f is 0-concave), the normal distribution is a log-concave measure.

2. The *uniform distribution on a convex set* $D \subset \mathbb{R}^s$ with density

$$f(x) = \begin{cases} 1/V(D) & x \in D, \\ 0 & x \notin D, \end{cases}$$

where $V(D)$ is the Lebesgue measure of D. The function $f(x)$ is quasi-concave on D, hence it generates a $1/s$-concave measure on D.

3. The density function of the *multivariate Dirichlet's distribution* is defined as

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha_1 + \cdots + \alpha_s)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_s)} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_s^{\alpha_s} & \text{if } x_i \geq 0,\ \sum_i x_i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here $\Gamma(\cdot)$ stands for the Gamma function. We define the open simplex

$$S = \left\{ x \in \mathbb{R}^s : \sum_{i=1}^{s} x_i = 1,\ x_i > 0,\ i = 1, \ldots, s \right\}.$$

The function $f(x)$ is $(\alpha_1 + \cdots + \alpha_s)^{-1}$-concave on S, and therefore, the resulting measure is β-concave with $\beta = (\alpha_1 + \cdots + \alpha_s + s - 1)^{-1}$ on the closed simplex $\overline{S}$.

4. The density function of the *m-dimensional Student's distribution* with parameter n

$$f(x) = \frac{\Gamma(\frac{m+n}{2})\sqrt{\det A}}{\Gamma(\frac{n}{2})\sqrt{(2\pi)^m}} \Big(1 + \frac{1}{n}(x-\mu)^T A (x-\mu) \Big)^{-(m+n)/2},$$

where A is a positive definite matrix. Since f is $(-\frac{2}{m+n})$-concave, the corresponding measure is $(-\frac{2}{n-m})$-concave.

5. The density function of the *m-dimensional F-distribution* with parameters $n_0, \ldots, n_m$, and $n = \sum_{i=1}^m n_i$ is defined as follows:

$$f(x) = \text{const} \prod_{i=1}^{m} x_i^{n_i/2-1} \Big(n_0 + \sum_{i=1}^{m} n_i x_i \Big)^{-n/2}, \quad x_i \ge 0,\ i = 1, \dots, m.$$

It is $[-(n_0/2+m)^{-1}]$-concave and the corresponding measure is $(-\frac{2}{n})$-concave.

6. The probability density function of the *Wishart distribution* is defined by

$$f(X) = \begin{cases} \dfrac{|X|^{\frac{N-q-2}{2}} e^{-\frac{1}{2} \operatorname{tr} A^{-1} X}}{2^{\frac{N-1}{2} q} \ \pi^{\frac{q(q-1)}{4}} \ |A|^{\frac{N-1}{2}} \prod\limits_{i=1}^{q} \Gamma\left(\frac{N-i}{2}\right)} & \text{for } X \succ 0 \\ 0 & \text{otherwise.} \end{cases}$$

Here X is assumed to be $q \times q$ matrix containing the variables and A is fixed positive definite $q \times q$ matrix. The symbol $\succ$ denotes the partial order on the positive definite cone.

We assume that there are $s = \frac{1}{2} q(q+1)$ independent variables and that $N \ge q+2$. The function f is log-concave.

7. The probability density function of the *beta distribution* is defined by

$$f(X) = \begin{cases} \dfrac{c(s_1, q) c(s_2, q)}{c(s_1 + s_2, q)} \, |X|^{\frac{1}{2}(s_1 - q - 1)} \, |I - X|^{\frac{1}{2}(s_2 - q - 1)} & \text{for } I \succ X \succ 0 \\ 0 & \text{otherwise.} \end{cases}$$

Here I stands for the identity matrix and the function $c(\cdot, \cdot)$ is defined as follows:

$$\frac{1}{c(k, q)} = 2^{qk/2} \pi^{q(q-1)/2} \prod_{i=1}^{q} \Gamma\left(\frac{k-i+1}{2} \right).$$

We have assumed that $s_1 \ge q+1$ and $s_2 \ge q+1$. The number of independent variables in X is $s = \frac{1}{2} q(q+1)$.

8. The *Cauchy distribution* regarded as a joint distribution of the random variables

$$Y_i = \frac{\sqrt{\nu} Z_i}{U} \quad i = 1, \dots, s,$$

where the random variables $Z_1, \dots, Z_s$ have the standard normal distribution, each of them is independent of U, and U has the χ-distribution with ν degrees of freedom. The probability density function is

$$f(x) = \frac{\Gamma\left(\frac{1}{2}(\nu + s)\right)}{(\pi\nu)^{\frac{s}{2}} \Gamma\left(\frac{1}{2}\nu\right) |R|^{\frac{1}{2}}} \left(1 + \frac{1}{\nu} x^T R^{-1} x \right)^{-\frac{1}{2}(\nu+s)}$$

for $x \in \mathbb{R}^s$. If $s = 1$ and $\nu = 1$ this reduces to the well-known univariate Cauchy density

$$f(x) = \frac{1}{\pi} \; \frac{1}{1+x^2} \; , \; -\infty < x < \infty.$$

The s-variate Cauchy density has the property that $f^{-\frac{1}{s}}$ is convex in $\mathbb{R}^s$ and thus the distribution is quasi-concave by virtue of Corollary 1.

9. The probability density function of *the Pareto distribution* is

$$f(x) = a(a+1)\dots(a+s-1)\left(\prod_{j=1}^{s} \Theta_j\right)^{-1} \left(\sum_{j=1}^{s} \Theta_j^{-1} x_j - s + 1\right)^{-(a+s)}$$

for $x_i > \Theta_i, \; i = 1, \dots, s$, and $f(x) = 0$ otherwise. Here $\Theta_i, \; i = 1, \dots, s$ are positive constants. Since $f^{-\frac{1}{s}}$ is convex in $\mathbb{R}^s$, Corollary 1 implies that the Pareto distribution is quasi-concave.

10. A univariate *gamma distribution* is given by a probability density of the form

$$f(z) = \begin{cases} \dfrac{\lambda^{\vartheta} z^{\vartheta-1} e^{-\lambda z}}{\Gamma(\vartheta)} & \text{for } z > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Here $\lambda > 0$ and $\vartheta > 0$ are constants. For $\lambda = 1$ the distribution is called the standard gamma distribution. If a random variable Y has the gamma distribution, then λY has the standard gamma distribution.

An s-variate gamma distribution can be defined by a certain linear transformation of s independent random variables $Z_1, \dots, Z_s$ that have the standard gamma distribution. Given an $s \times m$ matrix A $(1 \le m \le 2^s - 1)$ with 0-1 elements such that no column is the zero vector, setting $Z = (Z_1, \dots, Z_s)$, we define

$$Y = AZ.$$

The random vector Y has an s-variate standard gamma distribution. The univariate gamma density function is obviously log-concave. Thus, the *s-variate standard gamma distribution* is log-concave by virtue of Theorem 3.

11. Every distribution function of an s-dimensional *binary random vector* is α-concave on $\mathbb{Z}^s$ for all $\alpha \in [-\infty, \infty]$.
Indeed, let x, y be binary vectors, $\lambda \in (0,1)$ and let $z \ge \lambda x + (1-\lambda)y$. Since z is integer and x and y binary, then $z \ge x$ and $z \ge y$. Thus $F(z) \ge \max(F(x), F(y)) = \max(F(x), F(y))$. Consequently, F is ∞-concave.

12. *The binomial, the Poisson, the geometric, and the hypergeometric* one-dimensional probability distributions satisfy the conditions of Theorem 5 (see [280, p. 109]), and are, therefore, log-concave.

2.2.3 Convexity of Probabilistically Constrained Sets

Let us recall that a function g is called quasi-convex, if $-g$ is quasi-concave in the sense of Definition 1. One of the most general results in the convexity theory of probabilistic optimization is the following:

Theorem 7. *Let $g_j(\cdot,\cdot)$, $j \in J$ be quasi-concave functions of the variables $x \in \mathbb{R}^n$ and $z \in \mathbb{R}^s$. If $Z \in \mathbb{R}^s$ is a random variable that has α-concave probability distribution, then the function*

$$G(x) = \mathbb{P}[g_j(x,Z) \geq 0,\, j \in J]$$

is α-concave on the set

$$D = \{x \in \mathbb{R}^n : \exists z \in \mathbb{R}^s \text{ such that } g_j(x,z) \geq 0,\, j \in J\}.$$

As a consequence, under the assumptions of Theorem 7, we obtain convexity statements for sets described by probabilistic constraints.

Corollary 2. *Assume that $g_j(\cdot,\cdot)$, $j \in J$ are quasi-concave functions jointly in both arguments, and that $Z \in \mathbb{R}^s$ is a random variable that has an α-concave probability distribution. The the following set is convex and closed:*

$$\mathcal{X}_0 = \big\{x \in \mathbb{R}^n : \mathbb{P}[g_i(x,Z) \geq 0,\ i = 1,\dots,m] \geq p\big\}.$$

Observe that the closure of the set follows from the continuity of all α-concave functions.

Theorem 8. *Given random variables $Y_i \in \mathbb{R}^s$, assume that $g_j(\cdot,\cdot)$, $j \in J$ are quasi-concave functions jointly in both arguments, and that Z_i, $i = 1,\dots,m$ have α_i-concave distributions. Then the set with first order stochastic dominance constraint is convex and closed:*

$$\mathcal{X}_d = \big\{x \in \mathbb{R}^n : g_i(x,Z_i) \succeq_{(1)} Y_i,\ i = 1,\dots,m\big\}.$$

Proof. Let us fix i and $\eta \in \mathbb{R}$ and consider the function

$$\mathbb{P}\big[g_i(x,Z_i) - \eta \leq 0\big] = 1 - \mathbb{P}\big[g_i(x,Z_i) - \eta > 0\big].$$

Constraint $g_i(x,Z_i) \succeq_{(1)} Y_i$ can be formulated as follows:

$$\mathbb{P}\big[g_i(x,Z_i) - \eta > 0\big] \geq 1 - \mathbb{P}\big[Y_i \leq \eta\big] \quad \text{for all} \quad \eta \in \mathbb{R}.$$

Denote the set of x satisfying this inequality by $X_i(\eta)$. By Theorem 7 the function at the left hand side of the last inequality is quasi-concave. Thus the set $X_i(\eta)$ is convex and closed by Corollary 2. The set $\mathcal{X}_d$ is the intersection of the sets $X_i(\eta)$ for $i = 1,\dots,m$ and all $\eta \in \mathbb{R}$, and, therefore, it is convex and closed. □

There is an intriguing relation between the sets constrained by first and second order dominance relation to a benchmark random variable (see [108]). We denote the space of integrable random variables by $\mathcal{L}_1(\Omega, \mathcal{F}, \mathbb{P})$ and set

$$A_{(1)}(Y) = \{X \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathbb{P}) : X \succeq_{(1)} Y\},$$
$$A_{(2)}(Y) = \{X \in \mathcal{L}_1(\Omega, \mathcal{F}, \mathbb{P}) : X \succeq_{(2)} Y\}.$$

It is proved in [107] that the set $A_{(2)}(Y)$ is convex and closed in $\mathcal{L}_1(\Omega, \mathcal{F}, \mathbb{P})$. The set $A_{(1)}(Y)$ is closed, because convergence in $\mathcal{L}_1$ implies convergence in probability. It is not convex in general.

Theorem 9. *Assume that Y has a continuous probability distribution function. Then*

$$A_{(2)}(Y) = \overline{\operatorname{co}}\, A_{(1)}(Y),$$

where $\overline{\operatorname{co}}\, A_{(1)}(Y)$ stands for the closed convex hull of $A_{(1)}(Y)$.

If the underlying probability space is discrete and such that $\Omega = \{1, \ldots, N\}$, $\mathcal{F}$ is the set of all subsets of Ω and $\mathbb{P}[k] = 1/N$, $k = 1, \ldots, N$, we can remove the closure:

$$A_{(2)}(Y) = \operatorname{co} A_{(1)}(Y),$$

Let us consider the special case

$$g_i(x, Z) := \langle a_i(Z), x\rangle + b_i(Z).$$

These functions are not necessarily quasi-concave in both arguments. If $a_i(Z) = a_i$, $i = 1, \ldots, m$ we can apply Theorem 7 to conclude that the set $\mathcal{X}_0$ is convex.

Corollary 3. *The following set is convex:*

$$\mathcal{X}_l = \big\{x \in \mathbb{R}^n : \mathbb{P}[\langle a_i, x\rangle \le b_i(Z),\ i = 1, \ldots, m] \ge p\big\}$$

whenever $b_i(\cdot)$ are quasi-concave functions and Z has a quasi-concave probability distribution.

If the functions g_i are not separable, we can invoke Theorem 6 (see also [202]).

Corollary 4. *The following set is convex:*

$$\mathcal{X}_1 = \big\{x \in \mathbb{R}^n : \mathbb{P}[\langle a_i(Z), x\rangle \le b_i)] \ge p_i,\, i = 1, \ldots, m\big\} \tag{2.3}$$

whenever the vectors $a_i(Z)$ have a log-concave probability distribution.

In particular, we obtain that the set $\mathcal{X}_1$ is convex if $a_i(Z)$ have one of the multivariate distributions from Section 2.2.2, *e.g.*, the uniform, the normal, the Gamma distribution, *etc.*

2.2.4 Connectedness of Probabilistically Constrained Sets

It will be demonstrated later (Lemma 4) that the probabilistically constrained set $\mathcal{X}$ is union of cones and, thus, $\mathcal{X}$ could be disconnected. The following result provides a sufficient condition for $\mathcal{X}$ to be topologically connected.

Theorem 10. *Assume that the functions $g_i(\cdot, Z), i = 1, \ldots, m$ are quasi-concave and that the following condition is satisfied: for all $x^1, x^2 \in \mathbb{R}^n$ there exists a point $x^* \in \mathbb{R}^n$ such that*

$$g_i(x^*, z) \geq \min\{g_i(x^1, z), g_i(x^2, z)\} \quad \textit{for all } z \in \mathbb{R}^s, \ \textit{for all } i = 1, \ldots, m.$$

Then the set $\mathcal{X}_0$ is connected.

Proof. Let $x^1, x^2 \in \mathcal{X}_0$ be arbitrary given points. We construct a path joining the two pints, which is contained entirely in $\mathcal{X}_0$. Let x^* be the point that exists according to the assumption. We set

$$\pi(t) = \begin{cases} (1-2t)x^1 + 2tx^* & \text{for } 0 \leq t \leq 1/2 \\ 2(1-t)x^* + (2t-1)x^2 & \text{for } 1/2 < t \leq 1 \end{cases}$$

We observe that quasi-concavity of g_i, $i = 1, \ldots, m$ and the assumptions of the theorem imply for $0 \leq t \leq 1/2$ and for every i the following inequality:

$$g_i((1-2t)x^1 + 2tx^*, z) \geq \min\{g_i(x^1, z), g_i(x^*, z)\} = g_i(x^1, z).$$

Therefore,

$$\mathbb{P}[g(\pi(t)) \geq 0] \geq \mathbb{P}[g(x^1) \geq 0] \geq p \quad \text{for} \quad 0 \leq t \leq 1/2.$$

Similar argument applies for $1/2 < t \leq 1$. Consequently, $\pi(t) \in \mathcal{X}_0$, and this proves the assertion. □

In [155] a slightly more general version of this result is proved in order to deal with probabilistic constraints involving stochastic processes.

2.3 Random Right Hand Side

We pay special attention to problems with separable constraint functions. Consider the following problem:

$$\begin{aligned} \max \ & f(x) \\ \text{subject to } & \mathbb{P}\big[g(x) \geq Z\big] \geq p, \\ & x \in \mathcal{D}. \end{aligned} \tag{2.4}$$

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, are concave functions. Let $\mathcal{D} \subseteq \mathbb{R}^n$ be a closed convex set, and Z be an m-dimensional random vector. We denote $g = (g_1, \ldots, g_m)$. For two vectors a and b the inequality $a \leq b$ is understood componentwise.

2.3.1 Continuity and Differentiability Properties of Distribution Functions

When the probabilistic constraint involves inequalities with random variables on the right hand site only, we can express it as a constraint on a distribution function:

$$\mathbb{P}\big[g(x) \geq Z\big] \geq p \quad \Longleftrightarrow \quad F_Z(g(x)) \geq p$$

Thus, continuity and differentiability properties of distribution functions are relevant to the numerical solution of probabilistic optimization problems.

Theorem 11. *Suppose that Z has an α-concave distribution with $\alpha \in (-\infty, 0)$ and that the support of it* $\mathrm{supp} P_Z$ *has non-empty interior in $\mathbb{R}^s$. Then F_Z is locally Lipschitz-continuous on* $\mathrm{int}\ \mathrm{supp} P_Z$.

Proof. From the assumption that P_Z is α-concave for $\alpha < 0$, we infer that the function $F_Z^\alpha(\cdot)$ is convex. The assertion follows from the fact that convex functions are locally Lipschitz on the interior of their domain and from the Lipschitz continuity of the mapping $t \longmapsto t^{1/\alpha}$ away from 0. Fixing a point $z \in \mathrm{int}\ \mathrm{supp} P_Z$, there is a neigbourhood U of z contained in the interior of the support and such that F_Z^α is locally Lipschitz with Lipschitz constant L_1. Decreasing U if necessary, we can find a compact set K such that $U \subset K \subset \mathrm{int}\ \mathrm{supp} P_Z$. Thus, $\min_{z \in K} F_Z(z) = r > 0$. Let L_2 be the Lipschitz constant of $t \mapsto t^{1/\alpha}$ on the interval $[r, 1]$. We obtain

$$|F_Z(z_1) - F_Z(z_1)| \leq L_2 |F_Z^\alpha(z_1) - F_Z^\alpha(z_1)| \leq L_2 L_1 \|z_1 - z_2\|.$$

□

Theorem 12. *Suppose that all one-dimensional marginal distribution functions of an s-dimensional random vector Z are locally Lipschitz continuous. Then F_Z is locally Lipschitz-continuous as well.*

Proof. The statement can be proved by straightforward evaluation of the distribution function by marginals for $s = 2$ and mathematical induction on the dimension of the space. □

Assume that the measure P_Z has a density. It should be emphasized that the continuity and the essential boundedness of the density do not imply the Lipschitz continuity of the distribution function F_Z.

Theorem 13. *Assume that P_Z has a continuous density $\theta(\cdot)$ and that all one-dimensional marginal densities are continuous as well. Then F_Z is continuously differentiable.*

Proof. We demonstrate the statement for $s = 2$. The assertions then follows by induction. The existence of the partial derivatives follows from the continuity of the density θ by virtue of the theorem of Fubini:

$$\frac{\partial F_x}{\partial z_1}(z_1,z_2)=\int_{-\infty}^{z_2}\theta(z_1,t)dt \quad \text{and} \quad \frac{\partial F_x}{\partial z_2}(z_1,z_2)=\int_{-\infty}^{z_1}\theta(t,z_2)dt.$$

First, we observe that the mapping $(x_1,x_2)\mapsto \int\limits_a^{x_2}\theta(x_1,t)dt$ is continuous for every $a\in\mathbb{R}$ by the uniform continuity of $\theta(\cdot)$ on compact sets in $\mathbb{R}^2$. Given the points (x_1,x_2) and (y_1,y_2), we have:

$$\left|\frac{\partial F}{\partial x_1}\ (x_1,x_2)-\frac{\partial F}{\partial y_1}(y_1,y_2)\right|=\left|\int\limits_{-\infty}^{x_2}\theta(x_1,t)dt-\int\limits_{-\infty}^{y_2}\theta(y_1,t)dt\right|$$
$$\leq\left|\int\limits_{x_2}^{y_2}\theta(y_1,t)dt\right|+\left|\int\limits_{-\infty}^{x_2}[\theta(x_1,t)-\theta(y_1,t)]dt\right|\leq\varepsilon.$$

The last inequality is satisfied if the points (x_1,x_2) and (y_1,y_2) are sufficiently close by the continuity of $\theta(\cdot)$ and the uniform continuity of the function $(x_1,x_2)\mapsto\int\limits_a^{x_2}\theta(x_1,t)dt$.

The limit exists uniformly around x_1 because of the continuity of the one-dimensional marginal densities. □

2.3.2 *p*-Efficient Points

We concentrate on deriving an equivalent algebraic description of the feasible set. The level set of the distribution function of Z can be described as follows:

$$\mathcal{Z}=\big\{z\in\mathbb{R}^m : \mathbb{P}[Z\leq z]\geq p\big\}. \tag{2.5}$$

Clearly, problem (2.4) can be compactly rewritten as

$$\begin{aligned}&\max f(x)\\ &\text{subject to } g(x)\in\mathcal{Z},\\ &x\in\mathcal{D}.\end{aligned} \tag{2.6}$$

Lemma 2. *For every $p\in(0,1)$ the level set $\mathcal{Z}$ is non-empty and closed.*

Proof. The assertion follows from the monotonicity and the right continuity of the distribution function. □

Till the end of this section we denote the probability distribution function of Z by F omitting the subscript. The marginal probability distribution function of the ith component Z_i will be denoted by F_i.
We recall the concept of a p-efficient point.

Definition 6. *Let $p\in(0,1]$. A point $v\in\mathbb{R}^m$ is called a p-efficient point of the probability distribution function F, if $F(v)\geq p$ and there is no $z\leq v$, $z\neq v$ such that $F(z)\geq p$.*

The p-efficient points are minimal points of the level set $\mathcal{Z}$ with respect to the partial order in $\mathbb{R}^m$ generated by the non-negative cone. This notion was first introduced in [278]. Similar concept is used in [326]. The concept was studied and applied in the context of discrete distributions and linear problems in the papers [105, 106, 282] and in the context of general distributions in [104].

Obviously, for a scalar random variable Z and for every $p \in (0,1]$ there is exactly one p-efficient point: the smallest v such that $F(v) \geq p$. Since $F(v) \leq F_i(v_i)$ for every $v \in \mathbb{R}^m$ and $i = 1, \dots, m$, we obtain that the set of p-efficient points is bounded from below.

Lemma 3. *Let $p \in (0,1]$ and let l_i be the p-efficient point of the one-dimensional marginal distribution F_i, $i = 1, \dots, m$. Then every $v \in \mathbb{R}^m$ such that $F(v) \geq p$ must satisfy the inequality $v \geq l = (l_1, \dots, l_m)$.*

Let $p \in (0,1)$ and let v^j, $j \in J$, be *all* p-efficient points of Z, where J is an arbitrary set. Denoting the positive orthant in $\mathbb{R}^m$ by $\mathbb{R}^m_+$, we define the cones

$$K_j = v^j + \mathbb{R}^m_+, \quad j \in J.$$

The following result can be derived from Phelps theorem [267, Lemma 3.12] about the existence of conical support points, but an easy direct proof is provided.

Theorem 14. $\mathcal{Z} = \bigcup_{j \in J} K_j$.

Proof. If $y \in \mathcal{Z}$ then either y is p-efficient or there exists a vector w such that $w \leq y$, $w \neq y$, $w \in \mathcal{Z}$. By Lemma 3, one must have $l \leq w \leq y$. The set $Z_1 := \{z \in \mathcal{Z} : l \leq z \leq y\}$ is compact because the set $\mathcal{Z}$ is closed. Thus, there exists $w^1 \in Z_1$ with the minimal first coordinate. If w^1 is a p-efficient point, then $y \in w^1 + \mathbb{R}^m_+$, what had to be shown. Otherwise, we define $Z_2 := \{z \in \mathcal{Z} : l \leq z \leq w^1\}$, and choose a point $w^2 \in Z_2$ with the minimal second coordinate. Proceeding in the same way, we shall find the minimal element w^m in the set $\mathcal{Z}$ with $w^m \leq w^{m-1} \leq \cdots \leq y$. Therefore, $y \in w^m + \mathbb{R}^m_+$, and this completes the proof. □

By virtue of Theorem 14 we obtain (for $0 < p < 1$) the following *disjunctive semi-infinite* formulation of problem (2.6):

$$\begin{aligned} &\max f(x) \\ &\text{subject to } g(x) \in \bigcup_{j \in J} K_j, \\ &x \in \mathcal{D}. \end{aligned} \tag{2.7}$$

Its main advantage is an insight into the nature of the non-convexity of the feasible set. The main difficulty is the implicit character of the disjunctive constraint.

Let S stand for the simplex in $\mathbb{R}^{m+1}$, $S = \{\alpha \in \mathbb{R}^{m+1} : \sum_{i=1}^{m+1} \alpha_i = 1, \alpha_i \geq 0\}$. We define the convex hull of the p-efficient points:

$$E = \Big\{ \sum_{i=1}^{m+1} \alpha_i v^{j_i} : \alpha \in S,\ j_i \in J \Big\}.$$

The convex hull of $\mathcal{Z}$ has a semi-infinite disjunctive representation as well.

Lemma 4. $\operatorname{co}\mathcal{Z} = E + \mathbb{R}^m_+$.

Proof. By Theorem 14 every point $y \in \operatorname{co}\mathcal{Z}$ can be represented as a convex combination of points in the cones K_j. By the theorem of Caratheodory we can write $y = \sum_{i=1}^{m+1} \alpha_i (v^{j_i} + w^i)$, where $w^i \in \mathbb{R}^m_+$, $\alpha \in S$ and $j_i \in J$. The vector $w = \sum_{i=1}^{m+1} \alpha_i w^i$ belongs to $\mathbb{R}^m_+$. Therefore, $y \in \sum_{i=1}^{m+1} \alpha_i v^{j_i} + \mathbb{R}^m_+$. □

Theorem 15. *For every $p \in (0,1)$ the set $\operatorname{co}\mathcal{Z}$ is closed.*

Proof. Consider a sequence $\{z^k\}$ of points of $\operatorname{co}\mathcal{Z}$ which is convergent to a point $\bar{z}$. We have

$$z^k = \sum_{i=1}^{m+1} \alpha_i^k y_i^k,$$

with $y_i^k \in \mathcal{Z}$, $\alpha_i^k \geq 0$, and $\sum_{i=1}^{m+1} \alpha_i^k = 1$. By passing to a subsequence, if necessary, we can assume that the limits

$$\bar{\alpha}_i = \lim_{k\to\infty} \alpha_i^k$$

exist for all $i = 1, \ldots, m+1$. By Lemma 3 all points y_i^k are bounded below by some vector l. For simplicity of notation we may assume that $l = 0$.

Let $I = \{i : \bar{\alpha}_i > 0\}$. Clearly, $\sum_{i\in I} \bar{\alpha}_i = 1$. We obtain

$$z^k \geq \sum_{i\in I} \alpha_i^k y_i^k.$$

We observe that $0 \leq \alpha_i^k y_i^k \leq z^k$ for all $i \in I$ and all k. Since $\{z^k\}$ is convergent and $\alpha_i^k \to \bar{\alpha}_i > 0$, each sequence $\{y_i^k\}$, $i \in I$, is bounded. Therefore we can assume that each of them is convergent to some limit $\bar{y}_i$, $i \in I$. By virtue of Lemma 2 $\bar{y}_i \in \mathcal{Z}$. Passing to the limit in the last displayed inequality we obtain

$$\bar{z} \geq \sum_{i\in I} \bar{\alpha}_i \bar{y}_i \in \operatorname{co}\mathcal{Z}.$$

Due to Lemma 4, $\bar{z} \in \operatorname{co}\mathcal{Z}$. □

Theorem 16. *For every $p \in (0,1)$ the set of extreme points of $\operatorname{co}\mathcal{Z}$ is nonempty and it is contained in the set of p-efficient points.*

Proof. The set $\operatorname{co}\mathcal{Z}$ is included in $l + \mathbb{R}^m_+$, by virtue of Lemma 3 and Lemma 4. Therefore, it does not contain any line. Since it is closed by Theorem 15, it has at least one extreme point.

Let w be an extreme point of $\operatorname{co} \mathcal{Z}$. Suppose that w is not a p-efficient point. Then Theorem 14 implies that there exists a p-efficient point $v \le w$, $v \neq w$. Since $w + \mathbb{R}^m_+ \subset \operatorname{co} \mathcal{Z}$, the point w is a convex combination of v and $w + (w - v)$. Consequently, w cannot be extreme. □

For a general random vector the set of p-efficient points may be unbounded and not closed.

The representation becomes very handy for problem (2.26) when the vector Z has a discrete distribution on $\mathbb{Z}^s$. In [105] discrete distributions are investigated, where the random vector Z takes values on a grid. Without loss of generality we can assume that $Z \in \mathbb{Z}^s$. Figure 2.1 illustrates the structure of the probabilistically constrained set.

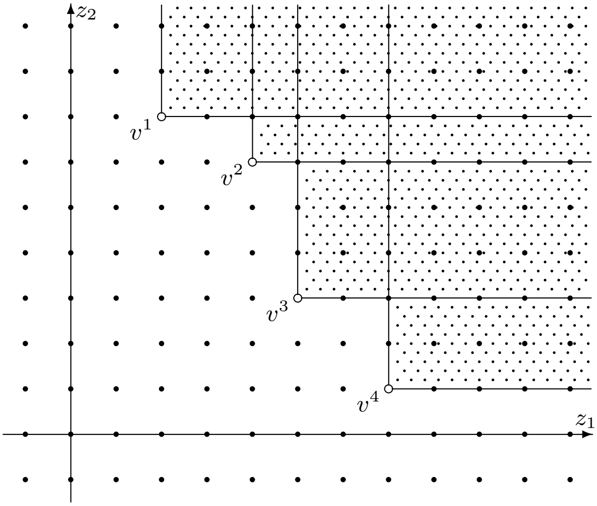


Figure 2.1. Example of the set $\mathcal{Z}$ with p-efficient points $v^1, \dots, v^4$

Theorem 17. *For each $p \in (0, 1)$ the set of p-efficient points of an integer random vector is non-empty and finite.*

Proof. The result follows from Dickson's Lemma [32, Corollary 4.48] and Lemma 3. For convenience we provide a short proof here. We shall at first show that at least one p-efficient point exists. Since $p < 1$, there must exist y such that $F(y) \ge p$. By Lemma 3, all v such that $F(v) \ge p$ are bounded below by the vector l of p-efficient points of one-dimensional marginals. Therefore, if y is not p-efficient, one of finitely many integer points v such that $l \le v \le y$ must be p-efficient.

Now we prove the finiteness of the set of p-efficient points. Suppose that there exists an infinite sequence of different p-efficient points v^j, $j = 1, 2, \dots$

Since they are integer, and the first coordinate v_1^j is bounded from below by l_1, with no loss of generality we may select a subsequence which is non-decreasing in the first coordinate. By a similar token, we can select further subsequences which are non-decreasing in the first k coordinates ($k = 1, \dots, s$). Since the dimension s is finite, we obtain a subsequence of different p-efficient points which is non-decreasing in all coordinates. This contradicts the definition of a p-efficient point. □

Our proof can be easily adapted to the case of non-uniform grids for which a uniform lower bound on the distance of grid points in each coordinate exists. In this way we obtain the following *disjunctive* formulation with a finite index set J for problem (2.6):

$$\begin{aligned} \min\ & f(x) \\ \text{subject to } & g(x) \in \textstyle\bigcup_{j\in J} K_j, \\ & x \in \mathcal{D}. \end{aligned} \tag{2.8}$$

The concept of α-concavity on a set can be used at this moment to find an equivalent representation of the set $\mathcal{Z}$ for the discrete distributions.

Theorem 18. *Let $\mathcal{Z}$ be the set of all possible values of an integer random vector Z. If the distribution function F of Z is α-concave on $\mathcal{Z} + \mathbb{Z}_+^s$, for some $\alpha \in [-\infty, \infty]$, then for every $p \in (0,1)$ one has*

$$\mathcal{Z} = \{y \in \mathbb{R}^s : \, y \ge z \ge \sum_{j\in J} \lambda_j v^j, \, \sum_{j\in J} \lambda_j = 1, \, \lambda_j \ge 0, \, z \in \mathbb{Z}^s\},$$

where v^j, $j \in J$, are the p-efficient points of F.

Proof. By the monotonicity of F we have $F(y) \ge F(z)$ if $y \ge z$. It is, therefore, sufficient to show that $\mathbb{P}(Z \le z) \ge p$ for all $z \in \mathbb{Z}^s$ such that $z \ge \sum_{j\in J} \lambda_j v^j$ with $\lambda_j \ge 0$, $\sum_{j\in J} \lambda_j = 1$. We consider five cases with respect to α.

Case 1: $\alpha = \infty$. It follows from the definition of α-concavity that

$$F(z) \ge \max\{F(v^j), j \in J : \lambda_j \ne 0\} \ge p.$$

Case 2: $\alpha = -\infty$. Since $F(v^j) \ge p$ for each index $j \in J$ such that $\lambda_j \ne 0$, the assertion follows as in Case 1.

Case 3: $\alpha = 0$. By the definition of α-concavity,

$$F(z) \ge \prod_{j\in J} [F(v^j)]^{\lambda_j} \ge \prod_{j\in J} p^{\lambda_j} = p.$$

Case 4: $\alpha \in (-\infty, 0)$. By the definition of α-concavity,

$$[F(z)]^\alpha \le \sum_{j\in J} \lambda_j [F(v^j)]^\alpha \le \sum_{j\in J} \lambda_j p^\alpha = p^\alpha.$$

Since $\alpha < 0$, we obtain $F(z) \geq p$.

Case 5: $\alpha \in (0, \infty)$. By the definition of α-concavity,

$$[F(z)]^\alpha \geq \sum_{j \in J} \lambda_j [F(v^j)]^\alpha \geq \sum_{j \in J} \lambda_j p^\alpha = p^\alpha.$$

□

The consequence of this theorem is that under α-concavity assumption all integer points contained in $\operatorname{co} \mathcal{Z} = E + \mathbb{R}_+^m$ satisfy the probabilistic constraint. This demonstrates the importance of the notion of α-concave distribution function introduced in Definition 3. For example, the set $\mathcal{Z}$ illustrated in Figure 2.1 does not correspond to any α-concave distribution function, because its convex hull contains integer points which do not belong to $\mathcal{Z}$. These are the points (3,6), (4,5) and (6,2).

Under the conditions of Theorem 18, problem (2.8) can be formulated in the following equivalent way:

$$\begin{aligned}
\max\quad & f(x) \\
\text{subject to}\quad & x \in \mathcal{D} \\
& g(x) \geq z, && (2.9) \\
& z \in \mathbb{Z}^m, && (2.10) \\
& z \geq \sum_{j \in J} \lambda_j v^j && (2.11) \\
& \sum_{j \in J} \lambda_j = 1 \\
& \lambda_j \geq 0,\ j \in J.
\end{aligned}$$

So, the probabilistic constraint has been replaced by algebraic equations and inequalities, together with the integrality requirement (2.10). This condition cannot be dropped, in general. However, if other conditions of the problem imply that $g(x)$ is integer, we may dispose of z totally, and replace constraints (2.9)–(2.11) with

$$g(x) \geq \sum_{j \in J} \lambda_j v^j.$$

This may be the case for example, when we have an additional constraint in the definition of $\mathcal{D}$ that $x \in \mathbb{Z}^n$, and $g(x) = Tx$, where T is a matrix of appropriate dimension with integer elements.

If Z takes values on a non-uniform grid, condition (2.10) should be replaced by the requirement that z is a grid point.

2.4 Optimality Conditions and Duality Theory

Let us split variables in problem (2.6):

$$\begin{aligned} \max\ & f(x) \\ & g(x) \geq z, \\ & x \in \mathcal{D}, \\ & z \in \mathcal{Z}. \end{aligned} \tag{2.12}$$

We assume that $p \in (0,1)$. Associating Lagrange multipliers $u \in \mathbb{R}^m_+$ with constraints $g(x) \geq z$, we obtain the Lagrangian function

$$L(x,z,u) = f(x) + \langle u, g(x) - z \rangle.$$

The dual functional has the form

$$\Psi(u) = \sup_{(x,z)\in\mathcal{D}\times\mathcal{Z}} L(x,z,u) = h(u) - d(u),$$

where

$$h(u) = \sup\{f(x) + \langle u, g(x)\rangle \mid x \in \mathcal{D}\}, \tag{2.13}$$

$$d(u) = \inf\{\langle u, z\rangle \mid z \in \mathcal{Z}\}. \tag{2.14}$$

For any $u \in \mathbb{R}^m_+$ the value of $\Psi(u)$ is an upper bound on the optimal value F^* of the original problem. The best Lagrangian upper bound will be given by the optimal value D^* of the problem:

$$\inf_{u\geq 0} \Psi(u). \tag{2.15}$$

We call (2.15) the dual problem to problem (2.6). For $u \not\geq 0$ one has $d(u) = -\infty$, because the set $\mathcal{Z}$ contains a translation of $\mathbb{R}^m_+$. The function $d(\cdot)$ is concave. Note that $d(u) = -\sigma_{\mathcal{Z}}(-u)$, where $\sigma_{\mathcal{Z}}(\cdot)$ is the support function of the set $\mathcal{Z}$. By virtue of Theorem 15 and [161, Chapter V, Proposition 2.2.1], we have

$$d(u) = \inf\{\langle u, z\rangle \mid z \in \operatorname{co}\mathcal{Z}\}. \tag{2.16}$$

Let us consider the *convex hull problem*:

$$\begin{aligned} \max\ & f(x) \\ & g(x) \geq z, \\ & x \in \mathcal{D}, \\ & z \in \operatorname{co}\mathcal{Z}. \end{aligned} \tag{2.17}$$

We make the following assumption.

Constraint Qualification Condition. *There exist points $x^0 \in \mathcal{D}$ and $z^0 \in \operatorname{co}\mathcal{Z}$ such that $g(x^0) > z^0$.*

If the Constraint Qualification Condition is satisfied, from the duality theory in convex programming [298, Corollary 28.2.1] we know that there exists

$\hat{u} \geq 0$ at which the minimum in (2.15) is attained, and $D^* = \Psi(\hat{u})$ is the optimal value of the convex hull problem (2.17).

We now study in detail the structure of the dual functional Ψ. We shall characterize the solution sets of the two subproblems (2.13) and (2.14), which provide values of the dual functional. Let us define the following sets:

$$\begin{aligned} V(u) &= \{v \in \mathbb{R}^m : \langle u, v\rangle = d(u) \text{ and } v \text{ is a } p\text{-efficient point}\}, \\ C(u) &= \{d \in \mathbb{R}^m_+ : d_i = 0 \text{ if } u_i > 0,\ i = 1, \ldots, m\}. \end{aligned} \tag{2.18}$$

Lemma 5. *For every $u > 0$ the solution set of (2.14) is non-empty. For every $u \geq 0$ it has the following form: $\hat{Z}(u) = V(u) + C(u)$.*

Proof. Let us at first consider the case $u > 0$. Then every recession direction d of $\mathcal{Z}$ satisfies $\langle u, d\rangle > 0$. Since $\mathcal{Z}$ is closed, a solution to (2.14) must exist. Suppose that a solution z to (2.14) is not a p-efficient point. By virtue of Theorem 14, there is a p-efficient $v \in \mathcal{Z}$ such that $v \leq z$, and $v \neq z$. Thus, $\langle u, v\rangle < \langle u, z\rangle$, which is a contradiction.

In the general case $u \geq 0$, the solution set of the problem to (2.14), if it is non-empty, always contains a p-efficient point. Indeed, if a solution z is not p-efficient, we must have a p-efficient point v dominated by z, and $\langle u, v\rangle \leq \langle u, z\rangle$ holds by the non-negativity of u. Consequently, $\langle u, v\rangle \leq \langle u, z\rangle$ for all p-efficient $v \leq z$, which is equivalent to $z \in \{v\} + C(u)$, as required.

If the solution set of (2.14) is empty then $V(u) = \emptyset$ and the assertion is true as well. □

The last result allows us to calculate the subdifferential of d in a closed form.

Lemma 6. *For every $u \geq 0$ one has $\partial d(u) = \operatorname{co} V(u) + C(u)$. If $u > 0$ then $\partial d(u)$ is non-empty.*

Proof. From (2.14) we obtain $d(u) = -\sigma_{\mathcal{Z}}(u)$, where $\sigma_{\mathcal{Z}}(\cdot)$ is the support function of $\mathcal{Z}$ and, consequently, of $\operatorname{co}\mathcal{Z}$. Recall that $\sigma_{\mathcal{Z}}(u) = \delta^*_{\mathcal{Z}}(u)$, where the latter is the conjugate of the indicator function of $\mathcal{Z}$. These facts follow from the structure of $\mathcal{Z}$ described Theorem 14, by virtue of Corollary 16.5.1 in [298]. Thus

$$\partial d(u) = \partial\delta^*_{\mathcal{Z}}(-u).$$

Recall that $\operatorname{co}\mathcal{Z}$ is closed, by Theorem 15. Using [298, Theorem 23.5], we observe that $s \in \partial\delta^*_{\mathcal{Z}}(-u)$ if and only if $\delta^*_{\mathcal{Z}}(-u) + \delta_{\operatorname{co}\mathcal{Z}}(s) = -\langle s, u\rangle$, where $\delta_{\operatorname{co}\mathcal{Z}}(\cdot)$ is the indicator function of $\operatorname{co}\mathcal{Z}$. It follows that $s \in \operatorname{co}\mathcal{Z}$ and $\delta^*_{\mathcal{Z}}(-u) = -\langle s, u\rangle$. Consequently,

$$\langle s, u\rangle = d(u). \tag{2.19}$$

Since $s \in \operatorname{co}\mathcal{Z}$ we can represent it as follows:

$$s = \sum_{j=1}^{m+1} \alpha_j e^j + w,$$

where e^j, $j = 1, \ldots, m+1$, are extreme points of $\operatorname{co} \mathcal{Z}$ and $w \geq 0$. Using Theorem 16 we conclude that e^j are p-efficient points. Moreover

$$\langle s, u \rangle = \sum_{j=1}^{m+1} \alpha_j \langle u, e^j \rangle + \langle u, w \rangle \geq d(u), \tag{2.20}$$

because $\langle u, e^j \rangle \geq d(u)$ and $\langle u, w \rangle \geq 0$. Combining (2.19) and (2.20) we conclude that $\langle u, e^j \rangle = d(u)$ for all j, and $\langle u, w \rangle = 0$. Thus $s \in \operatorname{co} V(u) + C(u)$.

Conversely, if $s \in \operatorname{co} V(u) + C(u)$ then (2.19) holds true. This implies that $s \in \partial d(u)$, as required.

The set $\partial d(u)$ is non-empty for $u > 0$ by virtue of Lemma 5. □

Now we analyze the function $h(\cdot)$. Define the set of maximizers in (2.13),

$$X(u) = \{x \in \mathcal{D} : f(x) + \langle u, g(x) \rangle = h(u)\}.$$

By the convexity of the set $\mathcal{D}$ and by the concavity of f and g, the solution set $X(u)$ is convex for all $u \geq 0$.

Lemma 7. *Assume that the set $\mathcal{D}$ is compact. The subdifferential of the function h is described as follows for every $u \in \mathbb{R}^m$:*

$$\partial h(u) = \operatorname{co} \{g(x) : x \in X(u)\}.$$

Proof. The function h is convex on $\mathbb{R}^m$. Since the set $\mathcal{D}$ is compact and f and g are concave, the set $X(u)$ is compact. Therefore, the subdifferential of $h(u)$ for every $u \in \mathbb{R}^m$ is the closure of $\operatorname{co} \{g(x) : x \in X(u)\}$ (see [161, Chapter VI, Lemma 4.4.2]). By the compactness of $X(u)$ and concavity of g, the set $\{g(x) : x \in X(u)\}$ is closed. Therefore, we can omit taking the closure in the description of the subdifferential of $h(u)$. □

This analysis provides the basis for the following necessary and sufficient optimality conditions for problem (2.15).

Theorem 19. *Assume that the Constraint Qualification Condition is satisfied and that the set $\mathcal{D}$ is compact. A vector $u \geq 0$ is an optimal solution of (2.15) if and only if there exists a point $x \in X(u)$, points $v^1, \ldots, v^{m+1} \in V(u)$ and scalars $\beta_1 \ldots, \beta_{m+1} \geq 0$ with $\sum_{j=1}^{m+1} \beta_j = 1$, such that*

$$g(x) - \sum_{j=1}^{m+1} \beta_j v^j \in C(u), \tag{2.21}$$

where $C(u)$ is given by (2.18).

Proof. Since $-C(u)$ is the normal cone to the positive orthant at $u \geq 0$, the necessary and sufficient optimality condition for (2.15) has the form

$$\partial \Psi(u) \cap C(u) \neq \emptyset \tag{2.22}$$

(*cf.* [298, Theorem 27.4]). Since $\operatorname{int}\operatorname{dom} d \neq \emptyset$ and $\operatorname{dom} h = \mathbb{R}^m$ we have $\partial\Psi(u) = \partial h(u) - \partial d(u)$. Using Lemma 6 and Lemma 7, we conclude that there exist

$$\begin{aligned}
&p\text{-efficient points } v^j \in V(u), \quad j = 1, \dots, m+1,\\
&\beta^j \geq 0, \quad j = 1, \dots, m+1, \quad \sum_{j=1}^{m+1} \beta_j = 1,\\
&x^j \in X(u), \quad j = 1, \dots, m+1,\\
&\alpha^j \geq 0, \quad j = 1, \dots, m+1, \quad \sum_{j=1}^{m+1} \alpha_j = 1,
\end{aligned}$$

such that

$$\sum_{j=1}^{m+1} \alpha_j g(x^j) - \sum_{j=1}^{m+1} \beta_j v^j \in C(u). \tag{2.23}$$

If the functions f and g were strictly concave, the set $X(u)$ would be a singleton. Then all x^j would be identical and the above relation would immediately imply (2.21). Otherwise, let us define

$$x = \sum_{j=1}^{m+1} \alpha_j x^j.$$

By the convexity of $X(u)$ we have $x \in X(u)$. Consequently,

$$f(x) + \sum_{i=1}^{m} u_i g_i(x) = h(u) = f(x^j) + \sum_{i=1}^{m} u_i g_i(x^j), \quad j = 1, \dots, m+1.$$

Multiplying the last equation by α_j and adding we obtain

$$f(x) + \sum_{i=1}^{m} u_i g_i(x) = \sum_{j=1}^{m+1} \alpha_j \Big[f(x^j) + \sum_{i=1}^{m} u_i g_i(x^j) \Big].$$

Since $g_i(x) \geq \sum_{j=1}^{m+1} \alpha_j g_i(x^j)$, substituting into the above equation, we obtain

$$f(x) \leq \sum_{j=1}^{m+1} \alpha_j f(x^j).$$

If $g_i(x) > \sum_{j=1}^{m+1} \alpha_j g_i(x^j)$ and $u_i > 0$ for some i, the above inequality becomes strict, in contradiction to the concavity of f. Thus, for all $u_i > 0$ we have $g_i(x) = \sum_{j=1}^{m+1} \alpha_j g_i(x^j)$, and it follows that

$$g(x) - \sum_{j=1}^{m+1} \alpha_j g(x^j) \in C(u).$$

Since $C(u)$ is a convex cone, we can combine the last relation with (2.23) and obtain (2.21), as required.

Now we prove the converse implication. Assume that we have $x \in X(u)$, points $v^1, \dots, v^{m+1} \in V(u)$ and scalars $\beta_1 \dots, \beta_{m+1} \geq 0$ with $\sum_{j=1}^{m+1} \beta_j = 1$, such that (2.21) holds true. By Lemma 6 and Lemma 7 we have

$$g(x) - \sum_{j=1}^{m+1} \beta_j v^j \in \partial\Psi(u).$$

Thus (2.21) implies (2.22), which is a necessary and sufficient optimality condition for (2.15). □

Since the set of p-efficient points is not known, we need a numerical method for solving the convex hull problem (2.17) or its dual (2.15).

Using these optimality conditions we obtain the following duality result.

Theorem 20. *Assume that the Constraint Qualification Condition for problem (2.12) is satisfied, the probability distribution of the vector Z is $\alpha-$concave for some $\alpha \in [-\infty, \infty]$, and the set $\mathcal{D}$ is compact. If a point $(\hat{x}, \hat{z})$ is an optimal solution of (2.12), then there exists a vector $\hat{u} \geq 0$, which is an optimal solution of (2.15) and the optimal values of both problems are equal. If $\hat{u}$ is an optimal solution of problem (2.15), then there exist a point $\hat{x}$, such that $(\hat{x}, g(\hat{x}))$ is a solution of problem (2.12), and the optimal values of both problems are equal.*

Proof. From the α-concavity assumption we obtain that problems (2.12) and (2.17) coincide. If $\hat{u}$ is optimal solution of problem (2.15), we obtain the existence of points $\hat{x} \in X(\hat{u})$, $v^1, \dots, v^{m+1} \in V(u)$ and scalars $\beta_1 \dots, \beta_{m+1} \geq 0$ with $\sum_{j=1}^{m+1} \beta_j = 1$, such that the optimality conditions in Theorem 19 are satisfied. Setting $\hat{z} = g(\hat{x})$ we have to show that $(\hat{x}, \hat{z})$ is an optimal solution of problem (2.12) and that the optimal values of both problems are equal. First we observe that this point is feasible. Set $s \in C(\hat{u}) : s = g(\hat{x}) - \sum_{j=1}^{m+1} \beta_j v^j$. From the definitions of $X(\hat{u})$, $V(\hat{u})$, and $C(\hat{u})$ we obtain

$$\begin{aligned} h(\hat{u}) = f(\hat{x}) + \langle \hat{u}, g(\hat{x}) \rangle &= f(\hat{x}) + \langle \hat{u}, \sum_{j=1}^{m+1} \beta_j v^j + s \rangle \\ &= f(\hat{x}) + \sum_{j=1}^{m+1} \beta_j d(\hat{u}) + \langle \hat{u}, s \rangle = f(\hat{x}) + d(\hat{u}). \end{aligned}$$

Thus, $f(\hat{x}) = h(\hat{u}) - d(\hat{u}) = D^* \geq F^*$, which proves that $(\hat{x}, \hat{z})$ is an optimal solution of problem (2.12) and $D^* = F^*$.

If $(\hat{x}, \hat{z})$ is a solution of (2.12), then by [298, Theorem 28.4] there is a vector $\hat{u} \geq 0$ such that $\hat{u}_i(\hat{z}_i - g_i(\hat{x})) = 0$ and $\partial f(\hat{x}) + \partial\langle \hat{u}, g(\hat{x}) - \hat{z}\rangle \cap -\partial\big[\delta_{\mathcal{D}}(\hat{x}) + \delta_{\mathcal{Z}}(\hat{z})\big] \neq \varnothing$, where $\delta_C(\cdot)$ denotes the indicator function of the set C. Thus, there are vectors

$$s \in \partial f(\hat{x}) + \partial\langle u, g(\hat{x})\rangle \cap -\partial\delta_{\mathcal{D}}(\hat{x}) \tag{2.24}$$

and

$$\hat{u} \in \partial\delta_{\mathcal{Z}}(\hat{z}). \tag{2.25}$$

The first inclusion (2.24) is the optimality condition for problem (2.13), and thus $x \in X(\hat{u})$. By virtue of [298, Theorem 23.5] the inclusion (2.25) is equivalent to $\hat{z} \in \partial\delta^*_{\mathcal{Z}}(\hat{u})$. Using Lemma 6 we obtain that $\hat{z} \in \partial d(\hat{u})$ =co$V(\hat{u})+C(\hat{u})$. Thus, there exists points $v^1, \ldots, v^{m+1} \in V(u)$ and scalars $\beta_1 \ldots, \beta_{m+1} \geq 0$ with $\sum_{j=1}^{m+1} \beta_j = 1$, such that $\hat{z} - \sum_{j=1}^{m+1} \beta_j v^j \in C(\hat{u})$. Using the complementarity condition $\hat{u}_i(\hat{z}_i - g_i(\hat{x})) = 0$ we conclude that the optimality conditions of Theorem 19 are satisfied. Thus $\hat{u}$ is an optimal solution of (2.15). □

For the special case of discrete distribution and linear constraints we can obtain a more specific necessary and sufficient condition for the existence of an optimal solution of (2.8).

The *linear* probabilistic optimization problem assumes that $g(x) = Tx$, where T is an $m \times n$ matrix, $f(x) = \langle c, x\rangle$ with $c \in \mathbb{R}^n$. Furthermore, $\mathcal{D}$ is a convex closed polyhedron in $\mathbb{R}^n$. It is usually formulated as follows:

$$\begin{aligned} &\min\ \langle c, x\rangle \\ &\text{subject to } \mathbb{P}[Tx \geq Z] \geq p \\ &\qquad Ax \geq b \\ &\qquad x \geq 0. \end{aligned} \tag{2.26}$$

Here A is an $s \times n$ matrix and $b \in \mathbb{R}^s$.

Assumption 2.1. *The set* $\Lambda := \{(u, w) \in \mathbb{R}^{m+s}_+ \mid A^T w + T^T u \leq c\}$ *is nonempty.*

Theorem 21. *Assume that the feasible set of (2.26) is non-empty and that* Z *has a discrete distribution. Then (2.26) has an optimal solution if and only if Assumption 2.1 holds.*

Proof. If (2.8) has an optimal solution, then for some $j \in J$ the linear optimization problem

$$\begin{aligned} &\min\ \langle c, x\rangle \\ &\text{subject to } Tx \geq v^j \\ &\qquad Ax \geq b \\ &\qquad x \geq 0 \end{aligned} \tag{2.27}$$

has an optimal solution. By duality in linear programming, its dual problem

$$\begin{aligned} &\max\ \langle v^j, u\rangle + \langle b, w\rangle \\ &\text{subject to } T^T u + A^T w \le c \\ &u, w \ge 0 \end{aligned} \tag{2.28}$$

has an optimal solution and the optimal values of both programs are equal. Thus, Assumption 2.1 must hold. On the other hand, if Assumption 2.1 is satisfied, all dual programs (2.28) for $j \in J$ have non-empty feasible sets, so the objective values of all primal problems (2.27) are bounded from below. Since one of them has a non-empty feasible set by assumption, an optimal solution must exist. □

2.5 Methods for Solving Probabilistic Programming Problems

When the constraint functions $g_i(x, Z)$, $i = 1, \dots m$ are not separable the optimization problem is difficult to handle. Numerical methods for these problems are based on combinatorial techniques (see Section 2.5.8), on response surface approximations (see Section 2.5.6), or via Monte Carlo methods (see Chapter 5 in this volume).

Numerical techniques for probabilistic problems with random right hand sides ($g(x, Z) := \tilde{g}(x) - Z$) are much better developed, particularly for linear function $\tilde{g}(x) = Tx$. If the distribution of Z is α-concave, it follows from Corollary 2 that the feasible set of this problem is convex. Therefore, methods of convex programming can be applied. The specificity here is in the implicit definition of the feasible set and in the difficulty to evaluate the constraint function

$$G(x) = \mathbb{P}[Tx \ge Z] = F_Z(Tx).$$

It is even more difficult to estimate its gradient if it exists. Specialized Monte Carlo integration techniques have been developed for some classes of distributions of Z, in particular for the normal distribution and for the gamma distribution (see [223, 224, 349–351]).

We review some of the known methods and present in more detail two recent methods for solving nonlinear probabilistic problems: the dual method in Section 2.5.3 and the primal-dual method in Section 2.5.4. Both of these methods are based on the duality analysis presented in the previous section.

2.5.1 A Cutting Plane Method

One of the first methods for probabilistic optimization is based on cutting planes techniques for the following problem:

$$\begin{aligned} &\min\ f(x) \\ &\text{subject to } \mathbb{P}[Tx \ge Z] \ge p \\ &Ax = b,\ x \ge 0. \end{aligned} \tag{2.29}$$

It is assumed that the constraint function $G(x) = \mathbb{P}[Tx \geq Z]$ is quasi-concave and it has continuous gradients. Additionally, we assume that there exists a bounded convex polyhedron C^1 containing the set of feasible solutions of problem (2.29).

Furthermore, the following constraint qualification condition is satisfied: there exists a vector x^0 such that

$$G(x^0) > p \ , \ x^0 \in \{x \mid Ax = b \ , \ x \geq 0\} \tag{2.30}$$

The algorithm works in two phases. In the first phase a feasible point x^0 satisfying the constraint qualification condition is found. This can be done by maximizing $G(x)$ subject to the constraints $Ax = b \ , \ x \geq 0$, by using any gradient method. Of course in our case, we do not need to carry out all steps in the gradient descent method, it is sufficient to find a point x^* such that $G(x^*) > p$.

The second phase consists of the following steps.

Step 1. Solve the problem:

$$\begin{gathered} \min f(x) \\ \text{subject to } x \in C^k. \end{gathered}$$

Let x^k be an optimal solution. If x^k is feasible, then stop, x^k is an optimal solution of problem (2.29).

Step 2. Let λ^k be the largest $\lambda \geq 0$ such that $x^0 + \lambda(x^k - x^0)$ is feasible and set

$$y^k = x^0 + \lambda^k(x^k - x^0).$$

If $G(y^k) = p$, then define

$$C^{k+1} = \left\{x \mid x \in C^k \ , \ \nabla G(y^k)(x - y^k) \geq 0\right\}.$$

Consider any other constraint that is active at y^k and set C^{k+1} to be the intersection of C^k and the set determined by this constraint. Go to Step 1.

2.5.2 The Logarithmic Barrier Function Method

If Z has a log-concave distribution on $\mathbb{R}^n$ then the constraint function of problem (2.29) is log-concave on the set $\{x \in \mathbb{R}^n : G(x) \geq p\}$. We can solve problem (2.29) by using logarithmic penalty functions. We take a decreasing sequence of positive numbers $\left\{s^k\right\}$ such that $\lim\limits_{k\to\infty} s^k = 0$ and solve the problem

$$\begin{aligned} \min\{f(x) - s^k \log(G(x) - p)\} \\ \text{subject to } Ax = b \\ x \geq 0. \end{aligned}$$

We obtain a point x^k as an optimal solution of this problem. The sequence $\{f(x^k)\}$ converges to the optimal value of problem (2.29) under the assumptions that $f(\cdot)$ is a continuous function, $G(\cdot)$ is continuous and log-concave, constraint qualification condition (2.30) is satisfied, and the set $\{x \in \mathbb{R}^n : Ax = b,\ x \geq 0\}$ is bounded.

A modern primal-dual interior point method using similar idea is developed in [329].

2.5.3 The Dual Method

This method has been proposed in [104] for solving nonlinear probabilistic problems of form (2.6). The idea of the method is to solve the dual problem (2.15) using the information about the subgradients of the dual functional Ψ to generate convex piecewise-linear approximations of Ψ. Suppose that the values of the functional Ψ at certain points u^j, $j = 1, \dots, k$, are available. Moreover, we assume that the corresponding solutions $v^1, \dots, v^k$ and $x^1, \dots, x^k$ of the two problems (2.16) and (2.13) are available as well. According to Lemma 5 we can assume that v^j, $j = 1, \dots, k$ are p-efficient points. By virtue of Lemma 7 and Lemma 6 the following function $\Psi_k(\cdot)$ is a lower bound of Ψ:

$$\Psi_k(u) := \max_{1 \leq j \leq k} \left[\Psi(u^j) + \langle g(x^j) - v^j, u - u^j \rangle \right].$$

Minimizing $\Psi_k(u)$ over $u \geq 0$, we obtain the next iterate u^{k+1}. For the purpose of numerical tractability, we shall impose an upper bound $b \in \mathbb{R}$ on the dual variables u_j. We define the feasible set of the dual problem as follows:

$$U := \{u \in \mathbb{R}^m : 0 \leq u_i \leq b,\ i = 1, \dots, m\}$$

where b is a sufficiently large number. We also use $\varepsilon > 0$ as a stopping test parameter.

The algorithm works as follows:

Step 0. Select a vector $u^1 \in U$. Set $\Psi_0(u^1) = -\infty$ and $k = 1$.
Step 1. Calculate

$$h(u^k) = \max\{f(x) + \langle u^k, g(x) \rangle \mid x \in \mathcal{D}\}, \tag{2.31}$$

$$d(u^k) = \min\{\langle u^k, z \rangle \mid z \in \operatorname{co} \mathcal{Z}\}. \tag{2.32}$$

Let x^k be the solution of problem (2.31) and v^k be the solution of problem (2.32).

Step 2. Calculate $\Psi(u^k) = h(u^k) - d(u^k)$. If $\Psi(u^k) \leq \Psi_{k-1}(u^k) + \varepsilon$ then stop; otherwise continue.

Step 3. Define

$$\Psi_k(u) = \max_{1 \leq j \leq k} \left[\Psi(u^j) + \langle g(x^j) - v^j, u - u^j \rangle \right],$$

and find a solution u^{k+1} of the problem

$$\min_{u\in U} \Psi_k(u).$$

Step 4. Increase k by one and go to Step 1.

Problem (2.31) is a convex nonlinear problem, and it can be solved by a suitable numerical method for nonlinear optimization. Problem (2.32) requires a dedicated numerical method. In particular applications, specialized methods may provide its efficient numerical solution. Alternatively, one can approximate the random vector Y by finitely many realizations (scenarios). More detailed discussion on this idea will follow in Section 2.5.7.

Theorem 22. *Suppose that $\varepsilon = 0$. Then the sequences $\Psi(u^k)$ and $\Psi_k(u^k)$, $k = 1, 2, \ldots$, converge to the optimal value of problem (2.15). Moreover, every accumulation point of the sequence $\{u^k\}$ is an optimal solution of (2.15).*

Proof. The convergence of the method follows from a standard argument about cutting plane methods for convex optimization (see, *e.g.*, [161, Theorem 4.2.3]). □

Let us discuss a way of recovering a primal solution from the sequences of points $\{u^k\}$, $\{x^k\}$ and $\{v^k\}$ generated by the method.

It follows from Theorem 22 that for every $\varepsilon > 0$ the dual method has to stop after finitely many iterations at some step k for which

$$\Psi(u^k) - \varepsilon \le \Psi_{k-1}(u^k) \le \min_{u\in U} \Psi(u). \tag{2.33}$$

Let us define the set of active cutting planes at u^k:

$$J = \big\{ j \in \{1, \ldots, k-1\} : \Psi(u^j) + \langle g(x^j) - v^j, u^k - u^j \rangle = \Psi_{k-1}(u^k) \big\}.$$

The subdifferential of $\Psi_{k-1}(\cdot)$ has the form

$$\partial \Psi_{k-1}(u) = \Big\{ s \in \mathbb{R}^m : s = \sum_{j\in J} \alpha_j \big(g(x^j) - v^j\big),\ \sum_{j\in J} \alpha_j = 1,\ \alpha_j \ge 0,\ j \in J \Big\}.$$

Since u^k is a minimizer of $\Psi_{k-1}(\cdot)$, there must exist a subgradient s such that

$$s \in C(u^k).$$

Thus there exist non-negative α_j totaling 1 such that

$$\sum_{j\in J} \alpha_j \big(g(x^j) - v^j\big) \in C(u^k). \tag{2.34}$$

By the definition of Ψ,

$$\Psi(u^j) = f(x^j) + \langle u^j, g(x^j) \rangle - \langle u^j, v^j \rangle.$$

Substituting this into the definition of the set J we obtain that

$$\Psi_{k-1}(u^k) = f(x^j) + \langle g(x^j) - v^j, u^k \rangle, \quad j \in J.$$

Multiplying both sides by α_j and summing up we conclude that

$$\Psi_{k-1}(u^k) = \sum_{j\in J} \alpha_j f(x^j) + \Big\langle \sum_{j\in J} \alpha_j \big(g(x^j) - v^j\big), u^k \Big\rangle.$$

This combined with (2.34) yields

$$\Psi_{k-1}(u^k) = \sum_{j\in J} \alpha_j f(x^j). \tag{2.35}$$

Define

$$\bar{x} = \sum_{j\in J} \alpha_j x^j, \quad \bar{z} = \sum_{j\in J} \alpha_j v^j.$$

Clearly, $\bar{x} \in \mathcal{D} \cap \operatorname{co} \mathcal{Z}$. Using the concavity of g and (2.34) we see that

$$g(\bar{x}) \geq \sum_{j\in J} \alpha_j g(x^j) \geq \sum_{j\in J} \alpha_j v^j = \bar{z}.$$

Thus the point $(\bar{x}, \bar{z})$ is feasible for the convex hull problem (2.17).

It follows from the concavity of f and (2.35) that

$$f(\bar{x}) \geq \sum_{j\in J} \alpha_j f(x^j) = \Psi_{k-1}(u^k).$$

By the stopping test (2.33),

$$f(\bar{x}) \geq \Psi(u^k) - \varepsilon. \tag{2.36}$$

Since the value of $\Psi(u)$ is an upper bound for the objective value at any feasible point (x, z) of the convex hull problem, we conclude that $(\bar{x}, \bar{z})$ is an ϵ-optimal solution of this problem.

The above construction can be carried out at every iteration k. In this way we obtain a certain sequence $(\bar{x}^k, \bar{v}^k)$, $k = 1, 2, \ldots$ Since the sequence $\{\bar{x}^k\}$ is contained in a compact set and each $(\bar{x}^k, \bar{z}^k)$ is feasible for the convex hull problem (2.17), the sequence $\{\bar{z}^k\}$ is included in a compact set as well. Thus the sequence $\{(\bar{x}^k, \bar{v}^k)\}$ has accumulation points. It follows from Theorem 22 and from (2.36) that every accumulation point of the sequence $\{(\bar{x}^k, \bar{v}^k)\}$ is a solution of the convex hull problem (2.17). Under the assumptions of Corollary 2 the accumulation point is a solution of problem (2.6).

2.5.4 The Primal-Dual Method

This approach was first suggested for linear probabilistic problems in [191]. Involving tools of non-smooth analysis the method was successfully developed for nonlinear probabilistic optimization and general distributions in [104].

The algorithm presented in the previous section is based on a cutting plane approximation of the entire dual functional. The method of the this section involves approximations of the functional $d(\cdot)$ only. The method consists of an iterative generation of p-efficient points and the solution of a restriction of problem (2.4). The restriction is based on the disjunctive representation of $\operatorname{co}\mathcal{Z}$ by the p-efficient points generated so far.

We assume that we know a compact set B containing all p-efficient points v such that there exists $x \in \mathcal{D}$ satisfying $v \leq g(x)$. It may be just a box with the lower bound at l, the vector of p-efficient points of all marginal distributions of Y, and with the upper bound above the maxima of $g_i(x)$ over $x \in \mathcal{D}$, $i = 1, \dots, m$. Such a box exists by the compactness of $\mathcal{D}$. We also use a stopping test parameter $\varepsilon > 0$.

We denote the simplex in $\mathbb{R}^k$ by S_k, *i.e.*,

$$S_k := \{\lambda \in \mathbb{R}^k : \lambda_i \geq 0, \sum_{i=1}^{k} \lambda_i = 1\}.$$

The primal-dual method follows the steps:

Step 0. Select a p-efficient point $v^1 \in B$ such that there exists $\tilde{x} \in \mathcal{D}$ satisfying $g(\tilde{x}) > v^1$. Set $J_1 = \{1\}$, $k = 1$.

Step 1. Solve the *master problem*

$$\max f(x) \tag{2.37}$$

$$g(x) \geq \sum_{j \in J_k} \lambda_j v^j, \tag{2.38}$$

$$x \in \mathcal{D}, \ \lambda \in S_k. \tag{2.39}$$

Let u^k be the vector of Lagrange multipliers associated with the constraint (2.38).

Step 2. Calculate $d_k(u^k) = \min_{j \in J_k} \langle u^k, v^j \rangle$.

Step 3. Find a p-efficient solution v^{k+1} of the subproblem:

$$\min_{z \in \mathcal{Z} \cap B} \langle u^k, z \rangle$$

and calculate $d(u^k) = \langle v^{k+1}, u^k \rangle$.

Step 4. If $d(u^k) \geq d_k(u^k) - \varepsilon$ then stop; otherwise set $J_{k+1} = J_k \cup \{k+1\}$, increase k by one, and go to Step 1.

The first p-efficient point v^1 can be found by solving the subproblem at Step 3 for some $u \geq 0$. All master problems will be solvable, if the first one is solvable, which is assumed at Step 0. Moreover, all master problems satisfy Slater's constraint qualification condition with the point $\tilde{x}$ and $\tilde{\lambda} = (1, 0, \ldots, 0)$. Therefore, it is legitimate to assume at Step 1 that we obtain a vector of Lagrange multipliers associated with (2.38). The subproblem at Step 3 is the same as (2.32) in the dual method. It requires a dedicated approach.

Theorem 23. *Let $\varepsilon = 0$. The sequence $\{f(x^k)\}$, $k = 1, 2, \ldots$ converges to the optimal value of the convex hull problem (2.17). Every accumulation point $\hat{x}$ of the sequence $\{x^k\}$ is an optimal solution of problem (2.17), with $z = g(\hat{x})$.*

Proof. We formulate the dual problem to the master problem (2.37)–(2.39).

The dual functional is defined as follows:

$$\Phi_k(u) = \sup\Big\{ f(x) + \langle u, g(x) - \sum_{j \in J_k} \lambda_j v^j \rangle : \ x \in \mathcal{D}, \ \lambda \in S_k \Big\} = h(u) - d_k(u),$$

where $h(u)$ is the same as in (2.13) and

$$d_k(u) = \inf_{\lambda \in S_k} \sum_{j \in J_k} \lambda_j \langle u, v^j \rangle.$$

It is clear that $d_k(u) = \min_{j \in J_k} \langle u, v^j \rangle \geq d(u)$, where $d(u)$ is as in (2.14). Thus the function $\Phi_k(u)$ is a lower bound of the dual functional $\Psi(u)$, *i.e.*,

$$\Phi_k(u^k) \leq \Psi(u^k).$$

Since $J_k \subset J_{k+1}$, for every feasible point (x, λ) of problem (2.37)–(2.39) at iteration k, the point $(x, (\lambda, 0))$ is feasible at iteration $k+1$. Therefore the sequence $\{f(x^k)\}$ is monotonically increasing. By duality, the sequence $\{\Phi_k(u^k)\}$ is monotonically increasing as well.

For $\delta > 0$ consider the set K_δ of iteration numbers k for which

$$\Phi_k(u^k) + \delta \leq \Psi(u^k).$$

Suppose that $k \in K_\delta$. We obtain the following chain of inequalities for all $j \leq k$:

$$\begin{aligned} \delta \leq \Psi(u^k) - \Phi_k(u^k) &= -d(u^k) + d_k(u^k) = -\min_{z \in \mathcal{Z} \cap B} \langle u^k, z \rangle + \min_{j \in J_k} \langle u^k, v^j \rangle \\ &\leq \langle u^k, v^j - v^{k+1} \rangle \leq \|u^k\| \cdot \|v^j - v^{k+1}\|. \end{aligned}$$

We shall show later that there exists $M > 0$ such that $\|u^k\| \leq M$ for all k. Therefore

$$\|v^{k+1} - v^j\| \geq \delta/M \quad \text{for all} \quad k \in K_\delta \quad \text{and all} \quad j = 1, \ldots, k.$$

It follows from the compactness of the set B that the set K_δ is finite for every $\delta > 0$. Thus, we can find a subsequence $\mathcal{K}$ such that

$$\Psi(u^k) - \Phi_k(u^k) \to 0, \quad k \in \mathcal{K}.$$

Since for all k

$$\Psi(u^k) \geq \min_{u \geq 0} \Psi(u) \geq \min_{u \geq 0} \Phi_k(u) = \Phi_k(u^k), \tag{2.40}$$

and the sequence $\{\Phi_k(u^k)\}$ is nondecreasing, we conclude that

$$\lim_{k \to \infty} \Phi_k(u^k) = \min_{u \geq 0} \Psi(u).$$

We also have $\Phi_k(u^k) = f(x^k)$ and thus the sequence $\{f(x^k)\}$ is convergent to the optimal value of the convex hull problem (2.17). Since $\{x^k\}$ is included in $\mathcal{D}$, it has accumulation points and every accumulation point $\hat{x}$ is a solution of (2.17), with $z = g(\hat{x})$.

It remains to show that the multipliers u^k are uniformly bounded. To this end observe that the Lagrangian

$$L_k(x, \lambda, u^k) = f(x) + \left\langle u^k, g(x) - \sum_{j=1}^{k} \lambda_j v^j \right\rangle$$

achieves its maximum in $\mathcal{D} \times S_k$ at x^k and some λ^k. The optimal value is equal to $f(x^k)$ and it is bounded above by the optimal value μ of the convex hull problem (2.17).

The point $\tilde{x}$ and $\tilde{\lambda} = (1, 0, \ldots, 0)$ is in $\mathcal{D} \times S_k$. Therefore

$$L_k(\tilde{x}, \tilde{\lambda}, u^k) \leq \mu.$$

It follows that

$$\langle u^k, g(\tilde{x}) - v^1 \rangle \leq \mu - f(\tilde{x}).$$

Recall that $g(\tilde{x}) - v^1 > 0$. Therefore u^k is an element of the compact set

$$U = \{u \in \mathbb{R}^m : \langle u, g(\tilde{x}) - v^1 \rangle \leq \mu - f(\tilde{x}),\ u \geq 0\}.$$

If we use $\varepsilon > 0$ at Step 4, then relations (2.40) guarantee that the current solution x^k is ε-optimal for the convex hull problem (2.17). □

Under the assumption that the distribution function of the random vector Y is α-concave for some $\alpha \in \overline{\mathbb{R}}$, the suggested algorithms provide an optimal solution of problem (2.4). Otherwise, we obtain an upper bound of the optimal value. Moreover, the solution point $\hat{x}$ determined by both algorithms satisfies the constraint $g(x) \in \operatorname{co} \mathcal{Z}$, and may not satisfy the probabilistic constraint.

We now suggest an approach to determine a primal feasible solution.

Both the dual and the primal-dual method end with a collection of p-efficient points. In the primal-dual algorithm, we consider the multipliers λ_j of the master problem (2.37)–(2.39). We define $C = \{j \in J : \lambda_j > 0\}$. In the dual algorithm, we consider the active cutting planes in the last approximation, and set $C = \{j \in J : \beta_j > 0\}$, where J and β_j are determined in the proof of Theorem 22.

In both cases, if C contains only one element, the point $\hat{x}$ is feasible and therefore optimal for the disjunctive formulation (2.7). If, however, there are more elements in C, we need to find a feasible point. A natural possibility is to consider the *restricted* disjunctive formulation:

$$\begin{aligned} &\max\ f(x) \\ &\text{subject to } g(x) \in \bigcup_{j\in C} K_j, \\ &x \in \mathcal{D}. \end{aligned} \tag{2.41}$$

It can be solved by simple enumeration of all cases for $j \in C$:

$$\begin{aligned} &\max\ f(x) \\ &\text{subject to } g(x) \geq v^j, \\ &x \in \mathcal{D}. \end{aligned} \tag{2.42}$$

An alternative strategy would be to solve the corresponding bounding problem (2.42) every time a new p-efficient point is generated. If μ_j denotes the optimal value of (2.42), the lower bound at iteration k is

$$\bar{\mu}^k = \max_{0\leq j\leq k} \mu_j.$$

A quantitative estimate of the errors and the comparison of both methods are difficult and require new techniques.

2.5.5 Nonparametric Estimates of Distribution Functions

In this subsection we shall assume that the probabilistic constraint is formulated as follows:

$$\mathbb{P}(Tx \leq Z) \geq p.$$

Furthermore, we assume that the random variables $Z_1, \ldots, Z_m$ are independent and each has a continuous distribution with density $h_i(\cdot)$. Using the marginal distribution functions $F_i(z) = \mathbb{P}(Z_i \leq z)$, problem (2.26) can be written in the following equivalent form:

$$\begin{aligned} &\min\langle c, x\rangle \\ &\text{subject to} \prod_{i=1}^{m} \big(1 - F_i(z_i)\big) \geq p \\ &T_i x = z_i, \quad i = 1, \ldots, m \\ &Ax = b \\ &x \geq 0. \end{aligned} \tag{2.43}$$

If for any feasible solution x of the this problem the probabilistic constraint is satisfied as a strict inequality, we can take logarithm on both sides of this constraint.

We define the auxiliary functions:

$$g_i(t) = \begin{cases} \frac{h_i(t)}{1-F_i(t)} & \text{if} \quad F_i(t) < 1 \\ 0 & \text{if} \quad F_i(t) = 1 \end{cases} \tag{2.44}$$

Assuming that the functions $h_i(t), i = 1, \ldots, m$ are log-concave implies that the functions $1 - F_i(t)$ are log-concave as well. Moreover, using the log-concavity of $1 - F_i(t)$, we can show that $g_i(t)$ is a decreasing function. Manipulating (2.44) we obtain

$$1 - F_i(y_i) = e^{-\int\limits_{-\infty}^{y_i} g_i(t)dt}.$$

The functions $g_i(t)$ are estimated from samples.

Let $g_i^{(N)}$ denote an original estimator of g_i for a given N. We take a sample $\{\mathbb{Z}_{Ni}\}$ from the population with distribution function F_i, and create a grid $t_{N,1} < t_{N,2} < \ \ldots \ < t_{N,m}$. The original estimator $g_i^{(N)}$ can then be defined as

$$g_i^{(N)}(t) = \frac{F_i^{(N)}(t_{N,j+1}) - F_i^{(N)}(t_{N,\,j})}{(t_{N,j+1} - t_{N,\,j})(1 - F_i^{(N)}(t_{N,j}))}, \quad t_{N,j} < t \le t_{N,j+1},$$

where $F_i^{(N)}$ is the empirical distribution function corresponding to $F_i, i = 1, \ldots, r$.

We choose a point $x_{N,j}$ from the interval $(t_{N,j}, t_{N,j+1}]$ and a weight $w(x_{N,j})$ associated with it. Then we solve the problem

$$\inf_{U_j} \sum_{j=1}^{m} \left(U_j - g_i^{(N)}(x_{N,j})\right)^2 w(x_{N,j})$$
$$\text{subject to } U_j \ge U_{j+1}, \quad j = 1, \ldots, m-1.$$

Let $\hat{g}_i^{(N)}(x_{N,j})$ be the optimal solution of this problem. We construct $\hat{g}_i^{(N)}(\cdot)$ as a nondecreasing step function assigning the optimal solution to all arguments in the interval $(t_{N,j}, t_{N+1,j}]$. Further, we construct the approximation to $F_i(y_i)$ by setting

$$\hat{F_i}^{(N)}(y_i) = 1 - e^{-\int\limits_{-\infty}^{y_i} \hat{g}_i^{(N)}(t)dt}.$$

Now let us observe that the function

$$\log\left(1 - \hat{F_i}^{(N)}(y_i)\right) = -\int\limits_{-\infty}^{y_i} \hat{g_i}^{(N)}(t)dt,$$

is piecewise linear and concave. Assume that the function consists of a finite number J_i of linear pieces given by the following equations:

$$a_{ij}^T z + b_{ij}, \quad j = 1, \ldots, J_i,\, i = 1, \ldots, m.$$

Problem (2.43) is equivalent to the following linear programming problem:

$$\begin{aligned} &\min\langle c, x\rangle \\ \text{subject to } & z_i \geq a_{ij}^T z + b_{ij}, \quad j = 1, \ldots, J_i \\ & T_i x = z_i, \quad i = 1, \ldots, m \\ & Ax = b \\ & x \geq 0. \end{aligned}$$

The solution of the latter problem is an approximate solution of the original problem.

2.5.6 A Response Surface Method

The method will be described for problem (2.26) under the assumption that the random vector Z has a continuous and log-concave distribution. This implies that the constraining function

$$G(x) = \mathbb{P}\left(g_1(x, Z) \geq 0\;,\; \ldots\;,\; g_m(x, Z) \geq 0\right)$$

is log-concave in x. The idea of the method is to approximate $G(\cdot)$ by a concave quadratic function $Q(x) = x^T T x + h^T x + q$ (T is negative definite), then solve the approximate problem, take a new feasible point, improve the approximation, solve the problem with the new approximation *etc.* One difficulty is to develop a stopping rule in order to decide whether a solution is acceptable as an optimal solution. Some ideas are discussed in [102]. The algorithm can be described as follows.

Step 1. Given a collection of points $J^k = \{x^0, \ldots, x^{k-1}\}$ of feasible points and their corresponding values $p_i = \log G(x^i)$, $i = 1, \ldots, k-1$, solve the least squares problem

$$\min \sum_{i=0}^{k-1} (p_i - \langle x^i, T^k x^i\rangle + \langle h^k x\rangle + q^k)^2.$$

with respect to T^k, h^k, q^k such that T^k is negative semi-definite.

Step 2. We construct a quadratic function

$$\langle x^i, T^k x^i\rangle + \langle h^k x\rangle + q^k$$

and solve the approximate problem

$$\begin{aligned}&\min\langle c, x\rangle\\ \text{subject to}&\langle x^i, T^k x^i\rangle + \langle h^k x\rangle + q^k \geq p\\ &Ax = b\\ &x \geq 0.\end{aligned}$$

Let x^k be an optimal solution.

Step 3. Check the stopping rule for x^k, and accept it as optimal solution, or return to Step 1.

2.5.7 Discrete Distribution

A straightforward way to solve problem (2.4) when Z has a discrete distribution is to find all p-efficient points and to process all corresponding problems (2.27) (see for example [282]). Specialized bounding-pruning techniques can be used to avoid solving all of them. For example, any feasible solution $(\tilde{u}, \tilde{w})$ of the dual (2.28) can be used to generate a lower bound for (2.27). If it is worse than the best solution found so far, we can delete the problem (2.27); otherwise it has to be included into a list of problems to be solved exactly.

For multi-dimensional random vectors Z the number of p-efficient points can be very large and their straightforward enumeration – very difficult. It would be desirable, therefore, to avoid the complete enumeration and to search for promising p-efficient points only. This is accomplished by the next method.

The cone generation method

This is a specialized method which uses the specificity of the discrete distributions. It is related to column generation methods, which have been known since the classical work [128] as extremely useful tools of large scale linear and integer programming [30, 89]. The method is based on the same idea as the primal-dual method for nonlinear constraints.

The algorithm works as follows:

Step 0. Select a p-efficient point v^0. Set $J_0 = \{0\}$, $k = 0$.

Step 1. Solve the *master problem*

$$\min\ \langle c, x\rangle \tag{2.45}$$

$$Ax \geq b,$$

$$Tx \geq \sum_{j\in J_k} \lambda_j v^j, \tag{2.46}$$

$$\sum_{j\in J_k} \lambda_j = 1,$$

$$x \geq 0,\ \lambda \geq 0. \tag{2.47}$$

Let u^k be the vector of simplex multipliers associated with the constraint (2.46).

Step 2. Calculate an upper bound for the dual functional

$$\overline{d}(u^k) = \min_{j \in J_k} \langle u^k, v^j \rangle.$$

Step 3. Find a p-efficient solution v^{k+1} of the subproblem

$$\min_{z \in \mathcal{Z}_p} \langle u^k, z \rangle$$

and calculate

$$d(u^k) = \langle v^{k+1}, u^k \rangle.$$

Step 4. If $d(u^k) = \overline{d}(u^k)$ then stop; otherwise set $J_{k+1} = J_k \cup \{k+1\}$, increase k by one and go to Step 1.

The first p-efficient point v^0 can be found by solving the subproblem in Step 3, for an arbitrary $u \geq 0$. All master problems will be solvable, if the first one is solvable, *i.e.*, if the set $\{x \in \mathbb{R}^n_+ : Ax \geq b,\ Tx \geq v^0\}$ is non-empty. If not, adding a penalty term $M\mathbb{1}^T t$ to the objective, and replacing (2.46) by

$$Tx + t \geq \sum_{j \in J_k} \lambda_j v^j,$$

with $t \geq 0$ and a very large M, is the usual remedy ($\mathbb{1}^T = [1\ 1\ \ldots\ 1]$). The calculation of the upper bound at Step 2 is easy, because one can simply select $j_k \in J_k$ with $\lambda_{j_k} > 0$ and set $\overline{d}(u^k) = (u^k)^T v^{j_k}$. At Step 3 one may search for p-efficient solutions only, due to Lemma 5.

The algorithm is finite. Indeed, the set J_k cannot grow indefinitely, because there are finitely many p-efficient points (Theorem 17). If the stopping test of Step 4 is satisfied, optimality conditions for the convex hull problem (2.17) are satisfied. Moreover $\hat{J}_k = \{j \in J_k : \langle v^j, u^k \rangle = d(u^k)\} \subseteq \hat{J}(u)$.

When the dimension of x is large and the number of rows of T small, an attractive alternative to the cone generation method is provided by *bundle methods* applied directly to the dual problem

$$\max_{u \geq 0} \Big[h(u) + d(u) \Big],$$

because at any $u \geq 0$ subgradients of h and d are readily available. For a comprehensive description of bundle methods the reader is referred to [161, 188].

Let us now focus our attention on solving the auxiliary problem in Step 3, which is explicitly written as

$$\min\{\langle u, z \rangle \mid F(z) \geq p\}, \tag{2.48}$$

where $F(\cdot)$ denotes the distribution function of Z.

Assume that the components Z_i, $i = 1, \ldots, s$, are independent. Then we can write the probabilistic constraint in the following form:

$$\ln(F(z)) = \sum_{i=1}^{s} \ln(F_i(z_i)) \geq \ln p.$$

Since we know that at least one of the solutions is a p-efficient point, with no loss of generality we may restrict the search to grid vectors z. Furthermore, by Lemma 3, we have $z_i \geq l_i$, where l_i are p-efficient points of Z_i. For integer grids we obtain a nonlinear knapsack problem:

$$\begin{aligned} \min & \sum_{i=1}^{s} u_i z_i \\ & \sum_{i=1}^{s} \ln(F_i(z_i)) \geq \ln p, \\ & z_i \geq l_i, \quad z_i \in \mathbb{Z}, \quad i = 1, \ldots, s. \end{aligned}$$

If b_i is a known upper bound on z_i, $i = 1, \ldots, s$, we can transform the above problem to a 0–1 linear programming problem:

$$\begin{aligned} \min & \sum_{i=1}^{s} \sum_{j=l_i}^{b_i} j u_i y_{ij} \\ & \sum_{i=1}^{s} \sum_{j=l_i}^{b_i} \ln(F_i(j)) y_{ij} \geq \ln p, \\ & \sum_{j=l_i}^{b_i} y_{ij} = 1, \quad i = 1, \ldots, s, \\ & y_{ij} \in \{0, 1\}, \quad i = 1, \ldots, s, \quad j = l_i, \ldots, u_i. \end{aligned}$$

In this formulation, $z_i = \sum_{j=l_i}^{b_i} j y_{ij}$.

For log-concave marginals $F_i(\cdot)$ the following compact formulation is possible. Setting $z_i = l_i + \sum_{j=l_i+1}^{b_i} \delta_{ij}$ with binary δ_{ij}, we can reformulate the problem as a 0–1 knapsack problem:

$$\begin{aligned} \min & \sum_{i=1}^{s} \sum_{j=l_i+1}^{b_i} u_i \delta_{ij} \\ & \sum_{i=1}^{s} \sum_{j=l_i+1}^{b_i} a_{ij} \delta_{ij} \geq r, \\ & \delta_{ij} \in \{0, 1\}, \quad i = 1, \ldots, s, \quad , j = l_i + 1, \ldots b_i, \end{aligned}$$

where $a_{ij} = \ln F_i(j) - \ln F_i(j-1)$ and $r = \ln p - \ln F(l)$. Indeed, by the log-concavity, we have $a_{i,j+1} \leq a_{ij}$, so there is always a solution with nonincreasing δ_{ij}, $j = l_i + 1, \ldots, b_i$. Very efficient solution methods exist for such knapsack problems [239].

If the grid $\mathcal{Z}$ is not integer we can map it to integers by numbering the possible realizations of each Z_i in an increasing order.

One advantage of the cone generation method is that we can separate the search for new p-efficient points (via (2.48)) and the solution of the 'easy' part of the problem: the master problem (2.45)–(2.47) in Step 1. Another advantage is that we do not need to generate and keep all p-efficient points.

Let us consider the optimal solution x^{low} of the convex hull problem (2.17) and the corresponding multipliers λ_j. Define $J^{\text{low}} = \{j \in J : \lambda_j > 0\}$.

If J^{low} contains only one element, the point x^{low} is feasible and therefore optimal for the disjunctive formulation (2.8). If, however, there are more positive λ's, we need to generate a feasible point. A natural possibility is to consider the *restricted disjunctive* formulation:

$$\begin{aligned} &\min \langle c, x\rangle \\ &\text{subject to } Tx \in \textstyle\bigcup_{j\in J^{\text{low}}} K_j, \\ &x \in \mathcal{D}. \end{aligned} \tag{2.49}$$

It can be solved by simple enumeration of all cases for $j \in J^{\text{low}}$:

$$\begin{aligned} &\min \langle c, x\rangle \\ &\text{subject to } Tx \geq v^j, \\ &x \in \mathcal{D}. \end{aligned} \tag{2.50}$$

In general, it is not guaranteed that any of these problems has a non-empty feasible set. To ensure that problem (2.49) has a solution, it is sufficient that the following stronger version of Assumption 2.1 holds true.

Assumption 2.2. *The set $\Lambda := \{(u, w) \in \mathbb{R}^{m+s}_{+} \mid A^T w + T^T u \leq c\}$ is nonempty and bounded.*

Indeed, then each of the dual problems (2.28) has an optimal solution, so by duality in linear programming each of the subproblems (2.50) has an optimal solution. We can, therefore, solve all of them and choose the best solution.

An alternative strategy would be to solve the corresponding upper bounding problem (2.50) every time a new p-efficient point is generated. If U_j denotes the optimal value of (2.50), the upper bound at iteration k is

$$\bar{U}^k = \min_{0\leq j\leq k} U_j. \tag{2.51}$$

This may be computationally efficient, especially if we solve the dual problem (2.28), in which only the objective function changes from iteration to iteration.

If the distribution function of Z is α-concave on the set of possible values of Z, Theorem 18 provides an alternative formulation of the upper bound problem (2.41):

$$\begin{aligned} \min\ & \langle c, x\rangle \\ \text{subject to}\ & x \in \mathcal{D} \\ & Tx \geq z, \\ & z \in \mathbb{Z}^m, \\ & z \geq \sum_{j\in J_k} \lambda_j v^j, \\ & \sum_{j\in J_k} \lambda_j = 1 \\ & \lambda_j \geq 0,\ j \in J_k. \end{aligned} \tag{2.52}$$

Problem (2.52) provides a more accurate bound than the bound (2.51), because the set of integer z dominated by convex combinations of p-efficient points in J_k is not smaller than J_k. In fact, we need to solve this problem only at the end, with J_k replaced by J^{low}.

Special algorithms for probabilistic set-covering problem are presented in [39]. Branch-and-Bound techniques are developed in [40] for the case when x is an integer vector. The methods use the algebraic description of the feasible set by p-efficient points and suggest different techniques for generating the relevant p-efficient points.

Bounds via binomial moments

If the components of Z are dependent it is difficult to evaluate the constraint function $G(\cdot)$, *e.g.*, for solving the subproblem (2.14) in the cone generation algorithm. Still, some bounds on its optimal solution may prove useful. A number of bounds are developed using only partial information on the distribution function of Z in the form of the marginal distributions:

$$F_{i_1\dots i_k}(z_{i_1},\dots,z_{i_k}) = \mathbb{P}\{Z_{i_1} \leq z_{i_1},\dots Z_{i_k} \leq z_{i_k}\}, \quad 1 \leq i_1 < \dots < i_k \leq m.$$

Since for each marginal distribution one has $F_{i_1\dots i_k}(z_{i_1},\dots,z_{i_k}) \geq F(z)$ the following relaxation of $\mathcal{Z}$ (defined by (2.5)) can be obtained.

Lemma 8. *For each $z \in \mathcal{Z}_p$ and for every $1 \leq i_1 < \dots < i_k \leq s$ the following inequality holds true:*

$$F_{i_1\dots i_k}(z_{i_1},\dots,z_{i_k}) \geq p.$$

We can determine probability bounds by solving certain linear optimization problem (see ([55, 279, 280]). The following result is known:

Theorem 24. *For any distribution function $F : \mathbb{R}^m \to [0,1]$ and any $1 \leq k \leq m$, at every $z \in \mathbb{R}^m$ the optimal value of the following linear programming problem:*

$$\begin{array}{rcl}
 & \max\ v_m & \\
v_0+v_1+v_2+v_3+\cdots+v_m & = & 1 \\
v_1+2v_2+3v_3+\cdots+mv_m & = & \sum_{1\le i\le m} F_i(z_i) \\
v_2+\binom{3}{2}v_3+\cdots+\binom{m}{2}v_m & = & \sum_{1\le i_1<i_2\le m} F_{i_1i_2}(z_{i_1},z_{i_2}) \\
 & \vdots & \\
v_k+\binom{k+1}{k}v_{k+1}+\cdots+\binom{m}{k}v_m & = & \sum_{1\le i_1<\ldots<i_k\le m} F_{i_1\ldots i_k}(z_{i_1},\ldots,z_{i_k}) \\
v_0\ge 0,\ v_1\ge 0,\ \ldots,\ v_m\ge 0. & &
\end{array} \tag{2.53}$$

provides an upper bound for $F(z_1,\ldots,z_m)$.

We can use this result to bound the objective function in problem (2.48).

Proposition 1. *Let $Z=(Z_1,\ldots,Z_m)$ be an integer random vector and let $F_{i_1,\ldots,i_k}$ denote its marginal distribution functions. Then for every $p\in(0,1)$ and for every $1\le k\le m$ the optimal value of the problem*

$$\begin{array}{rcl}
 & \min\ \langle u,z\rangle & \\
v_0+v_1+v_2+v_3+\cdots+v_m & = & 1 \\
v_1+2v_2+3v_3+\cdots+mv_m & = & \sum_{1\le i\le m} F_i(z_i) \\
v_2+\binom{3}{2}v_3+\cdots+\binom{m}{2}v_m & = & \sum_{1\le i_1<i_2\le m} F_{i_1i_2}(z_{i_1},z_{i_2}) \\
 & \vdots & \\
v_k+\binom{k+1}{k}v_{k+1}+\cdots+\binom{m}{k}v_m & = & \sum_{1\le i_1<\ldots<i_k\le m} F_{i_1\ldots i_k}(z_{i_1},\ldots,z_{i_k}) \\
\multicolumn{3}{l}{v_0\ge 0,\ v_1\ge 0,\ \ldots,\ v_{s-1}\ge 0,\ v_m\ge p,\quad z_1\ge l_1,\ z_2\ge l_2,\ \ldots,\ z_m\ge l_m,} \\
 & z\in\mathbb{Z}^m &
\end{array} \tag{2.54}$$

provides a lower bound on the optimal value $d(u)$ given by (2.48).

Proof. If $z\in\mathcal{Z}$, that is, $F(z)\ge p$, then the optimal value of (2.53) satisfies $v_m\ge p$. Thus z and the solution v of (2.53) are feasible for (2.54). Since the objective functions of (2.48) and (2.54) are the same, the result follows. □

Problem (2.54) is a nonlinear mixed-integer problem. Its advantage over the original formulation is that it uses marginal functions in an explicit way which allows for the development of specialized solution methods.

2.5.8 Probabilistic Valid Inequalities

A relation between probabilistic constraints and the theory of valid inequalities in integer programming has been developed in [311].

We shall just sketch some ideas in this direction. Let us assume that the distribution of Z is approximated by finitely many scenarios $z_1, \dots, z_S$ having probabilities $p_1, \dots, p_S$. Under mild assumptions problem (2.1) can be converted to a mixed-integer programming problem

$$\begin{aligned}
&\min f(x) && (2.55)\\
&\text{subject to } g(x, z_s) + M v_s \geq 0, \quad s = 1, \dots, S, && (2.56)\\
&\sum_{s=1}^{S} p_s v_s \leq 1 - p, && (2.57)\\
&x \in \mathcal{D},\\
&v_s \in \{0, 1\}, \quad s = 1, \dots, S, && (2.58)
\end{aligned}$$

where M is a vector with sufficiently large components, so $v_s = 1$ makes (2.56) trivial. Each binary variable v_s indicates whether the current solution x violates the constraint $g(x, z_s) \geq 0$ or not, and the probability constraint takes on the form of the knapsack inequality (2.57).

In many applications it is possible to determine a partial order '$\preceq$' in the set of scenarios z_s, representing their difficulty for the constraints $g(x, z_s) \geq 0$. In the simplest setting, we shall have $z_s \preceq z_\sigma$ (z_s is easier than z_σ) if

$$g(x, z_\sigma) \geq 0 \;\Rightarrow\; g(x, z_s) \geq 0, \quad \text{for all } x \in X.$$

Then the mixed-integer formulation (2.55)–(2.58) can be augmented with the *precedence constraint*:

$$v_s \leq v_\sigma \text{ if } z_s \preceq z_\sigma.$$

Probabilistic valid inequalities for the binary variables v_s are developed on the basis of this structure. For each scenario z_s we define the set of comparable scenarios which are at least as hard as z_s:

$$A_s = \{z_j : z_s \preceq z_j\}.$$

If we fail to satisfy the constraint $g(x, z_s) \geq 0$ for scenario s, we shall fail for all scenarios in A_s, *i.e.*, $v_j = 1$ for all $z_j \in A_s$. In this way probabilistic counterparts of the concepts of a cover and cover inequalities known from integer programming are introduced. In our setting, a set $C \subset \{1, \dots, S\}$ is an *induced cover* if

$$\mathbb{P}\left\{\bigcup_{s \in C} A_s\right\} > 1 - p.$$

If $v_s = 1$ for all $s \in C$, then we must have $v_j = 1$ for all z_j in the union of the sets A_s, $s \in C$, and the probability constraint (2.57) will be violated. Therefore the following *induced cover inequality* must hold true:

$$\sum_{s \in C} v_s \leq |C| - 1.$$

The second implication of the partial order is that we do not need to enforce inequality (2.56) for all scenarios. By a similar argument, if $g(x, z_s) \geq 0$, then $g(x, z_\sigma) \geq 0$ for all $z_\sigma \preceq z_s$. Thus, we may try to determine a set $\mathcal{L}$ of *critical scenarios*, similarly to the set of p-efficient points of the problem with the random right hand side.

These two basic ideas can be put together to formulate the following approximate problem:

$$\begin{aligned}
\min\ & f(x) \\
\text{subject to } & g(x, z_s) + Mv_s \geq 0 \quad s \in \mathcal{L}, \\
& \sum_{s=1}^{S} p_s v_s \leq 1 - p, \\
& x \in X, \\
& v_s \leq v_\sigma, \quad \text{if } z_s \preceq z_\sigma \\
& v_s \in [0,1], \quad s = 1, \ldots, S, \\
& \sum_{s \in C} v_s \leq |C| - 1.
\end{aligned}$$

For a detailed description of this solution technique we refer to [311].

2.6 Cash Matching with Probabilistic Liquidity Constraints

There are many publication addressing interesting applications of probabilistic constraints. We do not attempt to address the potential of probabilistic optimization for solving applied problems. We return to a version of our starting example because our duality theory finds an interesting interpretation in its context.

We consider the following cash matching problem. We have random liabilities L_t in periods $t = 1, \ldots, T$ and a basket of n bonds. The payment of bond i in period t is denoted by a_{it}. It is zero for t before the purchase of the bond and for t greater than the maturity time of the bond. At the time of purchase a_{it} is the negative of the price of the bond, at the following periods it is equal to the coupon payment, and at the time of maturity it is equal to the face value plus the coupon payment. Our initial capital equals c_0.

The objective is to design a bond portfolio such that the probability of covering the liabilities over the entire period $1, \ldots, T$ is at least p. Subject to this condition, we want to maximize the final cash on hand, guaranteed with probability p.

Let us introduce the cumulative liabilities

$$Z_t = \sum_{\tau=1}^{t} L_\tau, \quad t = 1, \ldots, T.$$

Denoting by x_i the amount invested in bond i, we observe that the cumulative cash flows up to time t, denoted c_t, can be expressed as follows:

$$c_t = c_{t-1} + \sum_{i=1}^{n} a_{it} x_i, \quad t = 1, \dots, T.$$

Using cumulative cash flows and cumulative liabilities permits the carry-over of capital from one stage to the next one, while keeping the random quantities at the right hand side of the constraints. The problem takes on the form

$$\begin{aligned} &\max c_T \\ &\text{subject to } \mathbb{P}\big[c_t \geq Z_t,\ t = 1, \dots, T\big] \geq p, \\ &\qquad c_t = c_{t-1} + \sum_{i=1}^{n} a_{it} x_i, \quad t = 1, \dots, T, \\ &\qquad x \geq 0. \end{aligned}$$

Let us observe that first constraint of this problem is a probabilistic liquidity constraint. If the vector Z has a quasi-concave distribution (*e.g.*, joint normal distribution), the resulting problem is convex. Thus both the dual method form Section 2.5.3 and the primal-dual method from Section 2.5.4 yield optimal solutions of the problem.

The convex hull problem (2.17) can be now written as follows:

$$\max c_T$$

$$\text{subject to } c_t = c_{t-1} + \sum_{i=1}^{n} a_{it} x_i, \quad t = 1, \dots, T, \tag{2.59}$$

$$c_t \geq \sum_{j=1}^{T+1} \lambda_j v_t^j, \quad t = 1, \dots, T, \tag{2.60}$$

$$\sum_{j=1}^{T+1} \lambda_j = 1, \tag{2.61}$$

$$\lambda \geq 0,\ x \geq 0. \tag{2.62}$$

In constraint (2.60) the vectors $v^j = (v_1^j, \dots, v_T^j)$, for $j = 1, \dots, T+1$, are p-efficient trajectories of the cumulative liabilities $Z = (Z_1, \dots, Z_T)$. Constraints (2.60)–(2.62) require that the cumulative cash flows are greater than or equal to a convex combination of p-efficient trajectories. Recall that by Lemma 4, no more than $T+1$ p-efficient trajectories are needed. Unfortunately, we do not know the optimal collection of these trajectories.

Let us assign non-negative Lagrange multipliers $u = (u_1, \dots, u_T)$ to the constraint (2.60), multipliers $w = (w_1, \dots, w_T)$ to the constraints (2.59), and a multiplier $\rho \in \mathbb{R}$ to the constraint (2.61). For the convenience of notation we introduce the constant $w_{T+1} = 1$. The dual problem becomes

$$\min\ c_0 w_1 - \rho \tag{2.63}$$

$$\text{subject to } w_t = w_{t+1} + u_t, \quad t = T, T-1, \dots, 1, \tag{2.64}$$

$$\sum_{t=1}^{T} w_t a_{it} \le 0, \quad i = 1, \dots, n, \tag{2.65}$$

$$\rho \le \sum_{t=1}^{T} u_t v_t^j, \quad j = 1, \dots, T+1. \tag{2.66}$$

We can observe that each dual variable u_t is the cost of borrowing a unit of cash for one time period, t. The amount u_t is to be paid at the end of the planning horizon. It follows from (2.64) that each multiplier w_t is the amount that has to be returned at the end of the planning horizon if a unit of cash is borrowed at t and held till T.

The constraints (2.65) represent the *non-arbitrage condition.* For each bond i we can consider the following operation: borrow money to buy the bond and lend away its coupon payments, according to the rates implied by w_t's. At the end of the planning horizon, we collect all loans and pay off the debt. The profit from this operation should be non-positive for each bond, and this is represented by (2.65).

Let us observe that each product $u_t v_t^j$ is the the amount that has to be paid at the end, for having a debt in the amount v_t^j in period t. Recall that v_t^j is the p-efficient cumulative liability up to time t. Denote the implied one-period liabilities by

$$\begin{aligned} L_t^j &= v_t^j - v_{t-1}^j, \quad t = 2, \dots, T, \\ L_1^j &= v_1^j. \end{aligned}$$

Changing the order of summation, we obtain

$$\sum_{t=1}^{T} u_t v_t^j = \sum_{t=1}^{T} u_t \sum_{\tau=1}^{t} L_\tau^j = \sum_{\tau=1}^{T} L_\tau^j \sum_{t=\tau}^{T} u_t = \sum_{\tau=1}^{T} L_\tau^j (w_\tau - 1).$$

It follows that the sum appearing at the right hand side of (2.66) is the extra cost of covering the jth p-efficient liability sequence by borrowed money, that is, the difference between the amount that has to returned at the end of the planning horizon, and the total liability. The variable ρ, therefore, represents the minimal cost of this form, for all p-efficient trajectories. This allows us to interpret the dual objective function (2.63) as the amount obtained at T for lending away our capital c_0 decreased by the extra cost of covering a p-efficient liability sequence by borrowed money. By duality this quantity is the same as c_T, which implies that both ways of covering the liabilities are equally profitable.

To observe the work of the methods we have used data on 72 government bonds and AAA corporate bonds ranging from 6-month treasury bills (which do not pay coupons, but sell at discount) to 5-year bonds paying coupons each 6-months. The liabilities were assumed to be normally distributed with expectation 2,000,000 and standard deviation 100,000. The initial capital was $c_0 = 20,000,000$ and the number of 6-month periods $T = 10$. The probability $p = 0.95$. To facilitate the numerical solution of the method, the distribution of the liabilities was approximated by $N = 100$ equally likely scenarios.

The dual and the primal-dual methods were used to solve the problem. The search for new p-efficient points in both methods was implemented as a simple binary optimization problem with a knapsack constraint. Other subproblems were solved by the CPLEX linear programming solver.

The dual method terminated after 34 iterations finding the optimal portfolio of 9 bonds of different maturities. The primal-dual method found exactly the same solution after just 3 iterations. In both cases the computation time on a 1.7GHz PC was less than one minute.

The key element of both methods is the subproblem for generating p-efficient points.

The problem at hand was linear, and therefore both methods were equally easy to implement. If the functions f and g are nonlinear, one iteration of the primal-dual method requires more computational effort than one iteration of the dual method.

Acknowledgement. This research is supported by the NSF awards 0354500 and 0303728.

3

Theoretical Framework for Comparing Several Stochastic Optimization Approaches

James C. Spall, Stacy D. Hill, and David R. Stark

The Johns Hopkins University, Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, Maryland 20723-6099 USA
{james.spall,stacy.hill,david.stark}@jhuapl.edu

Summary. In this chapter, we establish a framework for formal comparisons of several leading optimization algorithms, providing guidance to practitioners for when to use or not use a particular method. The focus in this chapter is five general algorithm forms: random search, simultaneous perturbation stochastic approximation, simulated annealing, evolution strategies, and genetic algorithms. We summarize the available theoretical results on rates of convergence for the five algorithm forms and then use the theoretical results to draw some preliminary conclusions on the relative efficiency. Our aim is to sort out some of the competing claims of efficiency and to suggest a structure for comparison that is more general and transferable than the usual problem-specific numerical studies.

3.1 Introduction

To address the shortcomings of classical deterministic algorithms, a number of powerful optimization algorithms with embedded randomness have been developed. The population-based methods of evolutionary computation, for example, are one class among many of the available *stochastic* optimization algorithms. Hence, a user facing a challenging optimization problem for which a stochastic optimization method is appropriate meets the daunting task of determining which algorithm is appropriate for a given problem. This choice is made more difficult by some dubious claims that have been made about some popular algorithms. An inappropriate approach may lead to a large waste of resources, both from the view of wasted efforts in implementation and from the view of the resulting suboptimal solution to the optimization problem of interest.

Hence, there is a need for objective analysis of the relative merits and shortcomings of leading approaches to stochastic optimization. This need has certainly been recognized by others, as illustrated, for example, in recent conferences on evolutionary computation, where numerous sessions are devoted to comparing algorithms. Nevertheless, virtually all comparisons have been numerical tests on specific problems. For example, a large fraction of the book [323] is devoted to numerical comparisons. Although sometimes enlightening, such comparisons are severely limited in the *general* insight they provide. Some comparisons for *noisy* evaluations of a simple spherical loss function are given in [15], Chapter 6; however, some of the competitors were implemented in non-standard forms, making the results difficult to interpret for an analyst using a more conventional implementation. Spall [341] also has a number of comparisons (theoretical and numerical) for the cases of noise-free and noisy loss evaluations. At the other end of the spectrum are the 'no free lunch' theorems, [399], which simultaneously consider all possible loss functions and thereby draw conclusions that have limited practical utility since one always has at least *some* knowledge of the nature of the loss function being minimized.

Our aim in this chapter is to lay a framework for a *theoretical* comparison of efficiency applicable to a broad class of practical problems where some (incomplete) knowledge is available about the nature of the loss function. We will consider five basic algorithm forms: random search, simultaneous perturbation stochastic approximation (SPSA), simulated annealing (SAN), and two forms of evolutionary computation (evolution strategy and genetic algorithms). The basic optimization problem corresponds to finding an optimal point θ^*:

$$\theta^* = \arg\min_{\theta \in \Theta} L(\theta),$$

where $L(\theta)$ is the loss function to be minimized, Θ is the domain over which the search will occur, and θ is a p-dimensional vector of parameters. We are mainly interested in the case where θ^* is a *unique* global minimum.

Although stochastic optimization approaches other than the five above exist, we are restricting ourselves to the five general forms in order to be able to make tangible progress (note that there are various specific implementations of each of these general algorithm forms). These five algorithms are general-purpose optimizers with powerful capabilities for serious multivariate optimization problems. Further, they have in common the requirement that they only need measurements of the objective function, not requiring derivative information (gradient or Hessian) for the loss function. It is the long-term expectation that this theoretical framework will provide guidance to those faced with an optimization problem and the associated difficult choice of selecting a suitable method. It is critical to make an informed choice *prior* to investing the considerable resources required given the inherent difficulties in implementing a particular algorithm in a large-scale practical problem (software implementation, data preparation, algorithm tuning, *etc.*).

Central to the approach of this contribution will be the known theoretical analysis on the rate of convergence of each of the candidate algorithms. Our approach will be built as much as possible on *existing* theory characterizing the rates of convergence for the algorithms to perform the comparative analysis. There appears to be no previous analysis putting the theoretical results on a common basis for performing an objective comparison. Of course, this approach has limitations in general because many algorithms have little – or possibly no – theoretical justification. Nonetheless, it is our expectation that performing a formal theoretical comparison of the chosen algorithms will shed light on relative performance of other similar algorithms as well, even if the similar algorithms lack the same current level of theoretical justification.

One might ask whether questions of relative efficiency are relevant in light of the 'no free lunch (NFL)' theorems of [399] and others. The NFL theorems state, in essence, that the expected performance of any pair of optimization algorithms across all possible problems is identical. In practice, of course, one is not interested in solving 'all possible problems,' as there is usually some prior information about the problems of interest and this prior information will affect the algorithm implementation. Hence, the NFL results may not adequately reflect the performance of candidate algorithms as they are actually applied. In other words, some algorithms *do* work better than others on problems of interest. Nevertheless, the NFL results are an important backdrop against which to view the results here, providing limits on the extent to which one algorithm can be claimed as 'better' than another.

In Sections 3.2 through 3.5, we discuss the known convergence rate results on the five algorithm forms under consideration. Section 3.6 then uses these results to provide a theoretical framework for comparison. We demonstrate these results in analyzing the relative efficiency as the problem dimension increases.

3.2 Simple Global Random Search

We first establish a rate of convergence result for the simplest ('blind') random search method where we repeatedly sample over the domain of interest, $\Theta \subseteq \mathbb{R}^p$. This can be done in recursive form or in 'batch' (non-recursive) form by simply laying down a number of points in Θ and taking as our estimate of θ^* that value of θ yielding the lowest L value. A recursive implementation of this idea is as follows.

Step 0 (Initialization). Pick an initial value of θ, say $\hat{\theta}_0$, according to prior information or some probability distribution on the domain Θ. Calculate $L(\hat{\theta}_0)$. Set $k = 0$.

Step 1. Generate a new independent value of θ, say $\theta_{\text{new}}(k)$, according to the chosen probability distribution. If $L(\theta_{\text{new}}(k)) < L(\hat{\theta}_k)$, set $\hat{\theta}_{k+1} = \theta_{\text{new}}(k)$. Else take $\hat{\theta}_{k+1} = \hat{\theta}_k$.

Step 2. Repeat Step 1 until the maximum allowable number of L evaluations has been reached or the user is otherwise satisfied with the current estimate of θ^*.

It is well known that the random search algorithm above will converge in some stochastic sense under modest conditions (see, *e.g.*, [338]). A typical convergence theorem is of the following form (proof in [341], Section 2.2).

Theorem 1. *Suppose that θ^* is the unique minimizer of L on the domain Θ in the sense that $L(\theta^*) = \inf_{\theta\in\Theta} L(\theta)$ and $\inf\{L(\theta) : \|\theta - \theta^*\| \geq \varepsilon\} > L(\theta^*) > -\infty$ for all $\varepsilon > 0$. Suppose further that for any $\varepsilon > 0$ and $\forall k$, there exists a $\delta(\varepsilon) > 0$ such that*

$$\mathbb{P}\{\theta_{\text{new}}(k) : L(\theta_{\text{new}}(k)) < L(\theta^*) + \varepsilon\} \geq \delta(\varepsilon).$$

Then, for the random search algorithm, $\hat{\theta}_k \to \theta^*$ a.s. (almost surely) as $k \to \infty$.

While the above theorem establishes convergence of the simple random search algorithm, it is also of interest to examine the *rate* of convergence. The rate is intended to tell the analyst how close $\hat{\theta}_k$ is likely to be to θ^* for a given cost of search. The cost of search here will be expressed in terms of number of loss function evaluations. Knowledge of the rate is critical in practical applications as simply knowing that an algorithm will eventually converge begs the question of whether the algorithm will yield a practically acceptable solution in any reasonable period. To evaluate the rate, let us specify a 'satisfactory region' $S(\theta^*)$ representing some neighborhood of θ^* providing acceptable accuracy in our solution (*e.g.*, $S(\theta^*)$ might represent a hypercube about θ^* with the length of each side representing a tolerable error in each coordinate of θ). An expression related to the rate of convergence of the above simple random search algorithm is then given by

$$\mathbb{P}\{\hat{\theta}_k \in S(\theta^*)\} = 1 - (1 - \mathbb{P}\{\theta_{\text{new}}(k) \in S(\theta^*)\})^k \tag{3.1}$$

We will use this expression in Section 3.6 to derive a convenient formula for comparison of efficiency with other algorithms.

3.3 Simultaneous Perturbation Stochastic Approximation

The next algorithm we consider is SPSA. This algorithm is designed for continuous variable optimization problems. Unlike the other algorithms here, SPSA is fundamentally oriented to the case of *noisy* function measurements and most of the theory is in that framework. This will make for a difficult comparison with the other algorithms, but Section 3.6 will attempt a comparison nonetheless. The SPSA algorithm works by iterating from an initial guess of the optimal θ, where the iteration process depends on a highly efficient 'simultaneous perturbation' approximation to the gradient $g(\theta) \equiv \partial L(\theta)/\partial\theta$.

Assume that measurements $y(\theta)$ of the loss function are available at any value of θ:

$$y(\theta) = L(\theta) + noise.$$

For example, in a Monte Carlo simulation-based optimization context, $L(\theta)$ may represent the mean response with input parameters θ, and $y(\theta)$ may represent the outcome of one simulation experiment at θ. In some problems, exact loss function measurements will be available; this corresponds to the *noise* $= 0$ setting (and in the simulation example, would correspond to a deterministic, non-Monte Carlo, simulation). Note that no direct measurements (with or without noise) of the gradient of $L(\theta)$ are assumed available.

It is assumed that $L(\theta)$ is a differentiable function of θ and that the minimum point θ^* corresponds to a zero point of the gradient, *i.e.*,

$$g(\theta^*) \;=\; \left.\frac{\partial L(\theta)}{\partial \theta}\right|_{\theta=\theta^*} \;=\; 0. \tag{3.2}$$

In cases where more than one point satisfies (3.2), there exists theory that ensures that the algorithm will converge to the global minimum, [220]. (As a consequence of the basic recursive form of the algorithm there is generally not a risk of converging to a maximum or saddlepoint of $L(\theta)$, *i.e.*, to non-minimum points where $g(\theta)$ may equal zero.) Another extension of SPSA to global optimization is discussed in [88]. The SPSA procedure has the general recursive SA form:

$$\hat{\theta}_{k+1} \;=\; \hat{\theta}_k \;-\; a_k \hat{g}_k(\hat{\theta}_k),$$

where $\hat{g}_k(\hat{\theta}_k)$ is the estimate of the gradient $g(\theta)$ at the iterate $\hat{\theta}_k$ based on the above-mentioned measurements of the loss function and $a_k > 0$ is a 'gain' sequence. This iterate can be shown to converge under reasonable conditions (*e.g.*, [341] Section 7.3, and [112] for local convergence; [220] for global convergence). The core gradient approximation is

$$\hat{g}_k\left(\hat{\theta}_k\right) \;=\; \frac{y\left(\hat{\theta}_k \;+\; c_k \Delta_k\right) \;-\; y\left(\hat{\theta}_k \;-\; c_k \Delta_k\right)}{2c_k} \begin{bmatrix} \Delta_{k1}^{-1} \\ \Delta_{k2}^{-1} \\ \vdots \\ \Delta_{kp}^{-1} \end{bmatrix}, \tag{3.3}$$

where c_k is some 'small' positive number and the user-generated p-dimensional random perturbation vector, $\Delta_k = [\Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kp}]^T$, contains $\{\Delta_{ki}\}$ that are independent and symmetrically distributed about 0 with finite inverse moments $\mathbb{E}(|\Delta_{ki}|^{-1})$ for all k, i. One particular distribution for Δ_{ki} that satisfies these conditions is the symmetric Bernoulli ±1 distribution; two common distributions that do *not* satisfy the conditions (in particular, the critical finite inverse moment condition) are uniform and normal. The essential basis for efficiency of SPSA in multivariate problems is apparent in (3.3), where only two measurements of the loss function are needed to estimate the p-dimensional gradient vector for any p; this contrasts with the standard finite difference method of gradient approximation, which requires $2p$ measurements.

Most relevant to the comparative analysis goals of this chapter is the asymptotic distribution of the iterate. This was derived in [339], with further developments in [88,112,340]. Essentially, it is known that under appropriate conditions,

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} \mathcal{N}(\mu, \varSigma) \text{ as } k \to \infty, \tag{3.4}$$

where $\beta > 0$ depends on the choice of gain sequences (a_k and c_k), μ depends on both the Hessian and the third derivatives of $L(\theta)$ at θ^* (note that in general, $\mu \neq 0$ in contrast to many well-known asymptotic normality results in estimation), and $\varSigma$ depends on the Hessian matrix at θ^* and the variance of the noise in the loss measurements. Given the restrictions on the gain sequences to ensure convergence and asymptotic normality, the fastest allowable value for the rate of convergence of $\hat{\theta}_k$ to θ^* is $k^{-1/3}$. This contrasts with the fastest allowable rate of $k^{-1/2}$ for gradient-based algorithms such as Robbins-Monro SA.

Unfortunately, (3.4) is not directly usable in our comparative studies here since the other algorithms being considered here appear to have formal results for convergence rates only for the case of *noise-free* loss measurements. The authors are unaware of any general asymptotic distribution result for the noise-free case (note that it is *not* appropriate to simply let the noise level go to zero in (3.4) in deriving a result for the noise-free case; it is likely that the rate factor β will also change if an asymptotic distribution exists). Some partial results, however, are available that are related to the rate of convergence. Gerencsér [137] established that the moments $\left[\mathbb{E}\left(\left\|\hat{\theta}_k - \theta^*\right\|^q\right)\right]^{1/q}$ converge to zero at a rate of $k^{-1/2}$ for any $q > 0$, when a_k has the standard $1/k$ decay rate. More recently, Gerencsér and Vágó [138] established that the noise-free SPSA algorithm has a geometric rate of convergence when *constant* gains $a_k = a$ are used. In particular, for functions having bounded third derivatives, they show for sufficiently small a,

$$\limsup_{k\to\infty} \frac{\left\|\hat{\theta}_k - \theta^*\right\|}{\eta^k} = 1 \quad \text{a.s.}$$

for some $0 < \eta < 1$. Gerencsér and Vágó [138] go further for quadratic loss functions by specifying η in terms of a and the Hessian matrix of L. Unfortunately, even in the quadratic case, η is not fully specified in terms of quantities associated with L and the algorithm itself (*i.e.*, η depends on unknown constants).

3.4 Simulated Annealing Algorithms

The simulated annealing (SAN) method [187,226] was originally developed for optimization over discrete finite sets. The Metropolis SAN method produces a sequence that converges in probability to the set of global minima of the loss function as T_k, the *temperature*, converges to zero at an appropriate rate.

Gelfand and Mitter [134] present a SAN method for continuous parameter optimization. They obtained discrete-time recursions (which are similar to a stochastic approximation algorithm) for Metropolis-type SAN algorithms that, in the limit, optimize continuous parameter loss functions. Spall ([341] Section 8.6) summarizes this connection of SAN to SA in greater detail. Suppose that $\hat{\theta}_k$ is such a Metropolis-type SAN sequence for optimizing L. To define this sequence, let $q_k(x,\cdot)$ be the p-dimensional Gaussian density function with mean x and variance $b_k^2\sigma_k^2(x)I_p$, where $\sigma_k^2(x) = \max\{1, a_k^\tau\|x\|\}$, τ is fixed in the range $0 < \tau < 1/4$, and $a_k = a/k$ for large k, with $a > 0$. (Observe that $\sup\left\{\sigma_k^2(x), x \in A\right\} \to 1$ as $k \to \infty$ for any bounded set A.) Also, let

$$s_k(x,y) = \begin{cases} \exp\left(-\frac{L(y)-L(x)}{T_k}\right) & \text{if } L(y) > L(x) \\ 1 & \text{otherwise,} \end{cases}$$

where $T_k(x) = b_k^2\sigma_k^2(x)\big/(2a_k)$. The function $s_k(x,y)$ is the *acceptance probability*, as in the usual Metropolis algorithm.

The SAN sequence can be obtained through simulation, in a manner similar to the discrete case:

Step 1. Let $\hat{\theta}_k$ be the current state.
Step 2. Generate a candidate solution $\tilde{\theta}$ according to (the one-step Markov transition) probability density $q_k(\hat{\theta}_k,\cdot)$.
Step 3. Let $\delta_k = L(\tilde{\theta}) - L(\hat{\theta}_k)$. (Then $s_k(\hat{\theta}_k,\tilde{\theta}) \le 1$, where $s_k(\hat{\theta}_k,\tilde{\theta}) = 1$ if $\delta_k \le 0$). Let $\hat{\theta}_{k+1} = \tilde{\theta}$ if $\delta_k \le 0$. Otherwise, consider an independent random variable U_k uniformly distributed on the interval $[0,1]$. Let $\hat{\theta}_{k+1} = \hat{\theta}_k$ if $s_k(\hat{\theta}_k,\tilde{\theta}) > U_k$.

The resulting sequence $\hat{\theta}_k$ has Markov transition probabilities

$$\mathbb{P}\left\{\hat{\theta}_{k+1} \in A \middle| \hat{\theta}_k = x\right\} = \int_A p_k(y|x)dy,$$

where

$$p_k(y|x) = q_k(x,y)s_k(x,y) + r_k(x)\delta(y-x)$$

and $\delta(\cdot)$ is the Dirac-delta function.

Let $\{W_k\}$ be an i.i.d. sequence of p-dimensional standard Gaussian random vectors and let the sequence $\xi_0, \xi_1, \ldots$ be defined by setting

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k(g(\hat{\theta}_k) + \xi_k) + b_k W_k \text{ a.s., } k > 0. \tag{3.5}$$

The reason for introducing this form for the recursion is to show that $\hat{\theta}_k$ converges in probability to the set of global minima of L. This can be shown if we can show that the sequence $\hat{\theta}_k^x$ is tight, where $\hat{\theta}_k^x$ denotes the solution to

(3.5) with initial condition $\hat{\theta}_0 = x$. If $\hat{\theta}_k^x$ is tight, then it can be established that $\hat{\theta}_k^x$ converges in probability, uniformly in x, for x belonging to a compact set K. The limiting distribution is given by the loss function L. In particular, it is the uniform measure on the set of global minima of L. Thus, the main reason for introducing ξ_k is to facilitate the proof of tightness of $\hat{\theta}_k^x$. The sequence $\hat{\theta}_k^x$ is tight under certain restrictions on the sequences a_k and b_k, namely that $a_k = a/k$ (as mentioned above) and $b_k = b/\sqrt{k \log\log k}$ for large k, where a and b are positive constants.

The algorithm is a Metropolis algorithm in the usual sense (*i.e.*, as in the discrete case where the temperature sequence is independent of the state) if almost all $\hat{\theta}_k$ lie in some fixed compact set for all $k > K$, for some $K > 0$, since eventually $\sigma_k^2(\hat{\theta}_k) = 1$. (This assertion follows directly from steps in the proof of Lemma 2(a) in [134], page 121). The sequence $\{\hat{\theta}_k\}$ converges in probability to the global minimum of the loss function. If there is a unique global minimizer θ^*, then the sequence converges in probability to θ^*. To be specific, suppose that $L(\theta)$ has a unique minimum at θ^* and let $S(\theta^*)$ be a neighborhood of θ^*. Gelfand and Mitter [134] show that $\mathbb{P}\{\hat{\theta}_k \in S(\theta^*)\} \to 1$ as $k \to \infty$.

Furthermore, like SPSA, SAN has an asymptotic normality result (but unlike SPSA, this result applies in the noise-free case). In particular, following [403], assume that $a_k = a/k$, $b_k = (b/(k^\gamma \log(k^{1-\gamma} + B_0))^{1/2}$, where B_0, a, and b are positive constants, $0 < \gamma < 1$. Let $H(\theta^*)$ denote the Hessian of $L(\theta)$ evaluated at θ^* and let I_p denote the $p \times p$ identity matrix. Yin [403] showed that

$$[\log(k^{1-\gamma} + B_0)]^{1/2}(\hat{\theta}_k - \theta^*) \to \mathcal{N}(0, \Sigma) \text{ in distribution,}$$

where $\Sigma H + H^T \Sigma + (b/a)I = 0$.

3.5 Evolutionary Computation

3.5.1 General Principles and Theory

Evolutionary computation (EC) represents a class of stochastic search and optimization algorithms that use a Darwinian evolutionary model. The principle feature of an EC algorithm is the search through a population of candidate solutions for the optimal value of a loss function. There are three general approaches in evolutionary computation, namely evolutionary programming (EP), evolution strategies (ES) and genetic algorithms (GA). All three approaches work with a population of candidate solutions and randomly alter the solutions over a sequence of generations according to evolutionary operations of competitive selection, mutation and sometimes recombination (reproduction). The fitness of each population element to survive into the next generation is determined by a selection scheme based on evaluating the loss function for each element of the population. The selection scheme is such that the most favorable elements of the population tend to survive into the next generation while the unfavorable elements tend to perish.

The principal differences in the three approaches are the selection of evolutionary operators used to perform the search and the computer representation of the candidate solutions. EP uses selection and mutation only to generate new solutions. While both ES and GA use selection, recombination and mutation, recombination is used more extensively in GA. A GA traditionally performs evolutionary operations using binary encoding of the solution space, while EP and ES perform the operations using real-coded solutions. The GA also has a real-coded form and there is some indication that the real-coded GA may often be more efficient and provide greater precision than the binary-coded GA ([341], Chapters 9 and 10). The distinction among the three approaches has begun to blur as new hybrid versions of EC algorithms have arisen.

The formal convergence of EC algorithms to the optimal θ^* has been considered in a number of references. Eiben *et al.* [118] derived a convergence in probability result for an elitist GA using the powerful tools of Markov chain theory assuming a finite search space. This result characterized the convergence properties of the GA in terms of the selection, mutation, and recombination probabilities. Rudolph [306] analyzed the basic GA in the binary search space, the canonical GA, without elitist selection. He found that the canonical GA will never converge to the global optimum, and that convergence for this GA comes only by saving the best solution obtained over the course of the search. For function optimization it makes sense to keep the best solution obtained over the course of the search, so convergence is guaranteed. For GA in the binary search space, convergence results that assume a finite search space seem of little practical use; since there are a finite number of points to search, random search and simple enumeration are also guaranteed to converge. However since convergence is a precondition for convergence rate calculations, convergence results assuming a finite search space are not entirely meaningless. Rudolph [309] summarizes the sufficient conditions on the mutation, recombination, and selection probabilities for convergence of EC algorithms in finite search spaces, with a simplified mathematical structure that does not rely on Markov chain theory. Reeves and Rowe [293] and Spall [341], Chapter 10 include a review and further references related to EC convergence theory.

Convergence analysis for EP, ES, and real-valued GA often relies on the Borel-Cantelli Lemma (see for example [21]). The convergence proofs for these algorithms assume that the mutation is applied with non-zero probability such that the joint distribution of new solutions has non-zero probability everywhere. The restrictions made on the mutation operator seem to make these proofs of only academic interest. Convergence properties of EP, ES and real-valued GA may also be derived using the theory of Markov chains. Rudolph [307] details the theory and offers sufficient conditions for convergence of EC algorithms. Other approaches have been taken including modeling EC algorithms as supermartingales as in [307]. Qi and Palmeiri [283] analyzed the real-valued GA assuming an infinite population size. They found that the solutions for a GA using only selection converges in distribution to the distribution concentrated at the global optimum. Also the mean loss value for a real-valued GA with selection and mutation converges to the global optimum. Hart [153] takes a different tack. He defines a class of EC algorithms called evolutionary pattern search algorithms that encompass the real-coded GA, EP, and ES and establishes a stationary point convergence result by extending the convergence theory of generalized pattern search algorithms. The convergence result does not guarantee convergence to the global optimum; it only guarantees that a stationary point is found. Stopping rules related to modifying the mutation probability for the algorithms are provided, however the stopping rules seem to require that the pattern search algorithm structure be adopted.

Global convergence results can be given for a broad class of problems, but the same cannot be said for convergence *rates*. The mathematical complexity of analyzing EC convergence rates is significant. Determining how many generations of the population are required in order to ensure a certain error in the solution is apparently an open problem for arbitrary loss functions. Vose [386,387] showed that assuming an infinite population size, and for every $0 < \delta < 14$, the number of generations required for the GA to come within a distance δ of θ^* is $O(-\log \delta)$. This result is not directly usable in our comparison, however, since it does not give a *quantifiable* expression for the number of generations required to guarantee that the best population element will be within some δ distance of θ^*.

Additional convergence rate results that exist are for restricted classes of loss functions that have some special properties that can be taken advantage of and usually with simplified ECs. In particular, except for the 'big O' result above, [386,387] (which allows for all three fundamental operations-selection, mutation, and recombination), most of the convergence rate results available are for EC algorithms using selection and mutation only, or using selection and recombination. Both [45] and [307] examine ES algorithms that include selection, mutation and recombination. The function analyzed in both cases is the classic spherical loss function $L(\theta) = \|\theta\|^2$. Convergence rates based on the spherical loss function are somewhat useful, if it is assumed that the sphere approximates a local basin of attraction. A number of other convergence rate results are also available for the spherical loss function; see for example [283] for a real-valued GA.

3.5.2 Convergence Rates for ES Algorithms

This section presents several means by which to determine the rate of convergence for the ES approach to EC. One of the more practically useful convergence rates for EC algorithms applies in a particular class of convex loss functions. The following theorem due to Rudolph [308] is an application of a more general result by Rappl [285]. The theorem is the starting place for the specific convergence rate result that will be used for comparison in Section 3.6.

Definition 1. *An algorithm has a geometric rate of convergence if and only if* $\mathbb{E}[L_k^* - L(\theta^*)] = O(\eta^k)$ *where* $\eta \in (0,1)$ *defines the convergence rate.*

Theorem 2 ([308]). *Let* $\bar{\Theta}_k \equiv \{\hat{\theta}_{k1}, \hat{\theta}_{k2}, \ldots, \hat{\theta}_{kN}\}$ *be the sequence of populations of size* N *generated by some ES at generation* $k(\hat{\theta}_{ki})$ *represents the* i^{th} *estimate for* θ *from the population of* N *elements). If* $\mathbb{E}[L_k^* - L(\theta^*)] < \infty$ *and* $\mathbb{E}[L_{k+1}^* - L(\theta^*)|\bar{\Theta}_k] \leq \eta[L_k^* - L(\theta^*)]$ *a.s. for all* $k \geq 0$ *where* $L_k^* = \min\{L(\hat{\theta}_{k1}), L(\hat{\theta}_{k2}), \ldots, L(\hat{\theta}_{kN})\}$*, then the ES algorithm converges a.s. geometrically fast to the optimum of the objective function.*

The condition $\mathbb{E}[L_{k+1}^* - L(\theta^*)|\bar{\Theta}_k] \leq \eta[L_k^* - L(\theta^*)]$ implies that the sequence decreases monotonically on average. This condition is needed since in the $(1,\lambda)$-ES that will be considered below, the loss value of the best parent in the current generation may be worse than the loss value of the best parent of the previous generation, although on average this will not be the case. Rudolph [308] shows that a $(1,\lambda)$-ES using selection and mutation only (where the mutation probability is selected from a uniformly distributed distribution on the unit hyperball), with certain classes of loss functions, satisfies the assumptions of the theorem. One such class is the (K,q)-*strongly convex* functions:

Definition 2. *Let* $L : \Theta \to \mathbb{R}$*. Then* L *is called* (K,q)*-strongly convex on* Θ *if for all* $x, y \in \Theta$ *and for each* $\alpha \in [0,1]$ *the inequalities*

$$\frac{K}{2}\alpha(1-\alpha)\|x-y\|^2 \leq \alpha L(x) + (1-\alpha)L(y) - L(\alpha x + (1-\alpha)y) \leq \frac{G}{2}\alpha(1-\alpha)\|x-y\|^2$$

hold with $0 < K \leq G \equiv Kq < \infty$.

For example, every quadratic function is (K,q)-strongly convex if the Hessian matrix is positive definite. In the case of twice differentiable functions, fairly simple tests are available for verifying that a function is (K,q)-strongly convex, from Nemirovsky and Yudin [241]. Let ν_1 be the smallest eigenvalue and let ν_2 be the largest eigenvalue of the Hessian matrix. If there exist positive constants K and G such that $0 < K \leq \nu_1 \leq \nu_2 \leq G < \infty$ for all θ then the function L is (K,q)-strongly convex with $q = G/K$. Other tests are possible that only assume the existence of the gradient $g(\theta)$ (see [146]).

The convergence rate result for a $(1,\lambda)$-ES using only selection and mutation on a (K,q)-strongly convex loss function is geometric with a rate of convergence

$$\eta = \left(1 - M_{\lambda,p}^2 q^2\right)$$

where $M_{\lambda,p} = \mathbb{E}[B_{\lambda:\lambda}] > 0$ and where $B_{\lambda:\lambda}$ denotes the maximum of λ independent identically distributed Beta random variables. The computation of

$M_{\lambda,p}$ is complicated since it depends on both the number of offspring λ and the problem dimension p. Asymptotic approximations are available. Assuming p is fixed and $\lambda \to \infty$ then $M_{\lambda,p} \approx (2p^{-1} \log \lambda)^{1/2}$. To extend this convergence rate from a $(1, \lambda)$-ES to a (N, λ)-ES, note that each of the N parents generate λ/N offspring. Then the convergence rate for the (N, λ)-ES where offspring are only obtained by mutation is

$$\eta \le 1 - \frac{2p^{-1} \log(\lambda/N)}{q^2}$$

for (K, q)-strongly convex functions.

Let us now discuss an alternative method based on approximating the behavior of an idealized (N, λ)-ES as a solution to an ordinary differential equation. Let $r = \left\|\bar{\theta}_k - \theta^*\right\|$, where $\bar{\theta}_k$ is the center of mass (sample mean) of $\{\hat{\theta}_{k1}, \hat{\theta}_{k2}, \ldots, \hat{\theta}_{kN}\}$. Consider a loss function of the spherical-based form $L(\theta) = f\left(\|\theta - \theta^*\|\right)$, where f is a strictly increasing function. Then, an approximate description of the ES is given by the differential equation

$$\frac{dr}{dt} = -\frac{c(t)}{p} r(t),$$

where each time increment (t) of unity represents one iteration of the ES and $c(t)$ is some function dependent on the ES coefficients, [46]. An idealized ES may be based on the assumption that $c(t)$ is a constant, say c^*. As discussed in [46], this is tantamount to knowing the value of r at every time, and normalizing the mutation scale factor at each time so that it is proportional to r. Obviously, this implementation of an ES is idealized because r will almost certainly not be known in practice. Nevertheless, it provides a basis for *some* theoretical analysis. Solving the above differential equation with constant $c(t) = c^*$ and then inverting to solve for t yields

$$t = \frac{p}{c^*} \log\left[\frac{r(0)}{r(t)}\right]. \tag{3.6}$$

Expression (3.6) provides a basis for an estimate of the number of time steps to reach a given distance $r(t)$. Ignoring negligible computation associated with the algorithm itself (*e.g.*, the random number generation), the total cost of the algorithm is then the number of function evaluations per iteration times the number of time steps.

3.5.3 Convergence Rates for GA Algorithms

Based on results in [306] and elsewhere, [341], Section 10.5 and [344] discuss how it is possible to cast the binary bit-based GA in the framework of Markov chains. This allows for a rate of convergence analysis. Consider a GA with a population size of N. Further, suppose that each population element is a binary string of length b bits. Hence, there are 2^b possible strings for an *individual* population element. Then the total number of *unique* possible populations is given by (see [348])

$$N_{\text{pop}} \equiv \frac{(N + 2^b - 1)!}{(2^b - 1)!N!}.$$

It is possible to construct a Markov transition matrix Π that provides the probability of transitioning from one population of size N to another population of the same size. This transition matrix has dimension $N_{\text{pop}} \times N_{\text{pop}}$. An individual element in the transition matrix can be computed according to the formulas in [344] (see also [348]). These elements depend in a non-trivial way on the population size, crossover rate, mutation rate, and number of elements considered 'elite.'

Of primary interest in analyzing the performance of GA algorithms using Markov chains is the probability of obtaining a population that contains the optimum θ^*. Let π_k be an $N_{\text{pop}} \times 1$ vector having j^{th} component, $\pi_k(j)$, equal to the probability that the k^{th} generation will result in population j. From basic Markov chain theory,

$$\pi_k^T = \pi_{k-1}^T \Pi = \pi_0^T \Pi^k$$

where π_0 is an initial probability distribution.

The stationary distribution of the GA is then given by

$$\bar{\pi}^T \equiv \lim_{k \to \infty} \pi_k^T = \lim_{k \to \infty} \pi_0^T \Pi^k.$$

Further, under standard ergodicity assumptions for Markov chains, $\bar{\pi}$ satisfies $\bar{\pi}^T = \bar{\pi}^T \Pi$. This equation provides a mechanism for solving directly for the stationary distribution (*e.g.*, [168], pages 123-124).

Unfortunately, from a practical view, the Markov chain approach has a significant deficiency. The dimension N_{pop} grows very rapidly with increases in the number of bits b and/or the population size N. An estimate of the size of N_{pop} can be obtained by Stirling's approximation as follows:

$$N_{\text{pop}} \approx \sqrt{2\pi} \left(1 + \frac{2^b - 1}{N}\right)^N \left(1 + \frac{N}{2^b - 1}\right)^{2^b - 1} \left(\frac{1}{2^b - 1} + \frac{1}{N}\right)^{1/2}$$

Thus far, our analysis using the above approach has been restricted to scalar θ systems (requiring fewer bits b than a multivariate system) and low N. Examples are given in [341], Section 10.5 and [344]. An approach for compressing the size of the transition matrix (to emphasize only the most likely states) is given in [343]. However, this approach is only useful in an adaptive sense as the algorithm is running; it is not designed for *a-priori* efficiency analysis.

3.6 Comparative Analysis

3.6.1 Problem Statement and Summary of Efficiency Theory for the Five Algorithms

This section uses the specific algorithm results in Sections 3.2 to 3.5 above in drawing conclusions on the relative performance of the five algorithms. There are obviously many ways one can express the rate of convergence, but it is expected that, to the extent they are based on the theory outlined above, the various ways will lead to broadly similar conclusions. We will address the rate of convergence by focusing on the question:

With some high probability $1 - \rho$ (ρ a small number), how many $L(\cdot)$ function evaluations, say n, are needed to achieve a solution lying in some 'satisfactory set' $S(\theta^*)$ containing θ^*?

With the random search algorithm in Section 3.2, we have a closed form solution for use in questions of this sort while with the SPSA, SAN, and EC algorithms of Sections 3.3 through 3.5, we must apply the existing asymptotic results, assuming that they apply to the finite-sample question above. For the GA, there is a finite sample solution using the Markov chain approach. For each of the five algorithms, we will outline below an analytical expression useful in addressing the question. After we have discussed the analytical expressions, we present a comparative analysis in a simple problem setting for varying p.

Random Search

We can use (3.1) to answer the question above. Setting the left-hand side of (3.1) to $1 - \rho$ and supposing that there is a constant sampling probability $P^* = \mathbb{P}\{\theta_{\text{new}}(k) \in S(\theta^*)\}$ for all k, we have

$$n = \frac{\log\ \rho}{\log\ (1 - P^*)}. \tag{3.7}$$

Although (3.7) may appear benign at first glance, this expression grows rapidly as p gets large due to P^* approaching 0. (A numerically stable approximation that is useful with small P^* is given in [341], page 62). Hence, (3.7) shows the extreme inefficiency of simple random search in higher-dimensional problems as illustrated in the study below. Note that while (3.7) is in terms of the iterate $\hat{\theta}_k$, a result related to the rate of convergence for $L(\hat{\theta}_k)$ is given in [265], page 24; this result is in terms of extreme value distributions and also confirms the inefficiency of simple random search algorithms in high-dimensional problems.

Simultaneous Perturbation Stochastic Approximation

As mentioned in Section 3.4, there is no known asymptotic normality result in the case of noise-free measurements of $L(\theta)$ (although Gerencsér and

Vágó, [138], show that the rate of convergence is geometric with an unknown constant governing the decay). Nonetheless, a *conservative* representation of the rate of convergence is available by assuming a noisy case with small levels of noise. Then we know from (4.4) that the approximate distribution of $\hat{\theta}_k$ with optimal decay rates for the gains a_k and c_k is $\mathcal{N}(\theta^* + \mu/k^{1/3}, \mathit{\Sigma}/k^{2/3})$. In principle, then, one can use this distribution to compute the probabilities associated with arbitrary sets $S(\theta^*)$, and these probabilities will be directly a function of k. In practice, due to the correlation in $\mathit{\Sigma}$, this may not be easy and so inequalities such as in [363], Chapter 2 can be used to provide bounds on $\mathbb{P}\{\hat{\theta}_k \in S(\theta^*\})$ in terms of the marginal probabilities of the $\hat{\theta}_k$ elements.

For purposes of insight, consider a case where the covariance matrix $\mathit{\Sigma}$ is diagonal. If $S(\theta^*)$ is a hypercube of the form $[s_1^-, s_1^+] \times [s_2^-, s_2^+] \times ... \times [s_p^-, s_p^+]$, then $\mathbb{P}\{\hat{\theta}_k \in S(\theta^*)\}$ is a product of the marginal normal probabilities associated with each element of $\hat{\theta}_k$ lying in its respective interval $[s_i^-, s_i^+]$, $i = 1, 2, \ldots, p$. Such diagonal covariance matrices arise when the loss function is separable in each of the components of θ. Then we can find the k such that the product of probabilities equals $1 - \rho$. To illustrate more specifically, suppose further that $\mathit{\Sigma} = \sigma^2 I$, the $\mu/k^{1/3}$ term in the mean is negligible, that $S(\theta^*)$ is centered around θ^*, and that $\delta s \equiv s_i^+ - s_i^-$ for all i. (*i.e.*, $s_i^+ - s_i^-$ does not depend on i). Then for a specified ρ, we seek the n such that $\mathbb{P}\{\hat{\theta}_k \in S(\theta^*)\} = \mathbb{P}\{\hat{\theta}_{ki} \in [s_i^-, s_i^+]\}^p = 1 - \rho$. From standard $\mathcal{N}(0, 1)$ distribution tables, there exists a displacement factor, say $d(p)$, such that the probability contained within $\pm d(p)$ units contains probability amount $(1 - \rho)^{1/p}$; we are interested in the k such that $2d(p)\sigma/k^{1/3} = \delta s$. From the fact that SPSA uses two $L(\theta^*)$ evaluations per iteration, the value n to achieve the desired probability for $\hat{\theta}_k \in S(\theta^*)$ is then

$$n = 2\left(\frac{2d(p)\sigma}{\delta s}\right)^3 .$$

Unfortunately, the authors are unaware of any convenient analytical form for determining $d(p)$, which would allow a 'clean' analytical comparison with the efficiency formula (3.7) above (a closed-form approximation to normal probabilities of intervals is given in [171], pages 55-57, but this approximation does not yield a closed form for $d(p)$).

Simulated Annealing

Because SAN, like SPSA, has an asymptotic normality result, the method above for characterizing the rate of convergence for SPSA may also be used here. Again, we shall consider the case where the covariance matrix is diagonal ($\mathit{\Sigma} = \sigma^2 I$). Assume also that $S(\theta^*)$ is a hypercube of the form $[s_1^-, s_1^+] \times [s_2^-, s_2^+] \times ... \times [s_p^-, s_p^+]$ centered around θ^*, and that $\delta s \equiv s_i^+ - s_i^-$, for all i. The (positive) constant B_0 is assumed small enough that it can be ignored. At each iteration after the first, SAN must evaluate $L(\theta^*)$ only once per iteration. So the value n to achieve the desired probability for $\hat{\theta}_k \in S(\theta^*)$ is

$$\log n^{1-\gamma} \;=\; \left(\frac{2d(p)\sigma}{\delta s}\right)^2 .$$

Evolution Strategy

As discussed in Section 3.5, the rate-of-convergence results for algorithms of the evolutionary computation type are not as well developed as for the other three algorithms of this chapter. Theorem 2 gives a general bound on $\mathbb{E}[L(\hat{\theta}_k)-L(\theta^*)]$ for application of a (N,λ)-ES form of EC algorithm to (K,q)-strongly convex functions. A more explicit form of the bound is available for the $(1,\lambda)$-ES. Unfortunately, even in the optimistic case of an explicit numerical bound on $\mathbb{E}[L(\hat{\theta}_k)-L(\theta^*)]$, we cannot readily translate the bound into a probability calculation for $\hat{\theta}_k \in S(\theta^*)$, as used above (and, conversely, the asymptotic normality result on $\hat{\theta}_k$ for SPSA and SAN cannot be readily translated into one on $L(\hat{\theta}_k)$ since $\partial L/\partial\theta = 0$ at θ^*, see, *e.g.*, [327], pages 122-124, although Lehmann in [204], pages 338-339 suggests a possible means of coping with this problem via higher-order expansions). So, in order to make *some* reasonable comparison, let us suppose that we can associate a set $S(\theta^*)$ with a given deviation from $L(\theta^*)$, *i.e.*, $S(\theta^*) = \{\theta : L(\hat{\theta}_k) - L(\theta^*) \leq \varepsilon\}$ for some prespecified tolerance $\varepsilon > 0$ (note that $S(\theta^*)$ is a function of ε). As presented in [308], $\mathbb{E}[L(\hat{\theta}_k)-L(\theta)] \leq \eta^k$ for sufficiently large k, where η is the convergence rate in Section 3.5. Then by Markov's inequality,

$$1 - \mathbb{P}\{\hat{\theta}_k \in S(\theta^*)\} \leq \frac{\mathbb{E}[L(\hat{\theta}_k)-L(\theta^*)]}{\varepsilon} \leq \frac{\eta^k}{\varepsilon} \tag{3.8}$$

indicating that $\mathbb{P}\{\hat{\theta}_k \in S(\theta^*)\}$ is bounded below by the ES bounds mentioned in Section 3.5. For EC algorithms in general (and ES in particular), there are λ evaluations of the loss function for each generation k so that $n = \lambda k$, where

$$k = \frac{\log\rho - \log(1/\varepsilon)}{\log\left[1 - \frac{2}{pq^2}\log(\lambda/N)\right]}. \tag{3.9}$$

We also report results related to the differential equation solution (3.6). As noted, this solution is tied to some restrictions, namely to loss functions of the spherical-based form $L(\theta) = f\left(\|\theta-\theta^*\|\right)$ and to an idealized ES with a mechanism for adaptively scaling the mutation magnitude according to the current distance $r = \left\|\bar{\theta}_k - \theta^*\right\|$. Further, as a deterministic approximation to a stochastic algorithm, there is no simple way to determine the probability ρ defined above. If, as mentioned above, we consider $S(\theta^*)$ in the form of a hypercube $[s_1^-, s_1^+] \times [s_2^-, s_2^+] \times \ldots \times [s_p^-, s_p^+]$, we can specify a r_{inside} that defines the radius of the largest hypersphere that is contained within the hypercube and r_{outside} that defines the radius of the smallest hypersphere that lies outside (*i.e.*, contains) the hypercube. The number of function evaluations needed to yield a solution in $S(\theta^*)$ is then bounded above and below by the number required for a solution to lie in these inside and outside hyperspheres. That is, substituting r_{inside} or r_{outside} for $r(t)$ in the right-hand side of (3.6) yields an upper and lower bound, respectively, to the number of time steps, which, by the appropriate multiplication, yields bounds to the number of function evaluations.

Genetic Algorithm

As mentioned in Section 3.5, while the GA has a relatively clean theory that applies in both finite and asymptotic samples, there are significant challenges in computing the elements of the Markov transition matrix $\mathit{\Pi}$. The number of possible states – corresponding to the number N_{pop} of possible populations – grows extremely rapidly with the number of population elements N or the number of bits b. The computation of the $N_{\text{pop}} \times N_{\text{pop}}$ transition matrix $\mathit{\Pi}$ quickly overwhelms even the most powerful current or foreseeable personal computers.

Nevertheless, in principle, the Markov structure is convenient for establishing a convergence rate for the GA. Recall that π_k is the $N_{\text{pop}} \times 1$ vector having j^{th} component, $\pi_k(j)$, equal to the probability that the k^{th} generation will result in population j. Let us denote by S_J the set of indices j such that population j contains at least one member lying inside $S(\theta^*)$. Hence, $S_J \subseteq \{1, 2, \ldots, N\}$. Then

$$n = N + (N - N_{\text{elite}}) \min \left\{ k : \sum_{j \in S_J} \pi_k(j) \geq 1 - \rho \right\},$$

where N_{elite} denotes the number of elite elements in the population being saved from one generation to the next and we have assumed that all non-elite function evaluations are not 'saved' from one generation to the next (*i.e.*, every generation entails $N - N_{\text{elite}}$ function evaluations).

3.6.2 Application of Convergence Rate Expressions for Varying *p*

We now apply the results above to demonstrate relative efficiency for varying p. Because the GA result is computationally explosive as p gets larger (requiring a larger bit string length and/or population size), we restrict the comparison here to the four algorithms: random search, SPSA, SAN and ES. Let $\mathit{\Theta} = [0, 1]^p$ (the p-dimensional hypercube with minimum and maximum θ values of 0 and 1 for each component). We want to guarantee with probability 0.90 that each element of θ is within 0.04 units of the optimal. Let the (unknown) optimal θ, θ^*, lie in $(0.04, 0.96)^p$. The individual components of θ^* are θ_i^*. Hence,

$$S(\theta^*) = [\theta_1^* - 0.04,\ \theta_1^* + 0.04] \times [\theta_2^* - 0.04,\ \theta_2^* + 0.04] \times \ldots \\ \times [\theta_p^* - 0.04,\ \theta_p^* + 0.04] \subset \Theta.$$

Table 3.1 is a summary of relative efficiency for the setting above for $p = 2, 5$, and 10; the efficiency is normalized so that all algorithms perform equally at $p = 1$, as described below. The numbers in Table 3.1 are the ratios of the number of loss measurements for the given algorithm over the number for the best algorithm at the specified p; the highlighted values 1.0 indicate the best algorithm for each of the values of p. To establish a fair basis for comparison, we fixed the various parameters in the expressions above (*e.g.*, σ in SPSA and SAN, λ for the ES, *etc.*) so that the algorithms produced identical efficiency results for $p = 1$ (requiring $n = 28$ measurements to achieve the objective outlined above). These parameters do not explicitly depend on p. We then use these parameter settings as p increases. Of course, in practice, algorithm parameters are typically tuned for each new problem, including changes in p. Hence, the results may not reflect practical relative efficiency, including the cost of the tuning process. Rather, they point towards general efficiency trends as a function of problem dimension in the absence of problem-specific tuning.

For the random sampling algorithm, suppose uniform sampling on Θ is used to generate $\theta_{\text{new}}(k)$ for all k. Then, $P^* = 0.08^p$. For SPSA, we fix σ such that the same number of function measurements in the $p = 1$ case ($n = 28$) is used for both random search and SPSA (so $\delta s = 0.08$ and $\sigma = 0.0586$). Likewise, for SAN, we fix σ to achieve the same objective (so $\delta s = 0.08$ and $\sigma = 0.031390$). Also, for convenience, take $\gamma = 1/2$. To compare the (N, λ)-ES algorithm with the random search, SPSA, and SAN algorithms, it is assumed that the loss function is restricted to the (K, q)-strongly convex functions or spherical-based forms discussed in Section 3.5. Also let $\lambda = 14$, $N = 7$, $\varepsilon = 8.3$, $q = 4$, and $\rho = 0.1$. The variables were constrained here so that for $p = 1$, we have the same n ($= 28$) as realized for the other algorithms. Table 3.1 summarizes the performance comparison results.

Table 3.1. Ratios of loss measurements needed relative to best algorithm at each p, for $1 \le p \le 10$

	$p = 1$	$p = 2$	$p = 5$	$p = 10$
Random search	**1.0**	11.6	8970	2×10^9
SPSA	**1.0**	1.5	**1.0**	**1.0**
SAN	**1.0**	**1.0**	2.2	4.1
ES (from (3.8), (3.9))	**1.0**	1.9	1.9	2.8
ES (from (3.6) w. inside hypersphere)	**1.0**	2.1	2.4	3.8
ES (from (3.6) w. outside hypersphere)	**1.0**	1.8	1.8	2.6

Table 3.1 illustrates the explosive growth in the relative (and absolute) number of loss evaluations needed as p increases for the random search algorithm. The other algorithms perform more comparably, but there are still some non-negligible differences. For example, at $p = 5$, SAN will take 2.2 times more loss measurements than SPSA to achieve the objective of having $\hat{\theta}_k$ inside $S(\theta^*)$ with probability 0.90. Of course, as p increases, all algorithms take more measurements; the table only shows *relative* numbers of function evaluations (considered more reliable than absolute numbers).

This large improvement of SPSA and SAN relative to random search may partly result from the more restrictive regularity conditions of SPSA and SAN (*i.e.*, for formal convergence, SPSA assumes a several-times-differentiable loss function) and partly from the fact that SPSA and SAN work with *implicit* gradient information via gradient approximations. (The reasons for improvement with ES are less clear due to the lack of an identifiable connection to the gradient.) Of course, to maintain a fair comparison, SPSA and SAN, like the other algorithms here, explicitly use only loss evaluations, no direct gradient information. On the other hand, there are some differences between SPSA and SAN. The different gradient approximations in SPSA and SAN may explain their relative efficiency. The 'Metropolis-type approximation appears to be much farther away from an exact gradient-based algorithm than a finite-difference approximation' ([134], page 128). By contrast, SPSA, recall, uses a (highly efficient) finite-difference-like approximation to the gradient.

The performance for ES is quite good. The restriction to strongly convex loss functions (from (3.8) and (3.9)) or spherical losses (from (3.6)), however, gives the ES in this setting a strong structure not available to the other algorithms. It remains unclear what practical theoretical conclusions can be drawn on a broader class of problems. More advanced sensitivity studies for various λ, N, and q have not yet been completed. Further, the inequality in (3.8) provides an optimistic assessment of the convergence rate. Ideally, a more general rate-of-convergence theory will provide a more broadly applicable basis for comparison.

4

Optimization of Risk Measures

Andrzej Ruszczyński[1] and Alexander Shapiro[2]

[1] Rutgers University, Piscataway, NJ 08854, USA,
rusz@rutcor.rutgers.edu

[2] Georgia Institute of Technology, Atlanta, Georgia 30332-0205, USA,
ashapiro@isye.gatech.edu

Summary. We consider optimization problems involving coherent measures of risk. We derive necessary and sufficient conditions of optimality for these problems, and we discuss the nature of the nonanticipativity constraints. Next, we introduce dynamic measures of risk, and formulate multistage optimization problems involving these measures. Conditions similar to dynamic programming equations are developed. The theoretical considerations are illustrated with many examples of mean-risk models applied in practice.

4.1 Introduction

Consider a stochastic system whose output variable Z is a real valued random variable. If it depends on some decision vector $x \in \mathbb{R}^n$, we can write the relation

$$Z(\omega) = f(x, \omega), \quad \omega \in \Omega.$$

Here $f : \mathbb{R}^n \times \Omega \to \mathbb{R}$, and $(\Omega, \mathcal{F})$ is a measurable space. To focus attention, we shall be interested in the case when smaller values of Z are 'better', for example, Z may represent random cost or losses. It will be obvious how our considerations can be adapted to the case of reverse preferences.

In order to find the 'best' values of the decision vector x we can formulate the stochastic optimization problem

$$\min_{x \in S} \left\{ \phi(x) \doteq \mathbb{E}_P[f(x, \omega)] \right\},$$

where $S \subset \mathbb{R}^n$ is a set of feasible decision vectors, and P is a probability measure (distribution) on the sample space $(\Omega, \mathcal{F})$. The theory of such stochastic optimization problems and numerical methods for their solution are well developed (see [312]).

There are two basic difficulties associated with the above formulation. First, it is assumed that the probability distribution P is known. In real life applications the probability distribution is never known exactly. In some cases it can be estimated from historical data by statistical techniques. However, in many cases the probability distribution neither can be estimated accurately nor remains constant. Even worse, quite often one subjectively assigns certain weights (probabilities) to a finite number of possible realizations (called *scenarios*) of the uncertain data. Such a simplified model can hardly be considered an accurate description of the reality.

The second basic question is why we want to optimize the *expected* value of the random outcome Z. In some situations the same decisions under similar conditions are made repeatedly over a certain period of time. In such cases one can justify optimization of the expected value by arguing that, by the Law of Large Numbers, it gives an optimal decision on *average*. However, because of the variability of the data, the average of the first few results may be very bad. For example, one may lose all his investments, and it does not help that the decisions were optimal on average.

For these reasons, quantitative models of *risk* and *risk aversion* are needed. There exist several approaches to model decision making under risk. The classical approach is based on the *expected utility theory* of von Neumann and Morgenstern [385]. One specifies a *disutility function*[3] $g : \mathbb{R} \to \mathbb{R}$ and formulates the problem

$$\min_{x \in S} \Big\{ \phi_g(x) \doteq \mathbb{E}_P\big[g(f(x,\omega))\big] \Big\}. \tag{4.1}$$

Unfortunately, it is extremely difficult to elicit the disutility function of a decision maker.

The second approach is to specify *constraints* on risk. The most common is the *Value at Risk* constraint, which involves the critical value $z_{\max}$ allowed for risk exposure, and the probability $p_{\max}$ of excessive outcomes:

$$P\big[Z \geq z_{\max}\big] \leq p_{\max}.$$

In the stochastic optimization literature such constraints are called *probabilistic* or *chance constraints* [280]. Variations of this concept are known as integrated chance constraints [152], Conditional Value at Risk [300], or expected shortfall [1].

A direct way to deal with the issue of uncertain probability distribution, is to identify a plausible family $\mathcal{A}$ of probability distributions and, consequently, to consider the min-max problem

$$\min_{x \in S} \left\{ \phi(x) \doteq \sup_{P \in \mathcal{A}} \mathbb{E}_P[f(x,\omega)] \right\}. \tag{4.2}$$

The idea of the worst-case (min-max) formulation is not new of course. It goes back to von Neumann's game theory and was already discussed, for example

[3] We consider here minimization problems, and that is why we speak about disutility. Any disutility function g corresponds to a utility function $u : \mathbb{R} \to \mathbb{R}$ defined by $u(-z) = -g(z)$. Note that the function u is concave and increasing (nondecreasing) if and only if the function g is convex and increasing (nondecreasing).

in the context of stochastic programming, in Žáčková [388] almost 40 years ago.

The attempts to overcome the drawbacks of the expected value optimization have also a long history. One can try to reach a compromise between the optimization on average and the minimization of a certain measure of the involved risk. This leads to the formulation

$$\min_{x \in S} \big\{ \phi(x) \doteq \rho[F(x)] \big\}, \tag{4.3}$$

where $\rho(Z)$ is a mean-risk measure, defined on a space of random variables $Z : \Omega \to \mathbb{R}$, and $[F(x)](\omega) = f(x, \omega)$. The classical mean-variance risk measure $\rho(Z) \doteq \mathbb{E}[Z] + c\mathbb{V}\mathrm{ar}[Z]$, where c is a non-negative constant, is going back to Markowitz [214].

There are several problems with the mean-variance risk measure. First, the expectation and variance are measured in different units. Second, the mean-variance model is not consistent with the classical relation of stochastic dominance, which formalizes risk-averse preferences [246].

In recent years risk analysis came under intensive investigation, in particular from the point of view of the optimization theory. In this chapter we discuss a general theory of optimization of risk measures. We show, in particular, that the above approaches of min-max formulation (4.2) and risk measure formulation (4.3), in a sense, are equivalent to each other.

We also introduce and analyze new models of dynamic optimization problems involving risk functions. We introduce the concept of conditional risk mappings, and we derive dynamic programming relations for the corresponding optimization models. In this way we provide an alternative approach to the recent works [17, 119, 266, 294], where various dynamic risk models are considered.

4.2 Risk Functions

In this section we give a formal definition of risk functions and we discuss their basic properties. Let $(\Omega, \mathcal{F})$ be a sample space, equipped with sigma algebra $\mathcal{F}$, on which considered uncertain outcomes (random functions $Z = Z(\omega)$) are defined. By a *risk function* we understand a function $\rho(Z)$ which maps Z into the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$. In order to make this concept precise we need to define a space $\mathcal{Z}$ of allowable random functions $Z(\omega)$ for which $\rho(Z)$ is defined. It seems that a natural choice of $\mathcal{Z}$ will be the space of all $\mathcal{F}$-measurable functions $Z : \Omega \to \mathbb{R}$. However, typically, this space is too large for development of a meaningful theory. In almost all interesting examples considered in this chapter we deal with the space[4] $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$. We will discuss an appropriate choice of the space $\mathcal{Z}$ later.

[4] Recall that $\mathcal{L}_p(\Omega, \mathcal{F}, P, \mathbb{R}^n)$ denotes the linear space of all $\mathcal{F}$-measurable functions $\psi : \Omega \to \mathbb{R}^n$ such that $\int_\Omega \|\psi(\omega)\|^p \, dP(\omega) < +\infty$. More precisely, an element of

We assume throughout this chapter that $\mathcal{Z}$ is a linear space of $\mathcal{F}$-measurable functions and considered risk functions $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ are *proper.* That is, $\rho(Z) > -\infty$ for all $Z \in \mathcal{Z}$ and the domain

$$\operatorname{dom}(\rho) \doteq \{Z \in \mathcal{Z} : \rho(Z) < +\infty\}$$

is non-empty. We consider the following axioms associated with a risk function ρ. For $Z_1, Z_2 \in \mathcal{Z}$ we denote by $Z_2 \succeq Z_1$ the pointwise partial order meaning $Z_2(\omega) \geq Z_1(\omega)$ for all $\omega \in \Omega$.

(A1) *Convexity:*

$$\rho(\alpha Z_1 + (1-\alpha)Z_2) \leq \alpha\rho(Z_1) + (1-\alpha)\rho(Z_2)$$

for all $Z_1, Z_2 \in \mathcal{Z}$ and all $\alpha \in [0,1]$.

(A2) *Monotonicity:* If $Z_1, Z_2 \in \mathcal{Z}$ and $Z_2 \succeq Z_1$, then $\rho(Z_2) \geq \rho(Z_1)$.

(A3) *Translation Equivariance:* If $a \in \mathbb{R}$ and $Z \in \mathcal{Z}$, then $\rho(Z+a) = \rho(Z)+a$.

(A4) *Positive Homogeneity:* If $\alpha > 0$ and $Z \in \mathcal{Z}$, then $\rho(\alpha Z) = \alpha\rho(Z)$.

These axioms were introduced, and risk functions satisfying (A1)–(A4) were called *coherent risk measures*, in Artzner *et al.* [16].

In order to proceed with the analysis we need to associate with the space $\mathcal{Z}$ a dual space $\mathcal{Z}^*$ of measures such that the scalar product

$$\langle \mu, Z \rangle \doteq \int_\Omega Z(\omega)\, d\mu(\omega) \tag{4.4}$$

is well defined for all $Z \in \mathcal{Z}$ and $\mu \in \mathcal{Z}^*$. That is, we assume that $\mathcal{Z}^*$ is a linear space of finite signed measures[5] μ on $(\Omega, \mathcal{F})$ such that $\int_\Omega |Z|\, d|\mu| < +\infty$ for all $Z \in \mathcal{Z}$. We assume that $\mathcal{Z}$ and $\mathcal{Z}^*$ are *paired* (locally convex topological vector) spaces. That is, $\mathcal{Z}$ and $\mathcal{Z}^*$ are equipped with respective topologies which make them locally convex topological vector spaces and these topologies are compatible with the scalar product (4.4), *i.e.*, every linear continuous functional on $\mathcal{Z}$ can be represented in the form $\langle \mu, \cdot \rangle$ for some $\mu \in \mathcal{Z}^*$, and every linear continuous functional on $\mathcal{Z}^*$ can be represented in the form $\langle \cdot, Z \rangle$ for some $Z \in \mathcal{Z}$. In particular, we can equip each space $\mathcal{Z}$ and $\mathcal{Z}^*$ with its weak topology induced by its paired space. This will make $\mathcal{Z}$ and $\mathcal{Z}^*$ paired locally convex topological vector spaces provided that for any $Z \in \mathcal{Z} \setminus \{0\}$

$\mathcal{L}_p(\Omega, \mathcal{F}, P, \mathbb{R}^n)$ is a class of such functions $\psi(\omega)$ which may differ from each other on sets of P-measure zero. For $n = 1$ we denote this space by $\mathcal{L}_p(\Omega, \mathcal{F}, P)$. Unless stated otherwise, while dealing with these spaces we assume that $p \in [1, +\infty)$, P is a probability measure on $(\Omega, \mathcal{F})$ and expectations are taken with respect to P. For $\psi \in \mathcal{L}_p(\Omega, \mathcal{F}, P)$, its norm $\|\psi\|_p \doteq \left(\int_\Omega |\psi(\omega)|^p\, dP(\omega)\right)^{1/p}$.

[5] Recall that a finite signed measure μ can be represented in the form $\mu = \mu^+ - \mu^-$, where μ^+ and μ^- are non-negative finite measures on $(\Omega, \mathcal{F})$. This representation is called the Jordan decomposition of μ. The measure $|\mu| = \mu^+ + \mu^-$ is called the total variation of μ.

there exists $\mu \in \mathcal{Z}^*$ such that $\langle \mu, Z \rangle \neq 0$, and for any $\mu \in \mathcal{Z}^* \setminus \{0\}$ there exists $Z \in \mathcal{Z}$ such that $\langle \mu, Z \rangle \neq 0$.

If $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$, we can consider its dual space $\mathcal{Z}^* \doteq \mathcal{L}_q(\Omega, \mathcal{F}, P)$, where $q \in (1, +\infty]$ is such that $1/p + 1/q = 1$. Here $\mathcal{Z}$, equipped with the respective norm, is a Banach space and $\mathcal{Z}^*$ is its dual Banach space. In order to make these spaces paired spaces we can equip $\mathcal{Z}$ with its strong (norm) topology and $\mathcal{Z}^*$ with its weak* topology. Moreover, if $p \in (1, +\infty)$, then $\mathcal{Z}$ and $\mathcal{Z}^*$ are reflexive Banach spaces. In that case, they are paired spaces when equipped with their strong topologies. Note also that in this case every measure $\mu \in \mathcal{Z}^*$ has a density $\zeta \in \mathcal{L}_q(\Omega, \mathcal{F}, P)$, *i.e.*, $d\mu = \zeta dP$. When dealing with these spaces we identify the corresponding measure with its density and for $Z \in \mathcal{L}_p(\Omega, \mathcal{F}, P)$ and $\zeta \in \mathcal{L}_q(\Omega, \mathcal{F}, P)$ we use the scalar product

$$\langle \zeta, Z \rangle \doteq \int_\Omega \zeta(\omega) Z(\omega) \, dP(\omega).$$

Unless stated otherwise we always assume the following.

(C) For every $A \in \mathcal{F}$ the space $\mathcal{Z}$ contains the indicator[6] function $\mathbb{I}_A$.

Since the space $\mathcal{Z}$ is linear, this implies that $\mathcal{Z}$ contains all step functions of the form $\sum_{i=1}^m \alpha_i \mathbb{I}_{A_i}$, where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$, $i = 1, \dots, m$. This holds true, in particular, for every space $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$.

The partial order in the space $\mathcal{Z}$, appearing in condition (A2), is defined by the cone

$$\mathcal{Z}_+ \doteq \{Z \in \mathcal{Z} : Z(\omega) \geq 0, \ \forall \, \omega \in \Omega\},$$

i.e., $Z_2 \succeq Z_1$ if and only if $Z_2 - Z_1 \in \mathcal{Z}_+$. Consider the cone $\mathcal{Z}_+^*$ of all nonnegative measures in the space $\mathcal{Z}^*$. For any $Z \in \mathcal{Z}_+$ and any $\mu \in \mathcal{Z}_+^*$, we have that $\langle \mu, Z \rangle \geq 0$. Moreover, because of assumption (C) above, we have that $\mathcal{Z}_+^*$ coincides with the dual cone of the cone $\mathcal{Z}_+$, which is defined as the set of $\mu \in \mathcal{Z}^*$ such that $\langle \mu, Z \rangle \geq 0$ for all $Z \in \mathcal{Z}_+$.

We can now formulate the basic (conjugate) duality result. Recall that the conjugate function $\rho^* : \mathcal{Z}^* \to \overline{\mathbb{R}}$ of a risk function ρ is defined as

$$\rho^*(\mu) \doteq \sup_{Z \in \mathcal{Z}} \big\{ \langle \mu, Z \rangle - \rho(Z) \big\}, \tag{4.5}$$

and the conjugate of ρ^* (the biconjugate function) as

$$\rho^{**}(Z) \doteq \sup_{\mu \in \mathcal{Z}^*} \big\{ \langle \mu, Z \rangle - \rho^*(\mu) \big\}.$$

By $\operatorname{lsc}(\rho)$ we denote the lower semicontinuous hull of ρ taken with respect to the considered topology of $\mathcal{Z}$. The following is the basic duality result of convex analysis (see, *e.g.*, [299, Theorem 5] for a proof).

[6] Recall that the indicator function $\mathbb{I}_A$ is defined as $\mathbb{I}_A(\omega) = 1$ for $\omega \in A$ and $\mathbb{I}_A(\omega) = 0$ for $\omega \notin A$.

Theorem 1 (Fenchel-Moreau). *Suppose that function $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ is convex and proper. Then $\rho^{**} = \operatorname{lsc}(\rho)$.*

It follows that if ρ is convex and proper, then the representation

$$\rho(Z) = \sup_{\mu \in \mathcal{Z}^*} \big\{ \langle \mu, Z \rangle - \rho^*(\mu) \big\} \tag{4.6}$$

holds true if ρ is lower semicontinuous. Conversely, if (4.6) is satisfied for some function $\rho^*(\cdot)$, then ρ is lower semicontinuous and convex. Note also that if ρ is proper, lower semicontinuous and convex, then its conjugate function ρ^* is proper. Let us also remark that if $\mathcal{Z}$ is a Banach space and $\mathcal{Z}^*$ is its dual (*e.g.*, $\mathcal{Z} = \mathcal{L}_p(\Omega, \mathcal{F}, P)$ and $\mathcal{Z}^* = \mathcal{L}_q(\Omega, \mathcal{F}, P)$) and ρ is convex, then ρ is lower semicontinuous in the weak topology if and only if it is lower semicontinuous in the strong (norm) topology.

If the set Ω is finite, say $\Omega = \{\omega_1, \dots, \omega_K\}$, then the technical level of the analysis simplifies considerably. Every function $Z \in \mathcal{Z}$ can be identified with the vector $(Z(\omega_1), \dots, Z(\omega_K))$. Thus the space $\mathcal{Z}$ is finite dimensional, $\mathcal{Z} = \mathbb{R}^K$, and can be paired with itself. Moreover, in the finite dimensional case, if ρ is proper and convex, then it is continuous (and hence lower semicontinuous) at every point in the interior of its domain. In particular, it is continuous at every point if it is real valued. In order to avoid technical details one can be tempted to restrict the discussion to finite sample spaces. However, apart from restricting the generality, this would result in losing some important essentials of the analysis. It turns out that some important properties enjoyed by risk functions in the case of finite Ω do not extend to continuous distributions (see the examples in the next section).

As it was discussed above, in order for the representation (4.6) to hold we only need the convexity (condition (A1)) and lower semicontinuity properties to be satisfied. Let us observe that (4.6) is equivalent to

$$\rho(Z) = \sup_{\mu \in \mathcal{A}} \big\{ \langle \mu, Z \rangle - \rho^*(\mu) \big\}, \tag{4.7}$$

where

$$\mathcal{A} \doteq \{\mu \in \mathcal{Z}^* : \rho^*(\mu) < +\infty\}$$

is the domain of ρ^*. It is not difficult to show that if representation (4.6) (or, equivalently, representation (4.7)) holds true, then condition (A2) is satisfied if and only if the set $\mathcal{A}$ contains only non-negative measures, and condition (A3) is satisfied if and only if $\mu(\Omega) = 1$ for every $\mu \in \mathcal{A}$ (*cf.* [314]). We obtain that if conditions (A1)–(A3) are satisfied and ρ is lower semicontinuous, then the representation (4.7) holds true with $\mathcal{A} \subset \mathcal{P}$, where $\mathcal{P}$ denotes the set of all probability measures in the space $\mathcal{Z}^*$.

Moreover, if $\mathcal{Z}$ is a Banach lattice[7] and ρ satisfies conditions (A1) and (A2), then ρ is continuous at every point[8] $Z \in \operatorname{int}(\operatorname{dom}(\rho))$ ([314]). Note that every space $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ is a Banach lattice. Also if ρ is positively homogeneous, then $\rho^*(\mu) = 0$ for $\mu \in \mathcal{A}$ and $\rho^*(\mu) = +\infty$ otherwise. Therefore we have the following. Recall that

$$\mathcal{P} \doteq \left\{\zeta \in \mathcal{L}_q(\Omega, \mathcal{F}, P) : \textstyle\int_\Omega \zeta(\omega)\, dP(\omega) = 1,\ \zeta \succeq 0\right\}$$

denotes the set of probability measures in the dual space $\mathcal{L}_q(\Omega, \mathcal{F}, P)$.

Theorem 2. *Suppose that* $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$*, risk function* $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ *is proper and conditions* (A1)–(A3) *are satisfied. Then for all* $Z \in \operatorname{int}(\operatorname{dom}(\rho))$ *it holds that*

$$\rho(Z) = \sup_{\zeta \in \mathcal{P}} \left\{\langle \zeta, Z\rangle - \rho^*(\zeta)\right\}. \tag{4.8}$$

If, moreover, ρ *is positively homogeneous, then there exists a non-empty convex closed set* $\mathcal{A} \subset \mathcal{P}$ *such that for all* $Z \in \operatorname{int}(\operatorname{dom}(\rho))$ *it holds that*

$$\rho(Z) = \sup_{\zeta \in \mathcal{A}} \langle \zeta, Z\rangle. \tag{4.9}$$

In this way we have established the equivalent representation of convex risk functions, which corresponds to the min-max model (4.2).

In various forms of generality the above dual representations of convex risk functions were derived in [16, 127, 301, 314]. If the set Ω is finite, say $\Omega = \{\omega_1, \dots, \omega_K\}$ with respective (positive) probabilities $p_1, \dots, p_K$, then the corresponding set

$$\mathcal{P} = \left\{\zeta \in \mathbb{R}^K : \textstyle\sum_{k=1}^K p_k \zeta_k = 1,\ \zeta \geq 0\right\}$$

is bounded, and hence the set $\mathcal{A}$ is also bounded. It follows that if Ω is finite and ρ is proper and conditions (A1)–(A4) are satisfied, then $\rho(\cdot)$ is real valued and representation (4.9) holds.

4.3 The Utility Model

It is also possible to relate the theory of convex risk functions with the utility model (4.1). Let $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$ and $\mathcal{Z}^* \doteq \mathcal{L}_q(\Omega, \mathcal{F}, P)$, and let $g : \mathbb{R} \to \overline{\mathbb{R}}$

[7] It is said that Banach space $\mathcal{Z}$ is a Banach lattice, with respect to the considered partial order defined by the cone $\mathcal{Z}_+$, if $\mathcal{Z}$ is a lattice, *i.e.*, for any $Z_1, Z_2 \in \mathcal{Z}$ the element $\max\{Z_1(\cdot), Z_2(\cdot)\}$ also belongs to $\mathcal{Z}$, and moreover if $|Z_1(\cdot)| \leq |Z_2(\cdot)|$, then $\|Z_1\| \leq \|Z_2\|$.

[8] We denote by $\operatorname{int}(\operatorname{dom}(\rho))$ the interior of the domain of ρ. That is, $Z \in \operatorname{int}(\operatorname{dom}(\rho))$ if there is a neighborhood $\mathcal{N}$ of Z such that $\rho(Z')$ is finite for all $Z' \in \mathcal{N}$.

be a proper convex lower semicontinuous function such that the expectation $\mathbb{E}[g(Z)]$ is well defined[9] for all $Z \in \mathcal{Z}$. We can view the function g as a disutility function. Consider the risk function

$$\rho(Z) \doteq \mathbb{E}[g(Z)] \tag{4.10}$$

and assume that ρ is proper. Since g is lower semicontinuous and convex, we have that

$$g(z) = \sup_{\alpha \in \mathbb{R}} \{\alpha z - g^*(\alpha)\},$$

where g^* is the conjugate of g. As g is proper, the conjugate function g^* is also proper. It follows that

$$\rho(Z) = \mathbb{E}\left[\sup_{\alpha \in \mathbb{R}} \{\alpha Z - g^*(\alpha)\}\right]. \tag{4.11}$$

We use the following interchangeability principle (see, *e.g.*, Rockafellar and Wets [302, Theorem 14.60]). It is said that a linear space $\mathcal{M}$ of $\mathcal{F}$-measurable functions $\psi : \Omega \to \mathbb{R}^m$ is *decomposable* if for every $\psi \in \mathcal{M}$ and $B \in \mathcal{F}$, and every bounded and $\mathcal{F}$-measurable function $W : \Omega \to \mathbb{R}^m$, the space $\mathcal{M}$ also contains the function $V(\cdot) \doteq \mathbb{I}_{\Omega \setminus B}(\cdot)\psi(\cdot) + \mathbb{I}_B(\cdot)W(\cdot)$. In the subsequent analysis we work with spaces $\mathcal{M} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P, \mathbb{R}^m)$ which are decomposable. Now let $\mathcal{M}$ be a decomposable space and $h : \mathbb{R}^m \times \Omega \to \overline{\mathbb{R}}$ be a random lower semicontinuous function[10]. Then

$$\mathbb{E}\left[\inf_{y \in \mathbb{R}^m} h(y, \omega)\right] = \inf_{Y \in \mathcal{M}} \mathbb{E}\big[H_Y\big], \tag{4.12}$$

where $H_Y(\omega) \doteq h(Y(\omega), \omega)$, provided that the right hand side of (4.12) is less than $+\infty$. Moreover, if the common value of both sides in (4.12) is not $-\infty$, then

$$\bar{Y} \in \operatorname{argmin}_{Y \in \mathcal{M}} \mathbb{E}[H_Y] \text{ ifandonlyif } \bar{Y}(\omega) \in \operatorname{argmin}_{y \in \mathbb{R}^m} h(y, \omega) \text{ for a.e. } \omega \in \Omega.$$

Clearly the above interchangeability principle can be applied to a maximization, rather than minimization, procedure simply by replacing function $h(y, \omega)$ with $-h(y, \omega)$.

Let us return to the dual formulation (4.11) of the risk function (4.10). By using the interchangeability formula (4.12) with $h(\alpha, \omega) \doteq -[\alpha Z(\omega) - g^*(\alpha)]$ we obtain

$$\rho(Z) = \sup_{\zeta \in \mathcal{Z}^*} \big\{\langle \zeta, Z \rangle - \mathbb{E}[g^*(\zeta)]\big\}. \tag{4.13}$$

[9] It is allowed here for $\mathbb{E}[g(Z)]$ to take value $+\infty$, but not $-\infty$ since the corresponding risk function is required to be proper.

[10] A function $h : \mathbb{R}^m \times \Omega \to \overline{\mathbb{R}}$ is said to be *random lower semicontinuous* if its epigraphical mapping is closed valued and measurable. Random lower semicontinuous functions are also called *normal integrands* (*cf.* [302, Definition 14.27]).

It follows that ρ is convex and lower semicontinuous, and representation (4.6) holds with

$$\rho^*(\zeta) = \mathbb{E}[g^*(\zeta)].$$

Moreover, if the function g is nondecreasing, then ρ satisfies the monotonicity condition (A2). However, the risk function ρ does not satisfy condition (A3) unless $g(z) \equiv z$, and ρ is not positively homogeneous unless g is positively homogeneous.

4.4 Examples of Risk Functions

In this section we discuss several examples of risk functions which are commonly used in applications. In the following, P is a (reference) probability measure on $(\Omega, \mathcal{F})$ and, unless stated otherwise, all expectations and probabilistic statements are made with respect to P.

Example 1 (Mean-variance risk function). Consider

$$\rho(Z) \doteq \mathbb{E}[Z] + c\mathbb{V}\mathrm{ar}[Z], \tag{4.14}$$

where $c \geq 0$ is a given constant. It is natural to use here the space $\mathcal{Z} \doteq \mathcal{L}_2(\Omega, \mathcal{F}, P)$ since for any $Z \in \mathcal{L}_2(\Omega, \mathcal{F}, P)$ the expectation $\mathbb{E}[Z]$ and variance $\mathbb{V}\mathrm{ar}[Z]$ are well defined and finite.

By direct calculation we can verify that

$$\mathbb{V}\mathrm{ar}[Z] = \big\|Z - \mathbb{E}[Z]\big\|^2 = \sup_{\zeta \in \mathcal{Z}} \big\{\langle \zeta, Z - \mathbb{E}[Z]\rangle - \tfrac{1}{4}\|\zeta\|^2\big\},$$

where the scalar products and the norms are in the sense of the (Hilbert) space $\mathcal{L}_2(\Omega, \mathcal{F}, P)$. Since $\langle \zeta, Z - \mathbb{E}[Z]\rangle = \langle \zeta - \mathbb{E}[\zeta], Z\rangle$ we can rewrite the last expression as follows:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[Z] &= \sup_{\zeta \in \mathcal{Z}} \big\{\langle \zeta - \mathbb{E}[\zeta], Z\rangle - \tfrac{1}{4}\|\zeta\|^2\big\} \\ &= \sup_{\zeta \in \mathcal{Z}} \big\{\langle \zeta - \mathbb{E}[\zeta], Z\rangle - \tfrac{1}{4}\mathbb{V}\mathrm{ar}[\zeta] - \tfrac{1}{4}\big(\mathbb{E}[\zeta]\big)^2\big\}.\end{aligned}$$

Consequently, the above maximization can be restricted to such $\zeta \in \mathcal{Z}$ that $\mathbb{E}[\zeta] = 0$, and hence

$$\mathbb{V}\mathrm{ar}[Z] = \sup_{\substack{\zeta \in \mathcal{Z} \\ \mathbb{E}[\zeta]=0}} \big\{\langle \zeta, Z\rangle - \tfrac{1}{4}\mathbb{V}\mathrm{ar}[\zeta]\big\}.$$

Therefore the risk function ρ, defined in (4.14), can be equivalently expressed for $c > 0$ as follows:

$$\begin{aligned}\rho(Z) &= \mathbb{E}[Z] + c \sup_{\substack{\zeta\in\mathcal{Z}\\ \mathbb{E}[\zeta]=0}} \Big\{ \langle \zeta, Z\rangle - \tfrac{1}{4}\mathbb{V}\mathrm{ar}\ [\zeta] \Big\} \\ &= \sup_{\substack{\zeta\in\mathcal{Z}\\ \mathbb{E}[\zeta]=1}} \Big\{ \langle \zeta, Z\rangle - \frac{1}{4c}\mathbb{V}\mathrm{ar}[\zeta] \Big\}.\end{aligned}$$

It follows that for any $c \ge 0$ the function ρ is convex and lower semicontinuous. Furthermore

$$\rho^*(\zeta) = \begin{cases} \frac{1}{4c}\mathbb{V}\mathrm{ar}[\zeta], & \text{if } \mathbb{E}[\zeta]=1, \\ +\infty, & \text{otherwise.} \end{cases}$$

The function ρ satisfies the translation equivariance condition (A3), because the domain of its conjugate contains only ζ such that $\mathbb{E}[\zeta]=1$. However, for any $c \ge 0$ the function ρ is not positively homogeneous and it does not satisfy the monotonicity condition (A2), because the domain of ρ^* contains density functions which are not non-negative.

Example 2 (Mean-deviation risk function of order p). For $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$, $\mathcal{Z}^* \doteq \mathcal{L}_q(\Omega, \mathcal{F}, P)$ and $c \ge 0$ consider

$$\rho(Z) \doteq \mathbb{E}[Z] + c\left(\mathbb{E}\big[|Z - \mathbb{E}[Z]|^p\big]\right)^{1/p}. \tag{4.15}$$

Note that $\left(\mathbb{E}\big[|Z|^p\big]\right)^{1/p} = \|Z\|_p$, where $\|\cdot\|_p$ denotes the norm of the space $\mathcal{L}_p(\Omega, \mathcal{F}, P)$. We have that

$$\|Z\|_p = \sup_{\|\zeta\|_q \le 1} \langle \zeta, Z\rangle,$$

and hence

$$\left(\mathbb{E}\big[|Z - \mathbb{E}[Z]|^p\big]\right)^{1/p} = \sup_{\|\zeta\|_q \le 1} \langle \zeta, Z - \mathbb{E}[Z]\rangle = \sup_{\|\zeta\|_q \le 1} \langle \zeta - \mathbb{E}[\zeta], Z\rangle.$$

It follows that representation (4.9) holds with the set $\mathcal{A}$ given by

$$\mathcal{A} = \left\{\zeta' \in \mathcal{Z}^* : \zeta' = 1 + \zeta - \mathbb{E}[\zeta],\ \|\zeta\|_q \le c\right\}.$$

We obtain here that ρ satisfies conditions (A1), (A3) and (A4).

The monotonicity condition (A2) is more involved. Suppose that $p = 1$. Then $q = +\infty$ and hence for any $\zeta' \in \mathcal{A}$ and P-almost every $\omega \in \Omega$ we have

$$\zeta'(\omega) = 1 + \zeta(\omega) - \mathbb{E}[\zeta] \ge 1 - |\zeta(\omega)| - \mathbb{E}[\zeta] \ge 1 - 2c.$$

It follows that if $c \in [0, 1/2]$, then $\zeta'(\omega) \ge 0$ for P-almost every $\omega \in \Omega$, and hence condition (A2) follows. Conversely, take $\zeta \doteq c(-\mathbb{I}_A + \mathbb{I}_{\Omega\setminus A})$, for some $A \in \mathcal{F}$, and $\zeta' = 1 + \zeta - \mathbb{E}[\zeta]$. We have that $\|\zeta\|_\infty = c$ and $\zeta'(\omega) = 1 - 2c + 2cP(A)$ for all $\omega \in A$ It follows that if $c > 1/2$, then $\zeta'(\omega) < 0$ for all $\omega \in A$, provided that $P(A)$ is small enough. We obtain that for $c > 1/2$

the monotonicity property (A2) does not hold if the following condition is satisfied:

$$\text{For any } \varepsilon > 0 \text{ there exists } A \in \mathcal{F} \text{ such that } \varepsilon > P(A) > 0. \tag{4.16}$$

That is, for $p = 1$ the mean-deviation function ρ satisfies (A2) if, and provided that condition (4.16) holds, only if $c \in [0, 1/2]$.

Suppose now that $p > 1$. For a set $A \in \mathcal{F}$ and $\alpha > 0$ let us take $\zeta \doteq -\alpha \mathbb{I}_A$ and $\zeta' = 1 + \zeta - \mathbb{E}[\zeta]$. Then $\|\zeta\|_q = \alpha P(A)^{1/q}$ and $\zeta'(\omega) = 1 - \alpha + \alpha P(A)$ for all $\omega \in A$. It follows that if $p > 1$, then for any $c > 0$ the mean-deviation function ρ does not satisfy (A2) provided that condition (4.16) holds.

Example 3 (Mean-upper-semideviation risk function of order p). Let $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$ and for $c \geq 0$ consider[11]

$$\rho(Z) \doteq \mathbb{E}[Z] + c\left(\mathbb{E}\left[\left[Z - \mathbb{E}[Z]\right]_+^p\right]\right)^{1/p}. \tag{4.17}$$

For any $c \geq 0$ this function satisfies conditions (A1), (A3) and (A4), and similarly to the derivations of Example 2 it can be shown that representation (4.9) holds with the set $\mathcal{A}$ given by

$$\mathcal{A} = \left\{\zeta' \in \mathcal{Z}^* : \zeta' = 1 + \zeta - \mathbb{E}[\zeta],\ \|\zeta\|_q \leq c,\ \zeta \succeq 0\right\}. \tag{4.18}$$

Since $|\mathbb{E}[\zeta]| \leq \mathbb{E}|\zeta| \leq \|\zeta\|_q$ for any $\zeta \in \mathcal{L}_q(\Omega, \mathcal{F}, P)$, we have that every element of the above set $\mathcal{A}$ is non-negative and has its expected value equal to 1. This means that the monotonicity condition (A2) holds true, if and, provided that condition (4.16) holds, only if $c \in [0, 1]$ (see [314]). That is, ρ is a coherent risk function if $c \in [0, 1]$.

Example 4 (Mean-upper-semivariance from a target). Let $\mathcal{Z} \doteq \mathcal{L}_2(\Omega, \mathcal{F}, P)$ and for weight $c \geq 0$ and target $\tau \in \mathbb{R}$ consider

$$\rho(Z) \doteq \mathbb{E}[Z] + c\,\mathbb{E}\left[\left[Z - \tau\right]_+^2\right].$$

We can now use (4.13) with $g(z) = z + c(z - \tau)_+^2$. Since

$$g^*(\alpha) = \begin{cases} (\alpha - 1)^2/4c + \tau(\alpha - 1), & \text{if } \alpha \geq 1, \\ +\infty, & \text{otherwise,} \end{cases}$$

we obtain that

$$\rho(Z) = \sup_{\zeta \in \mathcal{Z},\, \zeta(\cdot) \geq 1} \left\{\mathbb{E}[\zeta Z] - \tau \mathbb{E}[\zeta - 1] - \frac{1}{4c}\mathbb{E}[(\zeta - 1)^2]\right\}.$$

Consequently, representation (4.7) holds with[12] $\mathcal{A} = \{\zeta \in \mathcal{Z} : \zeta - 1 \succeq 0\}$ and

[11] We denote $[a]_+^p \doteq (\max\{0, a\})^p$.

[12] Recall that $\mathcal{A} \doteq \operatorname{dom}(\rho^*)$.

$$\rho^*(\zeta) = \tau\mathbb{E}[\zeta - 1] + \frac{1}{4c}\mathbb{E}[(\zeta-1)^2], \quad \zeta \in \mathcal{A}.$$

If $c > 0$, none of the conditions (A3) and (A4) is satisfied by this risk function.

Example 5 (Mean-upper-semideviation of order p from a target). Let $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$ and for $c \geq 0$ and $\tau \in \mathbb{R}$ consider

$$\rho(Z) \doteq \mathbb{E}[Z] + c\left(\mathbb{E}\left[\left[Z-\tau\right]_+^p\right]\right)^{1/p}. \tag{4.19}$$

For any $c \geq 0$ and τ this risk function satisfies conditions (A1) and (A2), but not (A3) and (A4), if $c > 0$. We have

$$\begin{aligned}\left(\mathbb{E}\left[\left[Z-\tau\right]_+^p\right]\right)^{1/p} &= \sup_{\|\zeta\|_q \leq 1} \mathbb{E}\big(\zeta[Z-\tau]_+\big)\\ &= \sup_{\|\zeta\|_q \leq 1,\, \zeta(\cdot)\geq 0} \mathbb{E}\big(\zeta[Z-\tau]_+\big)\\ &= \sup_{\|\zeta\|_q \leq 1,\, \zeta(\cdot)\geq 0} \mathbb{E}\big(\zeta[Z-\tau]\big)\\ &= \sup_{\|\zeta\|_q \leq 1,\, \zeta(\cdot)\geq 0} \mathbb{E}\big[\zeta Z - \tau\zeta\big].\end{aligned}$$

We obtain that representation (4.7) holds with $\mathcal{A} = \{\zeta \in \mathcal{Z}^* : \|\zeta\|_q \leq c,\ \zeta \succeq 0\}$ and $\rho^*(\zeta) = \tau\mathbb{E}[\zeta]$ for $\zeta \in \mathcal{A}$.

Example 6. Let $v : \mathbb{R} \to \overline{\mathbb{R}}$ be a proper lower semicontinuous convex function. For $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$ consider the function

$$\rho(Z) \doteq \inf_{\alpha\in\mathbb{R}} \mathbb{E}\big[Z + v(Z-\alpha)\big]. \tag{4.20}$$

Assume that functions $\psi_\alpha(z) \doteq z + v(z-\alpha)$, $\alpha \in \mathbb{R}$, are bounded from below by a P-integrable function, and hence $\rho(Z) > -\infty$ for all $Z \in \mathcal{Z}$. Since the function $(Z, \alpha) \mapsto \mathbb{E}\big[Z + v(Z-\alpha)\big]$ is convex, it follows that $\rho(\cdot)$ is convex. Also $\rho(Z+a) = \rho(Z) + a$ for any $a \in \mathbb{R}$ and $Z \in \mathcal{Z}$. This can be shown by making the change of variables $z \mapsto z + a$ in the calculation of $\rho(Z+a)$. That is, ρ satisfies conditions (A1) and (A3).

Let us calculate the conjugate of ρ:

$$\begin{aligned}\rho^*(\zeta) &= \sup_{Z\in\mathcal{Z}} \big\{\mathbb{E}[\zeta Z] - \rho(Z)\big\}\\ &= \sup_{Z\in\mathcal{Z},\,\alpha\in\mathbb{R}} \mathbb{E}\big[\zeta Z - Z - v(Z-\alpha)\big]\\ &= \sup_{Z\in\mathcal{Z},\,\alpha\in\mathbb{R}} \mathbb{E}\big[(Z+\alpha)\zeta - Z - \alpha - v(Z)\big]\\ &= \sup_{Z\in\mathcal{Z}} \big\{\mathbb{E}[\zeta Z - Z - v(Z)]\big\} + \sup_{\alpha\in\mathbb{R}} \big\{\alpha(\mathbb{E}[\zeta]-1)\big\}.\end{aligned} \tag{4.21}$$

By the interchangeability formula (4.12), the first term in (4.21) can be expressed as follows:

$$\sup_{Z\in\mathcal{Z}} \mathbb{E}\big[\zeta Z - Z - v(Z)\big] = \mathbb{E}\left[\sup_{z\in\mathbb{R}}\big\{z(\zeta-1) - v(z)\big\}\right] = \mathbb{E}\left[v^*(\zeta-1)\right],$$

where $v^*(\cdot)$ is the conjugate function of $v(\cdot)$. The supremum with respect to α in (4.21) is $+\infty$, unless $\mathbb{E}[\zeta]=1$. We conclude that

$$\rho^*(\zeta) = \begin{cases} \mathbb{E}\left[v^*(\zeta-1)\right], & \text{if } \mathbb{E}[\zeta]=1, \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.22}$$

The function ρ satisfies the monotonicity condition (A2) if and only if its domain contains only probability density functions. This is equivalent to the condition that $\mathbb{E}[v^*(\zeta-1)] = +\infty$ for any such $\zeta \in \mathcal{Z}^*$ that the event '$\zeta(\omega) < 0$' happens with positive probability. In particular, ρ satisfies (A2) if $v^*(t) = +\infty$ for $t < -1$. This is the same as requiring that the function $\phi(z) \doteq z + v(z)$ is monotonically nondecreasing on $\mathbb{R}$.

Example 7 (Conditional value at risk). For $\mathcal{Z} \doteq \mathcal{L}_1(\Omega,\mathcal{F},P)$, $\mathcal{Z}^* \doteq \mathcal{L}_\infty(\Omega, \mathcal{F},P)$ and constants $\varepsilon_1 \ge 0$ and $\varepsilon_2 \ge 0$ consider

$$\rho(Z) \doteq \mathbb{E}[Z] + \inf_{\alpha\in\mathbb{R}} \mathbb{E}\big(\varepsilon_1[\alpha - Z]_+ + \varepsilon_2[Z-\alpha]_+\big). \tag{4.23}$$

Note that the above function ρ is of the form (4.20) with

$$v(z) \doteq \varepsilon_1[-z]_+ + \varepsilon_2[z]_+. \tag{4.24}$$

We have here that the function $z+v(z)$ is positively homogeneous, and monotonically nondecreasing if and only if $\varepsilon_1 \le 1$. It follows that for any $\varepsilon_1 \in [0,1]$ and $\varepsilon_2 \ge 0$, the above function ρ is a coherent risk function satisfying conditions (A1)–(A4). Moreover,

$$v^*(t) = \begin{cases} 0, & \text{if } t \in [-\varepsilon_1, \varepsilon_2], \\ +\infty, & \text{otherwise.} \end{cases}$$

Consequently we have that, for any $\varepsilon_1 \ge 0$ and $\varepsilon_2 \ge 0$, representation (4.9) holds with

$$\mathcal{A} = \big\{\zeta \in \mathcal{Z}^* : 1-\varepsilon_1 \le \zeta(\omega) \le 1+\varepsilon_2,\ \text{a.e. } \omega\in\Omega,\ \mathbb{E}[\zeta]=1\big\}.$$

For $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ we can write ρ in the form

$$\rho(Z) = (1-\varepsilon_1)\mathbb{E}[Z] + \varepsilon_1 \mathrm{CV@R}_\kappa[Z],$$

where $\kappa \doteq \varepsilon_2/(\varepsilon_1+\varepsilon_2)$ and

$$\mathrm{CV@R}_\kappa[Z] \doteq \inf_{a\in\mathbb{R}} \left\{ a + \frac{1}{1-\kappa}\mathbb{E}\big([Z-a]_+\big) \right\}$$

is the so-called Conditional Value at Risk function, [300]. By the above analysis we have that $\mathrm{CV@R}_\kappa[Z]$ is a coherent risk function for any $\kappa \in (0,1)$ and the corresponding set $\mathcal{A}$ is given by

$$\mathcal{A} = \big\{ \zeta \in \mathcal{Z}^* : 0 \le \zeta(\omega) \le (1-\kappa)^{-1}, \text{ a.e. } \omega \in \Omega,\ \mathbb{E}[\zeta] = 1 \big\}.$$

4.5 Stochastic Dominance Conditions

In all examples considered in Section 4.4, the space $\mathcal{Z}$ was given by $\mathcal{L}_p(\Omega, \mathcal{F}, P)$ with $\mathcal{Z}^* \doteq \mathcal{L}_q(\Omega, \mathcal{F}, P)$ and, moreover, the risk functions $\rho(Z)$ discussed there were dependent only on the distribution of Z. That is, each risk function $\rho(Z)$, considered in Section 4.4, could be formulated in terms of the cumulative distribution function (cdf) $F_Z(z) \doteq P(Z \le z)$ associated with $Z \in \mathcal{Z}$. In other words these risk functions satisfied the following condition:

(D) If $Z_1, Z_2 \in \mathcal{Z}$ are such that $P(Z_1 \le z) = P(Z_2 \le z)$ for all $z \in \mathbb{R}$, then $\rho(Z_1) = \rho(Z_2)$.

We say that risk function $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ is *law invariant* if it satisfies the above condition (D). For law invariant risk functions it makes sense to discuss their monotonicity properties with respect to various stochastic orders defined for (real valued) random variables.

Many stochastic orders can be characterized by a class $\mathcal{G}$ of functions $g : \mathbb{R} \to \mathbb{R}$ as follows. For (real valued) random variables Z_1 and Z_2 it is said that Z_2 dominates Z_1, denoted $Z_2 \succeq_{\mathcal{G}} Z_1$, if $\mathbb{E}[g(Z_2)] \ge \mathbb{E}[g(Z_1)]$ for all $g \in \mathcal{G}$ for which the corresponding expectations do exist. This stochastic order is called the *integral stochastic order* with *generator* $\mathcal{G}$. We refer to [237, Chapter 2] for a thorough discussion of this concept. For example, the *usual stochastic order*, written $Z_2 \succeq_{\mathrm{st}} Z_1$, corresponds to the generator $\mathcal{G}$ formed by all non-decreasing functions $g : \mathbb{R} \to \mathbb{R}$. It is possible to show that $Z_2 \succeq_{\mathrm{st}} Z_1$ if and only if $F_{Z_2}(z) \le F_{Z_1}(z)$ for all $z \in \mathbb{R}$ (*e.g.*, [237, Theorem 1.2.8]). We say that the integral stochastic order is *increasing* if all functions in the set $\mathcal{G}$ are non-decreasing. The usual stochastic order is an example of increasing integral stochastic order.

We say that a law invariant risk function ρ is *consistent* with the integral stochastic order if $Z_2 \succeq_{\mathcal{G}} Z_1$ implies $\rho(Z_2) \ge \rho(Z_1)$ for all $Z_1, Z_2 \in \mathcal{Z}$, *i.e.*, ρ is monotone with respect to $\succeq_{\mathcal{G}}$. For an increasing integral stochastic order we have that if $Z_2(\omega) \ge Z_1(\omega)$ for a.e. $\omega \in \Omega$, then $g(Z_2(\omega)) \ge g(Z_1(\omega))$ for any $g \in \mathcal{G}$ and a.e. $\omega \in \Omega$, and hence $\mathbb{E}[g(Z_2(\omega))] \ge \mathbb{E}[g(Z_1(\omega))]$. That is, if $Z_2 \succeq Z_1$ in the almost sure sense, then $Z_2 \succeq_{\mathcal{G}} Z_1$. It follows that if ρ is law invariant and consistent with respect to an increasing integral stochastic order, then it satisfies the monotonicity condition (A2). In other words if ρ does not satisfy condition (A2), then it cannot be consistent with any increasing integral stochastic order. In particular, for $c > 0$ the mean-variance risk function, defined in (4.14), is not consistent with any increasing integral stochastic order, and for $p > 1$ the mean-deviation risk function, defined in (4.15), is not consistent with any increasing integral stochastic order provided that condition (4.16) holds.

Consider now the usual stochastic order. It is well known that $Z_2 \succeq_{\text{st}} Z_1$ if and only if there exists a probability space $(\Omega, \mathcal{F}, P)$ and random variables $\hat{Z}_1$ and $\hat{Z}_2$ on it such that[13] $\hat{Z}_1 \overset{D}{\sim} Z_1$ and $\hat{Z}_2 \overset{D}{\sim} Z_2$, and $\hat{Z}_2(\omega) \geq \hat{Z}_1(\omega)$ for all $\omega \in \Omega$ (*e.g.*, [237, Theorem 1.2.4]). In our context, this relation between the usual stochastic order and the almost sure order cannot be used directly, because we are not allowed to change freely the probability space $(\Omega, \mathcal{F}, P)$, which is an integral part of our definition of a risk function.

However, if our space $(\Omega, \mathcal{F}, P)$ is sufficiently rich, so that a *uniform*[14] random variable $U(\omega)$ exists on this space, we can easily link the monotonicity assumption (A2) with the consistency with the usual stochastic order. Suppose that the risk function ρ is law invariant and satisfies the monotonicity condition (A2). Recall that $Z_2 \succeq_{\text{st}} Z_1$ if and only if $F_{Z_2}(z) \leq F_{Z_1}(z)$ for all $z \in \mathbb{R}$. Consider random variables $\hat{Z}_1 \doteq F_{Z_1}^{-1}(U)$ and $\hat{Z}_2 \doteq F_{Z_2}^{-1}(U)$, where the inverse distribution function is defined as

$$F_Z^{-1}(t) \doteq \inf \{z : F_Z(z) \geq t\}.$$

We obtain that $\hat{Z}_2(\omega) \geq \hat{Z}_1(\omega)$ for all $\omega \in \Omega$, and by virtue of (A2), $\rho(\hat{Z}_2) \geq \rho(\hat{Z}_1)$. By construction, $\hat{Z}_1$ has the same distribution as Z_1, and $\hat{Z}_2$ has the same distribution as Z_2. Since the risk function is law invariant, we conclude that $\rho(Z_2) \geq \rho(Z_1)$. Consequently, the risk function ρ is consistent with the usual stochastic order. It follows that in a sufficiently rich probability space the monotonicity condition (A2) and the consistency with the usual stochastic order are equivalent (for law invariant risk functions).

It is said that Z_2 is bigger than Z_1 in *increasing convex order*, written $Z_2 \succeq_{\text{icx}} Z_1$, if $\mathbb{E}[g(Z_2)] \geq \mathbb{E}[g(Z_1)]$ for all increasing convex functions $g : \mathbb{R} \to \mathbb{R}$ such that the expectations exist. Clearly this is an integral stochastic order with the corresponding generator given by the set of increasing convex functions. It is the counterpart of the classical stochastic dominance relation, which is the increasing concave order (recall that we are dealing here with minimization rather than maximization procedures). Consider the setting of Example 6 with risk function ρ defined in (4.20). Suppose that the function $\phi(z) \doteq z + v(z)$ is monotonically nondecreasing on $\mathbb{R}$. Note that $\phi(\cdot)$ is convex, since $v(\cdot)$ is convex. We obtain that if $Z_2 \geq_{\text{icx}} Z_1$, then $\mathbb{E}[\phi(Z_2 - \alpha)] \geq \mathbb{E}[\phi(Z_2 -$

[13] The notation $X \overset{D}{\sim} Y$ means that random variables X and Y, which can be defined on different probability spaces, have the same cumulative distribution function.

[14] Random variable $U : \Omega \to [0, 1]$ is said to be uniform if $P(U \leq z) = z$ for every $z \in [0, 1]$.

$\alpha)]$ for any fixed $\alpha \in \mathbb{R}$, and hence (by taking minimum over $\alpha \in \mathbb{R}$) that $\rho(Z_2) \geq \rho(Z_1)$. That is, the risk function defined in (4.20) is consistent with the increasing convex order. We have in this way re-established the stochastic dominance consistency result of [248].

The mean-upper-semideviation risk function of order $p \geq 1$ (Example 3) is also consistent with the increasing convex order, provided that $c \in [0, 1]$. We can prove this for $p = 1$ as follows (see [246]).

Suppose that $Z_2 \succeq_{\text{icx}} Z_1$. First, using $g(z) \doteq z$ we see that

$$\mathbb{E}[Z_1] \leq \mathbb{E}[Z_2]. \tag{4.25}$$

Second, setting $g(z) \doteq \big(z - \mathbb{E}[Z_1]\big)_+$ we find that

$$\mathbb{E}\Big[\big(Z_1 - \mathbb{E}[Z_1]\big)_+\Big] \leq \mathbb{E}\Big[\big(Z_2 - \mathbb{E}[Z_1]\big)_+\Big].$$

Using (4.25) we can continue this estimate as follows:

$$\begin{aligned}\mathbb{E}\Big[\big(Z_1 - \mathbb{E}[Z_1]\big)_+\Big] &\leq \mathbb{E}\Big[\big(Z_2 - \mathbb{E}[Z_2] + \mathbb{E}[Z_2] - \mathbb{E}[Z_1]\big)_+\Big] \\ &\leq \mathbb{E}\Big[\big(Z_2 - \mathbb{E}[Z_2]\big)_+\Big] + \mathbb{E}[Z_2] - \mathbb{E}[Z_1].\end{aligned}$$

This can be rewritten as

$$\mathbb{E}[Z_1] + \mathbb{E}\Big[\big(Z_1 - \mathbb{E}[Z_1]\big)_+\Big] \leq \mathbb{E}[Z_2] + \mathbb{E}\Big[\big(Z_2 - \mathbb{E}[Z_2]\big)_+\Big], \tag{4.26}$$

which is the required relation $\rho(Z_1) \leq \rho(Z_2)$ for $c = 1$. Combining inequalities (4.25) and (4.26) with coefficients $1 - c$ and c, we obtain the required result for any $c \in [0, 1]$. The proof for $p > 1$ can be found in [247] (in the stochastic dominance setting).

4.6 Differentiability of Risk Functions

In this section we discuss differentiability properties of risk functions. In the analysis of optimization of risk measures we also have to deal with composite functions of the form

$$\phi(x) \doteq \rho(F(x)).$$

Here $F : \mathbb{R}^n \to \mathcal{Z}$ is a mapping defined by $[F(x)](\cdot) \doteq f(x, \cdot)$, associated with a function $f : \mathbb{R}^n \times \Omega \to \mathbb{R}$. Of course, in order for this mapping F to be well defined we have to assume that the random variable $Z(\omega) = f(x, \omega)$ belongs to $\mathcal{Z}$ for any $x \in \mathbb{R}^n$. We say that the mapping F is *convex* if the function $f_\omega(\cdot) \doteq f(\cdot, \omega)$ is convex for every $\omega \in \Omega$. It is not difficult to verify and is well known that the composite function $\phi(x)$ is convex if F is convex and ρ is convex and satisfies the monotonicity condition (A2). Let us emphasize that in order to preserve convexity of the composite function ϕ we need convexity of F and ρ *and* the monotonicity property (A2).

Consider a point $\bar{Z} \in \mathcal{Z}$ such that $\rho(\bar{Z})$ is finite valued. Since it is assumed that ρ is proper, this means that $\bar{Z} \in \operatorname{dom}(\rho)$. The following limit (provided that it exists)

$$\rho'(\bar{Z}, Z) \doteq \lim_{t \downarrow 0} \frac{\rho(\bar{Z} + tZ) - \rho(\bar{Z})}{t}$$

is called the *directional derivative* of ρ at $\bar{Z}$ in direction Z. If this limit exists for all $Z \in \mathcal{Z}$, it is said that ρ is directionally differentiable at $\bar{Z}$. It is said that ρ is Hadamard directionally differentiable at $\bar{Z}$, if ρ is directionally differentiable at $\bar{Z}$ and, moreover, the following limit holds:

$$\rho'(\bar{Z}, Z) = \lim_{\substack{Z' \to Z \\ t \downarrow 0}} \frac{\rho(\bar{Z} + tZ') - \rho(\bar{Z})}{t}.$$

It can be observed that $\rho'(\bar{Z}, Z)$ is just the one sided derivative of the function $g(t) \doteq \rho(\bar{Z} + tZ)$ at $t = 0$. If ρ is convex, then the function $g : \mathbb{R} \to \overline{\mathbb{R}}$ is also convex, and hence $\rho'(\bar{Z}, Z)$ exists, although it can take values $+\infty$ or $-\infty$.

It said that an element $\mu \in \mathcal{Z}^*$ is a *subgradient* of ρ at $\bar{Z}$ if

$$\rho(Z) \geq \rho(\bar{Z}) + \langle \mu, Z - \bar{Z} \rangle, \quad \forall\, Z \in \mathcal{Z}.$$

The set of all subgradients of ρ, at $\bar{Z}$, is called the *subdifferential* of ρ and denoted $\partial\rho(\bar{Z})$. It is said that ρ is subdifferentiable at $\bar{Z}$ if $\partial\rho(\bar{Z})$ is non-empty. By convex analysis we have that if ρ is convex and continuous at $\bar{Z}$, then it is subdifferentiable at $\bar{Z}$, and, moreover, if $\mathcal{Z}$ is a Banach space (equipped with its norm topology), then ρ is Hadamard directionally differentiable at $\bar{Z}$.

It is said that ρ is Gâteaux (Hadamard) differentiable at $\bar{Z}$ if it is (Hadamard) directionally differentiable at $\bar{Z}$ and there exists $\bar{\mu} \in \mathcal{Z}^*$ such that $\rho'(\bar{Z}, Z) = \langle \bar{\mu}, Z \rangle$ for all $Z \in \mathcal{Z}$. The functional $\bar{\mu}$ represents the derivative of ρ at $\bar{Z}$ and denoted $\nabla\rho(\bar{Z})$. If the space $\mathcal{Z}$ is finite dimensional, then the concept of Hadamard differentiability coincides with the usual concept of differentiability. By convex analysis we have the following.

Theorem 3. *Suppose that $\mathcal{Z}$ is a Banach space (e.g., $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$), and ρ is convex and finite valued and continuous at $\bar{Z}$. Then ρ is subdifferentiable and Hadamard directionally differentiable at $\bar{Z}$, and the following formulas hold:*

$$\partial\rho(\bar{Z}) = \operatorname{argmax}_{\mu \in \mathcal{Z}^*} \left\{ \langle \mu, \bar{Z} \rangle - \rho^*(\mu) \right\}, \tag{4.27}$$

$$\rho'(\bar{Z}, Z) = \sup\nolimits_{\mu \in \partial\rho(\bar{Z})} \langle \mu, Z \rangle. \tag{4.28}$$

Moreover, ρ is Hadamard differentiable at $\bar{Z}$ if and only if $\partial\rho(\bar{Z}) = \{\bar{\mu}\}$ is a singleton, in which case $\nabla\rho(\bar{Z}) = \bar{\mu}$.

As we mentioned earlier, if $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$ and ρ satisfies conditions (A1) and (A2), then ρ is continuous and subdifferentiable at every point of the interior of its domain, see [314]. In particular, if ρ is real valued, then ρ is continuous and subdifferentiable at every point of $\mathcal{Z}$ and formulas (4.27) and (4.28) hold. Moreover, if ρ is a real valued coherent risk function, then representation (4.9) holds and

$$\partial\rho(\bar{Z}) = \operatorname{argmax}_{\zeta\in\mathcal{A}} \langle\zeta, \bar{Z}\rangle. \tag{4.29}$$

Consider now the composite function $\phi(x) \doteq \rho(F(x))$. Since $f_\omega(\cdot)$ is real valued, we have that if $f_\omega(\cdot)$ is convex, then it is directionally differentiable at every point $\bar{x} \in \mathbb{R}^n$ and its directional derivative $f'_\omega(\bar{x}, x)$ is finite valued. By using the chain rule for directional derivatives and (4.28) we obtain the following differentiability properties of the composite function, at a point $\bar{x} \in \mathbb{R}^n$ (*cf.* [314]).

Proposition 1. *Suppose that $\mathcal{Z}$ is a Banach space, the mapping $F : \mathbb{R}^n \to \mathcal{Z}$ is convex, the function ρ is convex, finite valued and continuous at $\bar{Z} \doteq F(\bar{x})$. Then the composite function $\phi(x) = \rho(F(x))$ is directionally differentiable at $\bar{x}$, its directional derivative $\phi'(\bar{x}, x)$ is finite valued for every $x \in \mathbb{R}^n$ and*

$$\phi'(\bar{x}, x) = \sup_{\mu\in\partial\rho(\bar{Z})} \int_\Omega f'_\omega(\bar{x}, x)\, d\mu(\omega). \tag{4.30}$$

Moreover, if $\partial\rho(\bar{Z}) = \{\bar{\mu}\}$ is a singleton, then the composite function ϕ is differentiable at $\bar{x}$ if and only if $f_\omega(\cdot)$ is differentiable at $\bar{x}$ for $\bar{\mu}$-almost every ω, in which case

$$\nabla\phi(\bar{x}) = \int_\Omega \nabla f_\omega(\bar{x})\, d\bar{\mu}(\omega).$$

It is also possible to write the above differentiability formulas in terms of subdifferentials. Suppose that F is convex. Then for any[15] measure $\mu \in \mathcal{Z}^*_+$ the integral function $\psi_\mu(x) \doteq \int_\Omega f_\omega(x)\, d\mu(\omega)$ is also convex. Moreover, if the integral function $\psi_\mu(\cdot)$ is finite valued (and hence continuous) in a neighborhood of a point $\bar{x} \in \mathbb{R}^n$, then

$$\psi'_\mu(\bar{x}, x) = \int_\Omega f'_\omega(\bar{x}, x)\, d\mu(\omega),$$

and by Strassen's disintegration theorem the following interchangeability formula holds:

$$\partial\psi_\mu(\bar{x}) = \int_\Omega \partial f_\omega(\bar{x})\, d\mu(\omega). \tag{4.31}$$

The integral in the right hand side of (4.31) is understood as the set of all vectors of the form $\int_\Omega \delta(\omega)\, d\mu(\omega)$, where $\delta(\omega)$ is a μ-integrable selection[16] of $\partial f_\omega(\bar{x})$.

[15] Recall that $\mathcal{Z}^*_+$ denotes the set of non-negative measures $\mu \in \mathcal{Z}^*$.

[16] It is said that $\delta(\omega)$ is a selection of $\partial f_\omega(\bar{x})$ if $\delta(\omega) \in \partial f_\omega(\bar{x})$ for almost every ω.

Suppose that the assumptions of Proposition 1 hold, and monotonicity condition (A2) is satisfied and hence ϕ is convex and $\partial\rho(\bar{Z}) \subset \mathcal{Z}_+^*$. Now formula (4.30) means that $\phi'(\bar{x}, \cdot)$ is equal to the supremum of $\psi'_\mu(\bar{x}, \cdot)$ over $\mu \in \partial\rho(\bar{Z})$. The functions $\psi'_\mu(\bar{x}, \cdot)$ are convex and positively homogeneous, and hence $\partial\phi(\bar{x})$ is equal to the topological closure of the union of the sets $\partial\psi'_\mu(\bar{x})$ over $\mu \in \partial\rho(\bar{Z})$. Consequently, we obtain that formula (4.30) can be written in the following equivalent form:[17]

$$\partial\phi(\bar{x}) = \mathrm{cl}\left\{\bigcup_{\mu\in\partial\rho(\bar{Z})} \int_\Omega \partial f_\omega(\bar{x})\, d\mu(\omega)\right\}. \tag{4.32}$$

Note that since $\partial\rho(\bar{Z})$ is convex, it is straightforward to verify that the set inside the parentheses at the right hand side of (4.32) is convex.

Let us consider now some examples discussed in Section 4.4.

Example 8 (Mean-upper-semideviation risk function of order p). Consider the setting of Example 3. We have that the risk function ρ, defined in (4.17), is a convex real valued continuous function. It follows that for any $Z \in \mathcal{Z}$ the subdifferential $\partial\rho(Z)$ is non-empty and formula (4.29) holds with the set $\mathcal{A}$ given in (4.18). That is,

$$\partial\rho(Z) = \left\{1 + \zeta - \mathbb{E}[\zeta] : \zeta \in \Delta_Z\right\},$$

where

$$\Delta_Z \doteq \mathrm{argmax}_{\zeta\in\mathcal{Z}^*} \left\{\langle\zeta, Y\rangle : \|\zeta\|_q \le c,\ \zeta \succeq 0\right\} \text{ and } Y \doteq Z - \mathbb{E}[Z]. \tag{4.33}$$

If $p \in (1, +\infty)$, then the set Δ_Z can be described as follows. If the function $Z(\cdot)$ is constant, then $Y(\cdot) \equiv 0$ and hence $\Delta_Z = \{\zeta : \|\zeta\|_q \le c,\ \zeta \succeq 0\}$. Suppose that $Z(\cdot)$ is not constant[18] and hence $Y(\cdot)$ is not identically zero. Note that the 'argmax' in (4.33) is not changed if Y is replaced by $Y_+(\cdot) \doteq [Y(\cdot)]_+$. With Y_+ is associated a unique point $\zeta^* \in \mathcal{Z}^*$ such that $\|\zeta^*\|_q = 1$ and $\langle\zeta^*, Y\rangle = \|Y\|_p$. Since $Y_+ \succeq 0$, it follows that $\zeta^* \succeq 0$ and $\Delta_Z = \{c\zeta^*\}$. That is, for $p > 1$ and nonconstant $Z \in \mathcal{Z}$, the subdifferential $\partial\rho(Z)$ is a singleton, and hence ρ is differentiable at Z.

Suppose now that $p = 1$ and hence $q = +\infty$. In that case

$$\Delta_Z = \left\{\zeta \in \mathcal{Z}^* : \begin{array}{l} \zeta(\omega) = c \text{ if } Y(\omega) > 0,\ \zeta(\omega) = 0 \text{ if } Y(\omega) < 0, \\ 0 \le \zeta(\omega) \le c \text{ if } Y(\omega) = 0 \end{array}\right\}.$$

It follows that Δ_Z is a singleton, and hence ρ is differentiable at Z, if and only if $Y(\omega) \neq 0$ for P-almost every $\omega \in \Omega$.

[17] By $\mathrm{cl}(S)$ we denote the topological closure of the set $S \subset \mathbb{R}^n$.

[18] Of course, this and similar statements here should be understood up to a set of P-measure zero.

Example 9 (Mean-upper-semideviation of order p from a target). Consider the setting of Example 5. The risk function ρ, defined in (4.19), is real valued convex and continuous. We have that

$$\partial\rho(Z) = \operatorname{argmax}_{\zeta\in\mathcal{Z}^*}\big\{\langle\zeta, Z-\tau\rangle : \|\zeta\|_q \le c,\ \zeta\succeq 0\big\}.$$

Similarly to the previous example, we have here that if $p > 1$, then ρ is differentiable at Z if and only if $P\{Z(\omega)\ne\tau\} > 0$. If $p = 1$, then ρ is differentiable at Z if and only if $P\{Z(\omega)\ne\tau\} = 1$.

Example 10. Consider the setting of Example 6 with the risk function ρ defined in (4.20). Because of (4.22) and by (4.27) we have

$$\partial\rho(Z) = \operatorname{argmax}_{\zeta\in\mathcal{Z}^*,\,\mathbb{E}[\zeta]=1}\mathbb{E}\big[\zeta Z - v^*(\zeta-1)\big]. \tag{4.34}$$

Also the subdifferential of function $h(\zeta) \doteq \mathbb{E}\big[\zeta Z - v^*(\zeta-1)\big]$ is given by

$$\partial h(\zeta) = \big\{Z'\in\mathcal{Z} : Z'(\omega)\in Z(\omega) - \partial v^*(\zeta(\omega)-1),\ \omega\in\Omega\big\}.$$

By the first order optimality conditions we have then that $\bar\zeta\in\mathcal{Z}^*$ is an optimal solution of the right hand side problem of (4.34) if and only if there exists $\bar\lambda\in\mathbb{R}$ such that

$$Z(\omega)-\bar\lambda\in\partial v^*(\bar\zeta(\omega)-1),\ \textit{a.e.}\ \omega\in\Omega,\ \text{ and }\ \mathbb{E}[\bar\zeta]=1.$$

Since the inclusion $a\in\partial v^*(z)$ is equivalent to $z\in\partial v(a)$ we obtain

$$\partial\rho(Z) = \big\{\zeta\in\mathcal{Z}^* : \zeta(\omega)\in 1+\partial v(Z(\omega)-\bar\lambda),\ \textit{a.e.}\ \omega\in\Omega,\ \mathbb{E}[\zeta]=1\big\}. \tag{4.35}$$

Note that $\bar\lambda$ is an optimal solution of the dual problem

$$\min_{\lambda\in\mathbb{R}}\sup_{\zeta\in\mathcal{Z}^*}\mathbb{E}\big[\zeta Z - v^*(\zeta-1)-\lambda(\zeta-1)\big].$$

By interchanging the integral and max operators (see (4.12)), the above problem can be written in the following equivalent form:

$$\min\nolimits_{\lambda\in\mathbb{R}}\mathbb{E}\left[\sup\nolimits_{z\in\mathbb{R}}\big\{(Z-\lambda)z - v^*(z-1)+\lambda\big\}\right].$$

Example 11 (Conditional value at risk). Consider the setting of Example 7 with ρ defined in (4.23). We can use results of the previous example with function $v(z)$ defined in (4.24). We have here that $\bar\lambda$ is an optimal solution of the problem

$$\min\nolimits_{\lambda\in\mathbb{R}}\mathbb{E}\big[-\varepsilon_1[\lambda-Z]_+ + \varepsilon_2[Z-\lambda]_+\big]. \tag{4.36}$$

For $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ an optimal solution $\bar\lambda$ of (4.36) is given by a κ-quantile of Z (recall that $\kappa = \varepsilon_2/(\varepsilon_1+\varepsilon_2)$). That is, $\bar\lambda\in[a,b]$ where

$$a \doteq \inf\{t : P(Z\le t)\ge\kappa\}\ \text{ and }\ b\doteq\sup\{t : P(Z\le t)\ge\kappa\}.$$

By (4.35) we have

$$\partial\rho(Z) = \left\{ \zeta \in \mathcal{Z}^* : \begin{array}{ll} \zeta(\omega) = 1 - \varepsilon_1, & \text{if } Z(\omega) < \bar{\lambda} \\ \zeta(\omega) = 1 + \varepsilon_2, & \text{if } Z(\omega) > \bar{\lambda} \\ \zeta(\omega) \in [1 - \varepsilon_1, 1 + \varepsilon_2], & \text{if } Z(\omega) = \bar{\lambda} \\ \mathbb{E}[\zeta] = 1 & \end{array} \right\}. \tag{4.37}$$

Note that elements (functions) $\zeta \in \partial\rho(Z)$ are defined up to sets of P-measure zero and the above formula (4.37) holds for any κ-quantile $\bar{\lambda} \in [a, b]$. Also recall that for $\varepsilon_1 = 1$ the risk function $\rho(\cdot)$ coincides with $CV@R_\kappa[\,\cdot\,]$.

4.7 Optimization of Risk Functions

In this section we consider the optimization problem

$$\min_{x \in S} \left\{ \phi(x) \doteq \rho(F(x)) \right\}. \tag{4.38}$$

Recall that with the mapping $F : \mathbb{R}^n \to \mathcal{Z}$ is associated the function $f(x, \omega) = [F(x)](\omega)$. We assume throughout this section, and the following Sections 4.8 and 4.9, that

(i) S is a non-empty closed convex subset of $\mathbb{R}^n$,
(ii) the mapping $F : \mathbb{R}^n \to \mathcal{Z}$ is convex,
(iii) the risk function $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ is proper, lower semicontinuous and satisfies conditions (A1) and (A2).

It follows that the composite function $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex, and hence optimization problem (4.38) is a convex problem. Because of the Fenchel-Moreau theorem, we can employ representation (4.7) of the risk function ρ to write problem (4.38) in the following min-max form:

$$\min_{x \in S} \sup_{\mu \in \mathcal{A}} \Phi(x, \mu), \tag{4.39}$$

where $\mathcal{A} \doteq \mathrm{dom}(\rho^*)$ and

$$\Phi(x, \mu) \doteq \langle \mu, F(x) \rangle - \rho^*(\mu). \tag{4.40}$$

Note that because of the assumed monotonicity condition (A2), the set $\mathcal{A}$ contains only non-negative measures, *i.e.*, $\mathcal{A} \subset \mathcal{Z}^*_+$. If, moreover, assumption (A3) holds, then $\mathcal{A}$ is a subset of the set $\mathcal{P} \subset \mathcal{Z}^*$ of probability measures, and for $\mu \in \mathcal{P}$,

$$\langle \mu, F(x) \rangle = \mathbb{E}_\mu[F(x)] = \textstyle\int_\Omega f(x, \omega)\, d\mu(\omega).$$

If assumption (A4) also holds, then $\rho^*(\mu) = 0$ and hence $\Phi(x, \mu) = \mathbb{E}_\mu[F(x)]$ for any $\mu \in \mathcal{A}$. Therefore if ρ is a coherent risk function, then problem (4.38) can be written in the min-max form

$$\min_{x\in S}\sup_{\mu\in\mathcal{A}}\mathbb{E}_\mu[F(x)].$$

We have here that the function $\Phi(x,\mu)$ is concave in μ and, since F is convex, is convex in x. Therefore, under various regularity conditions, the 'min' and 'max' operators in (4.39) can be interchanged to obtain the problem

$$\max_{\mu\in\mathcal{A}}\inf_{x\in S}\Phi(x,\mu). \tag{4.41}$$

For example, the following holds (*cf.* [314]).

Proposition 2. *Suppose that $\mathcal{Z}$ is a Banach space, the mapping F is convex, the function ρ is proper, lower semicontinuous and satisfies assumptions* (A1)–(A3)*. Then the optimal values of problems* (4.39) *and* (4.41) *are equal to each other, and if their common optimal value is finite, then problem* (4.41) *has an optimal solution $\bar\mu$. Moreover, the optimal values of* (4.39) *and* (4.41) *are equal to the optimal value of the problem*

$$\min_{x\in S}\Phi(x,\bar\mu), \tag{4.42}$$

and if $\bar x$ is an optimal solution of (4.39)*, then $\bar x$ is also an optimal solution of* (4.42)*.*

We obtain that, under assumptions specified in the above proposition, there exists a probability measure $\bar\mu\in\mathcal{P}$ such that problem (4.38) is 'almost' equivalent to problem (4.42). That is, optimal values of problems (4.38) and (4.42) are equal to each other and the set of optimal solutions of problem (4.38) is contained in the set of optimal solutions of problem (4.42). Of course, the corresponding probability measure $\bar\mu$ is not known apriori and could be obtained by solving the dual problem (4.41).

We also have that if the optimal values of problems (4.39) and (4.41) are equal to each other, then $\bar x$ is an optimal solution of (4.39) and $\bar\mu$ is an optimal solution of (4.41) if and only if $(\bar x,\bar\mu)$ is a *saddle point* of $\Phi(x,\mu)$, *i.e.*,

$$\bar x\in\operatorname{argmin}_{x\in S}\Phi(x,\bar\mu)\ \text{ and }\ \bar\mu\in\operatorname{argmax}_{\mu\in\mathcal{A}}\Phi(\bar x,\mu).$$

Conversely, if $\Phi(x,\mu)$ possesses a saddle point, then the optimal values of problems (4.39) and (4.41) are equal. Because of convexity and lower semicontinuity of ρ we have that $\rho^{**}(\cdot)=\rho(\cdot)$, and by (4.40) we obtain that

$$\operatorname{argmax}_{\mu\in\mathcal{A}}\Phi(\bar x,\mu)=\partial\rho(\bar Z),$$

where $\bar Z\doteq F(\bar x)$. Moreover, if $\psi(\cdot)\doteq\mathbb{E}_{\bar\mu}[F(\cdot)]$ is finite valued in a neighborhood of $\bar x$, then the first order optimality condition for $\bar x$ to be a minimizer of $\psi(x)$ over $x\in S$ is that[19] $0\in N_S(\bar x)+\partial\psi(\bar x)$. Together with Strassen's disintegration theorem (see (4.31)) this leads to the following optimality conditions.

[19] By $N_S(\bar x)\doteq\{y\in\mathbb{R}^n:(x-\bar x)^Ty\le 0,\ \forall\, x\in S\}$ we denote the normal cone to S at $\bar x\in S$. By the definition $N_S(\bar x)=\emptyset$ if $\bar x\notin S$.

Proposition 3. *Suppose that $\mathcal{Z}$ is a Banach space, the risk function ρ satisfies conditions* (A1)–(A3)*, the set S and the mapping F are convex, and $\bar{x} \in X$ and $\bar{\mu} \in \mathcal{P}$ are such that $\mathbb{E}_{\bar{\mu}}[F(\cdot)]$ is finite valued in a neighborhood of $\bar{x}$. Denote $\bar{Z} \doteq F(\bar{x})$. Then $(\bar{x}, \bar{\mu})$ is a saddle point of $\Phi(x, \mu)$ if and only if*

$$0 \in N_S(\bar{x}) + \mathbb{E}_{\bar{\mu}}[\partial f_\omega(\bar{x})] \ \text{ and } \ \bar{\mu} \in \partial \rho(\bar{Z}). \tag{4.43}$$

Under the assumptions of Proposition 3, conditions (4.43) can be viewed as optimality conditions for a point $\bar{x} \in S$ to be an optimal solution of problem (4.38). That is, if there exists a probability measure $\bar{\mu} \in \partial\rho(\bar{Z})$ such that the first condition of (4.43) holds, then $\bar{x}$ is an optimal solution of problem (4.38), *i.e.*, (4.43) are sufficient conditions for optimality. Moreover, under the assumptions of Proposition 2, the existence of such a probability measure $\bar{\mu}$ is a necessary condition for optimality of $\bar{x}$.

4.8 Nonanticipativity Constraints

The optimization problem (4.38) can be written in the following equivalent form:

$$\min_{X \in \mathcal{M}_S,\, x \in \mathbb{R}^n} \rho(F_X) \ \text{ subject to } X(\omega) = x, \, \forall\, \omega \in \Omega, \tag{4.44}$$

where $\mathcal{M} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P, \mathbb{R}^n)$ and $[F_X](\omega) \doteq f(X(\omega), \omega)$, for $X \in \mathcal{M}$, and

$$\mathcal{M}_S \doteq \{X \in \mathcal{M} : X(\omega) \in S, \ a.e. \ \omega \in \Omega\}.$$

Although the above problem involves optimization over the functional space $\mathcal{M}$, the constraints $X(\omega) = x$, $\omega \in \Omega$, ensure that this problem is equivalent to problem (4.38). These constraints are called the *nonanticipativity* constraints.

Ignoring the nonanticipativity constraints we can write the following relaxation of problem (4.44):

$$\min_{X \in \mathcal{M}_S} \rho(F_X). \tag{4.45}$$

Let us note now that the interchangeability principle, similar to (4.12), holds for risk functions as well.

Proposition 4. *Let $\mathcal{Z} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P)$, $\rho : \mathcal{Z} \to \mathbb{R}$ be a real valued risk function satisfying conditions* (A1) *and* (A2)*, $f : \mathbb{R}^n \times \Omega \to \mathbb{R}$ be a random lower semicontinuous function and $\mathcal{G} : \Omega \rightrightarrows \mathbb{R}^n$ be a closed valued measurable multifunction*[20]*. Let $F^{\mathcal{G}}(\omega) \doteq \inf_{x \in \mathcal{G}(\omega)} f(x, \omega)$ and suppose that $F^{\mathcal{G}} \in \mathcal{Z}$. Then*

$$\rho(F^{\mathcal{G}}) = \inf_{X \in \mathcal{M}} \left\{ \rho\big(F_X\big) : X(\omega) \in \mathcal{G}(\omega) \ a.e. \ \omega \in \Omega \right\}. \tag{4.46}$$

[20] A multifunction $\mathcal{G} : \Omega \rightrightarrows \mathbb{R}^n$ maps a point $\omega \in \Omega$ into a set $\mathcal{G}(\omega) \subset \mathbb{R}^n$. It is said that $\mathcal{G}$ is *closed valued* if $\mathcal{G}(\omega)$ is a closed subset of $\mathbb{R}^n$ for any $\omega \in \Omega$. It is said that $\mathcal{G}$ is *measurable* if for any closed set $A \subset \mathbb{R}^n$ the inverse image set $\mathcal{G}^{-1}(A) \doteq \{\omega \in \Omega : \mathcal{G}(\omega) \in \mathcal{A}\}$ is $\mathcal{F}$-measurable.

The above interchangeability formula can be either derived from (4.12) by using the dual representation (4.7) or proved directly. Indeed, for any $X \in \mathcal{M}$ such that $X(\cdot) \in \mathcal{G}(\cdot)$ we have that $F^{\mathcal{G}}(\cdot) \le f(X(\cdot),\cdot)$, and hence it follows by assumption (A2) that $\rho(F^{\mathcal{G}}) \le \rho(F_X)$. This implies that $\rho(F^{\mathcal{G}})$ is less than or equal to the right hand side of (4.46). Conversely, suppose for the moment that the minimum of $f(x,\omega)$ over $x \in \mathcal{G}(\omega)$ is attained for a.e. $\omega \in \Omega$, and let $\bar{X}(\cdot) \in \arg\min_{x\in\mathcal{G}(\cdot)} f(x,\cdot)$ be a measurable selection such that $\bar{X} \in \mathcal{M}$. Then $\rho(F^{\mathcal{G}}) = \rho(F_{\bar{X}})$, and hence $\rho(F^{\mathcal{G}})$ is greater than or equal to the right hand side of (4.46). It also follows then that

$$\bar{X} \in \operatorname{argmin}_{X\in\mathcal{M}} \left\{\rho\big(F_X\big) : X(\omega) \in \mathcal{G}(\omega) \ a.e.\ \omega \in \Omega\right\}.$$

Such arguments can be also pushed through without assuming existence of optimal solutions by considering ε-optimal solutions with arbitrary $\varepsilon > 0$. Let us emphasize that the monotonicity assumption (A2) is the key condition for (4.46) to hold.

By employing (4.46) with $\mathcal{G}(\omega) \equiv S$ and denoting $F^S(\omega) \doteq \inf_{x\in S} f(x,\omega)$, we obtain that the optimal value of problem (4.45) is equal to $\rho(F^S)$, provided that $F^S \in \mathcal{Z}$. The difference between the optimal values of problems (4.38) and (4.45), that is

$$\mathrm{RVPI}_\rho \doteq \inf_{x\in S} \rho[F(x)] - \rho(F^S),$$

is called the *Risk Value of Perfect Information.* Since problem (4.45) is a relaxation of problem (4.38), we have that RVPI_ρ is non-negative. It is also possible to show that if ρ is real valued and satisfies conditions (A1)–(A4), and hence representation (4.9) holds, then

$$\inf_{\mu\in\mathcal{A}} \mathrm{EVPI}_\mu \le \mathrm{RVPI}_\rho \le \sup_{\mu\in\mathcal{A}} \mathrm{EVPI}_\mu,$$

where

$$\mathrm{EVPI}_\mu \doteq \inf_{x\in S} \mathbb{E}_\mu\big[f(x,\omega)\big] - \mathbb{E}_\mu\big[\inf_{x\in S} f(x,\omega)\big]$$

is the Expected Value of Perfect Information associated with the probability measure μ (*cf.* [314]).

4.9 Dualization of Nonanticipativity Constraints

In addition to the assumptions (i)–(iii) of Section 4.7, we assume in this section that $\mathcal{Z} \doteq \mathcal{L}_p(\Omega,\mathcal{F},P)$ and $\mathcal{Z}^* \doteq \mathcal{L}_q(\Omega,\mathcal{F},P)$, and that $\mathcal{M}^* \doteq \mathcal{L}_q(\Omega,\mathcal{F},P,\mathbb{R}^n)$ is the dual of the space $\mathcal{M} \doteq \mathcal{L}_p(\Omega,\mathcal{F},P,\mathbb{R}^n)$. Consider the Lagrangian

$$L_0(X,x,\lambda) \doteq \rho(F_X) + \mathbb{E}[\lambda^T(X-x)], \quad (X,x,\lambda) \in \mathcal{M}\times\mathbb{R}^n\times\mathcal{M}^*,$$

associated with the nonanticipativity constraints of problem (4.44). Note that problem (4.44) can be written in the following equivalent form:

$$\min_{X\in\mathcal{M}_S,\, x\in\mathbb{R}^n} \left\{ \sup_{\lambda\in\mathcal{M}^*} L_0(X,x,\lambda) \right\}. \tag{4.47}$$

By interchanging the 'min' and 'max' operators in (4.47) we obtain the (Lagrangian) dual of problem (4.44). Observe that $\inf_{x\in\mathbb{R}^n} L_0(X,x,\lambda)$ is equal to $-\infty$ if $\mathbb{E}[\lambda]\neq 0$, and to $L(X,\lambda)$ if $\mathbb{E}[\lambda]=0$, where

$$L(X,\lambda) \doteq \rho(F_X) + \mathbb{E}[\lambda^T X].$$

Therefore the (Lagrangian) dual of problem (4.44) takes on the form

$$\max_{\lambda\in\mathcal{M}^*} \left\{ \inf_{X\in\mathcal{M}_S} L(X,\lambda) \right\} \text{ subject to } \mathbb{E}[\lambda]=0. \tag{4.48}$$

By the standard theory of Lagrangian duality we have that the optimal value of the primal problem (4.44) is greater than or equal to the optimal value of the dual problem (4.48). Moreover, under appropriate regularity conditions, there is no duality gap between problems (4.44) and (4.48), *i.e.*, their optimal values are equal to each other. In particular, if the Lagrangian $L_0(X,x,\lambda)$ possesses a saddle point $((\bar{X},\bar{x}),\bar{\lambda})$, then $(\bar{X},\bar{x})$ and $\bar{\lambda}$ are optimal solutions of problems (4.44) and (4.48), respectively, and there is no duality gap between problems (4.44) and (4.48). Noting that $L_0(X,x,\lambda)$ is linear in x and λ, we obtain that $((\bar{X},\bar{x}),\bar{\lambda})$ is a saddle point if and only if the following conditions hold:

$$\begin{gathered} \bar{X}(\omega)=\bar{x},\ \textit{a.e.}\ \omega\in\Omega,\ \text{and}\ \mathbb{E}[\bar{\lambda}]=0, \\ \bar{X}\in \operatorname{argmin}_{X\in\mathcal{M}_S} L(X,\bar{\lambda}). \end{gathered} \tag{4.49}$$

Consider the function $\Phi(X)\doteq\rho(F_X):\mathcal{M}\to\overline{\mathbb{R}}$. Because of convexity of F and assumptions (A1) and (A2), this function is convex. Its subdifferential $\partial\Phi(X)\subset\mathcal{M}^*$ is defined in the usual way. By convexity, assuming that ρ is continuous at $\bar{Z}\doteq F(\bar{x})$, we can write the following optimality conditions for (4.49) to hold:

$$-\bar{\lambda}\in N_S(\bar{x})+\partial\Phi(\bar{X}). \tag{4.50}$$

Therefore we obtain that if problem (4.38) possesses an optimal solution $\bar{x}$, then the Lagrangian $L_0(X,x,\lambda)$ has a saddle point if and only if there exists $\bar{\lambda}\in\mathcal{M}^*$ satisfying condition (4.50) and such that $\mathbb{E}[\bar{\lambda}]=0$.

The following result holds (*cf.* [314]).

Proposition 5. *Suppose that ρ satisfies conditions* (A1)–(A3) *and mapping F is convex. Furthermore, suppose that problem* (4.38) *possesses an optimal solution $\bar{x}$ and ρ is subdifferentiable at $F(\bar{x})$. Then there exists $\bar{\lambda}$ such that $((\bar{X},\bar{x}),\bar{\lambda})$, where $\bar{X}(\omega)\equiv\bar{x}$, is a saddle point of the Lagrangian $L_0(X,x,\lambda)$, and hence there is no duality gap between problems* (4.38) *and* (4.48)*, and $(\bar{X},\bar{x})$ and $\bar{\lambda}$ are optimal solutions of problems* (4.44) *and* (4.48)*, respectively.*

Let us return to the question of decomposing problem (4.49). Suppose that ρ is real valued and conditions (A1)–(A3) are satisfied, and hence by Theorem 2 representation (4.8) holds. Then

$$\inf_{X\in\mathcal{M}_S} L(X,\lambda) = \inf_{X\in\mathcal{M}_S} \sup_{\zeta\in\mathcal{P}} \big\{\mathbb{E}[\zeta F_X + \lambda^T X] - \rho^*(\zeta)\big\}. \tag{4.51}$$

Suppose, further, that the 'inf' and 'sup' operators at the right hand side of equation (4.51) can be interchanged (note that the function inside the parentheses in the right hand side of (4.51) is convex in X and concave in ζ). Then

$$\begin{aligned}\inf_{X\in\mathcal{M}_S} L(X,\lambda) &= \sup_{\zeta\in\mathcal{P}} \inf_{X\in\mathcal{M}_S} \big\{\mathbb{E}[\zeta F_X + \lambda^T X] - \rho^*(\zeta)\big\} \\ &= \sup_{\zeta\in\mathcal{P}} \left\{\mathbb{E}\Big(\inf_{x\in S}[\zeta(\omega)f(x,\omega) + \lambda(\omega)^T x]\Big) - \rho^*(\zeta)\right\},\end{aligned}$$

where the last equality follows by the interchangeability principle. Therefore, we obtain that, under the specified assumptions, the optimal value of the dual problem (4.48) is equal to $\sup_{\mathbb{E}[\lambda]=0,\zeta\in\mathcal{P}} D(\lambda,\zeta)$, where

$$D(\lambda,\zeta) \doteq \mathbb{E}\left\{\inf_{x\in S}[\zeta(\omega)f(x,\omega) + \lambda(\omega)^T x]\right\} - \rho^*(\zeta). \tag{4.52}$$

If, moreover, there is no duality gap between problems (4.38) and (4.48), then the following duality relation holds:

$$\inf_{x\in S} \rho[F(x)] = \sup_{\substack{\lambda\in\mathcal{M}^*,\, \zeta\in\mathcal{P} \\ \mathbb{E}[\lambda]=0}} D(\lambda,\zeta).$$

Note the separable structure of the right hand side of (4.52). That is, in order to calculate $D(\lambda,\zeta)$ one needs to solve the minimization problem inside the parentheses at the right hand side of (4.52) separately for every $\omega\in\Omega$, and then to take the expectation of the optimal values calculated.

4.10 Two-Stage Programming

Suppose now that the function $f(x,\omega)$ is given in the form

$$f(x,\omega) \doteq \inf_{y\in\mathcal{G}(x,\omega)} g(x,y,\omega),$$

where $g : \mathbb{R}^n \times \mathbb{R}^m \times \Omega \to \mathbb{R}$ is a random lower semicontinuous function and $\mathcal{G} : \mathbb{R}^n \times \Omega \rightrightarrows \mathbb{R}^m$ is a closed valued measurable multifunction. Note that it follows that the optimal value function $f(x,\omega)$ is measurable, and moreover is random lower semicontinuous provided that $\mathcal{G}(\cdot,\omega)$ are locally uniformly bounded. We refer to the corresponding problem (4.38) as a *two-stage program.* For example, if the set S is polyhedral,

$$g(x, y, \omega) \doteq c^T x + q(\omega)^T y, \tag{4.53}$$

$$\mathcal{G}(x, \omega) \doteq \big\{ y : T(\omega)x + W(\omega)y = h(\omega),\ y \geq 0 \big\}, \tag{4.54}$$

and $\rho(Z) \equiv \mathbb{E}[Z]$, then problem (4.38) becomes a two-stage linear stochastic programming problem.

It is important to note that it is implicitly assumed here that for every $x \in S$ the optimal value $f(x, \omega)$ is *finite* for all $\omega \in \Omega$. In particular, this requires the second stage problem to be feasible (*i.e.*, $\mathcal{G}(x, \omega) \neq \emptyset$) for every $\omega \in \Omega$. That is, it requires the considered two-stage problem to have a relatively complete recourse.

Suppose that ρ satisfies conditions (A1) and (A2). Then by the interchangeability formula (4.46) we have that, for a fixed $x \in S$,

$$\rho(F(x)) = \inf_{\substack{Y \in \mathcal{M} \\ Y(\cdot) \in \mathcal{G}(x, \cdot)}} \rho\left[\Gamma_Y(x)\right],$$

where $[\Gamma_Y(x)](\omega) \doteq g(x, Y(\omega), \omega)$ and $\mathcal{M} \doteq \mathcal{L}_p(\Omega, \mathcal{F}, P, \mathbb{R}^m)$. Consequently the first stage problem (4.38) is equivalent to the problem

$$\min_{x \in S,\, Y \in \mathcal{M}} \rho\left[\Gamma_Y(x)\right] \ \text{ s.t. } \ Y(\omega) \in \mathcal{G}(x, \omega) \ a.e.\ \omega \in \Omega. \tag{4.55}$$

Note again that the key property ensuring equivalence of problems (4.38) and (4.55) is the monotonicity condition (A2).

If the set $\Omega = \{\omega_1, \dots, \omega_K\}$ is finite, we can identify space $\mathcal{L}_p(\Omega, \mathcal{F}, P, \mathbb{R}^m)$ with the finite dimensional space $\mathbb{R}^{mK}$ of vectors $Y = (y_1, \dots, y_K)$. In that case $\Gamma_Y(x) = (g(x, y_1, \omega_1), \dots, g(x, y_K, \omega_K)) \in \mathbb{R}^K$ and ρ is a function from $\mathbb{R}^K$ to $\mathbb{R}$. Then problem (4.55) can be written in the form

$$\min_{x \in \mathbb{R}^n,\, Y \in \mathbb{R}^{mK}} \rho\left[\Gamma_Y(x)\right] \ \text{ s.t. } \ x \in S,\ y_k \in \mathcal{G}(x, \omega_k),\ k = 1, \dots, K. \tag{4.56}$$

In particular, if the function g and mapping $\mathcal{G}$ are given in the form (4.53) and (4.54), respectively, then problem (4.56) takes the form

$$\begin{aligned} &\min_{x \in S,\, Y \in \mathbb{R}^{mK}} \rho\big(c^T x + q_1^T y_1, \dots, c^T x + q_K^T y_K\big) \\ &\text{subject to } \ T_k x + W_k y_k = h_k,\ y_k \geq 0,\ \ k = 1, \dots, K, \end{aligned} \tag{4.57}$$

where $q_k \doteq q(\omega_k)$, $T_k \doteq T(\omega_k)$, $W_k \doteq W(\omega_k)$ and $h_k \doteq h(\omega_k)$. If, further, condition (A3) is satisfied, then

$$\rho\big(c^T x + q_1^T y_1, \dots, c^T x + q_K^T y_K\big) = c^T x + \rho\big(q_1^T y_1, \dots, q_K^T y_K\big).$$

Assume now that condition (A4) also holds true. Then the set $\mathcal{A}$ of probability measures, constituting the domain of the conjugate function ρ^*, can be identified with a certain convex closed subset of the simplex in $\mathbb{R}^K$:

$$\mathcal{A} \subset \left\{ p \in \mathbb{R}^K : \textstyle\sum_{k=1}^K p_k = 1,\ p_k \geq 0,\ k = 1, \dots, K \right\}.$$

In this case we can rewrite problem (4.57) as follows:

$$\begin{gathered}\min_{x \in S,\, Y \in \mathbb{R}^{mK}} \left(c^T x + \max_{p \in \mathcal{A}} \sum_{k=1}^{K} p_k q_k^T y_k \right) \\ \text{subject to } T_k x + W_k y_k = h_k,\ y_k \geq 0, \quad k = 1, \dots, K.\end{gathered}$$

In the following sections of this chapter we shall extend this observation to multistage problems.

4.11 Conditional Risk Mappings

In order to construct dynamic models of risk we need to extend the concept of a risk function. In multi-stage (dynamic) stochastic programming the main theoretical tool is the concept of *conditional expectation.* That is, let $(\Omega, \mathcal{F}_2, P)$ be a probability space, $\mathcal{F}_1$ be a sigma subalgebra of $\mathcal{F}_2$, *i.e.*, $\mathcal{F}_1 \subset \mathcal{F}_2$, and $\mathcal{X}_i$, $i = 1, 2$, be spaces of all $\mathcal{F}_i$-measurable and P-integrable functions $Z : \Omega \to \mathbb{R}$. The conditional expectation $\mathbb{E}[\,\cdot\,|\mathcal{F}_1]$ is defined as a mapping from $\mathcal{X}_2$ into $\mathcal{X}_1$ such that

$$\int_A \mathbb{E}[Z|\mathcal{F}_1](\omega)\, dP(\omega) = \int_A Z(\omega) dP(\omega), \;\; \text{for all}\, A \in \mathcal{F}_1 \text{ and } Z \in \mathcal{X}_2.$$

The approach that we adopt here is aimed at extending this concept to risk mappings. Our presentation is based on [313]. Let $(\Omega, \mathcal{F}_2)$ be a measurable space, $\mathcal{F}_1$ be a sigma subalgebra of $\mathcal{F}_2$, and $\mathcal{Z}_i$, $i = 1, 2$, be linear spaces of $\mathcal{F}_i$-measurable functions $Z : \Omega \to \mathbb{R}$. We assume that $\mathcal{Z}_1 \subset \mathcal{Z}_2$ and each space $\mathcal{Z}_i$ is sufficiently large such that it includes all $\mathcal{F}_i$-measurable step functions, *i.e.*, condition (C) is satisfied. Also we assume that with each $\mathcal{Z}_i$ is paired a dual space $\mathcal{Z}_i^*$ of finite signed measures on $(\Omega, \mathcal{F}_i)$. In applications we typically use spaces $\mathcal{Z}_i \doteq \mathcal{L}_p(\Omega, \mathcal{F}_i, P)$ and $\mathcal{Z}_i^* \doteq \mathcal{L}_q(\Omega, \mathcal{F}_i, P)$ for some (reference) probability measure P. At this moment, however, this is not essential and is not assumed. Let $\rho : \mathcal{Z}_2 \to \mathcal{Z}_1$ be a mapping, referred to as *risk mapping.* Consider the following conditions[21]:

(M1) *Convexity:*

$$\rho(\alpha Z_1 + (1 - \alpha) Z_2) \preceq \alpha \rho(Z_1) + (1 - \alpha) \rho(Z_2)$$

for all $Z_1, Z_2 \in \mathcal{Z}_2$ and all $\alpha \in [0, 1]$.

(M2) *Monotonicity:* If $Z_1, Z_2 \in \mathcal{Z}_2$ and $Z_2 \succeq Z_1$, then $\rho(Z_2) \succeq \rho(Z_1)$.

(M3) *Translation Equivariance:* If $Y \in \mathcal{Z}_1$ and $Z \in \mathcal{Z}_2$, then

$$\rho(Z + Y) = \rho(Z) + Y.$$

[21] Recall that the relation $Z_1 \preceq Z_2$ denotes the inequality $Z_1(\omega) \leq Z_2(\omega)$ for all $\omega \in \Omega$.

(M4) *Positive Homogeneity:* If $\alpha > 0$ and $Z \in \mathcal{Z}_2$, then $\rho(\alpha Z) = \alpha\rho(Z)$.

Axioms (M1)–(M4) generalize the conditions introduced in [294] for dynamic risk measures in the case of a finite space Ω.

If the sigma algebra $\mathcal{F}_1$ is trivial, *i.e.*, $\mathcal{F}_1 = \{\emptyset, \Omega\}$, then any $\mathcal{F}_1$-measurable function is constant over Ω, and hence the space $\mathcal{Z}_1$ can be identified with $\mathbb{R}$. In that case ρ maps $\mathcal{Z}_2$ into the real line $\mathbb{R}$, and conditions (M1)–(M4) coincide with the respective conditions (A1)–(A4). In order to emphasize that the risk mapping ρ is associated with spaces $\mathcal{Z}_1$ and $\mathcal{Z}_2$ we sometimes write it as $\rho_{\mathcal{Z}_2|\mathcal{Z}_1}$. We say that the risk mapping ρ is a *conditional* risk mapping if it satisfies conditions (M1)–(M3).

Remark 1. Note that if $Y \in \mathcal{Z}_1$, then we have by condition (M3) that

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

If, moreover, ρ is positively homogeneous (*i.e.*, condition (M4) holds), then $\rho(0) = 0$. Therefore, if conditions (M1)–(M4) hold, then $\rho(Y) = Y$ for any $Y \in \mathcal{Z}_1$.

For $\omega \in \Omega$, we associate with a risk mapping ρ the function

$$\rho_\omega(Z) \doteq [\rho(Z)](\omega), \quad Z \in \mathcal{Z}_2.$$

Note that since it is assumed that all functions of the space $\mathcal{Z}_1$ are real valued, it follows that ρ_ω maps $\mathcal{Z}_2$ into $\mathbb{R}$, *i.e.*, $\rho_\omega(\cdot)$ is also real valued. Conditions (M1), (M2) and (M4) simply mean that function ρ_ω satisfies the respective conditions (A1), (A2) and (A4) for every $\omega \in \Omega$. Condition (M3) implies (but is not equivalent to) condition (A3) for the functions ρ_ω, $\omega \in \Omega$.

We say that the mapping ρ is lower semicontinuous if for every $\omega \in \Omega$ the corresponding function ρ_ω is lower semicontinuous. With each function $\rho_\omega : \mathcal{Z}_2 \to \mathbb{R}$ is associated its conjugate function $\rho^*_\omega : \mathcal{Z}^*_2 \to \overline{\mathbb{R}}$, defined in (4.5). Note that although ρ_ω is real valued, it can happen that $\rho^*_\omega(\mu) = +\infty$ for some $\mu \in \mathcal{Z}^*_2$.

By $\mathcal{P}_{\mathcal{Z}^*_i}$ we denote the set of all probability measures on $(\Omega, \mathcal{F}_i)$ which are in $\mathcal{Z}^*_i$. Moreover, with each $\omega \in \Omega$ we associate a set of probability measures $\mathcal{P}_{\mathcal{Z}^*_2|\mathcal{F}_1}(\omega) \subset \mathcal{P}_{\mathcal{Z}^*_2}$ formed by all $\nu \in \mathcal{P}_{\mathcal{Z}^*_2}$ such that for every $B \in \mathcal{F}_1$ it holds that

$$\nu(B) = \begin{cases} 1, & \text{if } \omega \in B, \\ 0, & \text{if } \omega \notin B. \end{cases} \tag{4.58}$$

Note that ω is fixed here and B varies in $\mathcal{F}_1$. Condition (4.58) simply means that for every ω and every $B \in \mathcal{F}_1$ we know whether B happened or not. In particular, if $\mathcal{F}_1 = \{\emptyset, \Omega\}$, then $\mathcal{P}_{\mathcal{Z}^*_2|\mathcal{F}_1}(\omega) = \mathcal{P}_{\mathcal{Z}^*_2}$ for all $\omega \in \Omega$.

We can now formulate the basic duality result for conditional risk mappings (*cf.* [313]) which can be viewed as an extension of Theorem 2. Recall that $\langle \mu, Z\rangle = \mathbb{E}_\mu[Z]$ for $\mu \in \mathcal{P}_{\mathcal{Z}^*_i}$ and $Z \in \mathcal{Z}_i$.

Theorem 4. *Let $\rho = \rho_{\mathcal{Z}_2|\mathcal{Z}_1}$ be a lower semicontinuous conditional risk mapping satisfying conditions* (M1)–(M3)*. Then for every $\omega \in \Omega$ it holds that*

$$\rho_\omega(Z) = \sup_{\mu \in \mathcal{P}_{\mathcal{Z}_2^*|\mathcal{F}_1}(\omega)} \left\{ \langle \mu, Z \rangle - \rho^*_\omega(\mu) \right\}, \quad \forall\, Z \in \mathcal{Z}_2. \tag{4.59}$$

Moreover, if ρ is positively homogeneous (i.e., condition (M4) *holds), then for every $\omega \in \Omega$ there exists a closed convex set $\mathcal{A}(\omega) \subset \mathcal{P}_{\mathcal{Z}_2^*|\mathcal{F}_1}(\omega)$ such that*

$$\rho_\omega(Z) = \sup_{\mu \in \mathcal{A}(\omega)} \langle \mu, Z \rangle, \quad \forall\, Z \in \mathcal{Z}_2. \tag{4.60}$$

Conversely, suppose that a mapping $\rho : \mathcal{Z}_2 \to \mathcal{Z}_1$ can be represented in form (4.59) *for some $\rho^* : \mathcal{Z}_2^* \times \Omega \to \overline{\mathbb{R}}$. Then ρ is lower semicontinuous and satisfies conditions* (M1)–(M3)*.*

Remark 2. As it was mentioned in the discussion following Theorem 1, if $\mathcal{Z}_2$ is a Banach lattice (*e.g.*, $\mathcal{Z}_2 \doteq \mathcal{L}_p(\Omega, \mathcal{F}_2, P)$) and ρ satisfies conditions (M1) and (M2), then for any $\omega \in \Omega$ the corresponding function $\rho_\omega : \mathcal{Z}_2 \to \mathbb{R}$ is continuous, and hence is lower semicontinuous. Therefore, in the case of $\mathcal{Z}_2 \doteq \mathcal{L}_p(\Omega, \mathcal{F}_2, P)$, the assumption of lower semicontinuity of ρ in the above theorem holds true automatically.

Remark 3. The concept of conditional risk mappings is closely related to the concept of conditional expectations. Let P be a probability measure on $(\Omega, \mathcal{F}_2)$ and suppose that every $Z \in \mathcal{Z}_2$ is P-integrable. For $Z \in \mathcal{Z}_2$, define

$$\rho(Z) \doteq \mathbb{E}[Z|\mathcal{F}_1].$$

Suppose, further, that the space $\mathcal{Z}_1$ is large enough so that it contains $\mathbb{E}[Z|\mathcal{F}_1]$ for all $Z \in \mathcal{Z}_2$. Then $\rho : \mathcal{Z}_2 \to \mathcal{Z}_1$ is a well defined[22] mapping. The conditional expectation mapping ρ satisfies conditions (M1)–(M3) and is linear, and hence is positively homogeneous. The representation (4.60) holds with $\mathcal{A}(\omega) = \{\mu(\omega)\}$ being a singleton and $\mu_\omega = \mu(\omega)$ being a probability measure on $(\Omega, \mathcal{F}_2)$ for every $\omega \in \Omega$. Moreover, for any $A \in \mathcal{F}_2$ it holds that

$$\mu_\omega(A) = \mathbb{E}[\mathbb{I}_A|\mathcal{F}_1](\omega) = [P(A|\mathcal{F}_1)](\omega).$$

That is, $\mu(\cdot)$ is the conditional probability of P with respect to $\mathcal{F}_1$. Note that $\mathbb{E}[Z|\mathcal{F}_1](\omega) = \mathbb{E}_{\mu_\omega}[Z]$.

The family of conditional risk mappings is closed under the operation of taking maximum. Let $\left\{\rho^\nu = \rho^\nu_{\mathcal{Z}_2|\mathcal{Z}_1}\right\}_{\nu \in \mathcal{I}}$ be a family of conditional risk mappings satisfying assumptions (M1)–(M3). Suppose, further, that for every $Z \in \mathcal{Z}_2$ the function

[22] Note that the function $\mathbb{E}[Z|\mathcal{F}_1](\cdot)$ is defined up to a set of P-measure zero, *i.e.*, two versions of $\mathbb{E}[Z|\mathcal{F}_1](\cdot)$ can be different on a set of P-measure zero.

$$[\rho(Z)](\cdot) \doteq \sup_{\nu \in \mathcal{I}} \left[\rho^{\nu}(Z)\right](\cdot)$$

belongs to the space $\mathcal{Z}_1$, and hence ρ maps $\mathcal{Z}_2$ into $\mathcal{Z}_1$. It is then straightforward to verify that the max-function ρ also satisfies assumptions (M1)–(M3). Moreover, if ρ^{ν}, $\nu \in \mathcal{I}$, are lower semicontinuous and/or positively homogeneous, then ρ is also lower semicontinuous and/or positively homogeneous. In particular, let $\rho^{\nu}(Z) \doteq \mathbb{E}_{\nu}[Z|\mathcal{F}_1]$, $\nu \in \mathcal{I}$, where $\mathcal{I}$ is a family of probability measures on $(\Omega, \mathcal{F}_2)$. Suppose that the corresponding max-function

$$[\rho(Z)](\cdot) \doteq \sup_{\nu \in \mathcal{I}} \mathbb{E}_{\nu}[Z|\mathcal{F}_1](\cdot) \tag{4.61}$$

is well defined, *i.e.*, ρ maps $\mathcal{Z}_2$ into $\mathcal{Z}_1$. Then ρ is a lower semicontinuous positively homogeneous conditional risk mapping. It is possible to show that, under certain regularity conditions, the converse is also true, *i.e.*, a positively homogeneous conditional risk mapping can be represented in form (4.61) (*cf.* [313]).

4.12 Multistage Optimization Problems

In this section we discuss optimization of risk measures in a dynamical setting. We use the following framework. Let $(\Omega, \mathcal{F})$ be a measurable space and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_T$ be a sequence of sigma algebras such that $\mathcal{F}_1 = \{\emptyset, \Omega\}$ and $\mathcal{F}_T = \mathcal{F}$. Let $\mathcal{Z}_1 \subset \mathcal{Z}_2 \subset \cdots \subset \mathcal{Z}_T$ be a corresponding sequence of linear spaces of $\mathcal{F}_t$ measurable functions, $t = 1, \dots, T$, and let $\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}} : \mathcal{Z}_t \to \mathcal{Z}_{t-1}$ be conditional risk mapings satisfying assumptions (M1)–(M3). Also consider a sequence of functions $Z_t \in \mathcal{Z}_t$, $t = 1, \dots, T$. By the definition of spaces $\mathcal{Z}_t$, each function $Z_t : \Omega \to \mathbb{R}$ is $\mathcal{F}_t$-measurable, and since the sigma algebra $\mathcal{F}_1$ is trivial, $Z_1(\omega)$ is constant and the space $\mathcal{Z}_1$ can be identified with $\mathbb{R}$.

Consider the composite mappings[23] $\rho_{\mathcal{Z}_{t-1}|\mathcal{Z}_{t-2}} \circ \rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}$. Let us observe that conditions (M1)–(M4) are preserved by such compositions. That is, if conditions (M1) and (M2) (and also (M3), (M4)) hold for mappings $\rho_{\mathcal{Z}_{t-1}|\mathcal{Z}_{t-2}}$ and $\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}$, then these conditions hold for their composition as well. Therefore the assumption that the mappings $\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}$, $t = 2, \dots, T$, satisfy conditions (M1)–(M3) implies that the risk functions

$$\rho_t \doteq \rho_{\mathcal{Z}_2|\mathcal{Z}_1} \circ \cdots \circ \rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}} : \mathcal{Z}_t \to \mathbb{R}, \;\; t = 2, \dots, T,$$

satisfy conditions (A1)–(A3). Moreover, consider the space $\mathcal{Z} \doteq \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_T$ and $Z \doteq (Z_1, \dots, Z_T) \in \mathcal{Z}$. Define function $\tilde{\rho} : \mathcal{Z} \to \mathbb{R}$ as follows:

$$\begin{aligned}\tilde{\rho}(Z) \doteq Z_1 + \rho_{\mathcal{Z}_2|\mathcal{Z}_1}\Big[Z_2 + \rho_{\mathcal{Z}_3|\mathcal{Z}_2}\Big(Z_3 + \dots \\ \dots + \rho_{\mathcal{Z}_{T-1}|\mathcal{Z}_{T-2}}\big[Z_{T-1} + \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}(Z_T)\big]\Big)\Big].\end{aligned} \tag{4.62}$$

[23] The composite mapping $\rho_{\mathcal{Z}_{t-1}|\mathcal{Z}_{t-2}} \circ \rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}} : \mathcal{Z}_t \to \mathcal{Z}_{t-2}$ maps $Z_t \in \mathcal{Z}_t$ into $\rho_{\mathcal{Z}_{t-1}|\mathcal{Z}_{t-2}}\big[\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}(Z_t)\big]$.

By assumption (M3) we have

$$Z_{T-1} + \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}(Z_T) = \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}(Z_{T-1} + Z_T),$$

and so on for $t = T-1, \dots, 2$. Therefore we obtain that

$$\tilde{\rho}(Z) = \rho_T(Z_1 + \cdots + Z_T). \tag{4.63}$$

Thus, condition (M3) allows us to switch between the cumulative formulation (4.63) and nested formulation (4.62).

Remark 4. As it was mentioned above, we have that if $\rho_{\mathcal{Z}_{t-1}|\mathcal{Z}_{t-2}}$ and $\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}$ are positively homogeneous risk mappings, then the composite mapping $\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-2}} \doteq \rho_{\mathcal{Z}_{t-1}|\mathcal{Z}_{t-2}} \circ \rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}$ is also a positively homogeneous risk mapping. By virtue of Theorem 4, with these mappings are associated closed convex sets $\mathcal{A}_{t-1,t-2}(\omega)$, $\mathcal{A}_{t,t-1}(\omega)$ and $\mathcal{A}_{t,t-2}(\omega)$, depending on $\omega \in \Omega$, such that the corresponding representation (4.60) holds, provided that these mappings are lower semicontinuous. It is possible to show (*cf.* [313]) that $\mathcal{A}_{t,t-2}(\omega)$ is formed by all measures $\mu \in \mathcal{Z}_t^*$ representable in the form

$$\mu(A) = \int_\Omega [\mu_2(\tilde{\omega})](A)\, d\mu_1(\tilde{\omega}), \quad A \in \mathcal{F}_t,$$

where $\mu_1 \in \mathcal{A}_{t-1,t-2}(\omega)$ and $\mu_2(\cdot) \in \mathcal{A}_{t,t-1}(\cdot)$ is a weakly* $\mathcal{F}_t$-measurable selection. Unfortunately, this formula is not very constructive and in general it could be quite difficult to calculate the dual representation of the composite mapping explicitly.

Remark 5. Consider the composite function $\tilde{\rho}(\cdot)$. As we mentioned in Remark 4, it could be difficult to write it explicitly. The situation simplifies considerably if we assume a certain type of 'between stages independence' condition. That is, suppose that $Z \in \mathcal{Z}$ is such that the functions $\big[\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}(Z_t)\big](\omega)$, $t = T, \dots, 2$, are constants, *i.e.*, independent of ω. Then by condition (M3) we have that

$$\tilde{\rho}(Z) = Z_1 + \rho_{\mathcal{Z}_2|\mathcal{Z}_1}(Z_2) + \cdots + \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}(Z_T). \tag{4.64}$$

We discuss this further in Example 12 of the following section.

Now let us formulate a multistage optimization problem involving risk mappings. Suppose that we are given functions $f_t : \mathbb{R}^{n_t} \times \Omega \to \mathbb{R}$ and multifunctions $\mathcal{G}_t : \mathbb{R}^{n_{t-1}} \times \Omega \rightrightarrows \mathbb{R}^{n_t}$, $t = 1, \dots, T$. We assume that the functions $f_t(x_t, \omega)$ are $\mathcal{F}_t$-random lower semicontinuous[24], and the multifunctions $\mathcal{G}_t(x_{t-1}, \cdot)$ are closed valued and $\mathcal{F}_t$-measurable. Note that since the sigma algebra $\mathcal{F}_1$ is trivial, the function $f_1 : \mathbb{R}^{n_1} \to \mathbb{R}$ does not depend on $\omega \in \Omega$, and

[24] Recall that it is said that function $f_t(x_t, \omega)$ is $\mathcal{F}_t$-random lower semicontinuous if its epigraphical mapping is closed valued and $\mathcal{F}_t$-measurable.

by the definition $\mathcal{G}_1(\omega) \equiv G_1$, where G_1 is a fixed closed subset of $\mathbb{R}^{n_1}$. Let $\mathcal{M}_t$, $t = 1, \dots, T$, be linear spaces of $\mathcal{F}_t$-measurable functions $X_t : \Omega \to \mathbb{R}^{n_t}$, and $\mathcal{M} \doteq \mathcal{M}_1 \times \cdots \times \mathcal{M}_T$. With functions f_t we associate mappings $F_t : \mathcal{M}_t \to \mathcal{Z}_t$ defined as follows

$$[F_t(X_t)](\omega) \doteq f_t(X_t(\omega), \omega), \;\; X_t \in \mathcal{M}_t.$$

Since $\mathcal{F}_1$ is trivial, the space $\mathcal{M}_1$ can be identified with $\mathbb{R}^{n_1}$, and hence $F_1(X_1) = f_1(X_1)$.

Consider the problem

$$\begin{array}{ll} \min\limits_{X \in \mathcal{M}} & \rho_T\big(F_1(X_1) + \cdots + F_T(X_T)\big), \\ \text{s.t.} & X_t(\omega) \in \mathcal{G}_t(X_{t-1}(\omega), \omega), \; \omega \in \Omega, \; t = 1, \dots, T. \end{array} \tag{4.65}$$

We refer to (4.65) as a multistage risk optimization problem. By (4.62) and (4.63) we can write the equivalent nested formulation:

$$\begin{aligned} \rho_T\big(F_1(X_1) + \cdots + F_T(X_T)\big) = F_1(X_1) + \rho_{\mathcal{Z}_2|\mathcal{Z}_1}\Big[F_2(X_2) + \\ \cdots + \rho_{\mathcal{Z}_{T-1}|\mathcal{Z}_{T-2}}\big[F_{T-1}(X_{T-1}) + \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}\big(F_T(X_T)\big)\big]\Big]. \end{aligned} \tag{4.66}$$

Since $X \in \mathcal{M}$, it is assumed that $X_t(\omega)$ are $\mathcal{F}_t$-measurable, and hence (4.65) is adapted to the filtration $\mathcal{F}_t$, $t = 1, \dots, T$.

As an example consider the following linear setting. Suppose that[25]

$$\begin{aligned} f_t(x_t, \omega) &\doteq c_t(\omega) \cdot x_t, \\ \mathcal{G}_t(x_{t-1}, \omega) &\doteq \big\{x_t \in \mathbb{R}^{n_t} : B_t(\omega)x_{t-1} + A_t(\omega)x_t = b_t(\omega), \; x_t \geq 0\big\}, \end{aligned} \tag{4.67}$$

where $c_t(\omega), b_t(\omega)$ are vectors and $B_t(\omega), A_t(\omega)$ are matrices of appropriate dimensions. It is assumed that the corresponding vector-valued functions

$$\xi_t(\omega) \doteq (c_t(\omega), B_t(\omega), A_t(\omega), b_t(\omega)), \;\; t = 1, \dots, T,$$

are adapted to the filtration $\mathcal{F}_t$, *i.e.*, $\xi_t(\omega)$ is $\mathcal{F}_t$-measurable, $t = 1, \dots, T$. Then the nested formulation of the corresponding multistage risk optimization problem can be written as follows:

$$\begin{aligned} \min_{x_1 \in G_1} \Big(c_1 \cdot x_1 + \rho_{\mathcal{Z}_2|\mathcal{Z}_1}\Big[\inf\nolimits_{x_2 \in \mathcal{G}_2(x_1, \omega)} \big(c_2(\omega) \cdot x_2 + \cdots \\ + \rho_{\mathcal{Z}_{T-1}|\mathcal{Z}_{T-2}}\big[\inf\nolimits_{x_{T-1} \in \mathcal{G}_2(x_{T-2}, \omega)} c_{T-1}(\omega) \cdot x_{T-1} \\ + \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}[\inf\nolimits_{x_T \in \mathcal{G}_T(x_{T-1}, \omega)} c_T(\omega) \cdot x_T]\big]\big)\Big]\Big). \end{aligned}$$

The precise meaning of the nested formulation of problem (4.65) is explained by dynamic programming equations as follows. Define the (cost-to-go) function

[25] In order to avoid notational confusion we denote here by $a \cdot b$ the standard scalar product of two vectors $a, b \in \mathbb{R}^n$.

$$\mathcal{Q}_T(x_{T-1},\omega) \doteq \left[\rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}\big(V_T(x_{T-1})\big)\right](\omega), \tag{4.68}$$

where

$$[V_T(x_{T-1})](\omega) \doteq \inf_{x_T\in\mathcal{G}_T(x_{T-1},\omega)} f_T(x_T,\omega). \tag{4.69}$$

And so on for $t = T-1,\dots,2$,

$$\mathcal{Q}_t(x_{t-1},\omega) \doteq \left[\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}\big(V_t(x_{t-1})\big)\right](\omega), \tag{4.70}$$

where

$$\big[V_t(x_{t-1})\big](\omega) \doteq \inf_{x_t\in\mathcal{G}_t(x_{t-1},\omega)} \Big\{f_t(x_t,\omega) + \mathcal{Q}_{t+1}(x_t,\omega)\Big\}. \tag{4.71}$$

Of course, equations (4.70) and (4.71) can be combined into one equation:

$$\big[V_t(x_{t-1})\big](\omega) = \inf_{x_t\in\mathcal{G}_t(x_{t-1},\omega)} \Big\{f_t(x_t,\omega) + \left[\rho_{\mathcal{Z}_{t+1}|\mathcal{Z}_t}\big(V_{t+1}(x_t)\big)\right](\omega)\Big\}. \tag{4.72}$$

Finally, at the first stage we solve the problem

$$\inf_{x_1\in G_1} \mathcal{Q}_2(x_1). \tag{4.73}$$

The optimal value and the set of optimal solutions of problem (4.73) provide the optimal value and the first-stage set of optimal solutions of the multistage program (4.65).

It should be mentioned that for the dynamic programming equations (4.72) to be well defined we need to ensure that $V_t(x_{t-1}) \in \mathcal{Z}_t$ for every considered x_{t-1}. Note that since the function $f_T(x_t,\omega)$ is $\mathcal{F}_T$-random lower semicontinuous and $\mathcal{G}_T(x_{T-1},\cdot)$ is closed valued and $\mathcal{F}_T$-measurable, it follows that $[V_T(x_{T-1})](\cdot)$ is $\mathcal{F}_T$-measurable (*e.g.*, [302, Theorem 14.37]). Still one has to verify that $V_T(x_{T-1}) \in \mathcal{Z}_T$ in order for $\mathcal{Q}_T(x_{T-1},\omega)$ to be well defined. It will follow then that $\mathcal{Q}_T(x_{T-1},\cdot)$ is $\mathcal{F}_{t-1}$-measurable. In order to continue the process for $t = T-1$, it should be verified further that the function $\mathcal{Q}_T(x_{T-1},\omega)$ is $\mathcal{F}_{T-1}$-random lower semicontinuous. And so on for $t = T-2,\dots,2$. Finally, for $t = 2$ the function $\mathcal{Q}_2(x_1,\cdot)$ is $\mathcal{F}_1$-measurable, and hence does not depend on ω. Let us emphasize that the key assumption ensuring equivalence of the two formulations of the risk optimization problem is the monotonicity condition (M2) (*cf.* [313]).

Remark 6. In some cases the function $[V_T(\cdot)](\omega)$, where V_T is defined in (4.69), is convex for all $\omega \in \Omega$. This happens, for example, if $f_T(\cdot,\omega)$ is convex for all $\omega \in \Omega$, and $\mathcal{G}_T$ is defined by linear constraints of form (4.67). If $[V_T(\cdot)](\omega)$ is convex, then conditions (M1) and (M2) ensure that the corresponding function $\mathcal{Q}_T(\cdot,\omega)$ is also convex. Similarly, the convexity property propagates to the functions $\mathcal{Q}_t(\cdot,\omega)$, $t = T-1,\dots,2$. In particular, in the linear case, where f_t and $\mathcal{G}_t$ are defined in (4.67), the functions $\mathcal{Q}_t(\cdot,\omega)$, $t = T,\dots,2$, are convex for all $\omega \in \Omega$. In convex cases it makes sense to talk about subdifferentials[26] $\partial\mathcal{Q}_t(x_{t-1},\omega)$. In principle, these subdifferentials can be written in a recursive form by using equations (4.70) and (4.71) and the analysis of Section 4.6.

[26] These subdifferentials are taken with respect to x_{t-1} for a fixed value $\omega \in \Omega$.

4.13 Examples of Risk Mappings and Multistage Problems

In this section we adopt the framework of Sections 4.11 and 4.12 with $\mathcal{Z}_t \doteq \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$ and $\mathcal{Z}_t^* \doteq \mathcal{L}_q(\Omega, \mathcal{F}_t, P)$, $t = 1, \dots, T$. As before, unless stated otherwise, all expectations and probability statements are made with respect to the probability measure P. As it was already mentioned in Section 4.11, the conditional expectation

$$\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}(Z_t) \doteq \mathbb{E}[Z_t|\mathcal{F}_{t-1}]$$

is a simple example of a conditional risk mapping. For that choice of conditional risk mappings, we have

$$\big(\rho_{\mathcal{Z}_{t-1}|\mathcal{Z}_{t-2}} \circ \rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}\big)(Z_t) = \mathbb{E}[Z_t|\mathcal{F}_{t-2}],$$

and $\rho_T(\cdot) = \mathbb{E}[\,\cdot\,]$. In that case (4.66) becomes the standard formulation of a multistage stochastic programming problem and (4.68)–(4.73) represent well known dynamic programming equations.

Now let us discuss analogues of some examples of risk functions considered in Section 4.4.

Example 12. Consider the following extension of the mean-upper-semideviation risk function (of order $p \in [1, +\infty)$) discussed in Example 3. For $Z_t \in \mathcal{Z}_t$ and $c_t \geq 0$ define

$$\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}(Z_t) \doteq \mathbb{E}[Z_t|\mathcal{F}_{t-1}] + c_t\, \sigma_p(Z_t|\mathcal{F}_{t-1}),$$

where

$$\sigma_p(Z_t|\mathcal{F}_{t-1}) \doteq \Big(\mathbb{E}\Big[\big[Z_t - \mathbb{E}[Z_t|\mathcal{F}_{t-1}]\big]_+^p \big|\mathcal{F}_{t-1}\Big]\Big)^{1/p}.$$

If the sigma algebra $\mathcal{F}_{t-1}$ is trivial, then $\mathbb{E}[\,\cdot\,|\mathcal{F}_{t-1}] = \mathbb{E}[\,\cdot\,]$ and $\sigma_p(Z_t|\mathcal{F}_{t-1})$ becomes the upper semideviation of Z_t of order p. For a while we keep t fixed and we use the notation ρ for the above mapping $\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}$.

By using the analysis of Example 3 it is possible to show that ρ satisfies conditions (M1),(M3) and (M4), and also condition (M2), provided that $c_t \in [0, 1]$. Indeed, clearly ρ is positively homogeneous, *i.e.*, condition (M4) holds. Condition (M3) can be verified directly. That is, if $Y \in \mathcal{Z}_{t-1}$ and $Z_t \in \mathcal{Z}_t$, then

$$\begin{aligned}\rho(Z_t + Y) &= \mathbb{E}[Z_t + Y|\mathcal{F}_{t-1}] + c_t\Big(\mathbb{E}\Big[\big(Z_t + Y - \mathbb{E}[Z_t + Y|\mathcal{F}_{t-1}]\big)_+^p \mid \mathcal{F}_{t-1}\Big]\Big)^{1/p} \\ &= \mathbb{E}[Z_t|\mathcal{F}_{t-1}] + Y + c_t\Big(\mathbb{E}\Big[\big(Z_t - \mathbb{E}[Z_t|\mathcal{F}_{t-1}]\big)_+^p \mid \mathcal{F}_{t-1}\Big]\Big)^{1/p} \\ &= \rho(Z_t) + Y.\end{aligned}$$

For $\omega \in \Omega$ consider the function $\rho_\omega(\cdot) = [\rho(\cdot)](\omega)$. Consider also the conditional probability of P with respect to $\mathcal{F}_{t-1}$, denoted $\mu(\omega)$ or μ_ω (see Remark 3). Recall that $\mathbb{E}[Z_t|\mathcal{F}_{t-1}](\omega) = \mathbb{E}_{\mu_\omega}[Z_t]$, and hence

$$\rho_\omega(Z_t) = \mathbb{E}_{\mu_\omega}[Z_t] + c_t\Big(\mathbb{E}_{\mu_\omega}\Big[\big(Z_t - \mathbb{E}_{\mu_\omega}[Z_t]\big)_+^p\Big]\Big)^{1/p}.$$

For a fixed ω the function ρ_ω coincides with the risk function analyzed in Example 3 with μ_ω playing the role of the corresponding probability measure. Consequently, ρ_ω is convex, *i.e.*, condition (M1) holds, and condition (M2) follows, provided that $c_t \in [0,1]$.

We have that $\mu_\omega \in \mathcal{P}_{\mathcal{Z}_t^*|\mathcal{F}_{t-1}}(\omega)$ and its conditional probability density $g_\omega = d\mu_\omega/dP$ has the following properties: $g_\omega \in \mathcal{Z}_t^*$, $g_\omega \geq 0$, for any $A \in \mathcal{F}_t$, the function $\omega \mapsto \int_A g_\omega(\tilde{\omega})\,dP(\tilde{\omega})$ is $\mathcal{F}_{t-1}$-measurable and, moreover, for any $B \in \mathcal{F}_{t-1}$ it holds that

$$\int_B \int_A g_\omega(\tilde{\omega})\,dP(\tilde{\omega})\,dP(\omega) = P(A \cap B).$$

By the analysis of Example 3 it follows that the representation

$$\rho_\omega(Z_t) = \sup_{\zeta_t \in \mathcal{A}_t(\omega)} \mathbb{E}[\zeta_t Z_t],$$

holds with

$$\mathcal{A}_t(\omega) = \big\{\zeta_t' \in \mathcal{Z}_t^* : \zeta_t' = g_\omega\big(1 + \zeta_t - \mathbb{E}[g_\omega \zeta_t]\big),\ \|\zeta_t\|_q \leq c_t,\ \zeta_t \succeq 0\big\}.$$

In order to write the corresponding multistage problem in form (4.65) we need to describe the composite function $\tilde{\rho}$ defined in (4.62). In general a description of $\tilde{\rho}$ is quite messy. Let us consider the following two particular cases. Suppose that $p = 1$ and all c_t are zero except one, say c_T. That is, $\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}(\cdot) \doteq \mathbb{E}[\,\cdot\,|\mathcal{F}_{t-1}]$, for $t = 2,\dots,T-1$. In that case

$$\begin{aligned}\rho_{\mathcal{Z}_{T-1}|\mathcal{Z}_{T-2}}\big[Z_{T-1} + \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}(Z_T)\big] &= \mathbb{E}\big[Z_{T-1} + \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}(Z_T)\big|\mathcal{F}_{T-2}\big]\\ &= \mathbb{E}\big[Z_{T-1} + Z_T\big|\mathcal{F}_{T-2}\big] + c_T\mathbb{E}\big[\big(Z_T - \mathbb{E}[Z_T|\mathcal{F}_{T-1}]\big)_+\big|\mathcal{F}_{T-2}\big],\end{aligned}$$

and

$$\tilde{\rho}(Z) = \mathbb{E}\Big[Z_1 + \cdots + Z_T + c_T\big[Z_T - \mathbb{E}[Z_T|\mathcal{F}_{T-1}]\big]_+\Big].$$

Another case where calculations are simplified considerably is under the 'between stages independence' condition (compare with Remark 5). That is, suppose that the objective functions f_t and the constraint mappings $\mathcal{G}_t$, $t = 2,\dots,T$, are given in the form $f_t(x_t, \xi_t(\omega))$ and $\mathcal{G}_t(x_{t-1}, \xi_t(\omega))$, respectively, where $\xi_t(\omega)$ are random vectors defined on a probability space $(\Omega, \mathcal{F}, P)$. That is the case, for example, if f_t and $\mathcal{G}_t$ are defined in the form (4.67). With some abuse of notation we simply write $f_t(x_t, \xi_t)$ and $\mathcal{G}_t(x_{t-1}, \xi_t)$ for the corresponding random functions and mappings. It will be clear from the context when ξ_t is viewed as a random vector and when as its particular realization.

Assume that the sigma algebra $\mathcal{F}_t$ is generated by $(\xi_1(\omega),\dots,\xi_t(\omega))$, $t = 1,\dots,T$. Assume also that ξ_1 is not random, and hence the sigma algebra $\mathcal{F}_1$ is trivial. Assume further the following condition, referred to as the *between stages independence* condition:

(I) For every $t \in \{2, \dots, T\}$, random vector ξ_t is (stochastically) independent of $(\xi_1, \dots, \xi_{t-1})$.

Then the minimum in the right hand side of (4.69) is a function of x_{T-1} and ξ_T, and hence is independent of the random vector $(\xi_1, \dots, \xi_{T-1})$. It follows then that the corresponding cost-to-go function $\mathcal{Q}_T(x_{T-1})$, defined in (4.69), is independent of ω. By continuing this process backwards we obtain that, under the between stages independence condition, the cost-to-go functions are independent of ω and the corresponding dynamic programming equations can be written in the form

$$\mathcal{Q}_t(x_{t-1}) = \rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}(V_t(x_{t-1})), \tag{4.74}$$

$$V_t(x_{t-1})(\xi_t) = \inf_{x_t \in \mathcal{G}_t(x_{t-1},\xi_t)} \big\{ f_t(x_t, \xi_t) + \mathcal{Q}_{t+1}(x_t) \big\}, \tag{4.75}$$

with

$$\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}(V_t(x_{t-1})) = \mathbb{E}\big[V_t(x_{t-1})\big] + c_t \left(\mathbb{E}\Big[\big(V_t(x_{t-1}) - \mathbb{E}[V_t(x_{t-1})] \big)_+^p \Big] \right)^{1/p}.$$

Also in that case the optimization in problem (4.65) should be performed over functions $X_t(\xi_t)$ and (compare with (4.64))

$$\begin{array}{l} \rho_T(F_1(X_1) + F_2(X_2) + \cdots + F_T(X_T)) = \\ \qquad F_1(X_1) + \rho_{\mathcal{Z}_2|\mathcal{Z}_1}(F_2(X_2)) + \cdots + \rho_{\mathcal{Z}_T|\mathcal{Z}_{T-1}}(F_T(X_T)). \end{array} \tag{4.76}$$

Example 13. Consider the framework of Example 6. Let $v : \mathbb{R} \to \mathbb{R}$ be a convex real valued function such that the function $z + v(z)$ is monotonically nondecreasing on $\mathbb{R}$. Define

$$\big[\rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}(Z_t)\big](\omega) \doteq \inf_{Y \in \mathcal{Z}_{t-1}} \mathbb{E}\big[Z_t + v(Z_t - Y)|\mathcal{F}_{t-1}\big](\omega). \tag{4.77}$$

Of course, a certain care should be exercised in verification that the right hand side of equation (4.77) gives a well defined mapping. For a while we will keep t fixed and use notation $\rho = \rho_{\mathcal{Z}_t|\mathcal{Z}_{t-1}}$. Since the function $(Z_t, Y) \mapsto \mathbb{E}\big[Z_t + v(Z_t - Y)|\mathcal{F}_{t-1}\big](\omega)$ is convex, it follows that $\rho_\omega(\cdot)$ is convex, *i.e.*, the condition (M1) holds. Since $z + v(z)$ is nondecreasing, condition (M2) holds as well. It is also straightforward to verify that condition (M3) holds here by making change of variables $Z_t \mapsto Z_t - Y$. Let us calculate the conjugate function ρ^*_ω. In a way similar to (4.21) we have for $\zeta_t \in \mathcal{Z}^*_t$,

$$\begin{aligned} \rho^*_\omega(\zeta_t) &= \sup_{Z_t \in \mathcal{Z}_t} \big\{ \mathbb{E}[\zeta_t Z_t] - \rho_\omega(Z_t) \big\} \\ &= \sup_{Z_t \in \mathcal{Z}_t} \big\{ \mathbb{E}[\zeta_t Z_t] + \sup_{Y \in \mathcal{Z}_{t-1}} \mathbb{E}\big[-Z_t - v(Z_t - Y)|\mathcal{F}_{t-1}\big](\omega) \big\} \\ &= \sup_{Z_t \in \mathcal{Z}_t} \big\{ \mathbb{E}[\zeta_t (Z_t + Y)] + \sup_{Y \in \mathcal{Z}_{t-1}} \mathbb{E}\big[-Z_t - Y - v(Z_t)|\mathcal{F}_{t-1}\big](\omega) \big\}, \end{aligned}$$

and hence

$$\begin{aligned} \rho^*_\omega(\zeta_t) = \sup_{Z_t \in \mathcal{Z}_t} & \big\{ \mathbb{E}[\zeta_t Z_t] - \mathbb{E}\big[Z_t + v(Z_t) | \mathcal{F}_{t-1}\big](\omega) \big\} \\ & + \sup_{Y \in \mathcal{Z}_{t-1}} \mathbb{E}\big[Y(\zeta_t - 1) | \mathcal{F}_{t-1}\big](\omega). \end{aligned} \tag{4.78}$$

Since $Y \in \mathcal{Z}_{t-1}$, and hence $Y(\omega)$ is $\mathcal{F}_{t-1}$-measurable, we have

$$\mathbb{E}\big[Y(\zeta_t - 1) | \mathcal{F}_{t-1}\big](\omega) = Y(\omega)\big(\mathbb{E}[\zeta_t | \mathcal{F}_{t-1}](\omega) - 1\big).$$

Therefore, the second maximum in the right hand side of (4.78) is equal to zero if $\mathbb{E}[\zeta_t | \mathcal{F}_{t-1}](\omega) = 1$, and to $+\infty$ otherwise. It follows that the domain of ρ^*_ω is included (this inclusion can be strict) in the set

$$\mathcal{A}^*_t(\omega) \doteq \big\{ \zeta_t \in \mathcal{Z}^*_t : \mathbb{E}[\zeta_t | \mathcal{F}_{t-1}](\omega) = 1 \big\}.$$

Note that for any $B \in \mathcal{F}_{t-1}$ and $\zeta_t \in \mathcal{A}^*_t(\omega)$ it holds that $\int_B \zeta_t \, dP$ is equal to 1 if $\omega \in B$, and to 0 if $\omega \notin B$, *i.e.*, $\mathcal{A}^*_t(\omega)$ is a subset of $\mathcal{P}_{\mathcal{Z}^*_t | \mathcal{F}_{t-1}}(\omega)$.

Consider the conditional probability of P with respect to $\mathcal{F}_{t-1}$, denoted $\mu(\omega)$ or μ_ω (see Remark 3). We have that $\mu_\omega \in \mathcal{P}_{\mathcal{Z}^*_t | \mathcal{F}_{t-1}}(\omega)$ and let $g_\omega = d\mu_\omega / dP$ be its conditional probability density (properties of g_ω were discussed in the previous example). Recall that $\mathbb{E}[Z_t | \mathcal{F}_{t-1}](\omega) = \mathbb{E}_{\mu_\omega}[Z_t]$, and hence

$$\mathbb{E}\big[Z_t + v(Z_t) | \mathcal{F}_{t-1}\big](\omega) = \mathbb{E}[g_\omega(Z_t + v(Z_t))].$$

This can be substituted into (4.78). Since by the interchangeability formula the maximum over Z_t at the right hand side of (4.78) can be taken inside the integral, we obtain

$$\rho^*_\omega(\zeta_t) = \begin{cases} \mathbb{E}\left[\sup_{z_t \in \mathbb{R}} \big\{ (\zeta_t - g_\omega) z_t - g_\omega v(z_t) \big\}\right], & \text{if } \zeta_t \in \mathcal{A}^*_t(\omega), \\ +\infty, & \text{otherwise.} \end{cases} \tag{4.79}$$

By Theorem 4 we have then that

$$\rho_\omega(Z_t) = \sup_{\zeta_t \in \mathcal{A}^*_t(\omega)} \big\{ \mathbb{E}[\zeta_t Z_t] - \rho^*_\omega(\zeta_t) \big\}.$$

In particular, let $\mathcal{Z}_t \doteq \mathcal{L}_1(\Omega, \mathcal{F}_t, P)$, $\mathcal{Z}^*_t \doteq \mathcal{L}_\infty(\Omega, \mathcal{F}_t, P)$ and take

$$v(z) \doteq \varepsilon_1[-z]_+ + \varepsilon_2[z]_+,$$

where $\varepsilon_1 \in [0,1]$ and $\varepsilon_2 \geq 0$ (compare with Example 7). This function $v(z)$ is convex positively homogeneous, and the corresponding function $z + v(z)$ is nondecreasing. The maximum inside the expectation in the right hand side of (4.79) is equal to zero if $-\varepsilon_1 g_\omega \leq \zeta_t - g_\omega \leq g_\omega \varepsilon_2$, and to $+\infty$ otherwise. It follows that the corresponding risk mapping ρ satisfies conditions (M1)–(M4), and

$$\rho_\omega(Z_t) = \sup_{\zeta_t \in \mathcal{A}_t(\omega)} \mathbb{E}[\zeta_t Z_t],$$

where $\eta_1 \doteq 1 - \varepsilon_1$, $\eta_2 \doteq 1 + \varepsilon_2$,

$$\mathcal{A}_t(\omega) = \left\{ \zeta_t \in \mathcal{Z}_t^* : \begin{array}{l} \eta_1 g_\omega(\tilde{\omega}) \le \zeta_t(\tilde{\omega}) \le \eta_2 g_\omega(\tilde{\omega}),\ a.e.\ \tilde{\omega} \in \Omega, \\ \mathbb{E}[\zeta_t | \mathcal{F}_{t-1}](\omega) = 1 \end{array} \right\}.$$

The between stages independence condition can be introduced here in a way similar to the previous example. Under this condition formulas (4.74), (4.75) and (4.76) will hold here as well.

Acknowledgement. The first author was partially supported by the NSF award DMS-0303545.

Part II

Robust Optimization and Random Sampling

5

Sampled Convex Programs and Probabilistically Robust Design

Giuseppe Calafiore[1] and Marco C. Campi[2]

[1] Dipartimento di Automatica e Informatica – Politecnico di Torino
C.so Duca degli Abruzzi, 10124, Italy
giuseppe.calafiore@polito.it

[2] Dipartimento di Elettronica per l'Automazione – Università di Brescia
via Branze 38, 25123 Brescia, Italy
campi@ing.unibs.it

Summary. This chapter deals with the sampled scenarios approach to robust convex programming and its applications to control analysis and synthesis. It has been shown in previous work [71] that by randomly sampling a sufficient number of constraints among the (possibly) infinite constraints of a robust convex program, one obtains a standard convex optimization problem whose solution is 'approximately feasible,' in a probabilistic sense, for the original robust convex program. This is a *generalization* property in the learning theoretic sense, since the satisfaction of a certain number of 'training' constraints entails the satisfaction of other 'unseen' constraints. In this contribution we provide a new efficient bound on the generalization rate of sampled convex programs, and discuss several applications of this paradigm to robust control analysis and design problems.

5.1 Introduction

Robust convex programming [35, 141] deals with optimization problems subject to a family of convex constraints that are parameterized by uncertainty terms. Solving a robust convex program (RCP) amounts to determining an optimal solution that is feasible for all possible constraints in the parameterized family. In more precise terms, an RCP may be formalized as

$$\begin{aligned} \text{RCP} : \min_{\theta \in \Theta} c^T \theta \text{ subject to} & \\ f(\theta, \delta) \leq 0, \quad \forall \delta \in \Delta, & \end{aligned} \tag{5.1}$$

where θ is the optimization variable, δ is the uncertainty parameter, $\Theta \subseteq \mathbb{R}^{n_\theta}$ is a convex and closed set, and $\Delta \subseteq \mathbb{R}^{n_\delta}$. The objective to be minimized can be taken as linear without loss of generality. Further, it is assumed that $f(\theta, \delta) : \Theta \times \Delta \to (-\infty, \infty]$ is continuous and convex in θ, for any fixed value of $\delta \in \Delta$. Notice that no assumption is instead made on the dependence of $f(\theta, \delta)$ on δ.

The constraints are here expressed by the condition $f(\theta, \delta) \leq 0$, where f is a scalar-valued function. Considering scalar-valued constraint functions is without loss of generality, since multiple constraints $f_1(\theta, \delta) \leq 0, \ldots, f_{n_f}(\theta, \delta) \leq 0$ can be reduced to a single scalar-valued constraint by the position $f(\theta, \delta) \doteq \max_{i=1,\ldots,n_f} f_i(\theta, \delta)$.

We remark that, despite convexity, robust convex programs are in general hard to solve numerically, see [35,37,141]. This is one of the motivations that led us to consider probabilistic relaxations of the problem, see [71] for an in-depth discussion.

Important special cases of robust convex programs are robust linear programs, [36], for which $f(\theta, \delta) = \max_{i=1,\ldots,n_f} f_i(\theta, \delta)$ and each $f_i(\theta, \delta)$ is affine in θ, and robust semidefinite programs, [12, 37, 141], for which $f(\theta, \delta) = \lambda_{\max}[F(\theta, \delta)]$, where

$$F(\theta, \delta) = F_0(\delta) + \sum_{i=1}^{n_\theta} \theta_i F_i(\delta), \quad F_i(\delta) = F_i^T(\delta),$$

and $\lambda_{\max}[\cdot]$ denotes the largest eigenvalue.

The RCP paradigm has found to date applications in many engineering endeavors, such as truss topology design [34], robust antenna array design, portfolio optimization [142], and robust estimation [121]. However, we shall be mainly concerned here with control systems, where RCPs arise naturally in the context of analysis and synthesis based on parameter-dependent Lyapunov functions, see, *e.g.*, [12,100,101,124], as well as in various problems of robust filtering [139,401] and set-membership state reachability and filtering [75,140]. It is perhaps also worth noticing that RCPs encompass deterministic min-max games of the type

$$\min_{\theta \in \Theta} \max_{\delta \in \Delta} f(\theta, \delta). \tag{5.2}$$

These problems can indeed be cast in the RCP format as

$$\min_{\theta \in \Theta, \gamma} \gamma \text{ subject to}$$
$$f(\theta, \delta) \leq \gamma, \quad \forall \delta \in \Delta.$$

In [69, 71], a probabilistic approach has been proposed to approximately solve problem (5.1). This approach is based on sampling at random a finite number N of constraints in the family $\{f(\theta, \delta) \leq 0, \; \delta \in \Delta\}$ and solving the corresponding standard convex problem. In particular, we explicitly define the *scenario* counterpart of RCP as

$$\mathrm{RCP}_N : \min_{\theta \in \Theta} c^T \theta \text{ subject to} \tag{5.3}$$
$$f(\theta, \delta^{(i)}) \leq 0, \; i = 1, \ldots, N,$$

where $\delta^{(1)}, \ldots, \delta^{(N)}$ are N independent identically distributed (i.i.d.) samples, drawn according to some given probability measure denoted as $\mathbb{P}$. A scenario design is given by an optimal solution $\hat{\theta}_N$ of RCP$_N$. Notice that $\hat{\theta}_N$ is a random variable that depends on the random extractions $\delta^{(1)}, \ldots, \delta^{(N)}$.

5.1.1 Properties of RCP$_N$

Let us first specify more precisely the probabilistic setup that we shall use in the following. We assume that the support Δ for δ is endowed with a σ-algebra $\mathcal{D}$ and that $\mathbb{P}$ is defined over $\mathcal{D}$. Moreover, we assume that $\{\delta \in \Delta : f(\theta, \delta) \leq 0\} \in \mathcal{D}$, $\forall \theta \in \Theta$. We have the following definition.

Definition 1 (violation probability). *Let $\theta \in \Theta$ be given. The* probability of violation *of θ is defined as*

$$V(\theta) \doteq \mathbb{P}\{\delta \in \Delta : \; f(\theta, \delta) > 0\}.$$

For example, if a uniform (with respect to Lebesgue measure) probability distribution is assumed, then $V(\theta)$ measures the volume of 'bad' parameters δ such that the constraint $f(\theta, \delta) \leq 0$ is violated. Clearly, a solution θ with small associated $V(\theta)$ is feasible for most of the problem instances, *i.e.* it is *approximately feasible* for the robust problem.

Definition 2 (ϵ-level solution). *Let $\epsilon \in (0, 1)$. We say that $\theta \in \Theta$ is an ϵ-level robustly feasible (or, more simply, an ϵ-level) solution, if $V(\theta) \leq \epsilon$.*

Our goal is to devise an algorithm that returns a ϵ-level solution, where ϵ is any fixed small level. It was shown in [71] that the solution returned by RCP$_N$ has indeed this characteristic, as summarized in the following theorem.

Theorem 1 (Corollary 1 of [71]). *Assume that, for any extraction of $\delta^{(1)}$, ..., $\delta^{(N)}$, the scenario problem RCP$_N$ attains a unique optimal solution $\hat{\theta}_N$.*

Fix two real numbers $\epsilon \in (0, 1)$ (level parameter) and $\beta \in (0, 1)$ (confidence parameter) and let

$$N \geq N_{\text{lin}}(\epsilon, \beta) \doteq \left\lfloor \frac{n_\theta}{\epsilon \beta} \right\rfloor \tag{5.4}$$

($\lfloor \rfloor$ denotes integer rounding towards zero). Then, with probability no smaller than $1 - \beta$, $\hat{\theta}_N$ is ϵ-level robustly feasible.

In this theorem, probability $1-\beta$ refers to the probability $\mathbb{P}^N$ ($=\mathbb{P} \times \mathbb{P} \cdots \times \mathbb{P}$, N times) of extracting a 'bad' multisample, *i.e.* a multisample $\delta^{(1)}, \ldots, \delta^{(N)}$ such that $\hat{\theta}_N$ does not meet the ϵ-level feasibility property. Figure 5.1 gives a visual interpretation of the result in Theorem 1.

The inequality (5.4) provides the minimum number of sampled constraints that are needed in order to attain the desired probabilistic levels of robustness in the solution. The function $N_{\text{lin}}(\epsilon, \beta)$ gives therefore a bound on the *generalization rate* of the scenario approach, which relates to the ability of the

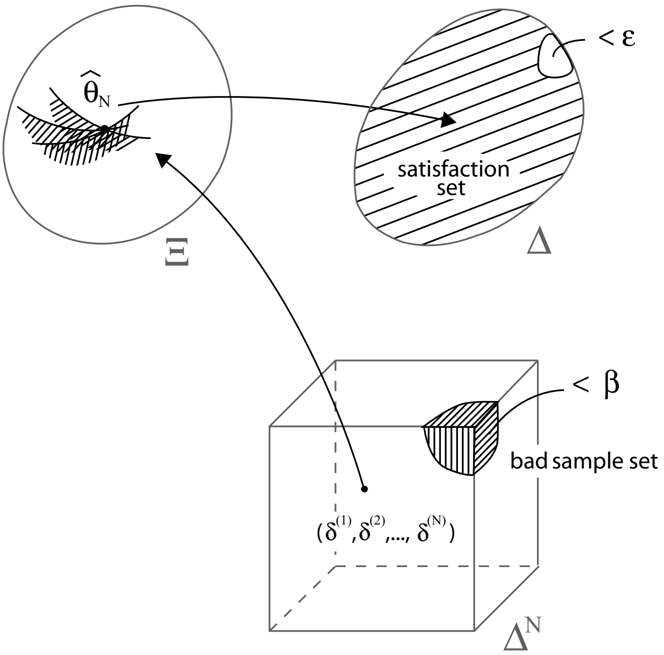


Figure 5.1. Interpretation of the scenario approach to robust convex programming: With probability at least $1-\beta$ the sampled scenarios $\delta^{(1)},\ldots,\delta^{(N)}$ lead to an optimal solution $\hat{\theta}_N$ which is feasible for all but at most a set of measure ϵ of the uncertainties

scenario solution of being feasible (with high probability) also with respect to constraints that were not explicitly taken into account in the solution of RCP_N (unseen scenarios). In formula (5.4), the suffix 'lin' underlines the fact that N grows linearly with respect to β^{-1}.

5.1.2 Related Works

The idea of seeking optimal designs that are robust in a probabilistic sense is of course not new. In the stochastic optimization literature, for instance, uncertainty in optimization is typically dealt with by introducing expectations, *i.e.* by attempting to solve a problem of the form $\min_{\theta\in\Theta} \mathbb{E}_{\delta}[f(\theta,\delta)]$ (cfr. the min-max problem formulation in (5.2)). Monte Carlo sampling techniques are commonly used in this context to numerically determine an approximate solution, see for instance [332]. More closely related to the RCP approach is the so-called *chance-constrained* optimization problem (CCP), where one seeks to optimize the objective under a constraint explicitly expressed in the form of a probability:

$$\mathrm{CCP}(\epsilon) : \min_{\theta \in \Theta} c^T \theta \text{ subject to} \qquad (5.5)$$
$$\mathbb{P}\{\delta \in \Delta : \ f(\theta, \delta) > 0\} \leq \epsilon.$$

It should be remarked that an exact numerical solution of the above probability constrained optimization problem is in general hopeless, see for instance the monograph [280]. A further negative news is that the probability constrained problem (5.5) is in general non-convex, even when the function $f(\theta, \delta)$ is convex in θ for all $\delta \in \Delta$. We direct the reader to the survey [281] and to Chapters 1 and 2 of this book for recent results related to chance-constrained optimization. Chapter 1, in particular, discusses numerically efficient sampling approximations of the chance-constrained problem, in the case when $f(\theta, \delta)$ is a bi-affine mapping. The relation between $\mathrm{CCP}(\epsilon)$ and RCP_N is discussed in Section 5.2.1 of this chapter.

The use of probabilistic robustness techniques in the domain of control design is instead relatively recent. The recent monograph [359] provides an historical perspective on the topic and a thorough survey of currently available randomized algorithms for approximately solving probabilistically constrained design problems in control. However, the randomized approach that we propose in this chapter is distinctively different from those discussed in [359] and in other related works such as [76, 129, 176, 249, 252, 273]. These latter references propose sequential stochastic algorithms for determining an approximately feasible design, based on random gradient descent or ellipsoidal iterations. The methodology described here is instead based on a one-shot solution of the sampled convex program by means, *e.g.*, of interior point techniques, and it is tailored to optimization. We shall not discuss sequential methods further here, but direct the reader to [359] and to the introduction in [71] for additional details on this subject.

5.2 An Improved Bound on the Generalization Rate

We next show that a better bound than (5.4) in fact holds for scenario convex problems. The new bound (Theorem 2 below) has both theoretical and practical importance. From the theoretical side, it shows that generalization is achieved with a number of samples that grows essentially as $O(\frac{n_\theta}{\epsilon} \ln \frac{1}{\beta})$. This implies that a much lower number of constraints is needed with respect to (5.4), which is important in practice when solving RCP_N numerically.

We start with a simplifying assumption that is made in order to avoid mathematical cluttering.

Assumption 5.1. *For all possible extractions $\delta^{(1)}, \ldots, \delta^{(N)}$, the optimization problem (5.3) is either unfeasible, or, if feasible, it attains a unique optimal solution.*

This assumption could actually be removed (*i.e.* we may allow for non-existence or non-uniqueness of the optimal solution) without harming the result, at the expense of complications in the proofs. These refined results may be obtained following a technique similar to the one developed in [71]. We now state the main result of this chapter.

Lemma 1 (Generalization rate of RCP$_N$). *Let Assumption 5.1 be satisfied, and let $\hat{\theta}_N$ be the optimal solution of (5.3), when the problem is feasible. Given probability level $\epsilon \in (0,1)$, define the event $\mathcal{B}$*

$$\mathcal{B} \doteq \left\{ (\delta^{(1)}, \ldots, \delta^{(N)}) : \mathrm{RCP}_N \textit{ is feasible, and } V(\hat{\theta}_N) > \epsilon \right\} \subseteq \Delta^N.$$

Then, it holds that

$$\mathbb{P}^N(\mathcal{B}) < \binom{N}{n_\theta}(1-\epsilon)^{N-n_\theta} \tag{5.6}$$

where n_θ is the number of decision variables in problem (5.3). In words, the probability of RCP$_N$ being feasible and providing a 'bad' solution (i.e. a solution with violation greater that ϵ) is bounded from above by the right hand side of (5.6).

In Lemma 1 and elsewhere, the measurability of $\mathcal{B}$, as well as that of other sets in Δ^N, is taken as an assumption. The proof of Lemma 1 needs some preliminaries, and it is hence reported in Section 5.6, to avoid breaking the continuity of discourse.

The following theorem is based on Lemma 1, and provides an explicit bound on the number of sampled scenarios that are needed to solve a robust convex problem to given probabilistic levels of accuracy and confidence.

Theorem 2. *Let Assumption 5.1 be satisfied. Fix two real numbers $\epsilon \in (0,1)$ (level parameter) and $\beta \in (0,1)$ (confidence parameter). If*

$$\begin{aligned} &N \geq N_{\mathrm{gen}}(\epsilon, \beta) \doteq \\ &\left\lceil \inf_{\nu\in(0,1)} \tfrac{1}{1-\nu} \left(\tfrac{1}{\epsilon} \ln \tfrac{1}{\beta} + n_\theta + \tfrac{n_\theta}{\epsilon} \ln \tfrac{1}{\nu\epsilon} + \tfrac{1}{\epsilon} \ln \tfrac{(n_\theta/\mathrm{e})^{n_\theta}}{n_\theta!} \right) \right\rceil \end{aligned} \tag{5.7}$$

($\lceil\,\rceil$ denotes the smallest integer greater than or equal to the argument) then, with probability no smaller than $1-\beta$, either the scenario problem RCP$_N$ *is unfeasible, and hence also RCP is unfeasible; or,* RCP$_N$ *is feasible, and then its optimal solution $\hat{\theta}_N$ is ϵ-level robustly feasible.*

In this theorem, probability $1-\beta$ refers to the N-fold probability $\mathbb{P}^N$ ($= \mathbb{P} \times \cdots \times \mathbb{P}$, N times). In other words, Theorem 2 states that if N (specified by (5.7)) random scenarios are drawn, the optimal solution of RCP$_N$ is ϵ-level feasible according to Definition 2, with high probability $1-\beta$.

Proof. We prove that, if N is chosen according to (5.7), then

$$\binom{N}{n_\theta}(1-\epsilon)^{N-n_\theta} \leq \beta \tag{5.8}$$

and hence, by Lemma 1,

$$\mathbb{P}^N\left(\left\{(\delta^{(1)},\ldots,\delta^{(N)}) : \mathrm{RCP}_N \text{ is feasible, and } V(\hat{\theta}_N) > \epsilon\right\}\right) < \beta.$$

Taking the complementary event, we would have

$$\mathbb{P}^N\Big(\Big\{(\delta^{(1)},\ldots,\delta^{(N)}) : \\ \mathrm{RCP}_N \text{ is unfeasible, or it is feasible and } V(\hat{\theta}_N) \le \epsilon\Big\}\Big) \ge 1-\beta$$

which is the claim of the theorem. Notice that the fact that if RCP_N is unfeasible then RCP is also unfeasible is obvious, since RCP_N exhibits only a subset of the constraints of RCP.

To prove that (5.7) implies (5.8), we proceed by simple algebraic manipulations. Any of the following inequality implies the next in a top-down fashion, where the first one comes from (5.7) where ν is a number in $(0,1)$:

$$\begin{aligned}
N &\ge \frac{1}{1-\nu}\left(\frac{1}{\epsilon}\ln\frac{1}{\beta} + n_\theta + \frac{n_\theta}{\epsilon}\ln\frac{1}{\nu\epsilon} + \frac{1}{\epsilon}\ln\left(\left(\frac{n_\theta}{e}\right)^{n_\theta}\frac{1}{n_\theta!}\right)\right) \\
(1-\nu)N &\ge \frac{1}{\epsilon}\ln\frac{1}{\beta} + n_\theta + \frac{n_\theta}{\epsilon}\ln\frac{1}{\nu\epsilon} + \frac{1}{\epsilon}\ln\left(\left(\frac{n_\theta}{e}\right)^{n_\theta}\frac{1}{n_\theta!}\right) \\
(1-\nu)N &\ge \frac{1}{\epsilon}\ln\frac{1}{\beta} + n_\theta + \frac{n_\theta}{\epsilon}\left(\ln\frac{n_\theta}{\nu\epsilon} - 1\right) - \frac{1}{\epsilon}\ln(n_\theta!) \\
N &\ge \frac{1}{\epsilon}\ln\frac{1}{\beta} + n_\theta + \frac{n_\theta}{\epsilon}\left(\ln\frac{n_\theta}{\nu\epsilon} - 1 + \frac{\nu N\epsilon}{n_\theta}\right) - \frac{1}{\epsilon}\ln(n_\theta!) \\
N &\ge \frac{1}{\epsilon}\ln\frac{1}{\beta} + n_\theta + \frac{n_\theta}{\epsilon}\ln N - \frac{1}{\epsilon}\ln(n_\theta!), \qquad (5.9)
\end{aligned}$$

where the last implication can be justified by observing that $\ln x \ge 1 - \frac{1}{x}$, for $x > 0$, and applying this inequality with $x = \frac{n_\theta}{\nu N\epsilon}$. Proceeding from (5.9), the next inequalities in the chain are

$$\begin{aligned}
\ln\beta &\ge -\epsilon N + \epsilon n_\theta + n_\theta\ln N - \ln(n_\theta!) \\
\beta &\ge \frac{N^{n_\theta}}{n_\theta!}e^{-\epsilon(N-n_\theta)} \\
\beta &\ge \frac{N(N-1)\cdots(N-n_\theta+1)}{n_\theta!}(1-\epsilon)^{N-n_\theta},
\end{aligned}$$

where, in the last implication, we have used the fact that

$$e^{-\epsilon(N-n_\theta)} \ge (1-\epsilon)^{N-n_\theta},$$

as it follows by taking logarithm of the two sides and further noting that $-\epsilon \ge \ln(1-\epsilon)$. The last inequality can be rewritten as

$$\beta \geq \binom{N}{n_\theta}(1-\epsilon)^{N-n_\theta},$$

which is (5.8). □

Bound (5.7) can be simplified and made explicit, as stated in the following corollary.

Corollary 1. *The results in Theorem 2 hold for*

$$N \geq N_{\log}(\epsilon,\beta) \doteq \left\lceil \frac{2}{\epsilon}\ln\frac{1}{\beta} + 2n_\theta + \frac{2n_\theta}{\epsilon}\ln\frac{2}{\epsilon} \right\rceil. \tag{5.10}$$

Proof. Observe that $(n_\theta/\mathrm{e})^{n_\theta} \leq n_\theta!$, and hence the last term in (5.7) is non-positive and can be dropped, leading to

$$N_{\mathrm{gen}}(\epsilon,\beta) \leq \left\lceil \frac{1}{1-\nu}\left(\frac{1}{\epsilon}\ln\frac{1}{\beta} + n_\theta + \frac{n_\theta}{\epsilon}\ln\frac{1}{\nu\epsilon}\right) \right\rceil, \tag{5.11}$$

where ν can be freely selected in $(0,1)$. The statement of the corollary is then obtained by selecting $\nu = 1/2$ in (5.11). We also note that further optimizing (5.11) with respect to ν always leads to a $\nu \leq 1/2$, with a corresponding improvement by at most of a factor 2. □

Remark 1 (sample complexity). Notice that bound (5.10) – and hence (5.7), which is tighter – substantially improves upon (5.4) in that dependence on $1/\beta$ is now logarithmic. This means that, in practice, confidence in the solution is not an issue in the scenario design approach, since values of β of the order of 10^{-10} or even smaller can be attained without substantially increasing the necessary number of samples. Table 5.1 shows a comparison of the these bounds for several values of ϵ and β.

Remark 2 (the role of convexity). Theorem 2 says that if we extract a *finite* number N of constraints, then the solution of the randomized problem, if feasible, satisfies most of the other unseen constraints. As we mentioned, this is a *generalization* property: the explicit satisfaction of some 'training' scenarios generalizes automatically to the satisfaction of other unseen scenarios. It is interesting to note that generalization calls for some kind of structure, and the only structure used here is convexity. So, convexity in the scenario approach is fundamental in two different respects: on the computational side, it allows for an efficient solution of the ensuing optimization problem, and on the theoretical side it allows for generalization.

Remark 3 (VC-dimension). Bound (5.10) depends on the problem structure through n_θ, the number of optimization variables, only. It is not difficult to conceive situations where the class of sets $\{\delta \in \Delta : f(\theta,\delta) > 0\} \subseteq \Delta$, parameterized in θ, has infinite VC-dimension (see, *e.g.*, [375] for a definition), even for small n_θ. Then, estimating $\mathbb{P}\{\delta \in \Delta : f(\theta,\delta) > 0\} = V(\theta)$ uniformly

Table 5.1. Comparison of sample-size bounds, for $n_\theta = 10$

	$\epsilon = 0.1$	$\epsilon = 0.01$	$\epsilon = 0.001$	$\epsilon = 0.0001$
$\beta = 0.01$	$N_{\text{lin}} = 10^4$ $N_{\text{log}} = 712$ $N_{\text{gen}} = 533$	$N_{\text{lin}} = 10^5$ $N_{\text{log}} = 11538$ $N_{\text{gen}} = 7940$	$N_{\text{lin}} = 10^6$ $N_{\text{log}} = 161249$ $N_{\text{gen}} = 105142$	$N_{\text{lin}} = 10^7$ $N_{\text{log}} = 2072821$ $N_{\text{gen}} = 1303039$
$\beta = 0.001$	$N_{\text{lin}} = 10^5$ $N_{\text{log}} = 758$ $N_{\text{gen}} = 562$	$N_{\text{lin}} = 10^6$ $N_{\text{log}} = 11999$ $N_{\text{gen}} = 8203$	$N_{\text{lin}} = 10^7$ $N_{\text{log}} = 165854$ $N_{\text{gen}} = 107683$	$N_{\text{lin}} = 10^8$ $N_{\text{log}} = 2118873$ $N_{\text{gen}} = 1327959$
$\beta = 0.0001$	$N_{\text{lin}} = 10^6$ $N_{\text{log}} = 804$ $N_{\text{gen}} = 589$	$N_{\text{lin}} = 10^7$ $N_{\text{log}} = 12459$ $N_{\text{gen}} = 8465$	$N_{\text{lin}} = 10^8$ $N_{\text{log}} = 170459$ $N_{\text{gen}} = 110219$	$N_{\text{lin}} = 10^9$ $N_{\text{log}} = 2164925$ $N_{\text{gen}} = 1352842$
$\beta = 0.00001$	$N_{\text{lin}} = 10^7$ $N_{\text{log}} = 850$ $N_{\text{gen}} = 617$	$N_{\text{lin}} = 10^8$ $N_{\text{log}} = 12920$ $N_{\text{gen}} = 8725$	$N_{\text{lin}} = 10^9$ $N_{\text{log}} = 175064$ $N_{\text{gen}} = 112748$	$N_{\text{lin}} = 10^{10}$ $N_{\text{log}} = 2210977$ $N_{\text{gen}} = 1377687$

with respect to θ is impossible and the VC-theory is of no use. Theorem 2 says that, if attention is restricted to $\hat{\theta}_N$, then estimating $V(\hat{\theta}_N)$ becomes possible at a low computational cost.

Remark 4 (A-priori and a-posteriori assessments). It is worth noticing that a distinction should be made between the *a-priori* and *a-posteriori* assessments that one can make regarding the probability of constraint violation. Indeed, *before* running the optimization, it is guaranteed by Theorem 2 that if $N \geq N_{\text{gen}}(\epsilon, \beta)$ samples are drawn, the solution of the randomized program will be ϵ-level robustly feasible, with probability no smaller than $1 - \beta$. However, the *a-priori* parameters ϵ, β are generally chosen to be not too small, due to technological limitations on the number of constraints that one specific optimization software can deal with.

On the other hand, once a solution has been computed (and hence $\theta = \hat{\theta}_N$ has been fixed), one can make an *a-posteriori* assessment of the level of feasibility using standard Monte Carlo techniques. In this case, a new batch of $\tilde{N}$ independent random samples of $\delta \in \Delta$ is generated, and the *empirical probability* of constraint violation, say $\hat{V}_{\tilde{N}}(\hat{\theta}_N)$, is computed according to the formula $\hat{V}_{\tilde{N}}(\hat{\theta}_N) = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} 1(f(\hat{\theta}_N, \delta^{(i)}) > 0)$, where $1(\cdot)$ is the indicator function. Then, the classical Hoeffding's inequality, [162], guarantees that

$$|\hat{V}_{\tilde{N}}(\hat{\theta}_N) - V(\hat{\theta}_N)| \leq \tilde{\epsilon}$$

holds with confidence greater than $1 - \tilde{\beta}$, provided that

$$\tilde{N} \geq \frac{\ln 2/\tilde{\beta}}{2\tilde{\epsilon}^2}$$

test samples are drawn. This latter *a-posteriori* verification can be easily performed using a very large sample size $\tilde{N}$, because no numerical optimization is involved in such an evaluation.

5.2.1 Chance-Constrained and Sampled Convex Programs

Consider the probability-constrained problem (5.5) The distinctive feature of CCP(ϵ) is that it is required that the neglected constraint set is chosen in an optimal way, *i.e.* among all sets of constraints with probability no larger than ϵ, the removed one is the one that allows for the greatest reduction in the design objective. In the optimization literature, this problem is called a 'chance-constrained' optimization problem, see, *e.g.*, [280, 374].

As we have already seen, RCP$_N$ returns with high probability a feasible solution of CCP(ϵ). In the next theorem, we establish a further connection between CCP(ϵ) and RCP$_N$.

Theorem 3. *Let $\epsilon, \beta \in (0,1)$ be given probability levels. Let $J_{\mathrm{CCP}(\epsilon)}$ denote the optimal objective value of the chance-constrained problem* CCP(ϵ) *in (5.5) when it is feasible (i.e. $J_{\mathrm{CCP}(\epsilon)} \doteq \inf_{\theta\in\Theta} c^T\theta$ subject to $V(\theta) \leq \epsilon$) and let $J_{\mathrm{RCP_N}}$ be the optimal objective value of the scenario problem* RCP$_N$ *in (5.3) when it is feasible (notice that $J_{\mathrm{RCP_N}}$ is a random variable, while $J_{\mathrm{CCP}(\epsilon)}$ is a deterministic value), with N any number satisfying (5.10). Then:*

1. *With probability at least $1-\beta$, if* RCP$_N$ *is feasible it holds that*
$$J_{\mathrm{RCP_N}} \geq J_{\mathrm{CCP}(\epsilon)};$$
2. *Assume* CCP(ϵ_1) *is feasible, where $\epsilon_1 = 1-(1-\beta)^{1/N}$. With probability at least $1-\beta$, it holds that*
$$J_{\mathrm{RCP_N}} \leq J_{\mathrm{CCP}(\epsilon_1)}.$$

Proof. The first claim is immediate, since from Theorem 2, with probability at least $1-\beta$, if RCP$_N$ is feasible, then its optimal solution $\hat{\theta}_N$ satisfies $V(\hat{\theta}_N) \leq \epsilon$, *i.e.* it is a feasible, albeit possibly not optimal, solution for problem CCP(ϵ), and hence $J_{\mathrm{RCP_N}} \geq J_{\mathrm{CCP}(\epsilon)}$.

To prove the second claim, notice that if θ is feasible for problem CCP(ϵ_1) with $\epsilon_1 = 1-(1-\beta)^{1/N}$, *i.e.*
$$\mathbb{P}\{\delta \in \Delta : \ f(\theta,\delta) > 0\} \leq 1-(1-\beta)^{1/N},$$
then for each of N independent extractions $\delta^{(1)},\ldots,\delta^{(N)}$ of δ it holds that
$$\mathbb{P}\{\delta^{(i)} \in \Delta : \ f(\theta,\delta^{(i)}) \leq 0\} \geq (1-\beta)^{1/N}, \quad i=1,\ldots,N,$$
and hence, by independence, the joint event $\{(\delta^{(1)},\ldots,\delta^{(N)}) \in \Delta^N : f(\theta,\delta^{(i)}) \leq 0,\ i=1,\ldots,N\}$ holds with probability at least $1-\beta$. This means that, with

probability at least $1-\beta$, a feasible point for $\mathrm{CCP}(\epsilon_1)$ is also a feasible point for RCP_N. We now have two possibilities, depending on whether $\mathrm{CCP}(\epsilon_1)$ attains an optimal solution (*i.e.* a $\hat{\theta}$ feasible for $\mathrm{CCP}(\epsilon_1)$ exists such that $c^T\hat{\theta} = J_{\mathrm{CCP}(\epsilon_1)}$) or not. In the first situation ($\hat{\theta}$ exists), taking $\theta = \hat{\theta}$ in the previous reasoning immediately implies that $J_{\mathrm{RCP_N}} \leq J_{\mathrm{CCP}}(\epsilon_1)$, as desired.

In the second situation ($\hat{\theta}$ does not exist), consider a point $\bar{\theta}$ which is feasible for $\mathrm{CCP}(\epsilon_1)$ and such that $c^T\bar{\theta} \leq J_{\mathrm{CCP}(\epsilon_1)} + \rho$, for some $\rho > 0$ (such a $\bar{\theta}$ exists since $J_{\mathrm{CCP}(\epsilon_1)} = \inf c^T\theta$ over θ's that are feasible for $\mathrm{CCP}(\epsilon_1)$). By the previous reasoning, this implies that, with probability at least $1-\beta$, the point $\bar{\theta}$ is also feasible for problem RCP_N, entailing

$$\mathbb{P}\left\{(\delta^{(1)},\ldots,\delta^{(N)}) \in \Delta^N : J_{\mathrm{RCP_N}} \leq J_{\mathrm{CCP}(\epsilon_1)} + \rho\right\} \geq 1-\beta. \tag{5.12}$$

For the purpose of contradiction, suppose now that result 2 in the theorem is violated so that $J_{\mathrm{RCP_N}} > J_{\mathrm{CCP}(\epsilon_1)}$ with probability larger than β. Since

$$\begin{aligned}&\left\{(\delta^{(1)},\ldots,\delta^{(N)}) \in \Delta^N : J_{\mathrm{RCP_N}} > J_{\mathrm{CCP}(\epsilon_1)}\right\}\\ &= \bigcup_{\nu>0}\left\{(\delta^{(1)},\ldots,\delta^{(N)}) \in \Delta^N : J_{\mathrm{CCP_N}} > J_{\mathrm{CCP}(\epsilon_1)} + \frac{1}{\nu}\right\},\end{aligned}$$

then

$$\begin{aligned}\beta &< \mathbb{P}^N\left\{(\delta^{(1)},\ldots,\delta^{(N)}) \in \Delta^N : J_{\mathrm{RCP_N}} > J_{\mathrm{CCP}(\epsilon_1)}\right\}\\ &= \lim_{\nu\to\infty}\mathbb{P}^N\left\{(\delta^{(1)},\ldots,\delta^{(N)}) \in \Delta^N : J_{\mathrm{RCP_N}} > J_{\mathrm{CCP}(\epsilon_1)} + \frac{1}{\nu}\right\}\end{aligned}$$

and we conclude that there exists a $\bar{\nu}$ such that $J_{\mathrm{RCP_N}} > J_{\mathrm{CCP}(\epsilon_1)} + \frac{1}{\bar{\nu}}$ with probability larger than β. But this contradicts (5.12) for $\rho = \frac{1}{\bar{\nu}}$, so concluding the proof. □

A few words help clarify result 2 in Theorem 3. First notice that $J_{\mathrm{CCP}(\epsilon)}$ is a non-increasing function of ϵ. Result 2 states that the optimal value $J_{\mathrm{RCP_N}}$ (where N has been selected so that the optimal solution is ϵ-level feasible with probability $1-\beta$) is, with probability at least $1-\beta$, no worse than $J_{\mathrm{CCP}(\epsilon_1)}$, for a certain $\epsilon_1 \leq \epsilon$ explicitly given. For a ready comparison between ϵ and ϵ_1, observe that relation $a^s \leq sa + (1-s)$ holds for any $a \geq 0$ and $0 \leq s \leq 1$ (as it easily follows by observing that the two sides coincide for $s = 0$ and $s = 1$ and that a^s is convex in s). Then, with the position $a \doteq 1-\beta; s \doteq 1/N$, we have

$$\epsilon_1 = 1-(1-\beta)^{1/N} \geq 1 - \left[\frac{1}{N}(1-\beta) + \left(1-\frac{1}{N}\right)\right] = \frac{\beta}{N},$$

which, used in result 2 of the theorem, gives $J_{\mathrm{RCP_N}} \leq J_{\mathrm{CCP}(\beta/N)}$, with N any number satisfying (5.10). For a crude evaluation, note that if $n_\theta > 1$ and β is assumed to be of the same order of ϵ, then the dominant term in (5.10) is

$\frac{2n_\theta}{\epsilon}\ln\frac{2}{\epsilon}$, leading to $\epsilon_1 \approx \frac{\beta}{N} \approx \frac{\beta}{2n_\theta \ln\frac{2}{\epsilon}}\epsilon$, where $\frac{\beta}{2n_\theta \ln\frac{2}{\epsilon}}$ is the rescaling factor between ϵ and ϵ_1.

5.3 Sampled Convex Programs in Robust Control

In this section, we consider design problems in control as an illustration of the robust optimization set-up discussed in the previous sections.

A wide variety of robust analysis and synthesis problems in control can be formulated as determining a vector of controller (or more generally 'design') parameters such that some performance specifications on the controlled system are satisfied, as the plant varies over a specified family of admissible plants. More precisely, many robust control problems can be expressed as optimization problems subject to closed-loop constraints that are parameterized by the uncertainties affecting the plant. In formal terms, if $\theta \in \Theta \subseteq \mathbb{R}^{n_\theta}$ is the 'design parameter' (which includes the actual controller parameters, plus possibly other additional variables such as parameters of Lyapunov functions, slack variables and scalings), and the family of admissible plants is parameterized by an 'uncertainty vector' $\delta \in \Delta \subseteq \mathbb{R}^{n_\delta}$, then the prototype control problem we refer to consists of minimizing a linear objective $c^T\theta$, subject to $f(\theta,\delta) \leq 0$, $\delta \in \Delta$, where $f(\theta,\delta) : \Theta \times \Delta \to (-\infty,\infty]$ is a function that specifies the closed-loop constraints. To make things more concrete, consider *e.g.* a robust $\mathcal{H}_\infty$ or $\mathcal{H}_2$ control problem. If the closed-loop system is denoted as $G_{cl}(\xi,\delta)$ (where ξ are design parameters), then we can take $\theta = (\xi,\gamma)$ and minimize γ subject to the constraints $\psi(\xi,\delta) \leq \gamma$, where

$$\psi(\xi,\delta) = \begin{cases} \|G_{cl}(\xi,\delta)\|, & \text{if } G_{cl}(\xi,\delta) \text{ is stable} \\ \infty, & \text{otherwise,} \end{cases}$$

and the norm is either the $\mathcal{H}_\infty$ or $\mathcal{H}_2$ norm. Here, $f(\theta,\delta) = \psi(\xi,\delta) - \gamma$, and $c^T\theta = \gamma$. When the function f is convex in θ (which happens in several, albeit not all practical control design cases) we are faced with an RCP problem of the kind discussed in the previous sections.

We shall discuss next several relevant control analysis and synthesis problems that can be naturally cast in the RCP format, and for which no deterministic polynomial-time algorithm is known that computes an exact solution. For these problems, the solution approach that we propose is to first relax the problem in a probabilistic sense and then solve the probabilistic problem via the sampled scenario approach.

5.3.1 Stability Analysis Using Parameter-Dependent Lyapunov Functions

Consider the family of linear systems described in state-space form as

$$\{\dot{x} = A(\delta)x, \quad \delta \in \Delta\}, \tag{5.13}$$

where $x \in \mathbb{R}^{n_x}$ is the state variable, and the parameter $\delta \in \Delta \subseteq \mathbb{R}^{n_\delta}$ parameterizing the system family is unknown, but constant in time. In the sequel, we shall refer to system families of the type (5.13) simply as 'uncertain systems'.

Let a symmetric matrix function $P(\xi, \delta)$ be chosen in a family parameterized by a vector of parameters $\xi \in \mathbb{R}^{n_\xi}$, and assume that $P(\xi, \delta)$ is linear in ξ, for all $\delta \in \Delta$. The dependence of $P(\xi, \delta)$ on the uncertainty δ, as well as the dependence of $A(\delta)$ on δ, are otherwise left generic. We introduce the following sufficient condition for robust stability, which follows directly from the standard Lyapunov theory.

Definition 3 (generalized quadratic stability – GQS). *Given a symmetric matrix function $P(\xi, \delta)$, linear in $\xi \in \mathbb{R}^{n_\xi}$ for all $\delta \in \Delta$, the uncertain system (5.13) is said to be quadratically stable with respect to $P(\xi, \delta)$ if there exists $\xi \in \mathbb{R}^{n_\xi}$ such that*

$$\begin{bmatrix} -P(\xi,\delta) & 0 \\ 0 & A^T(\delta)P(\xi,\delta) + P(\xi,\delta)A(\delta) \end{bmatrix} \prec 0, \quad \forall \delta \in \Delta \tag{5.14}$$

($\prec$ means negative definite). Such a $P(\xi, \delta)$ is called a Lyapunov matrix for the uncertain system (5.13).

For specific choices of the parameterization $P(\xi, \delta)$, the above GQS criterion clearly encompasses the popular quadratic stability (QS, [59, 61]) and affine quadratic stability (AQS, [131]) criteria, as well as the biquadratic stability condition of [367]. For instance, the quadratic stability condition is recovered by choosing $P(\xi, \delta) = P$ (*i.e.* ξ contains the free elements of $P = P^T$, and there is no dependence on δ), which amounts to determining a *single* Lyapunov matrix P that simultaneously satisfies (5.14). The AQS condition is instead obtained by choosing

$$P(\xi, \delta) = P_0 + \delta_1 P_1 + \cdots + \delta_{n_\delta} P_{n_\delta}, \tag{5.15}$$

where ξ represents the free elements in the matrices $P_i = P_i^T$, $i = 0, \ldots, n_\delta$. Notice that QS, AQS and GQS constitute a hierarchy of sufficient conditions for robust stability having decreasing conservatism. However, even the simplest (and most conservative) QS condition is hard to check numerically. Only in the case when the set $\{A(\delta),\ \delta \in \Delta\}$ is a polytope, the QS condition is exactly checkable numerically via convex optimization, [59, 61]. As a matter of fact, in this case a classical vertex result holds which permits to convert the infinite number of constraints entailed by (5.14) into a finite number of LMIs involving the vertices of the polytope. Notice however that in the classical case when $A(\delta)$ is an interval matrix, the number of vertices of the polytope grows as $2^{n_x^2}$, which means that QS cannot be checked with a computational effort that is polynomial in the problem size n_x.

The AQS condition is computationally hard even in the polytopic case with a fixed number of vertices, and therefore convex relaxations that lead to numerically tractable sufficient conditions for AQS have been proposed in the literature. For instance, in [131] a further multiconvexity requirement is imposed in order to obtain LMI sufficient conditions when Δ is a hypercube and $A(\delta)$ is affine, while in [124] the so-called $\mathcal{S}$-procedure is used for the same purpose. More recently, a generalization of the method, based on a class of Lyapunov functions that depend quadratically (instead of affinely) on δ has been proposed in [367], while the case of linear-fractional (LFT) dependence in $A(\delta)$ is studied in [264]. All these extensions are again particular cases of the GQS criterion defined above.

Now, notice that a key feature of condition (5.14) is that, for any *fixed* $\delta \in \Delta$ it represents a convex LMI condition in ξ, and therefore finding a feasible parameter ξ amounts indeed to solving a robust convex program. This is the key observation that makes the scenario paradigm well-suited for probabilistic analysis within the context of generalized quadratic stability. With pre-specified confidence, a matrix $P(\xi, \delta)$ generated by a scenario solution would be a Lyapunov matrix for all but a small fraction of the systems in the family (5.13).

Formalization as RCP$_N$

Notice that condition (5.14) is a feasibility condition expressed by a strict matrix inequality, while both problems RCP and RCP$_N$ are minimization problems subject to a non-strict inequality condition (in (5.1) we have $f(\theta,\delta) \leq 0$ as opposed to $f(\theta,\delta) < 0$). The precise formalization of the GQS problem within the scenario setting can be done in more than one way and it is to a certain extent a matter of taste. Here, as an illustration, we further develop this first example to indicate a possible way to cast it in the RCP$_N$ format. It is tacitly understood that similar formalizations apply to all other examples.

First, set an optimization program with the format of (5.1) as follows:

$$\text{RCP} : \min \alpha \text{ subject to}$$
$$-I \preceq \begin{bmatrix} -P(\xi,\delta) & 0 \\ 0 & A^T(\delta)P(\xi,\delta) + P(\xi,\delta)A(\delta) \end{bmatrix} \preceq \alpha I, \quad \forall \delta \in \Delta.$$

Then, assume a probability measure $\mathbb{P}$ over the uncertainties is given, and build the scenario counterpart of the problem

$$\text{RCP}_N : \min \alpha \text{ subject to}$$
$$-I \preceq \begin{bmatrix} -P(\xi,\delta^{(i)}) & 0 \\ 0 & A^T(\delta^{(i)})P(\xi,\delta^{(i)}) + P(\xi,\delta^{(i)})A(\delta^{(i)}) \end{bmatrix} \preceq \alpha I,$$
$$i = 1, \ldots, N,$$

where the scenarios $\delta^{(i)}$ are independently extracted at random according to $\mathbb{P}$. Here, the optimization variable is $\theta \doteq (\xi, \alpha)$. Note also that the lower bound $-I$ has been introduced without loss of generality since, otherwise, the solution may escape to infinity due to homogeneity of the constraint.

Applying Theorem 2 we can then conclude that, with probability at least $1-\beta$, either RCP_N is unfeasible, so that RCP and the original GQS is unfeasible, or the solution $(\bar{\xi},\bar{\alpha})$ of RCP_N is a ϵ-level solution for RCP. In the latter case, if $\bar{\alpha} \geq 0$, it is easily seen that GQS is again unfeasible. Finally, if $\bar{\alpha} < 0$, then $P(\bar{\xi},\delta)$ is a ϵ-level solution for GQS.

5.3.2 Generalized Quadratic Synthesis for Uncertain Systems

Consider the uncertain system

$$\dot{x} = A(\delta)x + B_1(\delta)w + B_2(\delta)u \tag{5.16}$$

$$z = C(\delta)x, \tag{5.17}$$

where $x \in \mathbb{R}^{n_x}$ is the state variable, $w \in \mathbb{R}^{n_w}$ is the exogenous input, $u \in \mathbb{R}^{n_u}$ is the control input, $z \in \mathbb{R}^{n_z}$ is the performance output, and all matrices are generic functions of $\delta \in \Delta$.

State-feedback stabilization

Suppose we want to stabilize (5.16) by means of a state-feedback control law $u = Kx$, where $K \in \mathbb{R}^{n_u,n_x}$ is a static feedback gain. The resulting closed-loop system is robustly stable if and only if $A_{cl}(\delta) \doteq A(\delta) + B_2(\delta)K$ is Hurwitz for all $\delta \in \Delta$. Using the enhanced LMI characterization proposed in [13] (Theorem 3.1), robust stabilizability of (5.16) is equivalent to the existence of matrices $V \in \mathbb{R}^{n_x,n_x}$, $R \in \mathbb{R}^{n_u,n_x}$, and a Lyapunov symmetric matrix function $P(\delta) \in \mathbb{R}^{n_x,n_x}$ such that

$$\begin{bmatrix} -(V+V^T) & V^TA^T(\delta) + R^TB_2^T(\delta) + P(\delta) & V^T \\ * & -P(\delta) & 0 \\ * & * & -P(\delta) \end{bmatrix} \prec 0, \quad \forall \delta \in \Delta \tag{5.18}$$

(asterisks denote entries that are easily inferred from symmetry). If a feasible solution is found, the robustly stabilizing feedback gain is recovered as $K = RV^{-1}$. A sufficient condition for robust stabilizability is hence readily obtained by considering a specific parameterized matrix function family $P(\xi,\delta)$ (linear in the parameter ξ, for any fixed $\delta \in \Delta$) in the above condition. The resulting problem is convex in the decision variable $\theta \doteq (\xi, V, R)$, for any fixed $\delta \in \Delta$, and it is therefore a robust convex problem. Notice again that this robust problem is hard to solve in general. As an exception, in the special case when $[A(\delta)\ B_2(\delta)]$ is affine in δ, Δ is a hypercube, and $P(\delta)$ is chosen in the affine form (5.15), the above robust condition can be transformed by a standard ‘vertexization’ argument into a finite set of LMIs involving the vertex matrices, and hence solved exactly (this latter special case is indeed the one presented in [13]). We remark however again that the number of vertices (and hence of LMI constraints) grows exponentially with the number of uncertain parameters n_δ, which makes this standard approach practically unviable in cases when n_δ is large.

This robust state-feedback stabilization problem is amenable to the scenario randomization approach similarly to the problem in Section 5.3.1. When $P(\delta) = P$ (*i.e.* we look for a parameter-independent Lyapunov matrix), we can alternatively use a standard Lyapunov inequality instead of (5.18). This is indeed the case shown in the numerical example presented in Section 5.4.1.

State-feedback robust $\mathcal{H}_2$ synthesis

For system (5.16)–(5.17), consider the problem of designing a state-feedback law $u = Kx$ such that the closed loop is robustly stable, and has guaranteed $\mathcal{H}_2$ performance level γ on the $w - z$ channel.

Adopting the LMI characterization of $\mathcal{H}_2$ performance proposed in [13, Theorem 3.3], we have that the closed-loop system with controller $K = RV^{-1}$ is robustly stable and has guaranteed $\mathcal{H}_2$ performance less than γ if there exist $Z = Z^T \in \mathbb{R}^{n_w,n_w}$, $R \in \mathbb{R}^{n_u,n_x}$ and $V \in \mathbb{R}^{n_x,n_x}$ and a Lyapunov symmetric matrix function $P(\delta) \in \mathbb{R}^{n_x,n_x}$ such that

$$\begin{bmatrix} -(V+V^T) & V^T A^T(\delta) + R^T B_2^T(\delta) + P(\delta) & V^T C(\delta) & V^T \\ * & -P(\delta) & 0 & 0 \\ * & * & -\gamma I & 0 \\ * & * & * & -P(\delta) \end{bmatrix} \prec 0, \quad \forall \delta \in \Delta$$

$$\begin{bmatrix} P(\delta) & B_1(\delta) \\ * & Z \end{bmatrix} \succ 0, \quad \operatorname{tr} Z < 1, \quad \forall \delta \in \Delta.$$

Again, we can recast the problem within the randomized setting by considering symmetric parameter-dependent Lyapunov matrix $P(\xi, \delta)$ linear in $\xi \in \mathbb{R}^{n_\xi}$. Notice also that the above matrix inequalities are linear in γ, and therefore the $\mathcal{H}_2$ level can also be minimized subject to these constraints.

5.3.3 Controller Synthesis for LPV Systems

Consider a linear parameter-varying (LPV) system of the form

$$\begin{bmatrix} \dot{x} \\ z \\ y \end{bmatrix} = \begin{bmatrix} A(\delta(t)) & B_1(\delta(t)) & B_2(\delta(t)) \\ C_1(\delta(t)) & 0 & D_{12}(\delta(t)) \\ C_2(\delta(t)) & D_{21}(\delta(t)) & 0 \end{bmatrix} \begin{bmatrix} x \\ w \\ u \end{bmatrix},$$

where $x \in \mathbb{R}^{n_x}$ is the state, $w \in \mathbb{R}^{n_w}$ is the exogenous input, $u \in \mathbb{R}^{n_u}$ is the control input, $z \in \mathbb{R}^{n_z}$ is the performance output, $y \in \mathbb{R}^{n_y}$ is the measured output, and $\delta(t) \in \mathbb{R}^{n_\delta}$ is a time-varying parameter, usually referred to as the *scheduling parameter*. In the LPV setting, the parameter $\delta(t)$ is known to be contained in a set Δ, whereas its actual value at time t, $\delta(t)$, is *a-priori* unknown but can be measured online. The LPV formulation has recently received considerable attention, since it forms the basis of systematic gain-scheduling approaches to nonlinear control design, see for instance [31, 48, 49, 320] and the survey [310].

The design objective is to determine a controller that processes at time t not only the measured output $y(t)$ but also the measured scheduling parameter $\delta(t)$, in order to determine the control input $u(t)$ for the system.

Quadratic control of LPV systems

Here, we consider a controller of the form

$$\begin{bmatrix} \dot{x}_k \\ u \end{bmatrix} = \begin{bmatrix} A_k(\delta(t)) & B_k(\delta(t)) \\ C_k(\delta(t)) & 0 \end{bmatrix} \begin{bmatrix} x_k \\ y \end{bmatrix}.$$

Suppose the controller has to be designed so that exponential stability is enforced while achieving a quadratic performance specification on the $w - z$ channel. The main difficulty of the problem resides in the fact that in natural applications of the LPV methodology the dependence of the matrices on the scheduling parameter is nonlinear. To address this issue, two main approaches have been proposed in the literature. One approach is based on embedding the nonlinear dependence into a simpler one (such as affine or linear-fractional), and then reduce the problem to some tractable finite-dimensional convex optimization problem, see for instance [18] and the references therein. Of course, this approach generally involves conservatism in the approximation. A second methodology is instead based on 'gridding' the parameter set, and hence transforming the solvability conditions of the original problem into a finite set of convex constraints, see for instance [10, 31, 400]. The problem with this approach is that the number of grid points (and of constraints, consequently) increases exponentially with the number of scheduling parameters, and may lead to numerically critical implementations. Recently, an alternative randomization-based technique for LPV design has been proposed in [129]. The motivation for this section comes from this latter approach. Indeed, the parameter-dependent inequalities derived in [129] are there solved using sequential stochastic gradient methods, while the same inequalities are here viewed as an instance of a robust convex feasibility problem, and hence directly amenable to the scenario solution.

To be specific, let the following (rather standard) assumptions (see [31, 129]) hold:

(i) $\left[D_{12}^T(\delta(t))C_1(\delta(t)) \;\; D_{12}^T(\delta(t))D_{12}(\delta(t)) \right] = \left[0 \; I \right]$, $\begin{bmatrix} B_1(\delta(t)) \\ D_{21}(\delta(t))D_{21}^T(\delta(t)) \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix}$, hold for all $\delta(t) \in \Delta$;

(ii) $\delta(t)$ is a piecewise continuous function of t, with a finite number of discontinuities in any interval.

Then, formalize the quadratic LPV control problem as follows: Given $\gamma > 0$, find matrices $A_k(\delta(t))$, $B_k(\delta(t))$, $C_k(\delta(t))$ such that the closed-loop system is exponentially stable, and

$$\sup_{w \in L_2 \setminus 0} \frac{\int_0^\infty z^T(t) z(t) dt}{\int_0^\infty w^T(t) w(t) dt} < \gamma,$$

for all $\delta(\cdot)$ such that $\delta(t) \in \varDelta$, $\forall t$. The solvability conditions for this problem are directly stated in terms of robust feasibility of three LMIs in [31] (Theorem 4.2). The synthesis LMIs reported below are an equivalent modification of those in [31].

The quadratic LPV L_2 control problem is solvable if and only if there exist $P = P^T \in \mathbb{R}^{n_x, n_x}$ and $Q = Q^T \in \mathbb{R}^{n_x, n_x}$ such that

$$\begin{bmatrix} A(\delta)P + PA^T(\delta) - \gamma B_2(\delta) B_2^T(\delta) & PC_1^T(\delta) & B_1(\delta) \\ * & -\gamma I & 0 \\ * & * & -I \end{bmatrix} \prec 0, \quad \forall \delta \in \varDelta$$

$$\begin{bmatrix} A^T(\delta)Q + QA(\delta) - C_2^T(\delta) C_2(\delta) & QB_1(\delta) & C_1^T(\delta) \\ * & -I & 0 \\ * & * & -\gamma I \end{bmatrix} \prec 0, \quad \forall \delta \in \varDelta$$

$$\begin{bmatrix} P & I \\ * & Q \end{bmatrix} \succ 0.$$

Moreover, if feasible $P \succ 0, Q \succ 0$ exist, then the LPV controller matrices are recovered as

$$\begin{aligned} A_k(\delta) &= A(\delta) - Q^{-1}C_2^T(\delta)C_2(\delta) - B_2(\delta)B_2^T(\delta)Z^{-1} + \gamma^{-1}Q^{-1}C_1^T(\delta)C_1(\delta) + \\ &\quad + \gamma^{-1}Q^{-1}(A^T(\delta)Q + QA(\delta) + \gamma^{-1}C_1^T(\delta)C_1(\delta) - C_2^T(\delta)C_2(\delta) + \\ &\quad + QB_1(\delta)B_1^T(\delta)Q)Q^{-1}Z^{-1}, \\ B_k(\delta) &= Q^{-1}C_2^T(\delta), \\ C_k(\delta) &= -B_2^T(\delta)Z^{-1}, \end{aligned}$$

where $Z \doteq (P - Q^{-1})/\gamma$.

Again, this LPV design problem (either finding a feasible design for fixed level γ, or minimizing γ subject to the above constraints) is stated in the form of an RCP, and it is hence amenable to the randomized scenario solution. In this specific context, the scenario approach can be viewed as a kind of gridding technique, where the grid points are randomly selected. The advantage resides in the fact that bound (5.10) can be used to determine the number of grid points, and this number is independent of the dimension of δ.

State-feedback synthesis for LPV systems

Similar to the approach in the previous section, we next consider a state-feedback design problem for a LPV system with guaranteed decay rate. Consider the LPV system

$$\dot{x} = A(\delta(t))x + B(\delta(t))u$$

and assume that the state is measured, and that the controller is of the form

$$u = K(\delta(t))x,$$

where

$$K(\delta(t)) = K_0 + \sum_{i=1}^{n_\delta} K_i \delta_i(t).$$

The control objective is to determine the matrices K_i, $i = 0, \ldots, n_\delta$, such that the controlled system has a guaranteed exponential decay rate $\nu > 0$. Specifically, defining the closed loop matrix $A_{cl}(\delta(t)) = A(\delta(t)) + B(\delta(t))K(\delta(t))$, the control objective is met if there exists a symmetric matrix $P \succ 0$ such that the matrix inequality

$$A_{cl}(\delta)P + PA_{cl}^T(\delta) + 2\nu P \prec 0 \tag{5.19}$$

holds for all $\delta \in \varDelta$. Introducing the new variables $Y_i \doteq K_i P$, $i = 0, \ldots, n_\delta$, the design requirements are satisfied if

$$\begin{bmatrix} \begin{array}{l} A(\delta)P + B(\delta)Y_0 + \sum_{i=1}^{n_\delta} B(\delta)Y_i\delta_i \\ +PA^T(\delta) + Y_0^T B^T(\delta) + \sum_{i=1}^{n_\delta} Y_i^T B^T(\delta)\delta_i + 2\nu P \end{array} & 0 \\ 0 & -P \end{bmatrix} \prec 0, \quad \forall \delta \in \varDelta.$$

5.4 Numerical Examples

We next report the results of some numerical experiments of control design performed using the scenario approach.

5.4.1 Robust State-Feedback Stabilization

Given the uncertain system

$$\dot{x} = A(\delta)x + Bu$$

we wish to design a state-feedback control law $u = Kx$ such that the closed-loop is quadratically stable, for all δ in the allowable uncertainty set $\varDelta$. This design specification is satisfied if and only if there exist $P \succ 0$ and Y such that

$$A(\delta)P + PA^T(\delta) + BY + Y^T B^T \prec 0, \quad \forall \delta \in \varDelta \tag{5.20}$$

(see for instance [59]). Due to homogeneity in these conditions, we can reformulate the problem in minimization form as the RCP

$$\min_{P,Y,\alpha} \alpha \text{ subject to}$$

$$-I \preceq \begin{bmatrix} -P & 0 \\ 0 & A(\delta)P + PA^T(\delta) + BY + Y^T B^T \end{bmatrix} \preceq \alpha I. \quad (5.21)$$

If the optimal α is negative, then the original design conditions are satisfied, and the controller is retrieved as $K = YP^{-1}$.

We here consider a simple numerical example, with

$$A(\delta) = \begin{bmatrix} \rho_2\delta_2 & 1 + \rho_1\delta_1 \\ -(1 + \rho_1\delta_1)^2 & 2(0.1 + \rho_2\delta_2)(1 + \rho_1\delta_1) \end{bmatrix}$$
$$B = \begin{bmatrix} 10 \\ 15 \end{bmatrix}$$

$\rho_1 = 0.9$, $\rho_2 = 0.5$, with $|\delta_1| \leq 1$, $|\delta_2| \leq 1$. The scenario counterpart of the problem is

$$\min_{P,Y,\alpha} \alpha \text{ subject to}$$

$$-I \preceq \begin{bmatrix} -P & 0 \\ 0 & A(\delta^{(i)})P + PA^T(\delta^{(i)}) + BY + Y^T B^T \end{bmatrix} \preceq \alpha I, \ i = 1, \ldots, N.$$

where $\delta^{(1)}, \ldots, \delta^{(N)}$ are i.i.d. uncertainty samples.

In this example we have $n_\theta = 3 + 2 + 1 = 6$ design variables (the free entries of symmetric P, plus the two entries of Y, and α). Setting $\epsilon = 0.1$ and $\beta = 0.01$, bound (5.7) requires at least $N = 336$ uncertainty samples. Assuming uniform probability measure over Δ, and solving numerically one instance of the scenario problem (by means of LMILab toolbox in Matlab) we obtained $\alpha = -0.0073$,

$$P = \begin{bmatrix} 0.0143 & 0.0312 \\ 0.0312 & 0.1479 \end{bmatrix}, \quad Y = \begin{bmatrix} -0.0093 & -0.0133 \end{bmatrix}$$

and hence the controller

$$K = YP^{-1} = [-0.8414 \ \ 0.0879]. \quad (5.22)$$

This controller was then tested *a-posteriori* via Monte Carlo. The estimated empirical probability of P, Y violating the design LMI (5.21) was 0.0145. A plot of the violation set is shown in Figure 5.2.

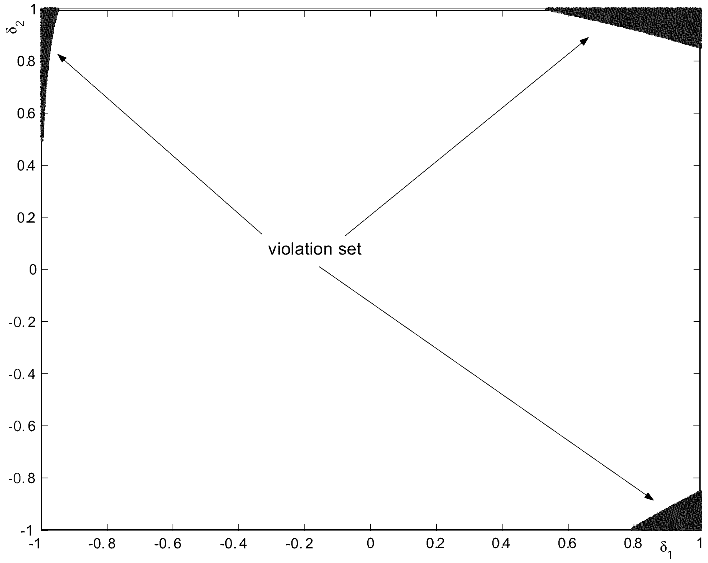


Figure 5.2. Violation set: the filled areas denote values of δ for which the LMI (5.21) with controller (5.22) is violated

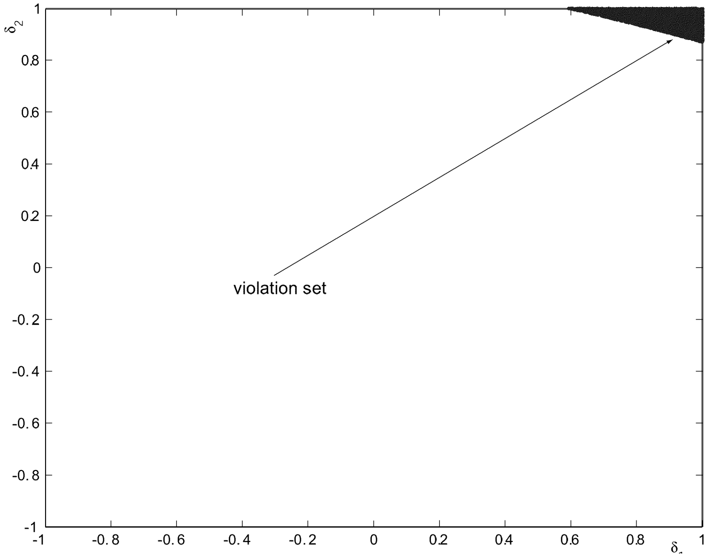


Figure 5.3. Violation set: the filled area denotes values of δ for which the LMI (5.20) with controller (5.22) is violated

Notice that it might also be meaningful to test *a-posteriori* the controller (5.22) against the original Lyapunov inequality (5.20). In this case, the *a-posteriori* Monte Carlo test yielded an estimated probability of 0.0065 of violating (5.20). A plot of this second violation set is shown in Figure 5.3.

Setting instead the *a-priori* probability levels to $\epsilon = 0.01$ and $\beta = 0.001$, bound (5.7) would require at least $N = 5170$ uncertainty samples. Solving one instance of the scenario problem, we found an optimal solution with $\alpha = -0.0055$ and

$$P = \begin{bmatrix} 0.0094 & 0.0218 \\ 0.0218 & 0.1292 \end{bmatrix}, \qquad Y = \begin{bmatrix} -0.0102 & -0.0147 \end{bmatrix}$$

and hence the controller

$$K = YP^{-1} = [-1.3581 \;\; 0.1147]. \tag{5.23}$$

This controller was again tested *a-posteriori* via Monte Carlo. The *a-posteriori* estimated probability of P, Y violating the design LMI (5.21) was 2.49×10^{-4}. This means in practice that the computed P is a Lyapunov matrix for all but a very small fraction of the closed-loop plants, see the violation set in Figure 5.4.

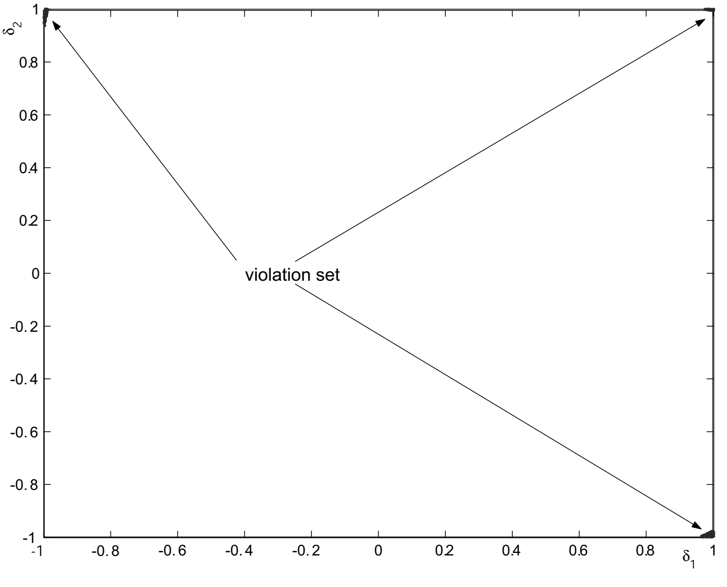


Figure 5.4. Violation set of LMI (5.21) with controller (5.23)

5.4.2 Synthesis of LPV Controller

We next present a numerical example of LPV state-feedback stabilization with guaranteed decay rate (see Section 5.3.3).

We consider a multivariable model given in [5] (see also the original paper [370] for a slightly different model and set of data) of the dynamics of the lateral motion of an aircraft. The state space equation is given by

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & L_p & L_\beta & L_r \\ g/V & 0 & Y_\beta & -1 \\ N_{\dot\beta}(g/V) & N_p & N_\beta + N_{\dot\beta} Y_\beta & N_r - N_{\dot\beta} \end{bmatrix} x + \begin{bmatrix} 0 & 0 \\ 0 & -3.91 \\ 0.035 & 0 \\ -2.53 & 0.31 \end{bmatrix} u,$$

where x_1 is the bank angle, x_2 its derivative, x_3 the sideslip angle, x_4 the yaw rate, u_1 the rudder deflection, u_2 the aileron deflection, and the coefficients in the A matrix have a physical interpretation as discussed in [5] and are subject to time variability.

The following nominal values for the parameters are taken: $L_p = -2.93$, $L_\beta = -4.75$, $L_r = 0.78$, $g/V = 0.086$, $Y_\beta = -0.11$, $N_{\dot\beta} = 0.1$, $N_p = -0.042$, $N_\beta = 2.601$ and $N_r = -0.29$. The actual values of the parameters fluctuate in time with a maximum variation of 15% from the nominal (central) values and are measured on-line.

Setting the desired decay rate to $\nu = 0.5$, and assuming uniform probability distribution over Δ, we applied the proposed scenario approach for the solution of this design problem. Similarly to Section 5.3.1, we introduced the RCP

$$\min \alpha \text{ subject to} \tag{5.24}$$

$$-I \preceq \begin{bmatrix} \begin{array}{l} A(\delta)P + B(\delta)Y_0 + \sum_{i=1}^{n_\delta} B(\delta)Y_i\delta_i \\ +PA^T(\delta) + Y_0^T B^T(\delta) + \sum_{i=1}^{n_\delta} Y_i^T B^T(\delta)\delta_i + 2\nu P \end{array} & 0 \\ 0 & -P \end{bmatrix} \preceq \alpha I,$$

$$\forall \delta \in \Delta.$$

and then solved its scenario counterpart with N computed as follows: the number of uncertainty terms is $n_\delta = 9$, so that $n_\theta = 83$. The probability levels were selected to be $\epsilon = 0.1$ and $\beta = 0.01$, yielding $N = 5232$ according to Theorem 2.

The solution of one instance of the scenario problem yielded optimal values $\alpha = -0.2294$, and

$$P = \begin{bmatrix} 0.4345 & -0.3404 & -0.0014 & -0.0061 \\ -0.3404 & 0.7950 & -0.0053 & -0.0007 \\ -0.0014 & -0.0053 & 0.4787 & 0.3604 \\ -0.0061 & -0.0007 & 0.3604 & 0.7507 \end{bmatrix}$$

$$K_0 = \begin{bmatrix} 2.3689 & 3.0267 & 0.2346 & -0.2593 \\ -0.0268 & -0.0028 & 1.6702 & -2.1804 \end{bmatrix} \times 10^3$$

$$K_1 = \begin{bmatrix} -1.9052 & -2.4343 & -0.1896 & 0.2091 \\ 0.0221 & 0.0021 & -1.3443 & 1.7538 \end{bmatrix} \times 10^4$$

$$K_2 = \begin{bmatrix} 1.5467 & 1.9763 & 0.1539 & -0.1698 \\ -0.0179 & -0.0017 & 1.0914 & -1.4238 \end{bmatrix} \times 10^4$$

$$K_3 = \begin{bmatrix} -1.8403 & -2.3515 & -0.1831 & 0.2020 \\ 0.0213 & 0.0020 & -1.2985 & 1.6944 \end{bmatrix} \times 10^3$$

$$K_4 = \begin{bmatrix} -269.0519 & -343.8210 & -26.8210 & 29.5593 \\ 3.1197 & 0.2849 & -190.1478 & 247.8405 \end{bmatrix}$$

$$K_5 = \begin{bmatrix} -325.6902 & -416.1430 & -31.9206 & 35.5400 \\ 3.7771 & 0.3475 & -229.8091 & 299.8153 \end{bmatrix}$$

$$K_6 = \begin{bmatrix} -0.7610 & -1.0314 & -0.3023 & 0.4185 \\ -2.1024 & -2.6955 & -0.5855 & 0.7303 \end{bmatrix} \times 10^4$$

$$K_7 = \begin{bmatrix} 8.4788 & 11.2324 & 1.2490 & -0.9159 \\ -0.0983 & -0.0090 & 5.9940 & -7.8125 \end{bmatrix}$$

$$K_8 = \begin{bmatrix} -0.8506 & -1.0279 & 0.1419 & -0.2416 \\ 2.1211 & 2.6972 & -0.5517 & 0.7533 \end{bmatrix} \times 10^4$$

$$K_9 = \begin{bmatrix} -1.7472 & -2.2325 & -0.1738 & 0.1922 \\ 0.0203 & 0.0019 & -1.2328 & 1.6084 \end{bmatrix} \times 10^3.$$

Different *a-posteriori* tests can be conducted on the computed solution. For instance, we may estimate by Monte Carlo the probability of violation of the constraint used in problem (5.24). This estimated empirical probability resulted to be equal to 8.65×10^{-5}.

Alternatively (and perhaps more meaningfully), we may test the solution against the original design inequality (5.19). In this case, using again $\tilde{N} = 6.103 \times 10^6$ parameter samples, we obtained an estimated empirical probability of violating (5.19) equal to zero, *i.e.* our design satisfied all the *a-posteriori* random tests.

5.5 Conclusions

In this chapter we presented a novel approach to robust control design, based on the concept of uncertainty scenarios. Within this framework, if the robustness requirements are imposed in a probabilistic sense, then a wide class of control analysis and synthesis problems are amenable to efficient numerical solution. This solution is computed solving a convex optimization problem having a finite number N of sampled constraints. An efficient lower bound is determined on the number N of scenarios that are required to obtain a design that guarantees an *a-priori* specified probabilistic robustness level.

The methodology is illustrated by several control design examples that present difficulties when tackled by means of standard worst-case techniques.

We believe that, due to its intrinsic simplicity, the scenario approach will be an appealing solution technique for many practical engineering design problems, also beyond the control applications mentioned in this chapter.

5.6 Appendix: Proof of Lemma 1

5.6.1 Preliminaries

We first recall a classical result due to Helly, see [298].

Lemma 2 (Helly). *Let $\{\mathcal{X}_i\}_{i=1,\dots,p}$ be a finite collection of convex sets in $\mathbb{R}^n$. If every sub-collection consisting of $n+1$ sets has a non-empty intersection, then the entire collection has a non-empty intersection.*

Next, we prove a key instrumental result. Consider the convex optimization program

$$\mathcal{P} : \min_{x \in \mathbb{R}^n} c^T x \text{ subject to}$$
$$x \in \bigcap_{i \in \{1,\dots,m\}} \mathcal{X}_i,$$

where $\mathcal{X}_i$, $i = 1, \dots, m$, are closed convex sets, and define the convex programs $\mathcal{P}_k$, $k = 1, \dots, m$, obtained from $\mathcal{P}$ by removing the k-th constraint:

$$\mathcal{P}_k : \min_{x \in \mathbb{R}^n} c^T x \text{ subject to}$$
$$x \in \bigcap_{i \in \{1,\dots,m\} \setminus k} \mathcal{X}_i.$$

In the following, we assume existence and uniqueness of the optimal solution x^* of $\mathcal{P}$, and of the optimal solution x_k^* of $\mathcal{P}_k$, $k = 1, \dots, m$.

We have the following definition.

Definition 4 (support constraint). *The k-th constraint $\mathcal{X}_k$ is a* support constraint *for $\mathcal{P}$ if $c^T x_k^* < c^T x^*$.*

The following theorem holds.

Theorem 4. *The number of support constraints for problem $\mathcal{P}$ is at most n.*

A proof of this result was first given by the authors of the present contribution in [71]. We here report an alternative and more compact proof based on an idea suggested to us by professor A. Nemirovski in a personal communication.

Proof. Let problem $\mathcal{P}$ have q support constraints $\mathcal{X}_{s_1}, \dots, \mathcal{X}_{s_q}$, where $\mathcal{S} \doteq \{s_1, \dots, s_q\}$ is a subset of q indices from $\{1, \dots, m\}$. We next prove (by contradiction) that $q \leq n$.

Let $J^* = c^T x^*$ and $J_k^* = c^T x_k^*$ denote the optimal objective values of $\mathcal{P}$ and $\mathcal{P}_k$, respectively. Consider the smallest objective improvement obtained by removing a support constraint

$$\eta_{\min} \doteq \min_{k \in \mathcal{S}} (J^* - J_k^*)$$

and, for some η with $0 < \eta < \eta_{\min}$, define the hyperplane

$$\mathcal{H} \doteq \{x : c^T x = J^* - \eta\}.$$

By construction, the q points x_k^*, $k \in \mathcal{S}$, lie in the half-space $\{x : c^T x < J^* - \eta\}$, while x^* lies in the half-space $\{x : c^T x > J^* - \eta\}$, and therefore $\mathcal{H}$ separates x_k^*, $k \in \mathcal{S}$, from x^*. Next, for all indices $k \in \mathcal{S}$, we denote with $\bar{x}_k^*$ the point of intersection between the line segment $\overline{x_k^* x^*}$ and $\mathcal{H}$.

Since $x_k^* \in \bigcap_{i \in \{1,\ldots,m\} \setminus k} \mathcal{X}_i$, $k \in \mathcal{S}$, and $x^* \in \bigcap_{i \in \{1,\ldots,m\}} \mathcal{X}_i$, then by convexity we have that $\bar{x}_k^* \in \bigcap_{i \in \{1,\ldots,m\} \setminus k} \mathcal{X}_i$, $k \in \mathcal{S}$, and therefore (since, by construction, $\bar{x}_k^* \in \mathcal{H}$)

$$\bar{x}_k^* \in \left(\bigcap_{i \in \{1,\ldots,m\} \setminus k} \mathcal{X}_i \right) \bigcap \mathcal{H}, \quad k \in \mathcal{S}.$$

For $i = 1, \ldots, m$, define the convex sets $\Omega_i \doteq \mathcal{X}_i \bigcap \mathcal{H}$, and consider any collection $\{\Omega_{i_1}, \ldots, \Omega_{i_n}\}$ of n of these sets.

Suppose now (for the purpose of contradiction) that $q > n$. Then, there must exist an index $j \notin \{i_1, \ldots, i_n\}$ such that $\mathcal{X}_j$ is a support constraint, and by the previous reasoning, this means that there exists a point $\bar{x}_j^*$ such that $\bar{x}_j^* \in \left(\bigcap_{i \in \{1,\ldots,m\} \setminus j} \mathcal{X}_i \right) \bigcap \mathcal{H}$. Thus, $\bar{x}_j^* \in \Omega_{i_1} \cap \cdots \cap \Omega_{i_n}$, that is the collection of convex sets $\{\Omega_{i_1}, \ldots, \Omega_{i_n}\}$ has at least a point in common. Now, since the sets Ω_i, $i = 1, \ldots, m$, belong to the hyperplane $\mathcal{H}$ (*i.e.* to $\mathbb{R}^{n-1}$, modulo a fixed translation) and all collections composed of n of these sets have a point in common, by Helly's lemma (Lemma 2) there exists a point $\tilde{x}$ such that $\tilde{x} \in \bigcap_{i \in \{1,\ldots,m\}} \Omega_i$. Such a $\tilde{x}$ would therefore be feasible for problem $\mathcal{P}$; moreover, it would yield an objective value $\tilde{J} = c^T \tilde{x} < c^T x^* = J^*$ (since $\tilde{x} \in \mathcal{H}$). This is a contradiction, because x^* would no longer be an optimal solution for $\mathcal{P}$, and hence we conclude that $q \leq n$. □

Remark 5 (Support constraints and active constraints). Notice that the set X_s of support constraints of problem $\mathcal{P}$ does not in general coincide with the set X_a of constraints that are *active* at the optimum. By active constraints, we mean those $\mathcal{X}_k$ for which $x^* \in \partial \mathcal{X}_k$, where $\partial \mathcal{X}_k$ denotes the boundary of the convex set $\mathcal{X}_k$. However, support constraints must be active constraints, *i.e.* $X_s \subseteq X_a$.

We are now ready to present a proof of Lemma 1.

5.6.2 Proof of Lemma 1

For clarity of exposition, we first assume that problem RCP_N is feasible for any selection of $\delta^{(1)}$, ..., $\delta^{(N)}$. The case where infeasibility can occur is obtained as an easy extension as indicated at the end of the proof.

Given N scenarios $\delta^{(1)}, \ldots, \delta^{(N)}$, select a subset $I = \{i_1, \ldots, i_{n_\theta}\}$ of n_θ indices from $\{1, \ldots, N\}$ and let $\hat{\theta}_I$ be the optimal solution of the program

$$\min_{\theta \in \Theta} c^T \theta \text{ subject to}$$
$$f(\theta, \delta^{(i_j)}) \leq 0, \ j = 1, \ldots, n_\theta.$$

Based on $\hat{\theta}_I$ we next introduce a subset Δ_I^N of the set Δ^N defined as

$$\Delta_I^N \doteq \{(\delta^{(1)}, \ldots, \delta^{(N)}) : \quad \hat{\theta}_I = \hat{\theta}_N\} \tag{5.25}$$

($\hat{\theta}_N$ is the optimal solution with all N constraints $\delta^{(1)}, \ldots, \delta^{(N)}$ in place).

Let now I range over the collection $\mathcal{I}$ of all possible choices of n_θ indices from $\{1, \ldots, N\}$ ($\mathcal{I}$ contains $\binom{N}{n_\theta}$ sets). We want to prove that

$$\Delta^N = \bigcup_{I \in \mathcal{I}} \Delta_I^N. \tag{5.26}$$

To show (5.26), take any $(\delta^{(1)}, \ldots, \delta^{(N)}) \in \Delta^N$. From the set of constraint $\delta^{(1)}, \ldots, \delta^{(N)}$ eliminate a constraint which is not a support constraint (this is possible in view of Theorem 4, since $N > n_\theta$). The resulting optimization problem with $N-1$ constraints admits the same optimal solution $\hat{\theta}_N$ as the original problem with N constraints. Consider now the set of the remaining $N-1$ constraints and, among these, remove a constraint which is not a support constraint for the problem with $N-1$ constraints. Again, the optimal solution does not change. If we keep going this way until we are left with n_θ constraints, in the end we still have $\hat{\theta}_N$ as optimal solution, showing that $(\delta^{(1)}, \ldots, \delta^{(N)}) \in \Delta_I^N$, where I is the set containing the n_θ constraints remaining at the end of the process. Since this is true for any choice of $(\delta^{(1)}, \ldots, \delta^{(N)}) \in \Delta^N$, (5.26) is proven.

Next, let

$$B \doteq \{(\delta^{(1)}, \ldots, \delta^{(N)}) : \quad V(\hat{\theta}_N) > \epsilon\}$$

and

$$B_I \doteq \{(\delta^{(1)}, \ldots, \delta^{(N)}) : \quad V(\hat{\theta}_I) > \epsilon\}.$$

We now have

$$\begin{aligned} B &= B \cap \Delta^N \\ &= B \cap (\cup_{I \in \mathcal{I}} \Delta_I^N) \quad \text{(apply (5.26))} \\ &= \cup_{I \in \mathcal{I}} (B \cap \Delta_I^N) \\ &= \cup_{I \in \mathcal{I}} (B_I \cap \Delta_I^N). \quad \text{(because of (5.25))} \end{aligned} \tag{5.27}$$

A bound for $\mathbb{P}^N(B)$ is now obtained by bounding $\mathbb{P}(B_I \cap \Delta_I^N)$ and then summing over $I \in \mathcal{I}$.

Fix any I, *e.g.* $I = \{1, \ldots, n_\theta\}$ to be more explicit. The set $B_I = B_{\{1,\ldots,n_\theta\}}$ is a cylinder with base in the cartesian product of the first n_θ constraint domains (this follows from the fact that condition $V(\hat{\theta}_{\{1,\ldots,n_\theta\}}) > \epsilon$ only involves the first n_θ constraints). Fix $(\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(n_\theta)}) \in$ base of the cylinder. For a point $(\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(n_\theta)}, \delta^{(n_\theta+1)}, \ldots, \delta^{(N)})$ to be in $B_{\{1,\ldots,n_\theta\}} \cap \Delta^N_{\{1,\ldots,n_\theta\}}$, constraints $\delta^{(n_\theta+1)}, \ldots, \delta^{(N)}$ must be satisfied by $\hat{\theta}_{\{1,\ldots,n_\theta\}}$, for, otherwise, we would not have $\hat{\theta}_{\{1,\ldots,n_\theta\}} = \hat{\theta}_N$, as it is required in $\Delta^N_{\{1,\ldots,n_\theta\}}$. But, $V(\hat{\theta}_{\{1,\ldots,n_\theta\}}) > \epsilon$ in $B_{\{1,\ldots,n_\theta\}}$. Thus, by the fact that the extractions are independent, we conclude that

$$\begin{aligned}\mathbb{P}^{N-n_\theta}\{(\delta^{(n_\theta+1)}, \ldots, \delta^{(N)}) :\ & (\bar{\delta}^{(1)}, \ldots, \bar{\delta}^{(n_\theta)}, \delta^{(n_\theta+1)}, \ldots, \delta^{(N)}) \\ & \in B_{\{1,\ldots,n_\theta\}} \cap \Delta^N_{\{1,\ldots,n_\theta\}}\} < (1-\epsilon)^{N-n_\theta}.\end{aligned}$$

The probability on the left hand side is nothing but the conditional probability that $(\delta^{(1)}, \ldots, \delta^{(N)}) \in B_{\{1,\ldots,n_\theta\}} \cap \Delta^N_{\{1,\ldots,n_\theta\}}$ given $\delta^{(1)} = \bar{\delta}^{(1)}, \ldots, \delta^{(n_\theta)} = \bar{\delta}^{(n_\theta)}$. Integrating over the base of the cylinder $B_{\{1,\ldots,n_\theta\}}$, we then obtain

$$\mathbb{P}^N(B_{\{1,\ldots,n_\theta\}} \cap \Delta^N_{\{1,\ldots,n_\theta\}}) < (1-\epsilon)^{N-n_\theta} \cdot \mathbb{P}^{n_\theta}(\text{base of } B_{\{1,\ldots,n_\theta\}}) \leq (1-\epsilon)^{N-n_\theta}.$$

From (5.27), we finally arrive to the desired bound for $\mathbb{P}^N(B)$

$$\mathbb{P}^N(B) \leq \sum_{I \in \mathcal{I}} \mathbb{P}^N(B_I \cap \Delta_I^N) < \binom{N}{n_\theta}(1-\epsilon)^{N-n_\theta}. \tag{5.28}$$

So far, we have assumed that RCP_N is feasible for any selection of $\delta^{(1)}, \ldots, \delta^{(N)}$. Relax now this assumption and call $\Delta_F^N \subseteq \Delta^N$ the set where RCP_N is indeed feasible. The same derivation can then be worked out by only focusing on the event Δ_F^N leading to the conclusion that (5.28) holds with $B \doteq \left\{(\delta^{(1)}, \ldots, \delta^{(N)}) \in \Delta_F^N : V(\hat{\theta}_N) > \epsilon\right\}$. □

6

Tetris: A Study of Randomized Constraint Sampling

Vivek F. Farias and Benjamin Van Roy

Department of Electrical Engineering
Stanford University
{bvr,vff}@stanford.edu

Summary. Approximate Dynamic Programming is a means of synthesizing near-optimal policies for large scale stochastic control problems. We examine here the LP approach to approximate Dynamic Programming [98] which requires the solution of a linear program with a tractable number of variables but a potentially large number of constraints. Randomized constraint sampling is one means of dealing with such a program and results from [99] suggest that in fact, such a scheme is capable of producing good solutions to the linear program that arises in the context of approximate Dynamic Programming. We present here a summary of those results, and a case study wherein the technique is used to produce a controller for the game of Tetris. The case study highlights several practical issues concerning the applicability of the constraint sampling approach. We also demonstrate a controller that matches - and in some ways outperforms - controllers produced by other state of the art techniques for large-scale stochastic control.

6.1 Introduction

Randomized constraint sampling has recently been proposed as an approach for approximating solutions to optimization problems when the number of constraints is intractable – say, a googol or even infinity. The idea is to define a probability distribution ψ over the set of constraints and to sample a subset consisting of some tractable number N of independent identically distributed constraints. Then, a *relaxed problem*, in which the same objective function is optimized but only the sampled constraints are imposed, is solved.

An immediate question raised is whether solutions to the relaxed problem provide meaningful approximations to solutions of the original optimization problem. This question is partially addressed by recent theory developed first in the context of linear programming [99] and then convex programming [71]. In particular, it has been shown that, for a problem with K variables, given a number of samples

$$N = O\left(\frac{1}{\epsilon}\left(K \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}\right)\right),$$

with probability at least $1 - \delta$, any optimal solution to the relaxed problem violates a set of constraints $\mathcal{V}$ with measure $\psi(\mathcal{V}) \leq \epsilon$. Hence, given a reasonable number of samples, one can ensure that treating the relaxed problem leads to an 'almost feasible' solution to the original problem. One interesting aspect of this result is that N does not depend on the number of constraints associated with the original optimization problem.

The aforementioned theoretical result leads to another question:

> In order that a solution to the relaxed problem be useful, does it suffice to know that the measure of the set $\mathcal{V}$ of constraints it violates, $\psi(\mathcal{V})$, is small?

It is not possible to address this question without more specific context. In some problems, every constraint is critical. In others, violating a small fraction of the constraints may be acceptable. Further, the context should influence the relative importance of constraints and therefore how the distribution ψ is selected.

Approximate dynamic programming offers one context in which randomized constraint sampling addresses a pressing need. The goal is to synthesize a suboptimal control policy for a large scale stochastic control problem. One approach that has received much recent attention entails solving a linear program with an intractable number of constraints [98,324,365]. For certain special cases, the linear program can be solved exactly [148,322] while [365,366] study constraint generation heuristics for general problems. Most generally, constraint sampling can be applied [99]. The linear programming approach to approximate dynamic programming provides a suitable context for assessing the effectiveness of constraint sampling. In particular, violation of constraints can be translated to a tangible metric – controller performance. The relationship is studied in [99], which offers motivation for why violation of a small fraction of constraints should not severely degrade controller performance. However, the theory is inconclusive. In this chapter, we present experimental results that further explore the promise of constraint sampling in this context.

Our study involves a specific stochastic control problem: the game of Tetris. In principle, an optimal strategy for playing Tetris might be computed via dynamic programming algorithms. However, because of the enormity of the state space, this is computationally infeasible. Instead, one might synthesize a suboptimal strategy using methods of approximate dynamic programming, as has been done in [43, 174, 369]. In this chapter, we experiment with the linear programming approach, which differs from others that have previously been applied to Tetris. This study sheds light on the effectiveness of both the linear programming approach to approximate dynamic programming as a means of producing controllers for hard stochastic control problems, and randomized constraint sampling as a way of dealing with an intractable number of constraints.

The remainder of this chapter is organized as follows: in Section 6.2, we make precise the notion of a stochastic control problem and present Tetris as an example of such a problem. In Section 6.3, we introduce the linear programming approach to dynamic programming. In the following section, we discuss how this linear programming approach might be extended to approximate dynamic programming and in doing so, discuss results from [98, 99] on the quality of approximation such an approach might achieve, and a practically implementable constraint sampling scheme. Finally in Section 6.5 we describe how a controller for Tetris was constructed using the LP approach for approximate dynamic programming along with constraint sampling.

6.2 Stochastic Control and Tetris

Consider a discrete-time dynamic system which, at each time t, takes on a state $x_t \in \mathcal{S}$ and takes as input an action $a_t \in \mathcal{A}_{x_t}$. We assume that the state space $\mathcal{S}$ is finite and that for each $x \in \mathcal{S}$, the set of actions $\mathcal{A}_x$ is finite. Let $p_a(x, y)$ denote the probability that the next state is y given that the current state is x and the current action is a.

A *policy* is a mapping $\pi : \mathcal{S} \to \mathcal{A}$ from state to action. A cost function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ assigns a cost $g(x, a)$ to each state-action pair (x, a). We pose as the objective to select a policy π that minimizes the expected sum of discounted future costs:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t g(x_t, a_t) \Big| x_0 = x, a_t = \pi(x_t)\right], \tag{6.1}$$

where $\alpha \in (0, 1)$ is the discount factor.

Tetris is a popular video game in which falling pieces are positioned by rotation and translation as they fall onto a wall made up of previously fallen pieces. Each piece is made up of four equally-sized bricks, and the Tetris board is a two-dimensional grid, ten-bricks wide and twenty-bricks high. Each piece takes on one of seven possible shapes. A point is received for each row constructed without any holes, and the corresponding row is cleared. The game terminates once the height of the wall exceeds 20. The objective is to maximize the expected number of points accumulated over the course of the game. A representative mid-game board configuration is illustrated in Figure 6.1.

Indeed, Tetris can be formulated as a stochastic control problem:

- The state x_t encodes the board configuration and the shape of the falling piece.
- The action a_t encodes the rotation and translation applied to the falling piece.
- It is natural to consider the reward (*i.e.*, negative cost) associated with a state-action pair to be the number of points received as a consequence of the action, and to consider as the objective maximization of the expected sum of rewards over the course of a game. However, we found that, with this formulation of reward, our approach (Section 6.5) did not yield

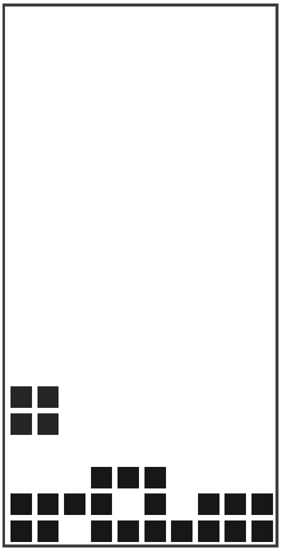


Figure 6.1. A representative Tetris board configuration

reasonable policies. We found that a different cost function, together with discounting, lead to effective policies. In particular, we set the cost $g(x_t, a_t)$ to be the height of the current Tetris wall, and let the objective be to minimize the expected sum of discounted future costs (6.1), with a discount factor $\alpha = 0.9$. Further, we set the cost of a transition to a termination state at $\frac{20}{1-\alpha}$ which is a trivial upper bound on the cost-to-go for a state under any policy. With this formulation, an optimal policy maximizes the number of rows cleared prior to termination with a greater emphasis on the immediate future, due to discounting.

Several interesting observations have been documented in the literature on Tetris. It was shown in [67] that the game terminates with probability one, under any policy. In terms of complexity, it is proven in [117] that for an off-line version of Tetris, where the player is offered knowledge of the shapes of the next K pieces to appear, optimizing various simple objectives is NP-complete, even to approximate. Though there is no formal connection between such results and the on-line model we consider, the results suggest that finding an optimal policy for on-line Tetris might also be difficult.

6.3 Dynamic Programming

For each policy π, define a cost-to-go function,

$$J_\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t g(x_t, a_t) \middle| x_0 = x, a_t = \pi(x_t)\right].$$

Given the optimal cost-to-go function

$$J^*(x) = \min_\pi J_\pi(x),$$

an optimal policy can be generated according to

$$\pi(x) \in \operatorname{argmin}_{a \in \mathcal{A}_x} \left(g(x,a) + \alpha \sum_{y \in \mathcal{S}} p_a(x,y) J^*(y) \right).$$

Define the dynamic programming operator $T : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$:

$$(TJ)(x) = \min_{a \in \mathcal{A}_x} \left(g(x,a) + \alpha \sum_{y \in \mathcal{S}} p_a(x,y) J(y) \right).$$

It is well-known that the optimal cost-to-go function J^* is the unique solution to Bellman's equation: $TJ = J$. Dynamic programming offers a suite of algorithms for solving this equation. One example involves a linear program:

$$\begin{array}{ll} \max_J & c^T J \\ \text{s.t.} & TJ \geq J \end{array}$$

Note that, as written above, the constraints are nonlinear. However, they can be converted to linear constraints since each constraint $(TJ)(x) \geq J(x)$ is equivalent to a set of linear constraints:

$$g(x,a) + \alpha \sum_{y \in \mathcal{S}} p_a(x,y) J(y) \geq J(x) \quad \forall a \in \mathcal{A}_x.$$

It is well-known that for any $|\mathcal{S}|$-dimensional vector $c > 0$, J^* is the unique optimal solution to this linear program (see, *e.g.*, [41]).

In principle, stochastic control problems like Tetris can be solved by dynamic programming algorithms. However, the computational requirements are prohibitive. For example, the above linear program involves one variable per state and one constraint per state-action pair. Clearly, Tetris presents far too many states ($\sim 2^{1400}$!) for such a solution method to be viable. One must therefore resort to approximations.

6.4 Approximate Dynamic Programming

In order to deal with an intractable state space, one might consider approximating the optimal cost-to-go function J^* by fitting a parameterized function approximator, in a spirit similar to statistical regression. A number of methods for doing this are surveyed in [44]. We will consider here cases where the approximator depends linearly on the parameters. Such an approximator can be thought of as a linear combination of pre-selected basis functions $\phi_1, \ldots, \phi_K : \mathcal{S} \mapsto \mathbb{R}$, taking the form $\sum_{k=1}^{K} r_k \phi_k$, where the parameters are weights $r_1, \ldots, r_K \in \mathbb{R}$. Generating such an approximation involves two steps:

1. Selecting basis functions $\phi_1, \ldots, \phi_K$.
2. Computing weights $r_1, \ldots, r_K$ so that $\sum_{k=1}^{K} r_k \phi_k \approx J^*$.

In our study of Tetris we will select basis functions based on problem specific intuition and compute weights by solving a linear program that with a reasonably small number of parameters but an intractable number of constraints. In this section, we discuss this linear program approach and the use of randomized constraint sampling in this context.

6.4.1 A Linear Program for Computing Basis Function Weights

It is useful to define a matrix $\Phi \in \mathbb{R}^{|\mathcal{S}| \times K}$ by

$$\Phi = \begin{bmatrix} | & & | \\ \phi_1 & \cdots & \phi_K \\ | & & | \end{bmatrix}$$

The linear program presented in Section 6.3, which computes J^*, motivates another linear program for computing weights $r \in \mathbb{R}^K$:

$$\begin{array}{ll} \max_r & c^T \Phi r \\ \text{s.t.} & T\Phi r \geq \Phi r. \end{array}$$

To distinguish the two, we will call this linear program the ALP (approximate linear program) and the one from Section 6.3 the ELP (exact linear program). Note that, while the ELP involves one variable per state, the ALP only has one variable per basis function. However, the ALP has as many constraints as the ELP. We will later discuss the use of constraint sampling to deal with this. For now, we will discuss results from [98] that support the ALP as a reasonable method for approximating the optimal cost-to-go function.

Let $\tilde{r}$ be an optimal solution to the ALP, and $\|J\|_{1,c} = \sum_x c(x) J(x)$ denote a weighted ℓ_1-norm. One result from [98] asserts that $\tilde{r}$ attains the minimum of $\|J^* - \Phi r\|_{1,c}$ within the feasible region of the ALP.

Lemma 1. *A vector r solves*

$$\begin{array}{ll} \max_r & c^T \Phi r \\ \text{s.t.} & T\Phi r \geq \Phi r \end{array}$$

if and only if it solves

$$\begin{array}{ll} \min_r & \|J^* - \Phi r\|_{1,c} \\ \text{s.t.} & T\Phi r \geq \Phi r. \end{array}$$

Recall that J^* is the optimal solution to the ELP for any $c > 0$. In the case of the ALP, however, the choice of c determines relative emphasis among states, with states corresponding to higher values of c likely to benefit from smaller approximation errors.

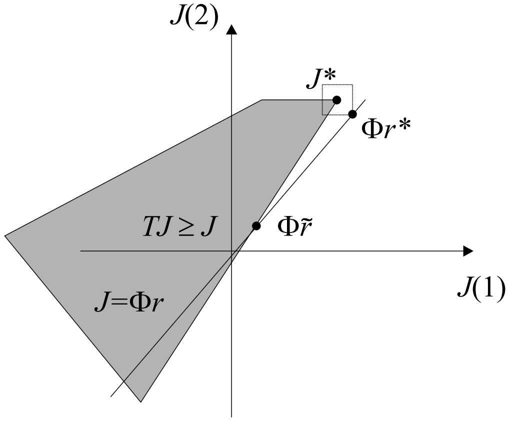


Figure 6.2. Graphical interpretation of the ALP

It is easy to see that if J^* is in the range of Φ then $\Phi\tilde{r} = J^*$. One might hope that if J^* is close to the range of Φ then the $\Phi\tilde{r}$ will be close to J^*. This is not promised by the above result, because of the restriction to the feasible region of the ALP. In particular, as illustrated in Figure 6.2, one might imagine $\Phi\tilde{r}$ being close to or far from J^* even though there is some (infeasible) r^* for which $\Phi r^* \approx J^*$. The following theorem (Theorem 4.1 from [98]) offers some assurance through a bound on how far $\Phi\tilde{r}$ can be from J^* in terms of the proximity of J^* to the range of Φ. The result requires that e, the vector with every component equal to 1, is in the range of Φ.

Theorem 1. *Let e be in the range of Φ and let c be a probability distribution. Then, if $\tilde{r}$ is an optimal solution to the approximate LP,*

$$\|J^* - \Phi\tilde{r}\|_{1,c} \leq \frac{2}{1-\alpha} \min_r \|J^* - \Phi r^*\|_\infty .$$

As discussed in [98], this bound is rather loose. In particular, for large state-spaces, $\|J^* - \Phi r^*\|_\infty$ is likely to be very large. Further, the bound does not capture the fact, that the choice of c has a significant impact on the error $\|J^* - \Phi\tilde{r}\|_{1,c}$. More sophisticated results in [98] address these issues by refining the above bound. To keep the discussion simple, we will not present those results here.

After computing a weight vector $\tilde{r}$, one might generate decisions according to a policy

$$\tilde{\pi}(x) = \mathrm{argmax}_{a \in \mathcal{A}_x} \left(g(x,a) + \alpha \sum_{y \in \mathcal{S}} p_a(x,y)(\varPhi \tilde{r})(y) \right).$$

Should this policy be expected to perform reasonably? This question is addressed by another result, adapted from Theorem 3.1 in [98].

Theorem 2. *Let J be such that $TJ \geq J$. Then*

$$\nu^T (J_{\tilde{\pi}} - J^*) \leq \frac{1}{1-\alpha} \|J - J^*\|_{1,c},$$

where $\nu(y) = \frac{1}{1-\alpha}(c(y) - \alpha \sum_x c(x) p_{\pi(x)}(x,y))$.

For each state x, the difference $J_{\tilde{\pi}}(x) - J^*(x)$ is the excess cost associated with suboptimal policy $\tilde{\pi}$ if the system starts in state x. It is easy to see that ν sums to 1. However, $\nu^T(J_{\tilde{\pi}} - J^*)$ is not necessarily a weighted average of excess costs associated with different states. Depending on the choice of c, individual elements of ν may be positive or negative. As such, the choice of c influences performance in subtle ways. [98] motivates choosing c to reflect the relative frequencies with which states are visited by good policies.

6.4.2 Randomized Constraint Sampling

If there are a reasonably small number of basis functions, the ALP involves a manageable number of variables but an intractable number of constraints. To deal with these constraints, we will use randomized constraint sampling, as proposed in [99]. In particular, consider the following relaxed linear program (RLP):

$$\begin{array}{ll} \max & c^T \varPhi r \\ \text{s.t.} & g_a(x) + \alpha \sum_{y \in \mathcal{S}} P_a(x,y)(\varPhi r)(y) \geq (\varPhi r)(x), \quad \forall (x,a) \in \mathcal{X}, \end{array}$$

where $\mathcal{X}$ is a set of N constraints each sampled independently according to a distribution ψ.

The use of constraint sampling is motivated to some extent by the following result from [98]. (An important generalization that applies to convex programs has been established in [71], see also the improved result in [70] and in Chapter 5 of this book.)

Theorem 3. *Consider a linear program with K variables and any number of constraints. Let ψ be a probability distribution over the constraints and let $\mathcal{X}$ be a set of N constraints sampled independently according to ψ, with $N \geq \frac{4}{\epsilon}\left(K \ln \frac{12}{\epsilon} + \ln \frac{2}{\delta}\right)$, $\epsilon \in (0,1)$ and $\delta \in (0,1)$. Let $r \in \mathbb{R}^K$ be an optimal solution to the linear program with all constraints relaxed except for those in $\mathcal{X}$, and let $\mathcal{V}$ be the set of constraints violated by r. Then, $\psi(\mathcal{V}) \leq \epsilon$ with probability at least $1 - \delta$.*

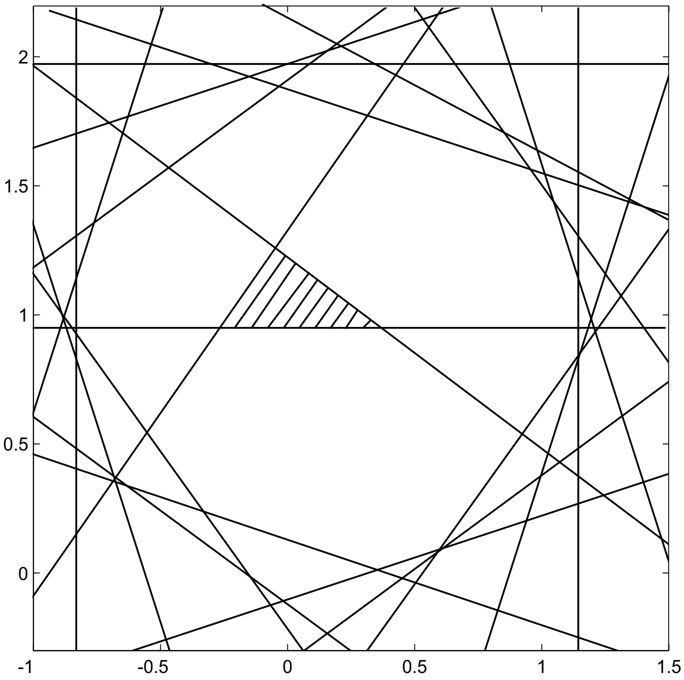


Figure 6.3. Graphical interpretation of the ALP

In spite of the above result, it is not clear whether the RLP will yield solutions close to those of the ALP. In particular, it might be the case that a few constraints affect the solution dramatically as Figure 6.3 amply illustrates. Fortunately, the structure of the ALP precludes such undesirable behavior, and we have the following result, which is adapted from [99].

Theorem 4. *Let ϵ and δ be scalars in $(0,1)$. let π^* be an optimal policy and $\mathcal{X}$ be a random set of N state-action pairs sampled independently according to the distribution*

$$\psi^*_\alpha(x) = (1-\alpha)\mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t \mathbf{1}\{x_t = x\}\middle| x_0 \sim c, a_t = \pi^*(x_t)\right].$$

Let $\hat{r}$ be a solution to the RLP. If

$$N \geq \frac{16\|J^* - \Phi\hat{r}\|_\infty}{(1-\alpha)\epsilon c^T J^*}\left(K \ln \frac{48\|J^* - \Phi\hat{r}\|_\infty}{(1-\alpha)\epsilon c^T J^*} + \ln \frac{2}{\delta}\right),$$

then, with probability at least $1-\delta$, we have

$$\|J^* - \Phi\hat{r}\|_{1,c} \leq \|J^* - \Phi\tilde{r}\|_{1,c} + \epsilon\|J^*\|_{1,c}$$

In the proof of this error bound, sampling according to ψ_{α}^{*} ensures that with high probability, $\|J^{*}-\Phi\hat{r}\|_{1,c} \approx \|J^{*}-\Phi\tilde{r}\|_{1,c}+$ a term that can be made arbitrarily small with N large. As such this is a weakness; sampling according to ψ_{α}^{*} requires knowledge of the optimal policy. Nevertheless, one might hope that for a distribution sufficiently 'close' to ψ_{α}^{*}, the bound of Theorem 4 still holds for a reasonable value of N. In any case, Theorem 4 offers some hope that the RLP is a tractable means for finding a meaningful approximation to J^{*}.

6.5 Synthesis of a Tetris Strategy

We've already seen that playing Tetris optimally is an example of a stochastic control problem with an intractable state-space. As a first step to coming up with a near-optimal controller for Tetris we select a linear approximation architecture for the tetris cost-to-go function. In particular, we will attempt to approximate the cost-to-go for a state using a linear combination of the following $K = 22$ basis functions:

- Ten basis functions, $\phi_0, \ldots, \phi_9$, mapping the state to the height h_k of each of the ten columns.
- Nine basis functions, $\phi_{10} \ldots \phi_{18}$, each mapping the state to the absolute difference between heights of successive columns: $|h_{k+1} - h_k|,\ k = 1, \ldots, 9$.
- One basis function, ϕ_{19} that maps state to the maximum column height: $\max_k h_k$.
- One basis function, ϕ_{20} that maps state to the number of 'holes' in the wall.
- One basis function, ϕ_{21} that is equal to one at every state.

Such an approximation architecture has been used with some success in [43, 174]. For example, in [43], the authors used an approximate dynamic programming technique – approximate policy iteration – to generate policies that averaged 3183 points a games which is comparable to an expert human player. The controller presented surpasses that performance.

In the spirit of the program presented in Section 6.4.2, we formulate the following RLP:

$$\begin{array}{lll} \max & \sum_{x \in \overline{\mathcal{X}}} (\Phi r)(x) & \\ \text{s.t.} & (T\Phi r)(x) \geq (\Phi r)(x), & \forall x \in \overline{\mathcal{X}}. \end{array}$$

where $\overline{\mathcal{X}}$ is a sample of states. Observe that in the above RLP, the sampling distribution takes on the role of c.

In our most basic set-up, we make use of a heuristic policy generated by guessing and adjusting weights for the basis functions until reasonable performance is achieved. The intent is to generate nearly i.i.d. samples of states, distributed according to the relative frequencies with which states are visited by the heuristic policy. To this end, some number N of games are played using the heuristic policy, and for some choice of M, states visited at times that are multiples of M are incorporated in the set $\overline{\mathcal{X}}$. Note that time, here, is measured in terms of the number of time-steps from the start of the first of the N games, rather than from the start of a current game. The reason for selecting states that are observed some M time-steps apart is to obtain samples that are near-independent. When consecutively visited states are incorporated in $\overline{\mathcal{X}}$, samples exhibit a high degree of statistical dependence. Consequently, a far greater total number of samples $|\overline{\mathcal{X}}|$ is required for the RLP to generate good policies. This is problematic, as computer memory limitations become an obstacle in solving linear programs with large numbers of constraints.

Now recall that in light of the results of Sections 6.4.1 and 6.4.2, we would like c to mimic the state distribution induced by the optimal policy as closely as possible. Thus, in addition to the basic set-up we have described above, we have also experimented with a bootstrapped version of the RLP. To understand the motivation for bootstrapping, suppose that the policy generated as an outcome of the RLP is superior to the initial heuristic used to sample states. Then, it is natural to consider producing a new sample of states based on the improved policy and solving the RLP again with this new sample. But why stop there? This procedure might be iterated to repeatedly amplify performance. This idea leads to our bootstrapping algorithm:

1. Begin with a simulator that uses a policy u_0.
2. Generate a sample $\overline{\mathcal{X}}_k$ of states using policy u_k.
3. Solve the RLP based on the sample $\overline{\mathcal{X}}_k$, to generate a policy u_{k+1}.
4. Increment k and go to step 2.

Other variants to this may include a more guarded update of the state-sampling distribution, wherein the sampling distribution used in a given iteration is the average of the distribution induced by the latest policy and the sampling distribution used in the previous iteration. That is, in Step 2 we might randomize between using samples generated by the current policy u_k, and the samples used in the generation of the previous collection, $\overline{\mathcal{X}}_{k-1}$.

In the next section, we discuss results generated by the RLP and bootstrapping.

6.5.1 Numerical Results

Our numerical experiments may be summarized as follows. All RLPs were limited to two million constraints, this figure being determined by available physical memory. Initial experiments with the simple implementation helped determine a suitable sampling interval, M. All subsequent RLPs were generated with a sampling interval of $M = 90$.

For a fixed simulator policy, five RLPs were generated, of which the best was picked for the next bootstrap iteration.

Table 6.1. Comparison with other algorithms

Algorithm	Performance	Computation time
TD-Learning	3183	Hours
Policy Gradient	5500	?
LP w/ Bootstrap	4274	hours

Figure 6.4 summarizes the performance of policies obtained from our experiments with the bootstrapping methodology. The 'median' figures are illustrative of the variance in the quality of policies obtained at a given iteration, while the 'best policy' figures correspond to the best performing policy at that iteration. Table 6.1 compares the performance of the best policy obtained in this process to that of other approaches used in the past [43, 174].

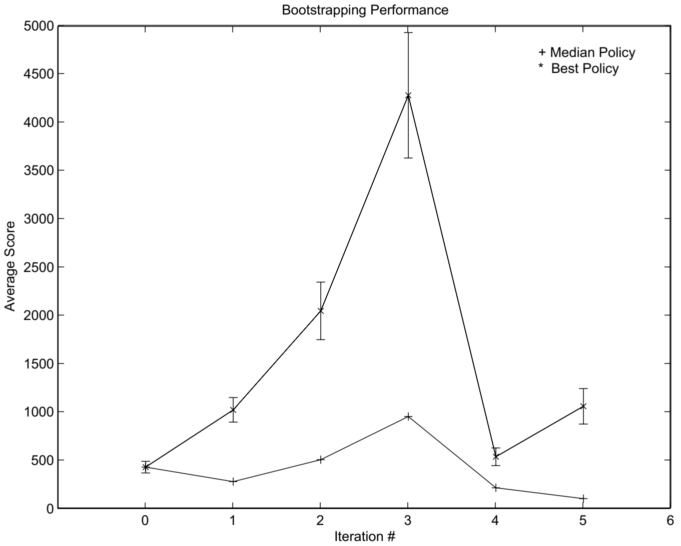


Figure 6.4. Bootstrapping performance

We now make some comments on the computation time required for our experiments. As mentioned previously, every RLP in our experiments had two million constraints. For general LPs this is a very large problem size. However, the RLP has special structure in that it has a small number of variables (22 in our case). We take advantage of this structure by solving the dual of the RLP. The dual has number of constraints equal to the number of basis functions (22 in our case) and so is effectively solved using a barrier method whose complexity is dominated by the number of constraints [60]. Using this, we are able to solve an RLP in minutes. Hence, the computation time is dominated by the time taken in generating state samples, which in our case translates to several hours for each RLP. These comments apply, of course, to RLPs for general large scale problems since the number of basis functions is typically several orders of magnitude smaller than the number of sampled states. We have found that solving smaller RLPs leads to larger variance in policy quality, and lower median policy performance.

Finally, one might expect successive iterations of the bootstrapping methodology to yield continually improving policies. However, in our experiments, we have observed that beyond three to four iterations, median as well as best policy performance degrade severely. Use of a more guarded update to the state sampling distribution as described in Section 4, does not seem to alleviate this problem. We are unable to explain this behavior.

6.6 Concluding Remarks

We have presented what we believe is a successful application of an exciting new technique for approximate dynamic programming that uses a 'constraint sampling' technique in a central way. Our experiments accentuate the importance of the question asked in the introduction, namely, what is the effect of the (small) number of violated constraints? The approximate LP of Section 6.4.1 provided an interesting setting in which we attempted to answer this question, and we concluded that the answer (in the case of approximate dynamic programming via the LP method) was intimately related to the sampling distribution used for sampling constraints. This connection was strongly borne out in our Tetris experiments; naive sampling distributions led to relatively poor policies.

As such, theorems in the spirit of Theorem 3, while highly interesting represent only a first step in the design of an effective constraint sampling scheme; they need to be complemented by results and schemes along the lines of those in Section 6.4.2 that assure us that the violated constraints cannot hurt us much. In the case of approximate dynamic programming, strengthening those results and developing an understanding of schemes that sample constraints effectively is an immensely interesting direction for future research.

Acknowledgement. The material in this chapter is based upon work supported by the National Science Foundation under Grant ECS-9985229 and by the Office of Naval Research under Grant MURI N00014-00-1-0637.

7

Near Optimal Solutions to Least-Squares Problems with Stochastic Uncertainty

Giuseppe Calafiore[1] and Fabrizio Dabbene[2]

[1] Dipartimento di Automatica e Informatica – Politecnico di Torino
C.so Duca degli Abruzzi, 10124, Italy
giuseppe.calafiore@polito.it

[2] IEEIT-CNR – Politecnico di Torino
C.so Duca degli Abruzzi, 10124, Italy
fabrizio.dabbene@polito.it

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

7.1 Introduction

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

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

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

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

$$f(x,\delta) \doteq \|A(\delta)x - y(\delta)\|^2, \tag{7.1}$$

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

$$x^*_{wc} = \arg\min_x \max_{\delta \in \Delta} f(x,\delta). \tag{7.2}$$

This worst-case framework is discussed for instance in the papers [79,121,319], and is closely related to Tikhonov-type regularization [361].

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

$$x^*_E = \arg\min_x \mathbb{E}_\delta[f(x,\delta)]. \tag{7.3}$$

We refer to problem (7.3) as the least-squares with stochastic uncertainty (LSSU) problem. Unfortunately, both problems (7.2) and (7.3) are numerically hard to solve. In [121] it is shown that the deterministic problem (7.2) is in general NP-hard. When the uncertainty enters the data in a rational manner, it is possible to compute a suboptimal solution that minimizes an upper bound on the optimal worst-case residual, using semi-definite relaxations, see [121]. In [79, 319] a solution with lower computational complexity is derived for the case of unstructured uncertainty entering in a simple additive form in the data A, y. However, no exact efficient method is known for the general structured nonlinear case. Similarly, in the stochastic problem (7.3), even the mere evaluation of the objective function, for fixed x, can be numerically prohibitive, since it amounts to the computation of a multi-dimensional integral.

In this chapter, we focus on the solution to the LSSU problem (7.3). Indeed, this problem falls in the general family of stochastic optimization programs, see for instance the survey [393]. Since, in general, one cannot compute exact expectations, a usual initial step in stochastic optimization is to use random sampling to construct an approximation of the original objective, and then compute a candidate solution with respect to this approximation. Known methods for stochastic programming then provide convergence results and confidence intervals for the optimal solution [159, 186, 211, 330]. A drawback of these results is that they are of asymptotic nature and do not provide explicit bounds on the *number of samples* (which impacts on the number of iterations) needed to reach a satisfactory solution.

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

The chapter is organized as follows. In Section 7.1.1 the notation is set and the main assumptions used throughout the chapter are stated. To illustrate the LSSU framework, in Section 7.2 we discuss a particular case of this problem where the expected value can be explicitly computed, and observe that the LSSU problem reduces to regularized deterministic LS, which can be solved via standard methods. The general case, when the expectation cannot be computed explicitly, is discussed in Section 7.3. In this section, we present the Learning Theory approach to stochastic optimization, and state the main result of this contribution in Theorem 2. Section 7.3.1 discusses a simple technique for numerical computation of the approximate solution. Section 7.4 discusses an alternative approach to confidence level solutions for LSSU, based on the stochastic gradient methods, recently proposed in [242]. Section 7.5 presents some numerical examples and comparisons. Conclusions are drawn in Section 7.6.

7.1.1 Notation and Assumptions

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

$$\mathbb{E}_\delta[g(\delta)] = \int_{\delta \in \varDelta} g(\delta) p_\delta(\delta) \mathrm{d}\delta.$$

Given N independent identically distributed (i.i.d.) samples $\delta^{(1)}, \ldots, \delta^{(N)}$ drawn according to $p_\delta(\delta)$, the *empirical expectation* operator on $g(\delta)$ is defined as

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

Consider the function

$$\phi(x) \doteq \mathbb{E}_\delta[f(x, \delta)], \tag{7.4}$$

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

$$x^* \doteq \arg \min_{x \in \mathbb{R}^n} \phi(x). \tag{7.5}$$

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

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

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

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

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

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

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

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

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

7.2 Closed-Form Solutions for Affine Uncertainty

In this section, to illustrate the framework of least-squares with stochastic uncertainty, we consider the special case when the uncertain parameter δ enters the data affinely. It can be easily shown that in this situation the expected value of the least-squares residual can be computed in closed-form. Therefore, the LSSU problem can be recast as a standard regularized LS problem. The case of generic nonlinear dependence of the data on the uncertain parameters, which is the key focus of this chapter, is then treated in Section 7.3.

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

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

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

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

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

$$\phi(x) = \mathbb{E}_\delta[f(x,\delta)] = \|A_0 x - y\|^2 + x^T Q x, \tag{7.6}$$

where

$$Q \doteq \sum_{i=1}^{\ell} \sigma_i^2 A_i^T A_i. \tag{7.7}$$

The objective function in (7.6) has the form of a regularized LS objective, and a minimizing solution (which always exists) can be easily computed in closed-form as detailed in the following theorem.

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

$$\phi(x) = \mathbb{E}_\delta[\|A(\delta)x - y\|^2]$$

are the solutions of the modified normal equations

$$(A_0^T A_0 + Q)x = A_0^T y, \tag{7.8}$$

where $Q \succeq 0$ is given in (7.7). A minimizing solution always exists. In particular, when $A_0^T A_0 + Q \succ 0$ the solution is uniquely given by

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

Proof. Differentiating the convex quadratic objective (7.6) with respect to x, the first order optimality conditions yield immediately (7.8). The only thing that needs to be proved is that these linear equations always admit a solution. Clearly, (7.8) has a solution if and only if $A_0^T y \in \mathcal{R}(A_0^T A_0 + Q)$, which is implied by $\mathcal{R}(A_0^T) \subseteq \mathcal{R}(A_0^T A_0 + Q)$. Now, since $\mathcal{R}(A_0^T) = \mathcal{R}(A_0^T A_0)$ (see for instance [173], Chapter 2), solvability of (7.8) is implied by the condition $\mathcal{R}(A_0^T A_0) \subseteq \mathcal{R}(A_0^T A_0 + Q)$. In turn, this latter condition is equivalent to

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

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

$$x^T (A_0^T A_0 + Q)x = x^T A_0^T A_0 x + x^T Q x = 0.$$

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

We remark that this result is quite standard, and can be easily extended to the case when the independence assumption on the δ_i's is removed, and the term y is considered uncertain too, see for instance [160]. However, in the case of generic nonlinear functional dependence of A, y on the uncertainty δ, and for generic density $p_\delta(\delta)$, the expectation of the residual cannot be computed in an efficient numerical way (nor in closed-form, in general). This motivates the developments of the next section.

7.3 Learning Theory Approach to Expected Value Minimization

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

A fundamental question at this point is whether the minimum of the empirical expectation converges in some suitable sense to the minimum of the true unknown expectation. Several *asymptotic* results of convergence are available in the stochastic optimization literature, see for instance [186, 330, 331]. Here, however, we depart from these usual approaches, typically based on central limit arguments, and use the Learning Theory framework [375] to provide both asymptotic *and* finite sample convergence results. This approach relies on the law of uniform convergence of empirical means to their expectations. These results are summarized below.

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

$$\hat{\phi}(x) \doteq \hat{E}_N[f(x,\delta)]. \tag{7.9}$$

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

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

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

$$\mathcal{F} \doteq \{f(x,\delta) : x \in \mathcal{X}\}, \quad f(x,\delta) = \|A(\delta)x - y(\delta)\|^2. \tag{7.10}$$

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

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

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

$$\mathbb{P}\sup_{x \in \mathcal{X}} \frac{|\phi(x) - \hat{\phi}(x)|}{V} > \epsilon \leq \alpha(N), \tag{7.11}$$

with $\alpha(N) \to 0$ as $N \to \infty$. Notice that the uniformity of bound (7.11) with respect to x is crucial, since x is *not* fixed and known in advance: the uniform 'closeness' of $\hat{\phi}(x)$ to $\phi(x)$ is the feature that allows us to perform the minimization on $\hat{\phi}(x)$ instead of on $\phi(x)$. Property (7.11) is usually referred to as the Uniform Convergence of the Empirical Mean (UCEM) property. A fundamental result of Learning Theory states that the UCEM property holds for a function class $\mathcal{F}$ whenever a particular measure of the complexity of the class, called the P-dimension of $\mathcal{F}$ (P-dim($\mathcal{F}$)), is finite. Moreover, this property holds independently of the probability distribution of the data. The interested reader can refer to the monographs [359, 375, 380] for formal definitions and further details.

The next lemma shows that the function class (7.10) under consideration has indeed finite P-dimension, and explicitly provides an upper bound on P-dim($\mathcal{F}$).

Lemma 1 (P-dimension of $\mathcal{F}$). *Consider the function family $\mathcal{F}$ defined in (7.10). Then*

$$\text{P-dim}(\mathcal{F}) \leq 9n.$$

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

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

for $c \in [0, M]$. Then, from Lemma 10.1 in [380], we have that P-dim($\mathcal{F}$) = VC-dim($\bar{\mathcal{F}}$), where VC-dim($\bar{\mathcal{F}}$) denotes the Vapnik-Chervonenkis dimension of the class $\bar{\mathcal{F}}$. Notice that the functions in $\bar{\mathcal{F}}$ are quadratic in the parameter vector $x \in \mathbb{R}^n$, therefore a bound on the VC-dimension can be derived from a result of Karpinski and Macintyre [181]:

$$\text{VC-dim}(\bar{\mathcal{F}}) \leq 2n \log_2(8e) < 9n.$$

□

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

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

$$N \geq \frac{128}{\epsilon^2} \left[\ln \frac{8}{\alpha} + 9n \left(\ln \frac{32\mathrm{e}}{\epsilon} + \ln \ln \frac{32\mathrm{e}}{\epsilon} \right) \right]. \tag{7.12}$$

Let x^* be a minimizer of $\phi(x)$ defined in (7.5), and let $\hat{x}_N$ be a minimizer of the empirical mean $\hat{\phi}(x)$. Then, if $\hat{x}_N \in \mathcal{X}$, it holds with probability at least $(1 - \alpha)$ that

$$\frac{\phi(\hat{x}_N) - \phi(x^*)}{V} \leq \epsilon,$$

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

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

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

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

$$\phi_g(x) \doteq \mathbb{E}_\delta[g(x,\delta)] = \frac{\phi(x) - K}{V},$$

and

$$\hat{\phi}_g(x) \doteq \hat{\mathbb{E}}_N[g(x,\delta)] = \frac{1}{N}\sum_{i=1}^{N} g(x,\delta^{(i)}) = \frac{\hat{\phi}(x) - \hat{K}}{V}, \tag{7.13}$$

where

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

Notice that a minimizer $\hat{x}$ of $\hat{\phi}(x)$ is also a minimizer of $\hat{\phi}_g(x)$. Then, Theorem 2 in [381] guarantees that, for $\alpha,\nu \in (0,1)$,

$$\mathbb{P}\sup_{g\in\mathcal{G}}\left|\mathbb{E}_\delta[g(\delta)] - \hat{\mathbb{E}}_N[g(\delta)]\right| > \nu \leq \alpha,$$

holds irrespective of the underlying distribution of δ, provided that

$$N \geq \frac{32}{\nu^2}\left[\ln\frac{8}{\alpha} + \text{P-dim}(\mathcal{G})\left(\ln\frac{16\text{ e}}{\nu} + \ln\ln\frac{16\text{ e}}{\nu}\right)\right].$$

Applying this theorem with $\nu = \epsilon/2$, and using the bound P-dim$(\mathcal{G})$ = P-dim$(\mathcal{F}) \leq \exists\backslash$ obtained in Lemma 1, we have that, for all $x \in \mathcal{X}$, it holds with probability at least $(1-\alpha)$ that

$$|\phi_g(x) - \hat{\phi}_g(x)| \leq \frac{\epsilon}{2}. \tag{7.14}$$

From (7.14), evaluated in $x = x^*$ it follows that

$$\phi_g(x^*) \geq \hat{\phi}_g(x^*) - \frac{\epsilon}{2} \geq \hat{\phi}_g(\hat{x}_N) - \frac{\epsilon}{2}, \tag{7.15}$$

where the last inequality follows since $\hat{x}_N$ is a minimizer of $\hat{\phi}_g$. From (7.14), evaluated in $x = \hat{x}_N$ it follows that

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

which substituted in (7.15), gives

$$\phi_g(x^*) \geq \phi_g(\hat{x}_N) - \epsilon.$$

From the last inequality and (7.13) it follows that

$$\phi(\hat{x}_N) - \phi(x^*) \le \epsilon V,$$

which concludes the proof. □

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

7.3.1 Numerical Computation of $\hat{x}_N$

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

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

$$\hat{\phi}(x) = \frac{1}{N}\sum_{i=1}^{N} \|A(\delta^{(i)})x - y(\delta^{(i)})\|^2 = \frac{1}{N}\|\mathcal{A}x - \mathcal{Y}\|^2$$

where

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

Therefore, an exact minimizer of $\hat{\phi}(x)$ can be readily computed as $\hat{x}_N = \mathcal{A}^{\dagger}\mathcal{Y}$, where $\mathcal{A}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathcal{A}$. Remark that, since $\mathcal{A}, \mathcal{Y}$ are functions of $\delta^{(i)}$, $i = 1, \ldots, N$, the resulting solution $\hat{x}_N$ is a random quantity, whose probability distribution is defined over the product space $\varDelta \times \varDelta \times \cdots \times \varDelta$ (N times). The solution $\hat{x}_N$ can be alternatively defined as the result given at the N-th iteration by the following standard recursive form of the LS algorithm, see, *e.g.*, [173].

Algorithm 7.1 *Assuming that $A(\delta^{(1)})$ is full-rank, an exact minimizer $\hat{x}_N$ of the empirical mean (7.9) can be recursively computed as*

$$\hat{x}_{k+1} = \hat{x}_k + K_{k+1}^{-1} A^T(\delta^{(k+1)}) \left(y(\delta^{(k+1)}) - A(\delta^{(k+1)})\hat{x}_k \right), \tag{7.16}$$

where

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

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

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

1. Given the *a-priori* set $\mathcal{X}$, fix the desired probabilistic levels α, ϵ, and determine the theoretical bound for N given in (7.12);
2. Compute $\hat{x}_N$. This computation needs random samples $\delta^{(i)}$, $i = 1, \ldots, N$ extracted according to $p_\delta(\delta)$ (see further comments on this point in Remark 2);
3. If $\hat{x}_N \in \mathcal{X}$, then with probability greater than $(1 - \alpha)$ this solution is an ϵ-suboptimal minimizer for $\phi(x)$, in the relative scale V.

Remark 2. For the implementation of the proposed method, two typical situations are possible. In a first situation, we explicitly know the uncertainties distribution $p_\delta(\delta)$ and the functional dependence $A(\delta), y(\delta)$. In this case one can generate the appropriate random samples $\delta^{(i)}$, $i = 1, \ldots, N$, using standard techniques for random sample generation (see for instance [359]). The probabilistic assessments in Theorem 2 are in this case explicitly referred to the probability measure p_δ. In other practical situations, the uncertainty δ is embedded in the data, and the corrupted data $A(\delta^{(i)}), y(\delta^{(i)})$ are directly available as observations. In this respect, we notice that the results in Theorem 2 hold irrespective of the underlying probability distribution, and hence they can be applied also in the cases where the measure p_δ exists but is unknown. In this case, $\hat{x}_N$ is computed using directly the corrupted data $A(\delta^{(i)}), y(\delta^{(i)})$ relative to the i-th experiment, for $i = 1, \ldots, N$, and the results of Theorem 2 hold with respect to the unknown probability measure p_δ.

Remark 3. Notice that, if the iterative Algorithm 7.1 is used for the computation of $\hat{x}_N$ then, in some particular instances of the problem, one may observe practical convergence in a number of iterations much smaller than N. This is in the nature of the results based on the Vapnik-Chervonenkis theory of learning, which provides theoretical bounds that hold *a-priori*, for any problem instance, and for all possible probability distributions of the uncertainties. Therefore, bound (7.12) holds always and *a-priori* (before even starting the estimation experiment), while practical convergence can only be assessed *a-posteriori*, on a specific instance of the problem. This issue is further discussed in the numerical examples section.

In the next section, we discuss an alternative approach to an approximate solution of the LSSU problem, based on stochastic gradient (SG) algorithms for stochastic optimization [241, 242]. In this latter approach, a candidate solution $\hat{x}$ is computed, with the property that its associated cost is a good approximation of the optimal value of the original problem, with high probability. The learning theory approach described previously basically works in two steps: a first step where the empirical mean $\hat{\phi}(x)$ is estimated, and a successive step where a minimizer for it is computed. In contrast, the SG method bypasses the empirical mean estimation step, and directly searches for a near optimal solution iteratively, following random gradient descent steps. The purpose of the next developments is to specialize the SG approach to the problem under study, and then use these results for comparison with those given in Theorem 2.

7.4 Stochastic Gradient Approach

The gradient of the function $f(x,\delta)$ defined in (7.1) is given by

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

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

Algorithm 7.2 *Let $N > 0$ be an a-priori fixed number of steps, and let λ_k, $k = 0,\ldots,N-1$ be a finite sequence of stepsizes, such that*

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

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

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

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

$$\begin{aligned} x_{k+1} &= [x_k - \lambda_k h(x_k, \delta^{(k)})]_{\mathcal{X}} \\ \hat{x}_k &= \frac{m_{k-1}}{m_k}\hat{x}_{k-1} + \frac{\lambda_k}{m_k}x_k, \\ m_k &= m_{k-1} + \lambda_k, \end{aligned} \tag{7.17}$$

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

From a classical result on stochastic optimization of Nemirowskii and Yudin [241], we have that for the solution computed by Algorithm 7.2 it holds that

$$E[\phi(\hat{x}_N)] - \phi^* \le \frac{R^2 + L^2 \sum_{k=0}^{N-1} \lambda_k^2}{2\sum_{k=0}^{N-1} \lambda_k}. \tag{7.18}$$

In particular, if we choose constant stepsizes $\lambda_k = \lambda = \frac{\gamma}{\sqrt{N}}$, then the right hand side of (7.18) becomes $\frac{R^2 + L^2\gamma^2}{2\gamma\sqrt{N}}$, which goes to zero as $O(1/\sqrt{N})$, for $N \to \infty$. If the constants R, L are known, then the optimal choice for γ is $\gamma = R/L$.

The following result, adapted from [242], gives a precise assessment of the quality of the solution obtained using the above algorithm, in terms of probabilistic levels.

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

$$N \geq \frac{1}{\alpha^2 \epsilon^2} \left(\frac{LR}{V} \right)^2 . \tag{7.19}$$

Let x^* be a minimizer of $\phi(x)$ defined in (7.4), and let $\hat{x}_N$ be the outcome of Algorithm 7.2, with stepsizes $\lambda_k = \lambda = \frac{R}{L\sqrt{N}}$. Then, it holds with probability at least $(1-\alpha)$ that

$$\frac{\phi(\hat{x}_N) - \phi(x^*)}{V} \leq \epsilon,$$

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

Notice that the update step (7.16) of Algorithm 7.1 and (7.17) of Algorithm 7.2 have a similar form. In particular, the recursive least-squares algorithm (Algorithm 7.1) can be interpreted as a stochastic gradient algorithm with matrix stepsizes defined by the gain matrices K_k^{-1}, as opposed to the scalar stepsizes λ_k appearing in (7.17). Interestingly however, the theoretical derivations follow two completely different routes, and lead to different bounds on the number N of steps required to attain the desired relative scale accuracy. In particular, bound (7.19) requires the knowledge of the parameters L, V, which can be hard to determine in practice, but does not depend directly on the problem dimension n. In contrast, bound (7.12) is independent of the L, V parameters, but depends on n, through the VC-dimension bound.

More importantly, we remark that bound (7.12) is almost independent of the probabilistic level α, since α appears under a logarithm, while bound (7.19) has a strong quadratic dependence on α. For this reason, we expect the bound (7.12) to be better than (7.19), when a high level of confidence is required.

We also remark that in [242] a modification of Algorithm 7.2 is also considered, which introduces a mechanism of 'averaging from a pool of experts'. With this modified approach, a sample bound

$$N \geq \frac{1}{2\epsilon^4} \ln \frac{1}{\alpha} \left(\frac{LR}{V} \right)^2$$

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

7.5 Numerical Examples

In the following sections, we illustrate the proposed approach on three numerical examples, and compare its performance with the stochastic gradient approach described in Section 7.4. In particular, Section 7.5.1 presents an example on polynomial interpolation, and Section 7.5.2 discusses a case with affine uncertainty. Also, an application to the problem of receding-horizon state estimation for uncertain systems is proposed in Section 7.5.3.

7.5.1 Polynomial Interpolation

We consider a problem of robust polynomial interpolation borrowed from [121]. For given integers $n \geq 1, m$, we seek a polynomial of degree $n-1$, $p(t) = x_1 + x_2 t + \cdots + x_n t^{n-1}$ that interpolates given points (a_i, y_i), $i = 1, \ldots, m$, that is

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

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

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

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

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

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

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

We therefore seek an interpolant that minimizes the average interpolation error

$$\mathbb{E}_\delta[\|A(\delta)x - y\|^2],$$

where

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

For a numerical example, we considered the data

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

with uncertainty level $\rho = 0.2$.

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

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

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

We wish to obtain a solution having relative scale error $\epsilon = 0.1$ with high confidence $(1 - \alpha) = 0.999$, using Algorithm 7.1. In this case, the theoretical bound (7.12) would require $N \geq 3,115,043$ samples of the uncertainty. However, as already remarked, while this is the *a-priori* bound, we can expect practical convergence for much smaller sample sizes. Indeed, in the example at hand, we observe practical convergence of Algorithm 7.1 already for $N \simeq 10,000$, see Figure 7.1.

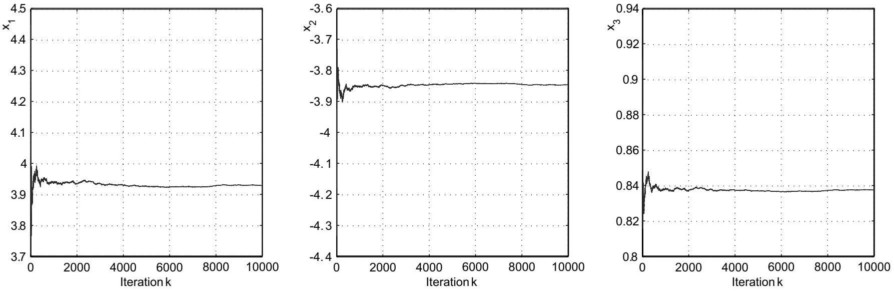


Figure 7.1. Convergence of Algorithm 7.1 for $N = 10,000$ iterations. Solution after N iterations: $\hat{x}_N = [3.926\ -3.840\ 0.837]^T$.

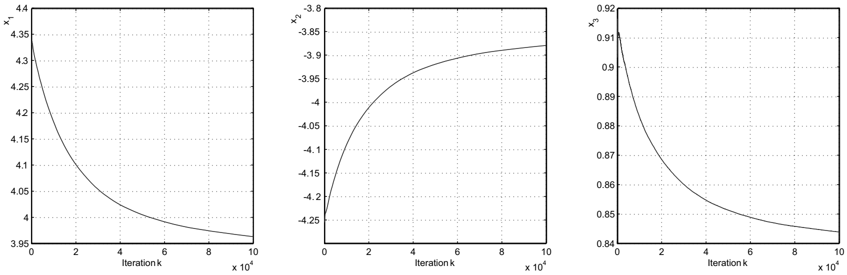


Figure 7.2. Evolution of Algorithm 7.2 for $N = 100,000$ iterations, $\lambda = 10^{-3}$. The algorithm has not yet converged. Solution after N iterations: $\hat{x}_N = [3.961\ -3.876\ 0.844]^T$.

We then compared the above results to the ones that can be obtained using the stochastic gradient approach of Algorithm 7.2. To this end, we first performed a preliminary step in order to obtain reasonable estimates of the parameters L, V. With the above choice of $\mathcal{X}$, we obtained the approximate bound $L/V \leq 0.25$. Therefore, the theoretical bound (7.19) would imply the (prohibitive) number of samples $N \geq 625,000,000$ to achieve the desired probabilistic levels. Also, from a practical point of view, we observed slower convergence with respect to Algorithm 7.1. Moreover, the behavior of the algorithm appeared to be very sensitive to the choice of the stepsize λ.

The evolution of the estimate for $N = 100,000$, and with $\lambda = 10^{-3}$ is shown in Figure 7.2.

7.5.2 An Example with Affine Uncertainty

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

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

with

$$A_0 = \begin{bmatrix} 3 & 1 & 4 \\ 0 & 1 & 1 \\ -2 & 5 & 3 \\ 1 & 4 & 5.2 \end{bmatrix}, A_1 = \begin{bmatrix} 0&0&0 \\ 0&0&0 \\ 0&0&0 \\ 0&0&1 \end{bmatrix}, A_2 = \begin{bmatrix} 0&0&1 \\ 0&1&0 \\ 0&0&0 \\ 0&0&0 \end{bmatrix}, A_3 = \begin{bmatrix} 0&0&0 \\ 0&0&0 \\ 1&0&0 \\ 0&0&0 \end{bmatrix},$$

and let δ_i be Gaussian random perturbations,[3] with zero mean and standard deviations $\sigma_1 = 0.067$, $\sigma_2 = 0.1$, $\sigma_3 = 0.2$. In this case, the exact solution from Theorem 1 is unique and results in

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

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

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

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

To seek a randomized solution having *a-priori* relative error $\epsilon = 0.1$ with high confidence $(1 - \alpha) = 0.999$, the theoretical bound (7.12) would require

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

$N \geq 3,115,043$ samples. Figure 7.3 shows the first $N = 20,000$ iterations of Algorithm 7.1, which resulted in the final solution

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

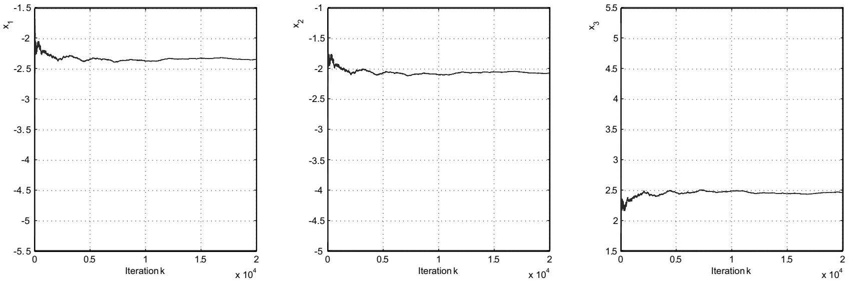


Figure 7.3. Evolution of Algorithm 7.1 for $N = 20,000$ iterations. Solution after N iterations: $\hat{x}_N = [-2.342\ -2.067\ 2.472]^T$.

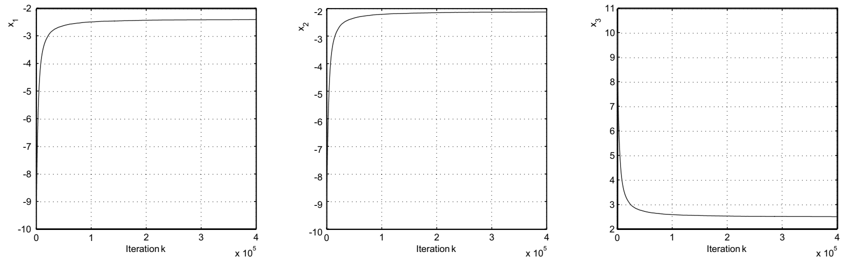


Figure 7.4. Evolution of Algorithm 7.2 for $N = 400,000$ iterations, $\lambda = 10^{-2}$. Solution after N iterations: $\hat{x}_N = [-2.390\ -2.114\ 2.519]^T$.

We next compared the performance of Algorithm 7.1 with that of Algorithm 7.2. For the above choice of $\mathcal{X}$, an estimated bound for the ratio L/V is $L/V \leq 0.11$, and therefore the theoretical bound (7.19) would imply $N \geq 484,000,000$ iterations to guarantee the desired probabilistic levels.

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

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

We notice that the SG algorithm failed to converge for larger values of λ. The evolution of the estimate is shown in Figure 7.4.

7.5.3 Receding-Horizon Estimation for Uncertain Systems

As a last example, we consider a problem of finite-memory state estimation for discrete-time uncertain linear systems. For systems *without* uncertainty, a least-squares solution framework for this problem has been recently proposed in [3]. The basic idea of this method is the following: assume that at a certain time t a prediction $\bar{x}_t$ for the state $x_t \in \mathbb{R}^n$ of the linear system

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

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

$$z_k = Cx_k + \eta_k,$$

and the process and measurement noises ξ_k, η_k, as well as the state x_t are assumed to have unknown statistics. The objective is to determine estimates $\hat{x}_t, \dots, \hat{x}_{t+h}$ of the system states. In [3], the key assumption is made that these estimates should satisfy the nominal state recursions (without noise), *i.e.* be of the form

$$\hat{x}_{t+k} = F^k \hat{x}_t, \quad k = 0, \dots, h. \tag{7.20}$$

From this assumption, it clearly follows that the only quantity that one needs to estimate is $\hat{x}_t$, since all the subsequent state estimates are then determined by (7.20). From (7.20), the estimated outputs are in turn given by

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

and therefore the natural criterion proposed in [3] is to minimize a least-squares error objective that takes into account the deviations of the estimates $\hat{z}_{t+k}$ from the actual measurements z_{t+k}, as well as an additional term that takes into account one's belief in the accuracy of the initial prediction $\bar{x}_t$. Collecting the output measurement in vector $Z_t \doteq [z_t^T, \dots, z_{t+h}^T]^T$, and the output estimates in vector $\hat{Z}_t(\hat{x}_t) \doteq [\hat{z}_t^T, \dots, \hat{z}_{t+T}^h]^T$, the optimization criterion is hence written as

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

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

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

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

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

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

where we defined

$$A(\delta) \doteq \begin{bmatrix} \mu I \\ K(\delta) \end{bmatrix}; \; y \doteq \begin{bmatrix} \mu \bar{x}_t \\ Z_t \end{bmatrix}; \; K(\delta) \doteq \begin{bmatrix} C(\delta) \\ C(\delta)F(\delta) \\ C(\delta)F^2(\delta) \\ \vdots \\ C(\delta)F^h(\delta) \end{bmatrix}.$$

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

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

A probable ϵ-near solution for this problem can be determined according to Theorem 2. Notice that, to the best of the authors' knowledge, no efficient exact method is available for solving this problem. Notice also that even when the uncertainty enters the system matrices $F(\delta), C(\delta)$ in a simple form (such as affine), the data matrix $A(\delta)$ has a very structured and nonlinear dependence on δ. Finally, we remark that applying the estimation procedure in a sliding-window fashion we obtain a finite-memory smoothing filter, in the sense that measurements over the forward time window $t, t+1, \ldots, t+h$ are used to determine an estimate $\hat{x}_t$ of the state at the initial time instant t.

To make a simple numerical example, we modified the model presented in [3], introducing uncertainty. Let therefore

$$F(\delta) = \begin{bmatrix} 0.9950 + \delta_1 & 0.0998 + \delta_2 \\ -0.0998 - \delta_2 & 0.9950 + \delta_3 \end{bmatrix} \tag{7.21}$$

$$C(\delta) = \begin{bmatrix} 1 + \delta_4 & 1 \end{bmatrix} \tag{7.22}$$

with $\delta^T \doteq [\delta_1, \ldots, \delta_4]$ and $\delta_1, \ldots, \delta_4$ independent and uniformly distributed in the intervals $\delta_1 \in [-0.1, 0]$, $\delta_2 \in [-0.01, 0.01]$, $\delta_3 \in [-0.1, 0]$, $\delta_4 \in [-0.1, 0.1]$. We selected estimation window $h = 10$, $\mu = 1$ and run Algorithm 7.1 up to $N = 10,000$ iterations, for each time instant t. Smoothed estimates have been computed over simulation time t from zero to 40. The simulation is run with initial state and initial prediction $x_0 = \bar{x}_0 = [1\ 1]^T$, and process and measurements noises are set to independent Gaussian with standard deviations equal to 0.02 and 0.01, respectively. Figure 7.5 shows the results obtained by the robust smoothing filter on this example. Notice the net improvement gained over the LS estimates of [3] which neglected uncertainty.

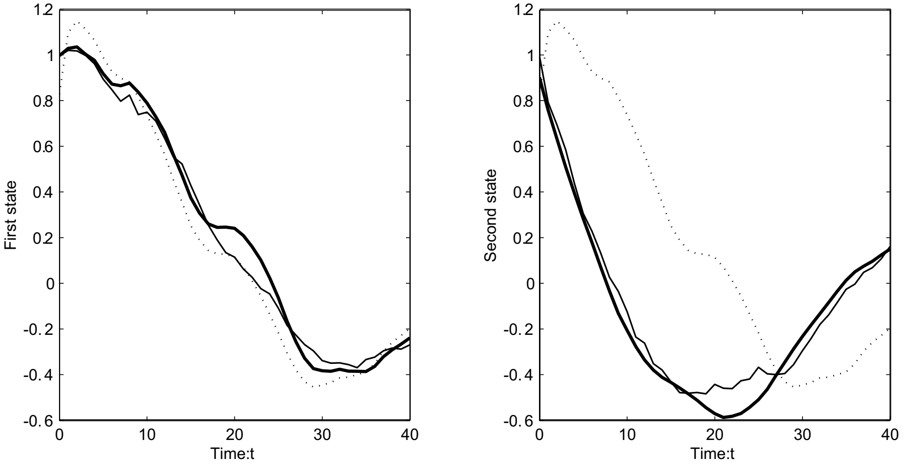


Figure 7.5. Smoothing estimates on the states of the uncertain system (7.21)–(7.22), obtained by means of the randomized robust filter. Bold lines show the state estimates obtained by the robust filter, light lines show a simulation of the actual states of the system, and dotted lines show the estimates obtained by the LS filter of [3] that neglects uncertainty. *Left figure:* first state; textitright figure: second state.

7.6 Conclusions

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

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

8

The Randomized Ellipsoid Algorithm for Constrained Robust Least Squares Problems

Stoyan Kanev and Michel Verhaegen

TU-Delft, DCSC, Mekelweg 2, 2628 CD Delft, the Netherlands,
{s.kanev,m.verhaegen}@dcsc.tudelft.nl

Summary. In this chapter a randomized ellipsoid algorithm is described that can be used for finding solutions to robust Linear Matrix Inequality (LMI) problems. The iterative algorithm enjoys the property that the convergence speed is independent on the number of uncertain parameters. Other advantages, as compared to the deterministic algorithms, are that the uncertain parameters can enter the LMIs in a general nonlinear way, and that very large systems of LMIs can be treated. Given an initial ellipsoid that contains the feasibility set, the proposed approach iteratively generates a sequence of ellipsoids with decreasing volumes, all containing the feasibility set. A method for finding an initial ellipsoid is also proposed based on convex optimization. For an important subclass of problems, namely for constrained robust least squares problems, analytic expressions are derived for the initial ellipsoid that could replace the convex optimization. The approach is finally applied to the problem of robust Kalman filtering.

8.1 Introduction

The linear least squares (LLS) problem arises in a wide variety of engineering applications, ranging from data fitting to controller and filter design. It is at the basis of the well-known Model Predictive Control strategy, an industrially very relevant control technique due to its ability to handle constrained on the inputs and outputs of the controlled system. The problem appears also in many single- and multi-objective controller design techniques such as LQR/LQG, pole-placement, $\mathcal{H}_2$, $\mathcal{H}_\infty$, PID *etc.* The Kalman filtering problem can also be rewritten as an LLS problem.

The LLS problem basically consists of finding the optimal solution to a set of equations $Ax \approx b$ in such a way that the error $\|b - Ax\|_2$ is minimized, where A is a given (usually tall) matrix and b is a given vector. Often, there is underlying structure in the data matrices (A, b), or they depend on some unknown structured matrix Δ that represents uncertainty. This is called the structured robust-least squares (SRLS) problem. It has been shown in [121] that whenever the data matrices depend in an affine way on uncertainty Δ the

SRLS problem is convex and can be solved using semidefinite programming (SDP). However, if the dependence on $\varDelta$ is not affine, the resulting problem is in general no longer solvable via SPD. In this chapter we consider a general dependence on the uncertainty in the data matrices. In addition, it is often desirable to include linear matrix inequality constraints in the SRLS problem. To this end we consider the structured constraint robust least-squares (SCRLS) problem

$$\text{(SCRLS)}: \begin{cases} \text{Find } x \in \mathbb{R}^N \text{ that achieves} \\ \gamma_{opt} = \min_x \max_{\varDelta \in \boldsymbol{\Delta}} \|b(\varDelta) - A(\varDelta)x\|_2^2, \text{ subject to} \\ F(x, \varDelta) \triangleq F_0(\varDelta) + \sum_{i=1}^{N} F_i(\varDelta)x_i \leq 0, \ \forall \varDelta \in \boldsymbol{\Delta} \end{cases} \tag{8.1}$$

where $x = [x_1, \ldots, x_N]^T$ denotes the vector of unknowns, $b(\varDelta) \in \mathbb{R}^p$, $A(\varDelta) \in \mathbb{R}^{p \times N}$, and $F_i(\varDelta)$ are known functions of the uncertainty $\varDelta$ that

1. Belong to some known uncertainty set $\boldsymbol{\Delta}$, and
2. Are coupled with some probability density function (p.d.f.) $f_{\boldsymbol{\Delta}}(\varDelta)$ inside the uncertainty set $\boldsymbol{\Delta}$.

The matrices $b(\varDelta)$, $A(\varDelta)$, and $F_i(\varDelta)$ may depend on the uncertainty $\varDelta$ in a general nonlinear way; it is only assumed that they remain bounded.

Remark 1. Whenever the uncertainty is fully deterministic or no *a-priori* information is available about its statistical properties, uniform distribution could be selected, *i.e.*

$$f_{\boldsymbol{\Delta}}(\varDelta) = \frac{1}{\text{vol}(\boldsymbol{\Delta})}, \ \forall \varDelta \in \boldsymbol{\Delta}.$$

where $\text{vol}(\boldsymbol{\Delta}) = \int_{\boldsymbol{\Delta}} dx$ denotes the volume of the uncertainty set.

Many practical controller/observer design problems are captured by the SCRLS problem (8.1). One such design problem, discussed in this chapter, is the Kalman filter design for uncertain systems.

For some specific uncertainty structures (*e.g.* in cases when A, b and F are all affine in $\varDelta$ and the uncertainty set is a polyhedron), the problem of finding a deterministic solution to the optimization problem (8.1) can be converted to an LMI optimization problem that can be solved numerically very efficiently. For general uncertainty structures, however, this problem is NP-hard, in which case one might be satisfied with computing an approximate solution in a probabilistic framework [359]. In this setting, given some desired accuracy and confidence, one computes near-optimal solutions in an iterative fashion using a randomized algorithm (RA). A RA basically generates at each iteration a random uncertainty sample from $\boldsymbol{\Delta}$ with the selected p.d.f. $f_{\boldsymbol{\Delta}}(\varDelta)$, for which the optimization variable x is updated. In this framework it is assumed that it is possible to generate samples of $\varDelta$ according to the selected p.d.f. $f_{\boldsymbol{\Delta}}(\varDelta)$. The reader is referred to [74] for more details on the available algorithms for uncertainty generation. An important property of these algorithms is their guaranteed convergence to a near-optimal solution for any desired accuracy and confidence in a finite number of iterations. Recently, improved bounds on the maximum number of iterations, that the RA can perform before convergence, have been derived. For more details on this topic, see [249, 359] and Chapter 11 of this book.

In this chapter the randomized ellipsoid algorithm (EA) [176] will be used to finding an approximate solution to (8.1). In summary, at each iteration the randomized EA performs two steps. In the first step a random uncertainty sample $\varDelta^{(i)} \in \boldsymbol{\Delta}$ is generated according to the given probability density function $f_{\boldsymbol{\Delta}}(\varDelta)$. With this generated uncertainty a suitably defined scalar convex function is parameterized so that at the second step of the algorithm an ellipsoid is computed, in which the solution set is guaranteed to lie. In this way the algorithm produces a sequence of ellipsoids with decreasing volumes, all containing the solution set. Using some existing facts, and provided that the solution set has a non-empty interior, it is established that this algorithm converges to a feasible solution in a finite number of iterations with probability one. It is also shown that even if the solution set has a zero volume, the EA converges to the solution set when the iteration number tends to infinity.

To initialize the algorithm an initial ellipsoid containing the solution set is needed. For general robust LMI problems a method is suggested for finding such an ellipsoid based on convex optimization. Furthermore, it is shown that for SCRLS problems one can derive analytic expression for the initial ellipsoid by making use of the structure of the problem.

8.2 Preliminaries

8.2.1 Notation

The notation used in the chapter is as follows. I_n denotes the identity matrix of dimension $n \times n$, $I_{n\times m}$ is a matrix of dimension $n \times m$ with ones on its main diagonal. A vector of dimension n with all elements equal to zero will be denoted as $\mathbf{0}_n$. The dimensions will often be omitted in cases where they can be implied from the context. For two matrices A and B of appropriate dimension, $\langle A, B\rangle \doteq \operatorname{tr}(A^T B)$. $\|.\|_F$ denotes the Frobenius norm. The Frobenius norm for a matrix $A \in \mathbb{R}^{m\times n}$ has the following useful properties:

$$\|A\|_F^2 = \langle A, A\rangle = \sum_{i=1}^{\min\{n,m\}} \sigma_i^2(A) = \sum_{i=1}^{n} \lambda_i(A^T A),$$

where $\sigma_i(A)$ are the singular values of the matrix A and $\lambda_i(A^T A)$ are the eigenvalues of the matrix $(A^T A)$. In addition to that, for any two matrices A and B of equal dimensions it holds that

$$\|A+B\|_F^2 = \|A\|_F^2 + 2\langle A, B\rangle + \|B\|_F^2. \tag{8.2}$$

$A \succ 0$ ($A \succeq 0$) means that A is positive definite (positive semi-definite). We also introduce the notation $\|x\|_Q^2 \doteq x^T Q x$ for $x \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{n\times n}$ with $Q \succeq 0$, which should not be mistaken with the standard notation for the vector p-norm ($\|x\|_p$). In LMIs, the symbols $\star$ will be used to indicate entries readily implied from symmetry. Futher, the volume of a closed set $\mathcal{A}$ is denoted as $\mathrm{vol}(\mathcal{A}) \doteq \int_{\mathcal{A}} dx$.

The notation $x \sim \mathcal{N}(\bar{x}, S)$ will be used to make clear that x is a random Gaussian vector with mean $\bar{x}$ and covariance SS^T. Finally, for a random variable x_k, $\hat{x}_{k+i|k}$ will denote the prediction of x_{k+i} made at time instant k (*i.e.* by using the input-output measurements up to time instant k).

Let $\mathcal{C}_n^+$ denote the cone of n-by-n symmetric non-negative definite matrices, *i.e.*

$$\mathcal{C}_n^+ \doteq \{A \in \mathbb{R}^{n\times n} :\ A = A^T,\ A \succeq 0\}.$$

For a symmetric matrix A we define the projection onto $\mathcal{C}_n^+$ as follows:

$$\mathbf{\Pi}^+ A \doteq \arg \min_{X \in \mathcal{C}_n^+} \|A - X\|_F.$$

Similarly, denoting

$$\mathcal{C}_n^- \doteq \{A \in \mathbb{R}^{n\times n} :\ A = A^T,\ A \preceq 0\},$$

the projection onto the cone of symmetric negative-definite matrices is defined as

$$\mathbf{\Pi}^- A \doteq \arg \min_{X \in \mathcal{C}_n^-} \|A - X\|_F. \tag{8.3}$$

These two projections have the following properties [76].

Lemma 1 (Properties of the projection). *For a symmetric matrix A, the following properties hold*

(P1) $\mathbf{\Pi}^+ A + \mathbf{\Pi}^- A = A$.

(P2) $\langle \mathbf{\Pi}^+ A, \mathbf{\Pi}^- A\rangle = 0$.

(P3) Let $A = U\Lambda U^T$, where U is an orthogonal matrix containing the eigenvectors of A, and Λ is a diagonal matrix with the eigenvalues λ_i, $i = 1, \ldots, n$, of A appearing on its diagonal. Then

$$\mathbf{\Pi}^+ A = U \mathit{diag}\{\lambda_1^+, \ldots, \lambda_n^+\} U^T,$$

with $\lambda_i^+ \doteq \max(0, \lambda_i)$, $i = 1, \ldots, n$. Equivalently,

$$\mathbf{\Pi}^- A = U \mathit{diag}\{\lambda_1^-, \ldots, \lambda_n^-\} U^T,$$

with $\lambda_i^- \doteq \min(0, \lambda_i)$, $i = 1, \ldots, n$.

(P4) $\mathbf{\Pi}^+ A$ *and* $\mathbf{\Pi}^- A$ *are continuous in A.*

8.2.2 Problem Formulation

Consider the (SCRLS) optimization problem (8.1). Since for any γ

$$\left\{x : \max_{\Delta} \|b(\Delta) - A(\Delta)x\|_2^2 \leq \gamma\right\} \Leftrightarrow \left\{x : \begin{bmatrix} I \; b(\Delta) - A(\Delta)x \\ \star \qquad \gamma \end{bmatrix} \succeq 0, \; \forall \Delta\right\}$$

the (SCRLS) problem can equivalently be rewritten in the form

$$(\mathcal{P}_{\mathcal{O}}) \; : \begin{cases} \min\limits_{x,\gamma} \gamma \\ \text{s.t. } U_\gamma(x,\Delta) \doteq \begin{bmatrix} F(x,\Delta) & 0 & 0 \\ \star & -I & b(\Delta) - A(\Delta)x \\ \star & \star & -\gamma \end{bmatrix} \preceq 0, \; \forall \Delta \in \boldsymbol{\Delta} \end{cases} \tag{8.4}$$

For a fixed $\gamma > 0$, the feasibility problem is defined as

$$(\mathcal{P}_{\mathcal{F}}) \; : \begin{cases} \text{Find } x \in \mathbb{R}^N \\ \text{such that } U_\gamma(x,\Delta) \preceq 0, \; \forall \Delta \in \boldsymbol{\Delta} \end{cases} \tag{8.5}$$

Note that $U_\gamma(x,\Delta)$ is affine in x and can be written in the form

$$U_\gamma(x,\Delta) = U_{\gamma,0}(\Delta) + \sum_{i=1}^{N} U_{\gamma,i}(\Delta)x_i.$$

We will first concentrate on the feasibility problem $(\mathcal{P}_{\mathcal{F}})$. Once we have an algorithm for solving it, the optimization problem $(\mathcal{P}_{\mathcal{O}})$ would only require a bisection algorithm on γ where at each iteration $(\mathcal{P}_{\mathcal{F}})$ is solved for a fixed γ.

8.3 The Randomized Ellipsoid Algorithm

This section presents the randomized ellipsoid algorithm, originally proposed in [176].

8.3.1 Feasibility Problem $(\mathcal{P}_{\mathcal{F}})$

Since the randomized algorithm presented here relies on the availability of algorithms for random uncertainty generation, the following assumption needs to be imposed.

Assumption 8.1. *It is assumed that random samples of Δ can be generated inside $\boldsymbol{\Delta}$ with the specified probability density $f_{\boldsymbol{\Delta}}(\Delta)$.*

For certain probability density functions there exist algorithms in the literature for generation of random samples of Δ. For instance, in [73] the authors consider the problem of generating (real and complex) vectors samples uniformly in the ball $\mathcal{B}(r) = \{x : \; \|x\|_p \leq r\}$. This is consequently extended for the matrix case, but only the 1-norm and the ∞-norm are considered. The important case of matrix 2-norm is considered later on in [74]. The reader is referred to [73, 74] for more details on the available algorithms for uncertainty generation.

The set of all feasible solutions to $(\mathcal{P}_{\mathcal{F}})$ is called the *solution set*, and is denoted as

$$\mathcal{S}_\gamma \doteq \{x \in \mathbb{R}^N : \ U_\gamma(x, \varDelta) \preceq 0, \ \forall \varDelta \in \mathbf{\Delta}\}.$$

Further, define the following function

$$v_\gamma(x, \varDelta) \doteq \|\mathbf{\Pi}^+[U_\gamma(x, \varDelta)]\|_F^2, \tag{8.6}$$

which is clearly non-negative for any $x \in \mathbb{R}^N$ and $\varDelta \in \mathbf{\Delta}$. The usefulness of the so-defined function $v_\gamma(x, \varDelta)$ stems from the following fact.

Lemma 2. *For a given pair $(\bar{x}, \bar{\varDelta}) \in \mathbb{R}^N \times \mathbf{\Delta}$ it holds that $U_\gamma(\bar{x}, \bar{\varDelta}) \in \mathcal{C}_q^-$ if and only if $v_\gamma(\bar{x}, \bar{\varDelta}) = 0$.*

Proof. Using the third property in Lemma 1 we note that $U_\gamma(\bar{x}, \bar{\varDelta}) \in \mathcal{C}_q^-$ holds if and only if

$$\mathbf{\Pi}^-[U_\gamma(\bar{x}, \bar{\varDelta})] = U_\gamma(\bar{x}, \bar{\varDelta})$$

Making use of the first property in Lemma 1 we then observe that

$$\mathbf{\Pi}^+[U_\gamma(\bar{x}, \bar{\varDelta})] = 0,$$

or equivalently, that $v_\gamma(\bar{x}, \bar{\varDelta}) = 0$. □

Using the result from Lemma 2 it follows that

$$\{x \in \mathbb{R}^N : \ v_\gamma(x, \varDelta) = 0, \ \forall \varDelta \in \mathbf{\Delta}\} \equiv \mathcal{S}_\gamma$$

holds. In this way the initial feasibility problem is reformulated as the problem of checking whether the solution to the following optimization problem

$$\min_{x \in \mathbb{R}^N} \sup_{\varDelta \in \mathbf{\Delta}} v_\gamma(x, \varDelta)$$

is equal to zero.

In the randomized ellipsoid algorithm, presented in this chapter, the gradient of the function $v_\gamma(\cdot, \cdot)$ is needed. The following result, which is also stated in [76], provides an analytic expression for it.

Lemma 3. *The function $v_\gamma(x, \varDelta)$, defined in equation (8.6), is convex and differentiable in x and its gradient is given by*

$$\nabla v_\gamma(x, \varDelta) = 2 \begin{bmatrix} \operatorname{tr}\left(U_{\gamma,1}(\varDelta)\mathbf{\Pi}^+[U_\gamma(x, \varDelta)]\right) \\ \vdots \\ \operatorname{tr}\left(U_{\gamma,N}(\varDelta)\mathbf{\Pi}^+[U_\gamma(x, \varDelta)]\right) \end{bmatrix} \tag{8.7}$$

Proof. By using the properties of the projection in Lemma 1 we observe that for some symmetric matrices R and $\varDelta R$ it can be written that

$$\begin{aligned}
\|\mathbf{\Pi}^+[R+\varDelta R]\|_F^2 &\overset{(\text{P1})}{=} \|R+\varDelta R-\mathbf{\Pi}^-[R+\varDelta R]\|_F^2\\
&\overset{(\text{P1})}{=} \|\mathbf{\Pi}^+R+\mathbf{\Pi}^-R+\varDelta R-\mathbf{\Pi}^-[R+\varDelta R]\|_F^2\\
&\overset{(8.2)}{=} \|\mathbf{\Pi}^+R\|_F^2+2\left\langle \mathbf{\Pi}^+R,\varDelta R\right\rangle+2\underbrace{\left\langle \mathbf{\Pi}^+R,\mathbf{\Pi}^-R\right\rangle}_{=0}\\
&\quad +\|\mathbf{\Pi}^-R+\varDelta R-\mathbf{\Pi}^-[R+\varDelta R]\|_F^2\\
&\quad +2\underbrace{\left\langle \mathbf{\Pi}^+R,-\mathbf{\Pi}^-[R+\varDelta R]\right\rangle}_{\geq 0}\\
&\overset{(\text{P2}),(\text{P3})}{\geq} \|\mathbf{\Pi}^+R\|_F^2+\left\langle 2\mathbf{\Pi}^+R,\varDelta R\right\rangle
\end{aligned}$$

In addition to that, noting that from (8.3) it follows that

$$\|A-\mathbf{\Pi}^+A\|_F^2=\min_{X\in\mathcal{C}_n^-}\|A-X\|_F^2, \tag{8.8}$$

we can write that

$$\begin{aligned}
\|\mathbf{\Pi}^+[R+\varDelta R]\|_F^2 &\overset{(\text{P1})}{=} \|R+\varDelta R-\mathbf{\Pi}^-[R+\varDelta R]\|_F^2\\
&\overset{(8.8)}{=} \min_{S\in\mathcal{C}_n^-}\|R+\varDelta R-S\|_F^2\\
&\leq \|R+\varDelta R-\mathbf{\Pi}^-R\|_F^2 \overset{(\text{P1})}{=} \|\mathbf{\Pi}^+R+\varDelta R\|_F^2\\
&\overset{(8.2)}{=} \|\mathbf{\Pi}^+R\|_F^2+\left\langle 2\mathbf{\Pi}^+R,\varDelta R\right\rangle+\|\varDelta R\|_F^2.
\end{aligned}$$

It thus follows that

$$\|\mathbf{\Pi}^+[R+\varDelta R]\|_F^2=\|\mathbf{\Pi}^+R\|_F^2+\left\langle 2\mathbf{\Pi}^+R,\varDelta R\right\rangle+O(\|\varDelta R\|_F^2).$$

Now, substitute $R=U_\gamma(x,\varDelta)$ and $\varDelta R=\sum_{i=1}^N U_{\gamma,i}(\varDelta)\varDelta x_i$ to obtain

$$v_\gamma(x+\varDelta x,\varDelta)\geq v_\gamma(x,\varDelta)+\sum_{i=1}^N\left\langle 2\mathbf{\Pi}^+[U_\gamma(x,\varDelta)]U_{\gamma,i}(\varDelta),\varDelta x_i\right\rangle \tag{8.9}$$

$$\begin{aligned}
&v_\gamma(x+\varDelta x,\varDelta)\\
&=v_\gamma(x,\varDelta)+\sum_{i=1}^N\left\langle 2\mathbf{\Pi}^+[U_\gamma(x,\varDelta)],U_{\gamma,i}(\varDelta)\right\rangle\varDelta x_i+O(\|\varDelta x\|_2^2),
\end{aligned}\tag{8.10}$$

The convexity follows from inequality (8.9), while the differentiability follows from equation (8.10). The gradient of $v_\gamma(x,\varDelta)$ is then given by (8.7). □

Now that the gradient of the function $v_\gamma(x,\varDelta)$ is derived analytically we are ready to proceed to the randomized approach that is based on the *Ellipsoid Algorithm* (EA) [59]. The starting point in EA is the computation of an initial ellipsoid that contains the solution set $\mathcal{S}_\gamma$. Then at each iteration of the EA two steps are performed. In the first step a random uncertainty sample $\Delta^{(i)} \in \mathbf{\Delta}$ is generated according to the given probability density function $f_{\mathbf{\Delta}}(\Delta)$. With this generated uncertainty the convex function $U_\gamma(x, \Delta^{(i)})$ is parameterized and used at the second step of the algorithm where an ellipsoid is computed, in which the solution set is guaranteed to lie. In this way the EA produces a sequence of ellipsoids with decreasing volumes, all containing the solution set. Using some existing facts, and provided that the solution set has a non-empty interior, it will be established that this algorithm converges to a feasible solution in a finite number of iterations with probability one. It is also shown that even if the solution set has a zero volume, the EA converges to the solution set when the iteration number tends to infinity. To initialize the algorithm, some methods are proposed for obtaining an initial ellipsoid that contains the solution set.

Define the ellipsoid

$$\mathcal{E}(\bar{x}, \bar{P}) = \{x \in \mathbb{R}^N : \ (x - \bar{x})^T \bar{P}^{-1}(x - \bar{x}) \leq 1\}$$

with center $\bar{x} \in \mathbb{R}^N$ and matrix $\bar{P} \in \mathcal{C}_N^+$ describing its shape and orientation. Assume that an initial ellipsoid $\mathcal{E}(x^{(0)}, P_0)$ is given that contains the solution set $\mathcal{S}_\gamma$.

We further assume that the dimension N of the vector of unknowns is is larger than one[1]. The problem of finding such an initial ellipsoid will be discussed in the next section. Define

$$H^{(0)} \doteq \{x \in \mathbb{R}^N : \ \nabla^T v_\gamma(x^{(0)}, \Delta)(x - x^{(0)}) \leq 0\}.$$

Due to the convexity of the function $v_\gamma(x, \Delta)$ we know that $H^{(0)}$ also contains the solution set $\mathcal{S}_\gamma$, and therefore $\mathcal{S}_\gamma \subseteq H^{(0)} \cap \mathcal{E}(x^{(0)}, P_0)$. We can then construct a new ellipsoid, $\mathcal{E}(x^{(1)}, P_1)$, as the *minimum volume* ellipsoid such that $\mathcal{E}(x^{(1)}, P_1) \supseteq H^{(0)} \cap \mathcal{E}(x^{(0)}, P_0) \supseteq \mathcal{S}_\gamma$, and such that the volume of $\mathcal{E}(x^{(1)}, P_1)$ is less than the volume of $\mathcal{E}(x^{(0)}, P_0)$. This, repeated iteratively, represents the main idea behind the Ellipsoid Algorithm [59].

Algorithm 8.1 (Randomized Ellipsoid Algorithm for ($\mathcal{P}_\mathcal{F}$))
Initialization: $i = 0$, $x^{(0)}$, $P_0 = P_0^T \succ 0$, $\varepsilon > 0$ *small, integer* $L > 0$.

Step 1. Set $i \leftarrow i + 1$.
Step 2. Generate a random sample $\Delta^{(i)}$ *with probability distribution* f_Δ.
Step 3. If $v_\gamma(x^{(i)}, \Delta^{(i)}) \neq 0$ *then take*

$$x^{(i+1)} = x^{(i)} - \frac{1}{N+1} \frac{P_i \nabla v_\gamma(x^{(i)}, \Delta^{(i)})}{\sqrt{\nabla^T v_\gamma(x^{(i)}, \Delta^{(i)}) P_i \nabla v_\gamma(x^{(i)}, \Delta^{(i)})}}$$

$$P_{i+1} = \frac{N^2}{N^2 - 1}\left(P_i - \frac{2}{N+1} \frac{P_i \nabla v_\gamma(x^{(i)}, \Delta^{(i)}) \nabla^T v_\gamma(x^{(i)}, \Delta^{(i)}) P_i^T}{\nabla^T v_\gamma(x^{(i)}, \Delta^{(i)}) P_i \nabla v_\gamma(x^{(i)}, \Delta^{(i)})}\right)$$

else take $x^{(i+1)} = x^{(i)}$, $P_{i+1} = P_i$.

[1] With $N = 1$ the algorithm simplifies to a bisection algorithm.

Step 4. Form the ellipsoid

$$\mathcal{E}(x^{(i+1)}, P_{i+1}) = \{x : \ (x - x^{(i+1)})^T P_{i+1}^{-1} (x - x^{(i+1)}) \leq 1\} \supseteq \mathcal{S}_\gamma.$$

Step 5. If $\left(\sqrt{\det(P)} < \varepsilon\right)$ *or* $\left(v_\gamma(x^{(i+j-L)}, \Delta^{(i+j-L)}) = 0 \text{ for } j = 0, 1, \ldots, L\right)$ *then Stop else Goto Step 1.*

The randomized EA is summarized in Algorithm 8.1. The algorithm terminates when the value of the function $v_\gamma(x^{(\cdot)}, \Delta^{(\cdot)})$ remains equal to zero for L successive iterations or when the volume of the ellipsoid (which is proportional to $\det(P)^{1/2}$) becomes smaller than a pre-defined small positive number ε. In the latter case no feasible solution is found (for instance due to the fact that the solution set has an empty interior, *i.e.* $\mathrm{vol}(\mathcal{S}_\gamma) = 0$). In such case γ has to be increased in the feasibility problem (8.5) and Algorithm 8.1 has to be started again until a feasible solution is found. Note that if the feasibility problem (8.5) is feasible for some γ^*, then it is also feasible for any $\gamma > \gamma^*$. It should also be noted that, due to the probabilistic nature of the algorithm, the fact that the algorithm terminates due to the cost function being equal to zero for a finite number L of successive iterations does not necessarily imply that a feasible solution is found (see also Remark 2 below). In practice, however, choosing L sufficiently large ensures the feasibility of the solution.

For proving the convergence of the algorithm, the following technical assumption needs to be additionally imposed.

Assumption 8.2. *For any $x^{(i)} \notin \mathcal{S}_\gamma$ there is a non-zero probability to generate a sample $\Delta^{(i)}$ for which $v_\gamma(x^{(i)}, \Delta^{(i)}) > 0$, i.e.*

$$\mathbb{P}\{v_\gamma(x^{(i)}, \Delta^{(i)}) > 0\} > 0.$$

This assumption is standard in the literature on randomized algorithms (see, *e.g.*, [76,273]) and is not restrictive in practice. Note that a sufficient condition for the assumption to hold is that the density function f_Δ is non-zero everywhere. The assumption implies that for any $x^{(i)} \notin \mathcal{S}_\gamma$ there exists a non-zero probability for the execution of a *correction step* (*i.e.* there is a non-zero probability for generation of $\Delta^{(i)} \in \mathbf{\Delta}$ such that $v_\gamma(x^{(i)}, \Delta^{(i)}) > 0$). Correction step means an iteration with $v_\gamma(x^{(i)}, \Delta^{(i)}) \neq 0$.

The convergence of the approach is established immediately, provided that Assumption 8.2 holds.

Lemma 4 (Convergence of Algorithm 8.1). *Consider Algorithm 8.1 without the stopping condition in Step 5 (or with $\varepsilon = 0$ and $L \to \infty$), and suppose that Assumption 8.2 holds. Suppose also that*

(i) $\mathrm{vol}(\mathcal{S}_\gamma) > 0$. *Then a feasible solution will be found in a finite number of iterations with probability one.*

(ii) $\mathrm{vol}(\mathcal{S}_\gamma) = 0$. *Then*

$$\lim_{i \to \infty} x^{(i)} = x^* \in \mathcal{S}_\gamma$$

with probability one.

Proof. Suppose that at the i-th iteration of Algorithm 8.1 $k(i)$ correction steps have been performed. Algorithm 8.1 generates ellipsoids with geometrically decreasing volumes so that for the i-th iteration we can write [59]

$$\mathrm{vol}(\mathcal{E}(x^{(i)}, P_i)) \le e^{-\frac{k(i)}{2N}} \mathrm{vol}(\mathcal{E}(x^{(0)}, P_0)),$$

Due to Assumption 8.2, for any $x^{(i)} \notin \mathcal{S}_\gamma$ there exists a non-zero probability for the execution of a correction step. Therefore, at any infeasible point $x^{k(i)}$ the algorithm will execute a correction step after a finite number of iterations with probability one. This implies that

$$\lim_{i\to\infty} \mathrm{vol}(\mathcal{E}(x^{(i)}, P_i)) = 0. \tag{8.11}$$

(i) If we then suppose that the solution set $\mathcal{S}_\gamma$ has a non-empty interior, *i.e.* $\mathrm{vol}(\mathcal{S}) > 0$, then from equation (8.11) and due to the fact that $\mathcal{E}(x^{(i)}, P_i) \supseteq \mathcal{S}_\gamma$ for all $i = 0, 1, \dots,$ it follows that in a finite number of iterations with probability one the algorithm will terminate at a feasible solution.

(ii) If we now suppose that $\mathrm{vol}(\mathcal{S}) = 0$, then due to the convexity of the function, and due to equation (8.11), the algorithm will converge to a point in $\mathcal{S}_\gamma$ with probability one. □

Remark 2. It needs to be noted, however, that Lemma 4 considers Algorithm 8.1 with $L \to \infty$, which in practice is never the case. For finite L the solution found by the algorithm can only be analyzed in a probabilistic sense. To be more specific, let some scalars $\epsilon \in (0, 1)$ and $\delta \in (0, 1)$ be given, and let x^* be the output of Algorithm 8.1 for $\varepsilon = 0$ and $L \ge \ln \frac{1}{\delta} / \ln \frac{1}{1-\epsilon}$. Then [130, 359]

$$\mathbb{P}\{\mathbb{P}\{v_\gamma(x^*, \mathit{\Delta}) > 0\} \le \epsilon\} \ge 1 - \delta.$$

Therefore, if we want with high confidence (*e.g.* $\delta = 0.01$) that the probability that x^* is an optimal solution is very high ($1 - \epsilon = 0.999$) then we need to select L larger than 4603. In practice, however, a much smaller value for L suffices.

The next lemma provides an upper bound on the maximum number of correction steps that can be executed by the randomized ellipsoid algorithm before a feasible solution is found.

Lemma 5. *Consider Algorithm 8.1, and suppose that Assumption 8.2 holds. Suppose further that the solution set has a non-empty interior, i.e.* $\mathrm{vol}(\mathcal{S}_\gamma) > 0$*. Then the number*

$$I_{EA} = 2N \left\lceil \ln \frac{\mathrm{vol}(\mathcal{E}(x^{(0)}, P_0))}{\mathrm{vol}(\mathcal{S}_\gamma)} \right\rceil \tag{8.12}$$

is an upper bound on the maximum number of correction steps that can be performed starting from any ellipsoid $\mathcal{E}(x^{(0)}, P_0) \supseteq \mathcal{S}_\gamma$*, where* $\lceil a \rceil$, $a \in \mathbb{R}$*, denotes the minimum integer number larger than or equal to* a*.*

Proof. It is shown in [59] that for the $k(i)$-th correction step one can write

$$\mathrm{vol}(\mathcal{E}(x^{(k(i))}, P_{k(i)})) \leq e^{-\frac{k(i)}{2N}} \mathrm{vol}(\mathcal{E}(x^{(0)}, P_0)).$$

Since the volume of the consecutive ellipsoids tends to zero, and since $\mathrm{vol}(\mathcal{S}_\gamma) > 0$, there exists an correction step number I_{EA} such that

$$e^{-\frac{k(i)}{2N}} \mathrm{vol}(\mathcal{E}(x^{(0)}, P_0)) \leq \mathrm{vol}(\mathcal{S}_\gamma), \text{ for } \{\forall i: \ k(i) \geq I_{EA}\}.$$

Therefore, we could obtain the number I_{EA} from the following relation

$$\frac{\mathrm{vol}(\mathcal{S}_\gamma)}{\mathrm{vol}(\mathcal{E}(x^{(0)}, P_0))} \geq e^{-\frac{k(i)}{2N}} \impliedby \{\forall i: \ k(i) \geq I_{EA}\}.$$

Now, by taking the natural logarithm on both sides one obtains

$$\ln \frac{\mathrm{vol}(\mathcal{S}_\gamma)}{\mathrm{vol}(\mathcal{E}(x^{(0)}, P_0))} \geq -\frac{k(i)}{2N} \impliedby \{\forall i: \ k(i) \geq I_{EA}\}$$

or

$$k(i) \geq 2N \ln \frac{\mathrm{vol}(\mathcal{E}(x^{(0)}, P_0))}{\mathrm{vol}(\mathcal{S}_\gamma)} \impliedby \{\forall i: \ k(i) \geq I_{EA}\}$$

Therefore, equation (8.12) is proven. □

8.3.2 Optimization Problem ($\mathcal{P}_\mathcal{O}$)

Above we focused our attention on the feasibility problem for a fixed value of γ in (8.5). Once we have developed an algorithm for the feasibility problem ($\mathcal{P}_\mathcal{F}$), a bisection algorithm on γ can be used to solve the initial optimization problem (8.4). This is summarized in Algorithm 8.2.

Algorithm 8.2 (Randomized Ellipsoid Algorithm for ($\mathcal{P}_\mathcal{O}$))
Initialization: real numbers $Tol > 0$ and $\gamma_{max} > 0$ (sufficiently large), accuracy $\epsilon \in (0,1)$ and confidence $\delta \in (0,1)$. Set $\gamma_1 = 1$, $\gamma_{LB} = 0$, $\gamma_{UB} \leftarrow \infty$, and $k = 1$.

```
Step 1. Find initial ellipsoid $\mathcal{E}_k^{(0)}(x_k^{(0)}, P_k^{(0)})$ for (8.5) with $\gamma = \gamma_k$ using the
    methods of Section 8.4.
Step 2. Set $\mathcal{E}_{opt}(x_{opt}, P_{opt}) = \mathcal{E}_k^{(0)}(x_k^{(0)}, P_k^{(0)})$.
Step 3. Run Algorithm 8.1 on problem (8.5) with $\gamma = \gamma_k$ and with initial
    ellipsoid $\mathcal{E}_{opt}(x_{opt}, P_{opt})$.
Step 4. Denote $\mathcal{E}_k^*(x_k^*, P_k^*)$ as the ellipsoid at the final iteration of Algorithm
    8.1.
Step 5. If $(\mathbb{P}\{\mathbb{P}\{v_\gamma(x_k^*, \Delta) > 0\} \leq \epsilon\} \geq 1 - \delta)$ then $(\gamma_{LB} = \gamma)$
    else $(\gamma_{UB} = \gamma$ and $\mathcal{E}_{opt}(x_{opt}, P_{opt}) = \mathcal{E}_k^*(x_k^*, P_k^*))$.
Step 6. Set $k \leftarrow k + 1$.
Step 7. If ($\gamma_{UB} = \infty$) then ($\gamma_k = 10\gamma_{k-1}$ and goto Step 1.)
    else (if $\gamma_{LB} = 0$ then $\gamma_k = 0.1\gamma_{k-1}$ else $\gamma_k = \frac{\gamma_{LB}+\gamma_{UB}}{2}$)
Step 8. If $\frac{(\gamma_{UB}-\gamma_{LB})}{\gamma_{UB}} > Tol$ and $\gamma_{LB} < \gamma_{max}$ goto Step 3.
Step 9. Exit the algorithm with $\gamma_{opt} = \gamma_{UB}$ achieved by $x_{opt}$.
```

The algorithm begins by checking whether a feasible solution to (8.5) for $\gamma = 1$ can be found by means of Algorithm 8.1. If not, γ is increased ten times to $\gamma = 10$ and Algorithm 8.1 is run again. In this way Algorithm 8.2 iterates between Step 1 and Step 7 until a feasible solution for some γ is found. After that Algorithm 8.2 begins to iterate between Step 3 and Step 8, so that at each cycle either γ_{UB} or γ_{LB} is set equal to the current γ, depending on whether this γ turns out to be feasible or not. In this way $[\gamma_{LB},\ \gamma_{UB}]$ is a constantly decreasing interval inside which the optimal γ lies. The algorithm is terminated once the length of this interval becomes smaller than the selected tolerance.

We next focus on the problem of computing an initial ellipsoid containing $\mathcal{S}_\gamma$, needed in the initialization of Algorithms 8.1 and 8.2.

8.4 Finding an Initial Ellipsoid $\mathcal{E}(x^{(0)}, P_0)$

In this section we consider the problem of finding initial ellipsoid that contains the solution set $\mathcal{S}_\gamma$. The idea that is exploited here is that for any fixed value $\hat{\Delta}$ of the uncertainty set $\mathbf{\Delta}$ it holds that the set

$$\mathcal{S}_\gamma(\hat{\Delta}) \doteq \{x :\ U_\gamma(x, \hat{\Delta}) \leq 0,\ \hat{\Delta} \in \mathbf{\Delta}\} \subseteq \mathcal{S}_\gamma.$$

Therefore, a reasonable option would be to search for the minimum volume ellipsoid (known as the *outer Löwner-John ellipsoid*) $\mathcal{E}(x^{(0)}, P)$ that contains the set $\mathcal{S}_\gamma(\hat{\Delta})$. This could be achieved by solving the optimization problem

$$\begin{array}{ll} \min\limits_{x^{(0)},Z} & \log\det Z^{-1} \\ \text{subject to} & \sup\limits_{x\in\mathcal{S}_\gamma(\hat{\Delta})} \|Zx - Zx^{(0)}\| \leq 1 \end{array}$$

and then taking $P = Z^{-2}$. However, as stated in [154], this is in general an NP-hard problem. For that reason we will not be interested here with finding any outer Löwner-John ellipsoid, but will rather propose a fast algorithm capable of finding some other outer ellipsoidal approximation of the set $\mathcal{S}_\gamma(\hat{\Delta})$.

We will first concentrate on general LMI optimization problems. In such cases one can find an ellipsoidal approximation of $\mathcal{S}_\gamma(\hat{\Delta})$ by means of solving a convex optimization problem. Subsequently the SCRLS problem will be considered and an analytical expression will be derived for $\mathcal{E}(x^{(0)}, P)$ in the case when the matrix $A(\hat{\Delta})$ has full column rank.

8.4.1 General Case

Here we consider the general case of finding an ellipsoid $\mathcal{E}(x^{(0)}, P_0)$ that contains the set $U_\gamma(x, \hat{\Delta}) \preceq 0$ without making use of its structure in (8.4), *i.e.* U_γ is first allowed to be any matrix function affine in x.

The following additional assumption needs to be imposed.

Assumption 8.3. *It is assumed that the set $\mathcal{S}_\gamma(\hat{\Delta})$ is bounded.*

Assumption 8.3 could be restrictive for some problems. If it does not hold, one can enforce it by including in $U_\gamma(x, \Delta) \leq 0$ additional hard constraints on the elements of the vector of unknowns x.

One way to find an outer approximation of the set $\mathcal{S}_\gamma(\hat{\Delta})$ is as follows [58]. Define the following barrier function for $\mathcal{S}_\gamma(\hat{\Delta})$

$$\phi(x) \doteq \begin{cases} \log\det(-U_\gamma(x, \hat{\Delta}))^{-1}, & \text{if } x \in \mathcal{S}_\gamma(\hat{\Delta}) \\ \infty, & \text{otherwise} \end{cases}$$

Denote the analytic center of $\mathcal{S}_\gamma(\hat{\Delta})$ as

$$x^* = \arg\min_x \phi(x).$$

Note that computing the analytic center is a convex optimization problem.

It is then shown in [58] that an outer approximating ellipsoid $\mathcal{E}(x^{(0)}, P_0)$ of the set $\mathcal{S}_\gamma(\hat{\Delta})$ is given by

$$x^{(0)} = x^*, \; P_0 = N(N-1)H^{-1}(x^*),$$

where $H(x) = [h_{ij}(x)] \in \mathcal{C}_N^+$ is the Hessian of $\phi(x)$ with elements

$$h_{ij}(x) = \mathrm{tr}[U_\gamma^{-1}(x, \hat{\Delta})U_{\gamma,i}(\hat{\Delta})U_\gamma^{-1}(x, \hat{\Delta})U_{\gamma,j}(\hat{\Delta})], \; i, j = 1, 2, \ldots, N.$$

8.4.2 The SCRLS Case

Let us now consider the SCRLS problem defined in (8.1). In this case we can reduce the computational complexity of the algorithm for finding an initial ellipsoid by making use of the structure of the least-squares problem. In particular, an initial ellipsoid can be found in this case by making use the fact that any ellipsoid that contains the set

$$\left\{x : \max_{\Delta \in \boldsymbol{\Delta}} \|b(\Delta) - A(\Delta)x\|_2^2 \leq \gamma \right\} \tag{8.13}$$

also contains the solution set

$$\mathcal{S}_\gamma = \left\{x : \max_{\Delta \in \boldsymbol{\Delta}} \|b(\Delta) - A(\Delta)x\|_2^2 \leq \gamma, \; F(x, \Delta) \succeq 0, \; \forall \Delta \in \boldsymbol{\Delta} \right\}.$$

On the other hand we note that for any $\hat{\varDelta} \in \mathbf{\Delta}$ the set

$$\mathcal{J}(\hat{\varDelta}) \doteq \left\{ x : \ \|b(\hat{\varDelta}) - A(\hat{\varDelta})x\|_2^2 \leq \gamma \right\}$$

contains the set defined in equation (8.13). Therefore, it will suffice to find an initial ellipsoid such that

$$\mathcal{E}(x^{(0)}, P_0) \supseteq \mathcal{J}(\hat{\varDelta})$$

for some $\hat{\varDelta} \in \mathbf{\Delta}$ in order to ensure that $\mathcal{E}(x^{(0)}, P_0)$ will also contain $\mathcal{S}_\gamma$. One possible choice for $\hat{\varDelta}$ is $\hat{\varDelta} = 0$ (provided that $0 \in \mathbf{\Delta}$, of course), but in practice any other (*e.g.* randomly generated) element $\hat{\varDelta}$ from the set $\mathbf{\Delta}$ can be used.

The following cases, related to the rank and dimension of the matrix A can be differentiated.

Case 1. $p = N$ and $A(\hat{\varDelta})$ is invertible. In this case

$$\mathcal{J}(\hat{\varDelta}) = \left\{ x : \ \left(x - A^{-1}(\hat{\varDelta})b(\hat{\varDelta})\right)^T \tfrac{A^T(\hat{\varDelta})A(\hat{\varDelta})}{\gamma} \left(x - A^{-1}(\hat{\varDelta})b(\hat{\varDelta})\right) \leq 1 \right\}$$

so that $\mathcal{E}(x^{(0)}, P_0) = \mathcal{E}\left(A^{-1}(\hat{\varDelta})b(\hat{\varDelta}), \tfrac{A^T(\hat{\varDelta})A(\hat{\varDelta})}{\gamma}\right)$.

Case 2. $p > N$ and $A(\hat{\varDelta})$ is left-invertible.

We can thus factorize $A(\hat{\varDelta})$ (*e.g.* by using the singular value decomposition) as

$$A(\hat{\varDelta}) = E \begin{bmatrix} A_1(\hat{\varDelta}) \\ 0 \end{bmatrix},$$

where E is a unitary matrix and $A_1(\hat{\varDelta})$ is a square non-singular matrix. Denoting

$$\begin{bmatrix} b_1(\hat{\varDelta}) \\ b_2(\hat{\varDelta}) \end{bmatrix} = E^T b(\hat{\varDelta}),$$

we can then write

$$\begin{aligned} \|(b(\hat{\varDelta}) - A(\hat{\varDelta})x)\|_2^2 &= \left\| E \begin{bmatrix} b_1(\hat{\varDelta}) - A_1(\hat{\varDelta})x \\ b_2(\hat{\varDelta}) \end{bmatrix} \right\|_2^2 \\ &= \|b_1(\hat{\varDelta}) - A_1(\hat{\varDelta})x\|_2^2 + \|b_2(\hat{\varDelta})\|_2^2 \leq \gamma. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathcal{J}(\hat{\varDelta}) &= \left\{ x : \ \|b(\hat{\varDelta}) - A(\hat{\varDelta})x\|_2^2 \leq \gamma \right\} \\ &= \left\{ x : \ \|b_1(\hat{\varDelta}) - A_1(\hat{\varDelta})x\|_2^2 \leq \gamma - \|b_2(\hat{\varDelta})\|_2^2 \right\} \\ &= \left\{ x : \ \left(x - A_1^{-1}(\hat{\varDelta})b_1(\hat{\varDelta})\right)^T \tfrac{A_1^T(\hat{\varDelta})A_1(\hat{\varDelta})}{\gamma - \|b_2(\hat{\varDelta})\|_2^2} \left(x - A_1^{-1}(\hat{\varDelta})b_1(\hat{\varDelta})\right) \leq 1 \right\}, \end{aligned}$$

so that $\mathcal{E}(x^{(0)}, P_0) = \mathcal{E}\left(A_1^{-1}(\hat{\varDelta})b_1(\hat{\varDelta}), \tfrac{A_1^T(\hat{\varDelta})A_1(\hat{\varDelta})}{\gamma - \|b_2(\hat{\varDelta})\|_2^2}\right)$.

Case 3. $A(\hat{\varDelta})$ is not full column rank. In this case we cannot obtain an analytic expression for the initial ellipsoid, which could be computed by means of solving the convex optimization problem described in Section 8.4.2.

8.5 Robust Kalman Filtering as SCRLS Problem

In this section we discuss how the randomized algorithm developed above can be used to solve the Kalman filtering problem in the presence of parametric uncertainty. Consider the following discrete-time linear system

$$\mathcal{S}: \left\{ \begin{aligned} z_{k+1} &= A^{\Delta} z_k + B_u^{\Delta} u_k + Q^{\Delta} \xi_k^x \\ y_k &= C_y^{\Delta} z_k + D_{yu}^{\Delta} u_k + R_y^{\Delta} \xi_k^y, \end{aligned} \right. \tag{8.14}$$

where $z \in \mathbb{R}^n$ is the state of the system, $u \in \mathbb{R}^m$ is the control action, $y \in \mathbb{R}^p$ is the measured output, and ξ^x and ξ^y are white Gaussian noises (*i.e.* random zero-mean processes with covariance matrices equal to the identity matrix).

8.5.1 The Kalman Filter

Consider the state-estimation problem from available input-output measurements. In [378], the Kalman filtering problem is formulated as a least-squares problem. To summarize this, let the system state z_k be first written in the general covariance representation as $z_k \sim \mathcal{N}(\hat{z}_{k|k-1}, S_{k|k-1})$, *i.e.* a random Gaussian process with mean $\hat{z}_{k|k-1}$ and covariance $P_{k|k-1} = S_{k|k-1} S_{k|k-1}^T > 0$.

$$\hat{z}_{k|k-1} = z_k + S_{k|k-1} n_k$$

where n_k is a zero-mean stochastic variable with covariance matrix equal to the identity matrix. It is assumed that $\hat{z}_{0|-1}$ and $P_{0|-1}$ are given. Combining this representation of the state z_k with the system equations (8.14) results in

$$\underbrace{\begin{bmatrix} \hat{z}_{k|k-1} \\ y_k - D_{yu} u_k \\ -B_u u_k \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} I_n & 0 \\ C_y & 0 \\ A & -I \end{bmatrix}}_{F^{\Delta}} \underbrace{\begin{bmatrix} z_k \\ z_{k+1} \end{bmatrix}}_{\mathbf{b}} + \underbrace{\begin{bmatrix} S_{k|k-1} & & \\ & R_y & \\ & & Q \end{bmatrix}}_{L^{\Delta}} \underbrace{\begin{bmatrix} n_k \\ \xi_k^y \\ \xi_k^x \end{bmatrix}}_{\Xi}. \tag{8.15}$$

Above, the vector Y contains only signals available at time instant k. Given the state estimate $\hat{z}_{k|k-1}$ from the previous time instant, and the square-root covariance matrix $\hat{S}_{k|k-1}$, an unbiased estimate of the state then be obtained by solving

$$\begin{bmatrix} \hat{z}_{k|k} \\ \hat{z}_{k+1|k} \end{bmatrix} = \arg \min_{\mathbf{b}} \max_{\Delta \in \mathbf{\Delta}} \|(L^{\Delta})^{-1}(Y - F^{\Delta}\mathbf{b})\|_2^2,$$

which is clearly a special case of the SCRLS problem (8.1). It can thus be solved by making use of the proposed randomized ellipsoid algorithm. Hence, we next concentrate on the problem of computation of the square-root covariance matrix $S_{k+1|k}$ that will be required for the optimization at time instant $(k+1)$.

8.5.2 Computation of the Square-Root Covariance Matrix $S_{k+1|k}$

In this subsection we are concerned with finding the minimum-trace state covariance matrix $P_{k+1|k}$ that is compatible, again in a probabilistic sense, with all possible values of the uncertainty. Its square-root $S_{k+1|k}$ is then to be used in the optimization problem at the next time instant $(k+1)$. To this end we following the same reasoning as in [378]. Assuming that the state estimate at time instant k is represented as

$$\hat{z}_{k|k-1} = z_k + S_{k|k-1} n_k,$$

we want to obtain a similar expression for time instant $k+1$

$$\hat{z}_{k+1|k} = z_{k+1} + S_{k+1|k} \tilde{n}_k, \tag{8.16}$$

with $\tilde{n}_k$ zero mean and covariance matrix equal to the identity matrix.

Pre-multiplying equation (8.15) by the non-singular matrix

$$T_l = \begin{bmatrix} C_y^\Delta & -I & 0 \\ A^\Delta & 0 & -I \\ I & 0 & 0 \end{bmatrix},$$

results in the equation

$$\begin{aligned} &\begin{bmatrix} C_y^\Delta \hat{z}_{k|k-1} + D_{yu}^\Delta u_k - y_k \\ A^\Delta \hat{z}_{k|k-1} + B_u^\Delta u_k \\ \hat{z}_{k|k-1} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} z_k \\ z_{k+1} \end{bmatrix} + \begin{bmatrix} C_y^\Delta S_{k|k-1} & R^\Delta & 0 \\ A^\Delta S_{k|k-1} & 0 & Q^\Delta \\ S_{k|k-1} & 0 & 0 \end{bmatrix} \begin{bmatrix} n_k \\ -\xi_k^y \\ -\xi_k^x \end{bmatrix}. \end{aligned} \tag{8.17}$$

Let now T_r be an orthogonal transformation matrix (*i.e.* $T_r T_r^T = I$) such that

$$\begin{bmatrix} C_y^\Delta S_{k|k-1} & R^\Delta & 0 \\ A^\Delta S_{k|k-1} & 0 & Q^\Delta \\ S_{k|k-1} & 0 & 0 \end{bmatrix} T_r T_r^T \begin{bmatrix} n_k \\ -\xi_k^y \\ -\xi_k^x \end{bmatrix} = \begin{bmatrix} \tilde{R}^\Delta & 0 & 0 \\ \tilde{G}^\Delta & S_{k+1|k}^\Delta & 0 \\ \bullet & \bullet & \bullet \end{bmatrix} \begin{bmatrix} \nu_k \\ \tilde{n}_k \\ \tilde{\xi}_k \end{bmatrix},$$

where the symbols $\bullet$ denote entries of no importance for the sequel. Note that the first row in (8.17) is independent on the variables z_k and z_{k+1} and ν_k can therefore be directly expressed, *i.e.* $\nu_k = (\tilde{R}^\Delta)^{-1}(C_y^\Delta \hat{z}_{k|k-1} + D_{yu}^\Delta u_k - y_k)$. Substituting this expression in the second row and subsequently moving the term $\tilde{G}^\Delta \nu_k$ to the left side of the equation, one gets an expression of the form (8.16). Thus $S_{k+1|k}^\Delta$ is the square-root covariance matrix which, however, depends on the uncertainty Δ. This motivates us to consider the following optimization problem for the covariance matrix

$$\begin{aligned} \min_{\gamma,\ P_{k+1|k}} \quad & \gamma \\ \text{subject to } & \gamma \geq \operatorname{tr}(P_{k+1|k}) \\ & P_{k+1|k} \succeq S^{\Delta}_{k+1|k}(S^{\Delta}_{k+1|k})^T,\ \forall \Delta \in \mathbf{\Delta}. \end{aligned} \tag{8.18}$$

For simplicity of notations we denote $M(\Delta) = S^{\Delta}_{k+1|k}(S^{\Delta}_{k+1|k})^T$.

To solve the optimization problem (8.18), we will again make use of the randomized EA approach. To this end we consider the problem of minimizing γ under the constraint that the matrix inequality

$$\begin{bmatrix} \gamma - \operatorname{tr}(P_{k+1|k}) & \\ & P_{k+1|k} - M(\Delta) \end{bmatrix} \succeq 0, \tag{8.19}$$

holds for all $\Delta \in \mathbf{\Delta}$. This problem can be rewritten in the form of the SCRLS problem, to which Algorithm 8.2 can be applied. The optimization variables here are the $\frac{1}{2}n(n+1)$ free entries of the n-by-n matrix $P_{k+1|k}$. For its initialization, an initial ellipsoid can be computed using the general (optimization-based) method described in Section 8.4. In fact, for the problem (8.19) it can be shown that Assumption 8.3 holds for any $\Delta^* \in \mathbf{\Delta}$ and $\gamma > \operatorname{tr}(M(\Delta^*))$.

Theorem 1. *Let m_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, denote the elements of the matrix $M(\Delta^*)$ for some $\Delta^* \in \mathbf{\Delta}$, and suppose that $\gamma > \operatorname{tr}(M(\Delta^*))$. Define the scalars*

$$\underline{p}_{ii} = m_{ii},\ i = 1, 2, \ldots, n,$$

$$\overline{p}_{ii} = \gamma - \sum_{j \neq i} m_{jj},\ \text{for } i = 1, 2, \ldots, n,$$

$$\overline{p}_{ij} = m_{ij} + \gamma - \operatorname{tr}(M(\Delta^*)),\ i, j = 1, 2, \ldots, n,\ j \neq i,$$

$$\underline{p}_{ij} = m_{ij} - \gamma + \operatorname{tr}(M(\Delta^*)),\ i, j = 1, 2, \ldots, n,\ j \neq i.$$

Let also $P_{k+1|k} = [p_{ij}]$ be any symmetric matrix for which (8.19) holds for all $\Delta \in \mathbf{\Delta}$. Then

$$\underline{p}_{ij} \leq p_{ij} \leq \overline{p}_{ij}, \tag{8.20}$$

for all $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$.

Proof. From (8.19) it follows that $P_{k+1|k} \succeq M(\Delta)$ for all $\Delta \in \mathbf{\Delta}$. Therefore it must also hold that $P_{k+1|k} \succeq M(\Delta^*) = [m_{ij}]$ for any fixed $\Delta^* \in \mathbf{\Delta}$. Therefore

$$p_{ii} \geq m_{ii} = \underline{p}_{ii},\ i = 1, 2, \ldots, n,$$

so that the lower bounds in (8.20) on the diagonal elements p_{ii} of $P_{k+1|k}$ has been shown.

On the other hand, $P_{k+1|k}$ should be such that $\operatorname{tr}(P_{k+1|k}) \leq \gamma$. This implies that

$$\gamma \geq p_{ii} + \sum_{j \neq i} p_{jj} \geq p_{ii} + \sum_{j=1,\ldots,n} m_{jj}, \ \forall i = 1, 2, \ldots, n, \tag{8.21}$$

so that

$$p_{ii} \leq \gamma - \sum_{j=1,\ldots,n} m_{jj} = \overline{p}_{ii}, \tag{8.22}$$

that completes the proof for the upper bounds on the diagonal elements p_{ii} of $P_{k+1|k}$.

In order to find lower and upper bounds on the non-diagonal entries we notice that $P_{k+1|k} \succeq M(\Delta^*)$ implies

$$\begin{bmatrix} p_{ii} & p_{ij} \\ p_{ji} & p_{jj} \end{bmatrix} \succeq \begin{bmatrix} m_{ii} & m_{ij} \\ m_{ji} & m_{jj} \end{bmatrix}, \ \forall i \neq j.$$

Using the Schur complement the above inequality is equivalent to

$$\left| \begin{array}{l} p_{jj} - m_{jj} \geq 0, \\ (p_{ii} - m_{ii}) - (p_{ij} - m_{ij})(p_{jj} - m_{jj})^{-1}(p_{ji} - m_{ji}) \geq 0. \end{array} \right.$$

From the second inequality, making use of the symmetry of the matrices $M(\Delta^*)$ and $P_{k+1|k}$ (*i.e.* $m_{ij} = m_{ji}$ and $p_{ij} = p_{ji}$), it follows that

$$\begin{aligned} |p_{ij} - m_{ij}| &\leq \sqrt{(p_{ii} - m_{ii})(p_{jj} - m_{jj})} \\ &\leq \sqrt{(\overline{p}_{ii} - m_{ii})(\overline{p}_{jj} - m_{jj})}, \end{aligned}$$

Substitution of equation (8.22) then results in

$$\begin{aligned} |p_{ij} - m_{ij}| &\leq \sqrt{(\overline{p}_{ii} - m_{ii})(\overline{p}_{jj} - m_{jj})} \\ &= \sqrt{(\gamma - \mathrm{tr}(M(\Delta^*)))^2} \\ &= |\gamma - \mathrm{tr}(M(\Delta^*))|. \end{aligned}$$

And since $\gamma \geq \mathrm{tr}(M(\Delta^*))$, we have shown that $P_{k+1|k} \succeq M(\Delta^*)$ implies

$$\begin{aligned} p_{ij} &\leq m_{ij} + \gamma - \mathrm{tr}(M(\Delta^*)) = \overline{p}_{ij}, \\ p_{ij} &\geq m_{ij} - \gamma + \mathrm{tr}(M(\Delta^*)) = \underline{p}_{ij}, \end{aligned}$$

so that also the upper and lower bounds on the non-diagonal elements of $P_{k+1|k}$ have been derived. □

A Faster Algorithm for Finding $P_{k+1|k}$

A more conservative, but computationally faster way to compute the covariance matrix $P_{k+1|k}$ so that it is compatible with all possible values of the uncertainty is to try to find it so that

$$P_{k+1|k} \succeq M(\Delta), \ \forall \Delta \in \boldsymbol{\Delta},$$

i.e. without the minimization over the trace of $P_{k+1|k}$ in (8.18). To this end we propose Algorithm 8.3 for computation of $P_{k+1|k}$.

Algorithm 8.3 (A faster algorithm for computation of $P_{k+1|k}$)
Initialization: small $\varepsilon > 0$, integer $K > 0$.

Step 1. Take $P^{(0)}_{k+1|k} = \varepsilon I$ and set $i = 1$.
Step 2. Set $i \leftarrow i + 1$.
Step 3. Generate a random sample $\Delta^{(i)}$ with probability distribution f_Δ.
Step 4. Compute

$$P^{(i)}_{k+1|k} = P^{(i-1)}_{k+1|k} - \left[P^{(i-1)}_{k+1|k} - M(\Delta^{(i)})\right]^{-} \tag{8.23}$$

Step 5. If $\|P^{(i)}_{k+1|k} - P^{(i-K)}_{k+1|k}\|_F = 0$ then take $P_{k+1|k} = P^{(i)}_{k+1|k}$ Stop else Goto Step 2.

The following result shows that by computing $P_{k+1|k}$ using Algorithm 8.3 ensures that $P_{k+1|k} \succeq M(\Delta^{(i)})$ (at least) for the generated uncertainty samples $\Delta^{(i)}$.

Lemma 6. *Suppose that L iterations of Algorithm 8.3 are performed. Then the matrix $P_{k+1|k} = P^{(L)}_{k+1|k}$ is such that*

(i) $P_{k+1|k} \succ 0$, and
(ii) $P_{k+1|k} \succeq M(\Delta^{(i)})$, for $i = 1, 2, \ldots, L$.

Proof.
(i) Noting that

$$\left[P^{(i-1)}_{k+1|k} - M(\Delta^{(i)})\right]^{-} \preceq 0,$$

it follows from equation (8.23) that

$$P^{(i)}_{k+1|k} \succeq P^{(i-1)}_{k+1|k} \tag{8.24}$$

for all $i = 1, \ldots, L$, and thus $P_{k+1|k} \succeq P^{(0)}_{k+1|k} \succ 0$.
(ii) Note that

$$\begin{aligned} P^{(i)}_{k+1|k} &= P^{(i-1)}_{k+1|k} - \left[P^{(i-1)}_{k+1|k} - M(\Delta^{(i)})\right]^{-} \\ &= P^{(i-1)}_{k+1|k} - M(\Delta^{(i)}) + M(\Delta^{(i)}) - \left[P^{(i-1)}_{k+1|k} - M(\Delta^{(i)})\right]^{-} \\ &= M^{\Delta^{(i)}} + \left[P^{(i-1)}_{k+1|k} - M(\Delta^{(i)})\right]^{+} \\ &\succeq M(\Delta^{(i)}) \end{aligned}$$

which, together with (8.24), implies (ii). □

Clearly, here we can also choose the parameter K of Algorithm 8.3 so as to ensure that

$$\mathbb{P}\{\mathbb{P}\{P_{k+1|k} < M(\Delta)\} \leq \epsilon\} \geq 1 - \delta$$

for some desired accuracy ϵ and confidence δ.

8.6 Conclusion

This chapter is focused on a very often encountered problem in robust control design and robust filtering, namely the robust constrained least-squares problem for general uncertainty structures. Computing deterministic solutions to such problems may be computationally prohibitive in practice, in which cases probabilistic, near-optimal solutions might be a good alternative. Such solutions can be computed using randomized algorithm, such as the randomized ellipsoid algorithm discussed in this chapter. The advantage of using this algorithm over other existing algorithms is, besides its improved convergence [176], that it can easily be initialized for SCRLS problems: analytic expressions have been provided for constructing an initial ellipsoid that contains the solution set. Furthermore, as an example, it has been shown how the robust Kalman filtering problem can be addressed in this framework by generalizing its least-squares formulation to the uncertainty case. Additional discussion is provided on the computation of the minimum-trace state covariance matrix at each time instant so that it is compatible, in a probabilistic sense, with all possible vales of the uncertainty.

9

Randomized Algorithms for Semi-Infinite Programming Problems

Vladislav B. Tadić[1], Sean P. Meyn[2], and Roberto Tempo[3]

[1] Department of Automatic Control and Systems Engineering
University of Sheffield, Sheffield S1 3JD, UK
v.tadic@sheffield.ac.uk

[2] Department of Electrical and Computer Engineering and the Coordinated
Science Laboratory, University of Illinois, Urbana-Champaign, IL 61801, USA
meyn@uiuc.edu

[3] IEIIT-CNR, Politecnico di Torino, 10129 Torino, Italy
tempo@polito.it

Summary. This chapter studies the development of Monte Carlo methods to solve semi-infinite, nonlinear programming problems. An equivalent stochastic optimization problem is proposed, which leads to a class of randomized algorithms based on stochastic approximation. The main results of the chapter show that almost sure convergence can be established under relatively mild conditions.

9.1 Introduction

In this chapter, we consider semi-infinite programming problems consisting of a possibly uncountable number of constraints. As a special case, we also study the determination of a feasible solution to an uncountable number of inequality constraints. Computational problems of this form arise in optimal and robust control, filter design, optimal experiment design, reliability, and numerous other engineering problems in which the underlying model contains a parameterized family of inequality constraints. More specifically, these parameters may represent time, frequency, or space, and hence may vary over an uncountable set.

The class of problems considered here are known as *semi-infinite programming problems* since the number of constraints is infinite, but there is a finite number of variables (see, *e.g.*, [145,270,292] and references cited therein). Several deterministic numerical procedures have been proposed to solve problems of this kind. Standard approach is to approximately solve the optimization problem through discretization using a deterministic grid (for a recent survey see [291], and also [158,269]). The algorithms based on this approach typically suffer from the curse of dimensionality so that their computational complexity is generally exponential in the problem dimension, see, *e.g.*, [364].

This chapter explores an alternative approach based on Monte Carlo methods and randomized algorithms. The use of randomized algorithms has become widespread in recent years for various problems, and is currently an active area of research. In particular, see [359] for a development of randomized algorithms for uncertain systems and robust control; [44, 54] for applications to reinforcement learning, and approximate optimal control in stochastic systems; [47, 315] for topics in mathematical physics; [234, 238] for applications in computer science and computational geometry; [144] for a treatment of Monte Carlo methods in finance; and [297] for a recent survey on Markov Chain Monte Carlo methods for approximate sampling from a complex probability distribution, and related Bayesian inference problems.

The main idea of this chapter is to reformulate the semi-infinite programming problem as a stochastic optimization problem that may be solved using stochastic approximation methods [196, 209]. The resulting algorithm can be easily implemented, and is provably convergent under verifiable conditions. The general results on Monte Carlo methods as well as the theoretical results reported in this chapter suggest that the computational complexity of the proposed algorithms is considerably reduced in comparison with existing deterministic methods (see also [296]).

The chapter is organized as follows. Section 9.2 contains a description of the general semi-infinite programming problems considered in this contribution, and an equivalent stochastic programming representation of these semi-infinite programming problems is proposed. Randomized algorithms to solve the stochastic programming problems are introduced in Section 9.3, and convergence proofs are contained in the Appendix.

9.2 Semi-Infinite Nonlinear Programming

The semi-infinite programming problems studied in this chapter are based on a given Borel-measurable function $g : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$. This function is used to define the constraint region,

$$D = \{x \in \mathbb{R}^p : g(x, y) \leq 0, \forall y \in \mathbb{R}^q\}. \tag{9.1}$$

The problems considered in this chapter concern computation of a point in D, and optimization of a given function f over this set. These problems are now described more precisely.

9.2.1 Two General Computational Problems

Constraint set feasibility

Is D non-empty? And, if so, how do we compute elements of D? That is, we seek algorithms to determine a solution $x^* \in \mathbb{R}^p$ to the following uncountably-infinite system of inequalities:

$$g(x, y) \leq 0, \quad \forall y \in \mathbb{R}^q \tag{9.2}$$

Optimization over D

For a given continuous function $f\colon \mathbb{R}^p \to \mathbb{R}$, how do we compute an optimizer over D? That is, a solution $x^* \in \mathbb{R}^p$ to the semi-infinite nonlinear program,

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to} \quad g(x,y) \le 0, \quad \forall y \in \mathbb{R}^q. \end{aligned} \tag{9.3}$$

These two problems cover the following general examples.

Min-max problems

Consider the general optimization problem in which a function $f\colon \mathbb{R}^p \to \mathbb{R}$ is to be minimized, of the specific form

$$f(x) = \max_y g(x,y), \quad x \in \mathbb{R}^p,$$

where $g\colon \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is continuous. Under mild conditions, this optimization problem can be formulated as a semi-infinite optimization problem (for details see [270]).

Sup-norm minimization

Suppose that $H\colon \mathbb{R}^p \to \mathbb{R}^r$ is a given measurable function to be approximated by a family of functions $\{G_i : i = 1, \ldots, p\}$ and their linear combinations. Let $G = [G_1|\cdots|G_p]$ denote the $p \times r$ matrix-valued function on $\mathbb{R}^q$, and consider the minimization of the function

$$f(x) = \sup_y \|H(y) - G(y)x\|, \quad x \in \mathbb{R}^p.$$

A vector x^* minimizing f provides the best approximation of H in the supremum norm. The components of x^* are then interpreted as basis weights. This is clearly a special case of the min-max problem in which $g(x,y) = \|H(y) - G(y)x\|$.

Common Lyapunov functions

Consider a set of parameterized real Hurwitz matrices $\mathcal{A} = \{A(y) : y \in Y \subseteq \mathbb{R}^q\}$. In this case, the feasibility problem is checking the existence of a symmetric positive definite matrix $P \succ 0$ which satisfies the Lyapunov strict inequalities

$$PA(y) + A^T(y)P \prec 0, \quad \forall y \in Y \subseteq \mathbb{R}^q.$$

This is equivalent to verify the existence of $P \succ 0$ which satisfies

$$PA(y) + A^T(y)P + Q \preceq 0, \quad \forall y \in Y \subseteq \mathbb{R}^q,$$

where $Q \succ 0$ is arbitrary. Clearly, the existence of a feasible solution $P \succ 0$ implies that the quadratic function $V(x) = x^T Px$ is a common Lyapunov function for the family of asymptotically stable linear systems

$$\dot{z} = A(y)z, \quad \forall y \in Y \subseteq \mathbb{R}^q.$$

This feasibility problem can be reformulated as follows: determine the existence of a symmetric positive definite matrix $P \succ 0$ to the system of scalar inequalities

$$\lambda_{\max}(PA(y) + A^T(y)P + Q) \preceq 0, \quad \forall y \in Y \subseteq \mathbb{R}^q,$$

where $\lambda_{\max}(A)$ denotes the largest eigenvalue of a real symmetric matrix A. This observation follows from the fact that $\lambda_{\max}(A) \leq 0$ if and only if $A \preceq 0$. We also notice that $\lambda_{\max}$ is a convex function and, if $\lambda_{\max}$ is a simple eigenvalue, it is also differentiable, see details in [206]. In the affirmative case when this common Lyapunov problem is feasible, clearly the objective is to find a solution $P \succ 0$.

9.2.2 Equivalent Stochastic Programming Representation

Algorithms to solve the semi-infinite programming problems (9.2), (9.3) may be constructed based on an equivalent stochastic programming representation. This section contains details on this representation and some key assumptions.

We adopt the following notation throughout the chapter: the standard Euclidean norm is denoted $\|\cdot\|$, while $d(\cdot,\cdot)$ stands for the associated metric. For an integer $r \geq 1$ and $z \in \mathbb{R}^r$, $\rho \in (0,\infty)$, the associated closed balls are defined as

$$B^r_\rho(z) = \{x' \in \mathbb{R}^r : \|z - z'\| \leq \rho\}, \quad B^r_\rho = B^r_\rho(0),$$

while $\mathcal{B}^r$ denotes the class of Borel-measurable sets on $\mathbb{R}^r$.

Throughout the chapter, a probability measure μ on $\mathcal{B}^q$ is fixed, where $q \geq 1$ is the integer used in (9.1). It is assumed that its support is full in the sense that

$$\mu(A) > 0 \text{ for any non-empty open set } A \subset \mathbb{R}^q. \tag{9.4}$$

In applications one will typically take μ of the form $\mu(dy) = p(y)\,dy$, $y \in \mathbb{R}^q$, where p is continuous, and strictly positive. A continuous function $h \colon \mathbb{R} \to \mathbb{R}_+$ is fixed with support equal to $(0,\infty)$, in the sense that

$$h(t) = 0 \text{ for all } t \in (-\infty, 0], \text{ and } h(t) > 0 \text{ for all } t \in (0,\infty). \tag{9.5}$$

For example, the function $h(t) = (\max\{0,t\})^2$, $t \in \mathbb{R}$, is a convex, C^1 solution to (9.5).

Equivalent stochastic programming representations of (9.2) and (9.3) are based on the probability distribution μ, the function h, and the following condition on the function g that determines the constraint region D:

$$g(x,\cdot) \text{ is continuous on } \mathbb{R}^q \text{ for each } x \in \mathbb{R}^p.$$

The following conditional average of g is the focus of the algorithms and results in this chapter,

$$\psi(x) = \int h(g(x,y))\mu(dy), \quad x \in \mathbb{R}^p. \tag{9.6}$$

The equivalent stochastic programming representation of the semi-infinite problems (9.2) or (9.3) is based on the following theorem.

Theorem 1. *Under the assumptions of this section, the constraint region may be expressed as*

$$D = \{x \in \mathbb{R}^p : \psi(x) = 0\}.$$

Proof. Suppose that $x \in D$. By definition, the following equation then holds for all $y \in \mathbb{R}^q$:

$$h(g(x,y)) = 0.$$

This and (9.6) establish the inclusion

$$D \subseteq \{x \in \mathbb{R}^p : \psi(x) = 0\}.$$

Conversely, if $x \notin D$, then there exists $y \in \mathbb{R}^q$ such that $g(x,y) > 0$. Continuity of the functions g and h implies that there exist constants $\delta, \varepsilon \in (0,\infty)$ such that

$$h(g(x,y')) \geq \varepsilon, \quad \text{for all } y' \in B_\delta^q(y).$$

This combined with the support assumption (9.4) implies that

$$\psi(x) \geq \int_{B_\delta^q(y)} h(g(x,y'))\mu(dy') \geq \varepsilon\mu(B_\delta^q(y)) > 0,$$

which gives the reverse inclusion

$$D^c \subseteq \{x \in \mathbb{R}^p : \psi(x) \neq 0\},$$

where D^c denotes the complement of D. □

Let Y be an $\mathbb{R}^q$-valued random variable defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ whose probability measure is μ, *i.e.*,

$$\mathbb{P}(Y \in B) = \mu(B), \quad B \in \mathcal{B}^q.$$

It follows that ψ may be expressed as the expectation

$$\psi(x) = \mathbb{E}(h(g(x,Y))), \quad x \in \mathbb{R}^p, \tag{9.7}$$

and the following corollaries are then a direct consequence of Theorem 1.

Corollary 1. *A vector $x \in \mathbb{R}^p$ solves the semi-infinite problem (9.2) if and only if it solves the equation*

$$\mathbb{E}(h(g(x, Y))) = 0.$$

Corollary 2. *The semi-infinite optimization problem problem (9.3) is equivalent to the constrained stochastic optimization problem*

$$\begin{aligned}&\text{minimize } f(x)\\ &\text{subject to} \quad \mathbb{E}(h(g(x, Y))) = 0.\end{aligned}$$

Corollaries 1 and 2 motivate the development of Monte Carlo methods to solve the semi-infinite problems (9.2) and (9.3). The search for a feasible or optimal $x \in D$ may be performed by sampling $\mathbb{R}^q$ using the probability measure μ. Specific algorithms are proposed in the next section.

9.3 Algorithms

We begin with consideration of the constraint satisfaction problem (9.2).

9.3.1 Systems of Infinitely Many Inequalities

Theorem 1 implies that solutions of (9.2) are characterized as global minima for the function ψ. If the functions h and g are differentiable, then minimization of ψ may be performed using a gradient algorithm, described as follows: given a vanishing sequence $\{\gamma_n\}_{n \geq 1}$ of positive reals, we consider the recursion

$$x_{n+1} = x_n - \gamma_{n+1} \nabla \psi(x_n), \quad n \geq 0.$$

Unfortunately, apart from some special cases (see, *e.g.*, [273]), it is impossible to determine analytically the gradient $\nabla \psi$.

On the other hand, under mild regularity conditions, (9.7) implies that the gradient may be expressed as the expectation

$$\nabla \psi(x) = \mathbb{E}(h'(g(x, Y)) \nabla_x g(x, Y)), \quad x \in \mathbb{R}^p, \tag{9.8}$$

where h' denotes the derivative of h. This provides motivation for the 'stochastic approximation' of $\nabla \psi$ given by

$$h'(g(x, Y)) \nabla_x g(x, Y),$$

and the following stochastic gradient algorithm to search for the minima of ψ:

$$X_{n+1} = X_n - \gamma_{n+1} h'(g(X_n, Y_{n+1})) \nabla_x g(X_n, Y_{n+1}), \quad n \geq 0. \tag{9.9}$$

In this recursion, $\{\gamma_n\}_{n\geq 1}$ again denotes a sequence of positive reals. The i.i.d. sequence $\{Y_n\}_{n\geq 1}$ has common marginal distribution μ, so that

$$\mathbb{P}(Y_n \in B) = \mu(B), \quad B \in \mathcal{B}^q, \ n \geq 1.$$

Depending upon the specific assumptions imposed on the functions h and g, analysis of the stochastic approximation recursion (9.9) may be performed following the general theory of, say, [38, 54, 196, 355].

The asymptotic behavior of the algorithm (9.9) is analyzed under the following assumptions:

A1 $\gamma_n > 0$ *for each* $n \geq 1$, $\sum_{n=1}^{\infty} \gamma_n = \infty$, *and* $\sum_{n=1}^{\infty} \gamma_n^2 < \infty$.

A2 *For each* $\rho \in [1, \infty)$, *there exists a Borel-measurable function* $\phi_\rho : \mathbb{R}^q \to [1, \infty)$ *such that*

$$\int \phi_\rho^4(y)\mu(dy) < \infty,$$

and for each $x, x', x' \in B_\rho^p$, $y \in \mathbb{R}^q$,

$$\begin{aligned} \max\{|h(g(x,y))|, |h'(g(x,y))|, \|\nabla_x g(x,y)\|\} &\leq \phi_\rho(y), \\ |h'(g(x',y)) - h'(g(x',y))| &\leq \phi_\rho(y)\|x' - x'\|, \\ \|\nabla_x g(x',y) - \nabla_x g(x',y)\| &\leq \phi_\rho(y)\|x' - x'\|. \end{aligned}$$

A3 $\nabla\psi(x) \neq 0$ *for all* $x \notin D$.

Assumption A1 holds if the step-size sequence is of the usual form $\gamma_n = n^{-c}$, $n \geq 1$, where the constant c lies in the interval $(1/2, 1]$.

Assumption A2 corresponds to the properties of the functions g and h. It ensures that ψ is well-defined, finite and differentiable, and that $\nabla\psi$ is locally Lipschitz continuous. This assumption is satisfied under appropriate assumptions on the function g, provided the function h is carefully chosen. Consider the special case in which h is the piecewise quadratic, $h(t) = (\max\{0, t\})^2$, $t \in \mathbb{R}$. Then, Assumption A2 holds under the following general assumptions on g: for each $\rho \in [1, \infty)$, there exists a Borel-measurable function $\varphi_\rho : \mathbb{R}^q \to [1, \infty)$ such that

$$\int \varphi_\rho^4(y)\mu(dy) < \infty,$$

and for each $x, x', x' \in B_\rho^p$, $y \in \mathbb{R}^q$,

$$\begin{aligned} \max\{g^2(x,y), \|\nabla_x g(x,y)\|\} &\leq \varphi_\rho(y), \\ |g(x',y) - g(x',y)| &\leq \varphi_\rho(y)\|x' - x'\|, \\ \|\nabla_x g(x',y) - \nabla_x g(x',y)\| &\leq \varphi_\rho(y)\|x' - x'\|. \end{aligned}$$

In the special case in which g is linear in x, so that there exist Borel-measurable functions $a : \mathbb{R}^q \to \mathbb{R}^p$, $b : \mathbb{R}^q \to \mathbb{R}$, with

$$g(x, y) = a^T(y)x + b(y), \quad x \in \mathbb{R}^p,\ y \in \mathbb{R}^q,$$

a bounding function φ_ρ may be constructed for each $\rho \in [1, \infty)$ provided

$$\int \|a(y)\|^4 \mu(dy) < \infty, \quad \int |b(y)|^4 \mu(dy) < \infty.$$

Assumption A3 corresponds to the properties of the stationary points of ψ. Consider the important special case in which $g(\cdot, y)$ is convex for each $y \in \mathbb{R}^q$. We may assume that h is convex and non-decreasing (notice that h is non-decreasing if it is convex and satisfies (9.5)), and it then follows that the function ψ is also convex in this special case. Moreover, since ψ is non-negative valued, convex, and vanishes only on D, it follows that $\nabla\psi(x) \neq 0$ for $x \in D^c$, so that Assumption A3 holds.

Theorem 2 states that stability of the algorithm implies convergence under the assumptions imposed here. General conditions to verify stability, so that $\sup_{0\leq n} \|X_n\| < \infty$ holds w.p.1., are included in [38, 54, 196].

Theorem 2. *Suppose that Assumptions A1–A3 hold. Then, on the event* $\{\sup_{0\leq n} \|X_n\| < \infty\}$, *we have convergence:*

$$\lim_{n\to\infty} d(X_n, D) = 0 \quad \textit{w.p.1.}$$

Proof. A proof of this theorem is included in Appendix 9.5.1.

In many practical situations, a solution to the semi-infinite problem (9.2) is known to lie in a predetermined bounded set Q. In this case one may replace the iteration (9.9) with the following projected stochastic gradient algorithm:

$$X_{n+1} = \Pi_Q(X_n - \gamma_{n+1}h'(g(X_n, Y_{n+1}))\nabla_x g(X_n, Y_{n+1})), \quad n \geq 0. \tag{9.10}$$

It is assumed that the constraint set $Q \subset \mathbb{R}^p$ is compact and convex, and $\Pi_Q(\cdot)$ is the projection on Q (*i.e.*, $\Pi_Q(x) = \arg\inf_{x'\in Q} \|x - x'\|$ for $x \in \mathbb{R}^p$). The step-size sequence and i.i.d. sequence $\{Y_n\}_{n\geq 1}$ are defined as above.

Under additional assumptions on g and h, we can prove that the algorithm (9.10) converges.

Theorem 3. *Let* $\{X_n\}_{n\geq 0}$ *be generated by (9.10), and suppose that Assumptions A1 and A2 hold. Suppose that* $D \cap Q \neq \emptyset$, *h is convex, and* $g(\cdot, y)$ *is convex for each* $y \in \mathbb{R}^q$. *Then*

$$\lim_{n\to\infty} d(X_n, D) = 0 \quad w.p.1.$$

Proof. A proof of this theorem is included in Appendix 9.5.2.

9.3.2 Algorithms for Semi-Infinite Optimization Problems

In this section, algorithms for the semi-infinite programming problem (9.3) are proposed, and their asymptotic behavior is analyzed.

Suppose that h is differentiable and that $g(\cdot, y)$ is differentiable for each $y \in \mathbb{R}^q$. Due to Theorem 1, the semi-infinite problem (9.3) is equivalent to the following constrained optimization problem:

$$\begin{aligned}&\text{minimize } f(x)\\ &\text{subject to} \quad \psi(x) = 0.\end{aligned} \tag{9.11}$$

Suppose that the gradient $\nabla\psi$ could be computed explicitly. Then, under general conditions on the function f and the set D, the constrained optimization problem (9.3) could be solved using the following penalty-function approach (see, *e.g.*, [42, 270]). Let $\{\delta_n\}_{n\geq 1}$ be an increasing sequence of positive reals satisfying $\lim_{n\to\infty} \delta_n = \infty$. Since $\psi(x) \geq 0$ for all $x \in \mathbb{R}^p$, this sequence can be used as penalty parameters for (9.11) in the following gradient algorithm:

$$x_{n+1} = x_n - \gamma_{n+1}(\nabla f(x_n) + \delta_{n+1}\psi(x_n)), \quad n \geq 0,$$

where $\{\gamma_n\}_{n\geq 1}$ is a sequence of positive reals.

However, since the gradient is typically unavailable, we may instead use (9.8) to obtain the estimate of $\nabla\psi$, given by $h'(g(x,Y))\nabla_x g(x,Y)$, and it is then quite natural to use the following stochastic gradient algorithm to search for the minima of the function f over D:

$$X_{n+1} = X_n - \gamma_{n+1}(\nabla f(X_n) + \delta_{n+1} h'(g(X_n, Y_{n+1}))\nabla_x g(X_n, Y_{n+1})), \quad n \geq 0, \tag{9.12}$$

where $\{\gamma_n\}_{n\geq 1}$ and $\{Y_n\}_{n\geq 1}$ have the same meaning as in the case of the algorithm (9.9).

The following assumptions are required in the analysis of the algorithm (9.12):

B1 $\gamma_n > 0$ *for each* $n \geq 1$, $\sum_{n=1}^{\infty} \gamma_n = \infty$, *and* $\sum_{n=1}^{\infty} \gamma_n^2 \delta_n^2 < \infty$.

B2 f *is convex and* ∇f *is locally Lipschitz continuous.*

B3 h *is convex and* $g(\cdot, y)$ *is convex for each* $y \in \mathbb{R}^q$. *For all* $\rho \in [1, \infty)$, *there exists a Borel-measurable function* $\phi_\rho : \mathbb{R}^q \to [1, \infty)$ *and such that*

$$\int \phi_\rho^4(y)\mu(dy) < \infty,$$

and, for all $x, x', x' \in B_\rho^p$, $y \in \mathbb{R}^q$,

$$\max\{|h(g(x,y))|, |h'(g(x,y))|, \|\nabla_x g(x,y)\|\} \leq \phi_\rho(y),$$

$$|h'(g(x',y)) - h'(g(x',y))| \leq \phi_\rho(y)\|x' - x'\|,$$

$$\|\nabla_x g(x',y) - \nabla_x g(x',y)\| \leq \phi_\rho(y)\|x' - x'\|.$$

B4 *$\eta^* \doteq \inf_{x \in D} f(x) > -\infty$, and the set of optimizers given by $D^* \doteq \{x \in D : f(x) = \eta^*\}$ is non-empty.*

Assumption B3 ensures that ψ is well-defined, finite, convex and differentiable. It also implies that $\nabla\psi$ is locally Lipschitz continuous. Assumption B4 is satisfied if D is bounded or f is coercive (*i.e.* the sublevel set $\{x : f(x) \leq N\}$ is bounded for each $N \geq 1$).

Theorem 4. *Let $\{X_n\}_{n\geq 0}$ be generated by (9.12), and suppose that Assumptions B1–B4 hold. Then, on the event $\{\sup_{0\leq n} \|X_n\| < \infty\}$,*

$$\lim_{n\to\infty} d(X_n, D^*) = 0 \quad \textit{w.p.1.}$$

Proof. A proof of this theorem is included in Appendix 9.5.3.

9.4 Conclusion

The main contribution of this chapter is to reformulate a given semi-infinite program as a stochastic optimization problem. One can then apply Monte Carlo and stochastic approximation methods to generate efficient algorithms, and provide a foundation for analysis. Under standard convexity assumptions, and additional relatively mild conditions, the proposed algorithms provide a convergent solution with probability one.

The next step is to test these algorithms in practical, non-trivial applications. In particular, we are interested in application to specific optimization problems, and to robust control. It is of interest to see how these randomization approaches compare to their deterministic counterparts. We also expect that the algorithms may be refined and improved within a particular application context.

9.5 Appendix

Here we provide proofs of the main results of the chapter. The following notation is fixed in this Appendix: let $\mathcal{F}_0 = \sigma\{X_0\}$, while $\mathcal{F}_n = \sigma\{X_0, Y_1, \ldots, Y_n\}$ for $n \geq 1$.

9.5.1 Proof of Theorem 2

We begin with the representation

$$\begin{aligned} X_{n+1} &= X_n - \gamma_{n+1}\nabla\psi(X_n) + \xi_{n+1}, \\ \psi(X_{n+1}) &= \psi(X_n) - \gamma_{n+1}\|\nabla\psi(X_n)\|^2 + \varepsilon_{n+1}, \quad n \geq 1, \end{aligned} \tag{9.13}$$

where the error terms in (9.13) are defined as

$$\xi_{n+1} = \gamma_{n+1}(\nabla\psi(X_n) - h'(g(X_n, Y_{n+1})\nabla_x g(X_n, Y_{n+1})),$$

$$\begin{aligned}
\varepsilon_{1,n+1} &\doteq (\nabla\psi(X_n))^T \xi_{n+1}, \\
\varepsilon_{2,n+1} &\doteq \int_0^1 (\nabla\psi(X_n + t(X_{n+1} - X_n)) - \nabla\psi(X_n))^T (X_{n+1} - X_n) dt, \\
\varepsilon_{n+1} &\doteq \varepsilon_{1,n+1} + \varepsilon_{2,n+1}, \quad n \geq 0.
\end{aligned}$$

The first step in our analysis of (9.13) is to establish the asymptotic properties of $\{\xi_n\}_{n\geq 1}$ and $\{\varepsilon_n\}_{n\geq 1}$.

Lemma 1. *Suppose that Assumptions A1 and A2 hold. Then, $\sum_{n=1}^{\infty} \xi_n$, $\sum_{n=1}^{\infty} \varepsilon_n$ converge w.p.1 on the event $\{\sup_{0\leq n} \|X_n\| < \infty\}$.*

Fix $\rho \in [1, \infty)$, and let $K_\rho < \infty$ serve as an upper bound on $\|\nabla\psi\|$, and a Lipschitz constant for $\nabla\psi$ on the set B_ρ^p. Due to A1,

$$\begin{aligned}
\|\xi_{n+1}\| I_{\{\|X_n\| \leq \rho\}} &\leq 2K_\rho \gamma_{n+1} \phi_\rho^2(Y_{n+1}) \\
|\varepsilon_{1,n+1}| I_{\{\|X_n\| \leq \rho\}} &\leq K_\rho \|\xi_{n+1}\| I_{\{\|X_n\| \leq \rho\}}, \\
|\varepsilon_{2,n+1}| I_{\{\|X_n\| \leq \rho\}} &\leq K_\rho \|X_{n+1} - X_n\|^2 I_{\{\|X_n\| \leq \rho\}} \\
&\leq 2K_\rho^3 \gamma_{n+1}^2 + 2K_\rho \|\xi_{n+1}\| I_{\{\|X_n\| \leq \rho\}}, \quad n \geq 0.
\end{aligned}$$

Consequently,

$$\mathbb{E}\left(\sum_{n=0}^{\infty} \|\xi_{n+1}\|^2 I_{\{\|X_n\| \leq \rho\}}\right) \leq 4K_\rho^2 \sum_{n=1}^{\infty} \gamma_n^2 \mathbb{E}(\phi_\rho^4(Y_n)) < \infty, \tag{9.14}$$

$$\mathbb{E}\left(\sum_{n=0}^{\infty} |\varepsilon_{1,n+1}|^2 I_{\{\|X_n\| \leq \rho\}}\right) \leq K_\rho^2 \mathbb{E}\left(\sum_{n=0}^{\infty} \|\xi_{n+1}\|^2 I_{\{\|X_n\| \leq \rho\}}\right) < \infty, \tag{9.15}$$

$$\begin{aligned}
\mathbb{E}\left(\sum_{n=0}^{\infty} |\varepsilon_{2,n+1}|^2 I_{\{\|X_n\| \leq \rho\}}\right) \leq& 2K_\rho \mathbb{E}\left(\sum_{n=0}^{\infty} \|\xi_{n+1}\|^2 I_{\{\|X_n\| \leq \rho\}}\right) \\
&+ 2K_\rho^3 \sum_{n=1}^{\infty} \gamma_n^2 < \infty.
\end{aligned} \tag{9.16}$$

Since X_n is measurable with respect to $\mathcal{F}_n$ and independent of Y_{n+1}, we have

$$\begin{aligned}
\mathbb{E}\left(\xi_{n+1} \|\xi_{n+1}\|^2 I_{\{\|X_n\| \leq \rho\}} | \mathcal{F}_n\right) &= 0 \textit{ w.p.1.}, \\
\mathbb{E}\left(\varepsilon_{1,n+1} \|\xi_{n+1}\|^2 I_{\{\|X_n\| \leq \rho\}} | \mathcal{F}_n\right) &= 0 \textit{ w.p.1.}, \quad n \geq 0.
\end{aligned}$$

Then, Doob's martingale convergence theorem (see, *e.g.*, [243]) and (9.14)–(9.16) implies that $\sum_{n=1}^{\infty} \xi_n$, $\sum_{n=1}^{\infty} \varepsilon_{1,n}$, $\sum_{n=1}^{\infty} \varepsilon_{2,n}$ converge w.p.1 on the event $\{\sup_{0\leq n} \|X_n\| \leq \rho\}$. Since ρ can be arbitrary large, it can easily be deduced that $\sum_{n=1}^{\infty} \xi_n$, $\sum_{n=1}^{\infty} \varepsilon_n$ converge w.p.1 on the event $\{\sup_{0\leq n} \|X_n\| < \infty\}$. □

Proof of Theorem 2. We again fix $\rho \in [1, \infty)$, and let $K_\rho < \infty$ serve as an upper bound on $\|\nabla\psi\|$, and a Lipschitz constant for $\nabla\psi$ on the set B_ρ^p. Fix an arbitrary sample $\omega \in \Omega$ from the event on which $\sup_{0\leq n}\|X_n\| \leq \rho$, and both $\sum_{n=1}^\infty \xi_n$ and $\sum_{n=1}^\infty \varepsilon_n$ are convergent (for the sake of notational simplicity, ω does not explicitly appear in the relations which follow in the proof). Due to Lemma 1 and the fact that $\nabla\psi$ is continuous, it is sufficient to show $\lim_{n\to\infty}\|\nabla\psi(X_n)\| = 0$. We proceed by contradiction.

If $\|\nabla\psi(X_n)\|$ does not vanish as $n \to \infty$, then we may find $\varepsilon > 0$ such that

$$\limsup_{n\to\infty}\|\nabla\psi(X_n)\| > 2\varepsilon.$$

On the other hand, (9.13) yields

$$\sum_{i=0}^{n-1}\gamma_{i+1}\|\nabla\psi(X_i)\|^2 = \psi(X_0) - \psi(X_n) + \sum_{i=1}^{n}\xi_i \leq K_\rho + \sum_{i=1}^{n}\xi_i, \quad n \geq 1.$$

Consequently,

$$\sum_{n=0}^{\infty}\gamma_{n+1}\|\nabla\psi(X_n)\|^2 < \infty, \tag{9.17}$$

and A1 then implies that $\liminf_{n\to\infty}\|\nabla\psi(X_n)\| = 0$. Otherwise, there would exist $\delta \in (0,\infty)$ and $j_0 \geq 1$ (both depending on ω) such that $\|\nabla\psi(X_n)\| \geq \delta$, $n \geq j_0$, which combined with A1 would yield

$$\sum_{n=0}^{\infty}\gamma_{n+1}\|\nabla\psi(X_n)\|^2 \geq \delta^2\sum_{n=j_0+1}^{\infty}\gamma_n = \infty.$$

Let $m_0 = n_0 = 0$ and

$$m_{k+1} = \{n \geq n_k : \|\nabla\psi(X_n)\| \geq 2\varepsilon\},$$

$$n_{k+1} = \{n \geq m_{k+1} : \|\nabla\psi(X_n)\| \leq \varepsilon\}, \quad k \geq 0.$$

Obviously, $\{m_n\}_{k\geq 0}$, $\{n_k\}_{k\geq 0}$ are well-defined, finite, and satisfy $m_k < n_k < m_{k+1}$ for $k \geq 1$. Moreover,

$$\|\nabla\psi(X_{m_k})\| \geq 2\varepsilon, \quad \|\nabla\psi(X_{n_k})\| \leq \varepsilon, \quad k \geq 1,$$

and

$$\|\nabla\psi(X_n)\| \geq \varepsilon, \quad \text{for } m_k \leq n < n_k,\ k \geq 1. \tag{9.18}$$

Due to (9.17), (9.18),

$$\varepsilon^2\sum_{k=1}^{\infty}\sum_{i=m_k}^{n_k-1}\gamma_{i+1} \leq \sum_{k=1}^{\infty}\sum_{i=m_k}^{n_k-1}\gamma_{i+1}\|\nabla\psi(X_i)\|^2 \leq \sum_{n=0}^{\infty}\gamma_{n+1}\|\nabla\psi(X_n)\|^2 < \infty.$$

Therefore,

$$\lim_{k\to\infty} \sum_{i=m_k+1}^{n_k} \gamma_i = 0, \tag{9.19}$$

while (9.13) yields, for each $k \geq 1$,

$$\begin{aligned} \varepsilon \leq \|\nabla\psi(X_{n_k}) - \nabla\psi(X_{m_k})\| \leq K_\rho \|X_{n_k} - X_{m_k}\| \\ = K_\rho \left\| -\sum_{i=m_k}^{n_k-1} \gamma_{i+1}\nabla\psi(X_i) + \sum_{i=m_k+1}^{n_k} \xi_i \right\| \\ \leq K_\rho^2 \sum_{i=m_k+1}^{n_k} \gamma_i + \left\| \sum_{i=m_k+1}^{n_k} \xi_i \right\|. \end{aligned} \tag{9.20}$$

However, this is not possible, since (9.19) and the limit process $k \to \infty$ applied to (9.20) yield $\varepsilon \leq 0$. Hence, $\lim_{n\to\infty} \|\nabla\psi(X_n)\| = 0$. This completes the proof.

9.5.2 Proof of Theorem 3

Let $C = D \cap Q$, and let $\Pi_C(\cdot)$ denote the projection operator onto the set C. Moreover, the sequence $\{\xi_n\}_{n\geq 0}$ has the same meaning as in the previous section, while

$$\begin{aligned} Z_{n+1} &= X_n - \gamma_{n+1} h'(g(X_n, Y_{n+1}))\nabla_x g(X_n, Y_{n+1}), \\ \varepsilon_{1,n+1} &= 2(X_n - \Pi_C(X_n))^T \xi_{n+1}, \\ \varepsilon_{2,n+1} &= \|Z_{n+1} - X_n\|^2, \\ \varepsilon_{n+1} &= \varepsilon_{1,n+1} + \varepsilon_{2,n+1}, \quad n \geq 0. \end{aligned}$$

Since ψ is convex (under the conditions of Theorem 3) and $\Pi_C(\cdot)$, $\Pi_Q(\cdot)$ are non-expansive, we have

$$\begin{aligned} (X_n - \Pi_C(X_n))^T \nabla\psi(X_n) &\geq \psi(X_n) - \psi(\Pi_C(X_n)) = \psi(X_n), \\ \|X_{n+1} - \Pi_C(X_{n+1})\| &\leq \|X_{n+1} - \Pi_C(X_n)\| \\ &= \|\Pi_Q(Z_{n+1}) - \Pi_Q(\Pi_C(X_n))\| \\ &\leq \|Z_{n+1} - \Pi_C(X_n)\| \end{aligned}$$

for $n \geq 0$. Then, it is straightforward to demonstrate that for all $n \geq 0$,

$$Z_{n+1} = X_n - \gamma_{n+1}\nabla\psi(X_n) + \xi_{n+1}, \tag{9.21}$$

and moreover,

$$\begin{aligned}\|X_{n+1}-&\varPi_C(X_{n+1})\|^2\\ &\leq\|(X_n-\varPi_C(X_n))+(Z_{n+1}-X_n)\|^2\\ &=\|X_n-\varPi_C(X_n)\|^2+2(X_n-\varPi_C(X_n))^T(Z_{n+1}-X_n)\\ &\quad+\|Z_{n+1}-X_n\|^2\\ &=\|X_n-\varPi_C(X_n)\|^2-2\gamma_{n+1}(X_n-\varPi_C(X_n))^T\nabla\psi(X_n)+\varepsilon_{n+1}\\ &\leq\|X_n-\varPi_C(X_n)\|^2-2\gamma_{n+1}\psi(X_n)+\varepsilon_{n+1}.\end{aligned}\tag{9.22}$$

Lemma 2. *Suppose that Assumptions A1 and A2 hold. Then,* $\sum_{n=1}^{\infty}\xi_n$, $\sum_{n=1}^{\infty}\varepsilon_n$ *converge w.p.1.*

Proof. Let $K\in[1,\infty)$ denote an upper bound of $\|\cdot\|$, $\|\varPi_C(\cdot)\|$, $\|\nabla\psi\|$ on Q. Due to A2,

$$\begin{aligned}\|\xi_{n+1}\|&\leq 2K\phi_K^2(Y_{n+1}),\\ |\varepsilon_{1,n+1}|&\leq 4K\|\xi_{n+1}\|,\\ |\varepsilon_{2,n+1}|&\leq 2K^2\gamma_{n+1}^2+2K\|\xi_{n+1}\|^2,\quad n\geq 0,\end{aligned}$$

and this implies the following bounds:

$$\begin{aligned}\mathbb{E}\left(\sum_{n=1}^{\infty}\|\xi_n\|^2\right)&\leq 4K^2\sum_{n=1}^{\infty}\gamma_n^2\mathbb{E}(\phi_K^2(Y_n))<\infty,\\ \mathbb{E}\left(\sum_{n=1}^{\infty}\|\varepsilon_{1,n}\|^2\right)&\leq 16K^2\mathbb{E}\left(\sum_{n=1}^{\infty}\|\xi_n\|^2\right)<\infty,\\ \mathbb{E}\left(\sum_{n=1}^{\infty}\|\varepsilon_{2,n}\|^2\right)&\leq 2K^2\sum_{n=1}^{\infty}\gamma_n^2+2\mathbb{E}\left(\sum_{n=1}^{\infty}\|\xi_n\|^2\right)<\infty.\end{aligned}$$

Then, using the same arguments as in the proof of Lemma 1, it can easily be deduced that $\sum_{n=1}^{\infty}\xi_n$, $\sum_{n=1}^{\infty}\varepsilon_n$ converge *w.p.1.* □

Proof of Theorem 3. Let $K\in[1,\infty)$ denote a simultaneous upper bound for $\|\cdot\|$, $\|\varPi_C(\cdot)\|$, $\|\nabla\psi\|$ on the set Q, and a Lipschitz constant for ψ on the same set. Moreover, let ω be an arbitrary sample from the event where $\sum_{n=1}^{\infty}\xi_n$, $\sum_{n=1}^{\infty}\varepsilon_n$ converge (for the sake of notational simplicity, ω does not explicitly appear in the relations which follow in the proof). Due to Lemma 2 and the fact that ψ is continuous and strictly positive on $Q\backslash D$, it is sufficient to show $\lim_{n\to\infty}\psi(X_n)=0$. Suppose the opposite. Then, there exists $\varepsilon\in(0,\infty)$ (depending on ω) such that $\limsup_{n\to\infty}\psi(X_n)>2\varepsilon$. On the other hand, (9.22) yields

$$\begin{aligned}\sum_{i=0}^{n-1}\gamma_{i+1}\psi(X_i)\leq&\|X_0-\varPi_C(X_0)\|^2-\|X_n-\varPi_C(X_n)\|^2\\ &+\sum_{i=1}^{n}\xi_i\leq K+\sum_{i=1}^{n}\xi_i,\quad n\geq 1.\end{aligned}$$

Consequently,

$$\sum_{n=0}^{\infty} \gamma_{n+1} \psi(X_n) < \infty. \tag{9.23}$$

Then, A1 implies $\liminf_{n\to\infty} \psi(X_n) = 0$. Otherwise, there would exist $\delta \in (0, \infty)$ and $j_0 \geq 1$ (both depending on ω) such that $\psi(X_n) \geq \delta$, $n \geq j_0$, which combined with A1 would yield

$$\sum_{n=0}^{\infty} \gamma_{n+1} \psi(X_n) \geq \delta \sum_{n=j_0+1}^{\infty} \gamma_n = \infty.$$

Let $m_0 = n_0 = 0$ and

$$\begin{aligned} m_{k+1} &= \{n \geq n_k : \psi(X_n) \geq 2\varepsilon\}, \\ n_{k+1} &= \{n \geq m_{k+1} : \psi(X_n) \leq \varepsilon\} \end{aligned}$$

for $k \geq 0$. Obviously, $\{m_n\}_{k\geq 0}$, $\{n_k\}_{k\geq 0}$ are well-defined, finite and satisfy $m_k < n_k < m_{k+1}$ for $k \geq 1$. Moreover,

$$\psi(X_{m_k}) \geq 2\varepsilon, \quad \psi(X_{n_k}) \leq \varepsilon \tag{9.24}$$

for $k \geq 1$, and

$$\psi(X_n) \geq \varepsilon, \quad \text{for } m_k \leq n < n_k,\ k \geq 0. \tag{9.25}$$

Due to (9.23), (9.25),

$$\varepsilon^2 \sum_{k=1}^{\infty} \sum_{i=m_k}^{n_k - 1} \gamma_{i+1} \leq \sum_{k=1}^{\infty} \sum_{i=m_k}^{n_k-1} \gamma_{i+1} \psi(X_i) \leq \sum_{n=0}^{\infty} \gamma_{n+1} \psi(X_n) < \infty.$$

Therefore,

$$\lim_{k\to\infty} \sum_{i=m_k+1}^{n_k} \gamma_i = 0, \tag{9.26}$$

while (9.21), (9.24) yield

$$\begin{aligned} \varepsilon \leq |\psi(X_{n_k}) - \psi(X_{m_k})| \leq K \|X_{n_k} - X_{m_k}\| \\ = K \left\| - \sum_{i=m_k}^{n_k-1} \gamma_{i+1} \nabla\psi(X_i) + \sum_{i=m_k+1}^{n_k} \xi_i \right\| \\ \leq K \sum_{i=m_k+1}^{n_k} \gamma_i + \left\| \sum_{i=m_k+1}^{n_k} \xi_i \right\| \end{aligned} \tag{9.27}$$

for $k \geq 1$. However, this is not possible, since (9.26) and the limit process $k \to \infty$ applied to (9.27) yield $\varepsilon \leq 0$. Hence, $\lim_{n\to\infty} \|\nabla\psi(X_n)\| = 0$. This completes the proof. □

9.5.3 Proof of Theorem 4

Let $f_n(x) = f(x) + \delta_n \psi(x)$ for $x \in \mathbb{R}^p$, $n \geq 1$, while $\Pi_{D^*}(\cdot)$ denotes the projection on the set D^* (*i.e.*, $\Pi_{D^*}(x) = \arg\inf_{x' \in D^*} \|x - x'\|$ for $x \in \mathbb{R}^p$). The following error sequences are defined for $n \geq 0$,

$$\begin{aligned} \xi_{n+1} &= \gamma_{n+1}\delta_{n+1}(\nabla\psi(X_n) - h'(g(X_n, Y_{n+1}))\nabla_x g(X_n, Y_{n+1})), \\ \varepsilon_{1,n+1} &= 2(X_n - \Pi_{D^*}(X_n))^T \xi_{n+1}, \\ \varepsilon_{2,n+1} &= \|X_{n+1} - X_n\|^2, \\ \varepsilon_{n+1} &= \varepsilon_{1,n+1} + \varepsilon_{2,n+1}\,. \end{aligned}$$

It is straightforward to verify that the following recursion holds for $n \geq 0$,

$$X_{n+1} = X_n - \gamma_{n+1}\nabla f_n(X_n) + \xi_{n+1},$$

and this implies the following bounds:

$$\begin{aligned} &\|X_{n+1} - \Pi_{D^*}(X_{n+1})\|^2 \\ &\quad \leq \|X_{n+1} - \Pi_{D^*}(X_n)\|^2 \\ &\quad = \|X_n - \Pi_{D^*}(X_n)\|^2 + 2(X_n - \Pi_{D^*}(X_n))^T(X_{n+1} - X_n) \\ &\qquad + \|X_{n+1} - X_n\|^2 \\ &\quad = \|X_n - \Pi_{D^*}(X_n)\|^2 \\ &\qquad - 2\gamma_{n+1}(X_n - \Pi_{D^*}(X_n))^T \nabla f_n(X_n) + \varepsilon_{n+1}, \quad n \geq 0. \end{aligned}$$

Lemma 3. *Suppose that Assumptions B3 and B4 hold. Moreover, let $\{x_k\}_{k\geq 0}$ be a bounded sequence from $\mathbb{R}^p$, while $\{n_k\}_{k\geq 0}$ is an increasing sequence of positive integers. Suppose that*

$$\liminf_{k\to\infty} \|x_k - \Pi_{D^*}(x_k)\| > 0.$$

Then,

$$\liminf_{k\to\infty} (x_k - \Pi_{D^*}(x_k))^T \nabla f_{n_k}(x_k) > 0.$$

Proof. Since $\{f_{n_k}\}_{k\geq 0}$ are convex and $f_{n_k}(\Pi_{D^*}(x_k)) = f(\Pi_{D^*}(x_k)) = \eta^*$ for $k \geq 0$, we have

$$(x_k - \Pi_{D^*}(x_k))^T \nabla f_{n_k}(x_k) \geq f_{n_k}(x_k) - f(\Pi_{D^*}(x_k)) = f_{n_k}(x_k) - \eta^*, \quad k \geq 0.$$

Therefore, it is sufficient to show that $\liminf_{k\to\infty} f_{n_k}(x_k) > \eta^*$. Suppose the opposite. Then, there exists $\varepsilon \in (0, \infty)$, $\tilde{x} \in \mathbb{R}^p$ and a subsequence $\{\tilde{x}_k, \tilde{n}_k\}_{k\geq 0}$ of $\{x_k, n_k\}_{k\geq 0}$ such that $\lim_{k\to\infty} \tilde{x}_k = \tilde{x}$, $\limsup_{k\to\infty} f_{\tilde{n}_k}(\tilde{x}_k) \leq \eta^*$ and $\|\tilde{x}_k - \Pi_{D^*}(\tilde{x}_k)\| \geq \varepsilon$ for $k \geq 0$. Consequently,

$$f(\tilde{x}) = \lim_{k\to\infty} f(\tilde{x}_k) \leq \limsup_{k\to\infty} f_{\tilde{n}_k}(\tilde{x}_k) \leq \eta^*, \tag{9.28}$$

$$d(\tilde{x}, D^*) = \|\tilde{x} - \Pi_{D^*}(\tilde{x})\| = \lim_{k\to\infty} \|\tilde{x}_k - \Pi_{D^*}(\tilde{x}_k)\| \geq \varepsilon.$$

Then, it can easily be deduced that $\tilde{x} \notin D$ (otherwise, (9.28) would imply $\tilde{x} \in D^*$). Therefore,

$$\lim_{k\to\infty} f_{\tilde{n}_k}(\tilde{x}_k) \geq \lim_{k\to\infty} \delta_{\tilde{n}_k}\psi(\tilde{x}_k) = \infty > \eta^*.$$

However, this is not possible. Hence, $\liminf_{k\to\infty} f_{n_k}(x_k) > \eta^*$. This completes the proof. □

Lemma 4. *Suppose that Assumptions B1 and B4 hold. Then,* $\lim_{n\to\infty}\|X_{n+1} - X_n\| = 0$ *and* $\sum_{n=1}^{\infty}\varepsilon_n$ *converges w.p.1 on the event* $\{\sup_{0\leq n}\|X_n\| < \infty\}$*.*

Proof. Let $\rho \in [1,\infty)$, while $K_\rho \in [\rho,\infty)$ denotes an upper bound of $\|\Pi_{D^*}(\cdot)\|$, $\|\nabla\psi\|$, ∇f on B_ρ^p. Due to B4,

$$\begin{aligned} \|\xi_{n+1}\|I_{\{\|X_n\|\leq\rho\}} &\leq 2K_\rho\gamma_{n+1}\delta_{n+1}\phi_\rho^2(Y_{n+1}), \\ |\varepsilon_{1,n+1}|I_{\{\|X_n\|\leq\rho\}} &\leq 4K_\rho\|\xi_{n+1}\|I_{\{\|X_n\|\leq\rho\}}, \\ |\varepsilon_{2,n+1}|I_{\{\|X_n\|\leq\rho\}} &\leq 2K_\rho^2\gamma_{n+1}^2 + 2\|\xi_{n+1}\|^2 I_{\{\|X_n\|\leq\rho\}}, \\ \|X_{n+1} - X_n\|I_{\{\|X_n\|\leq\rho\}} &\leq K_\rho\gamma_{n+1}(1+\delta_{n+1}) + \|\xi_{n+1}\|I_{\{\|X_n\|\leq\rho\}} \end{aligned} \tag{9.29}$$

for $k \geq 0$. Consequently,

$$\mathbb{E}\left(\sum_{n=0}^{\infty}\|\xi_{n+1}\|^2 I_{\{\|X_n\|\leq\rho\}}\right) \leq 4K_\rho^2\sum_{n=1}^{\infty}\gamma_n^2\delta_n^2\mathbb{E}(\phi_\rho^4(Y_n)) < \infty \tag{9.30}$$

$$\mathbb{E}\left(\sum_{n=0}^{\infty}|\varepsilon_{1,n+1}|^2 I_{\{\|X_n\|\leq\rho\}}\right) \leq 16K_\rho^2\mathbb{E}\left(\sum_{n=0}^{\infty}\|\xi_{n+1}\|^2 I_{\{\|X_n\|\leq\rho\}}\right) < \infty, \tag{9.31}$$

$$\begin{aligned} \mathbb{E}\left(\sum_{n=0}^{\infty}|\varepsilon_{2,n+1}|^2 I_{\{\|X_n\|\leq\rho\}}\right) \leq & 2\mathbb{E}\left(\sum_{n=0}^{\infty}\|\xi_{n+1}\|^2 I_{\{\|X_n\|\leq\rho\}}\right) \\ & + 2K_\rho^2\sum_{n=1}^{\infty}\gamma_n^2(1+\delta_n^2) < \infty. \end{aligned} \tag{9.32}$$

Owing to B1 and (9.29), (9.30), $\lim_{n\to\infty}\|X_{n+1} - X_n\| = 0$ w.p.1 on the event $\{\sup_{0\leq n}\|X_n\| \leq \rho\}$. On the other hand, using (9.31), (9.32) and the same arguments as in the proof of Lemma 9.14, it can be demonstrated that $\sum_{n=1}^{\infty}\varepsilon_n$ converges w.p.1 on the same event. Then, we can easily deduce that w.p.1 on the event $\{\sup_{0\leq n}\|X_n\| < \infty\}$, $\lim_{n\to\infty}\|X_{n+1} - X_n\| = 0$ and $\sum_{n=1}^{\infty}\varepsilon_n$ converges. This completes the proof. □

Proof of Theorem 4. Let $\rho \in [1,\infty)$, while $K_\rho \in [\rho,\infty)$ denotes an upper bound of $\|\Pi_{D^*}(\cdot)\|$ on B_ρ^p. Moreover, let ω be an arbitrary sample from the event where $\sup_{0\leq n}\|X_n\| \leq \rho$, $\lim_{n\to\infty}\|X_{n+1} - X_n\| = 0$ and $\sum_{n=1}^{\infty}\varepsilon_n$ converges (for the sake of notational simplicity, ω does not explicitly appear in the relations which follow in the proof). Due to Lemma 4, it is sufficient to show $\lim_{n\to\infty}\|X_n - \mathit{\Pi}_{D^*}(X_n)\| = 0$. Suppose the opposite. Then, there exists $\varepsilon \in (0,\infty)$ (depending on ω) such that $\limsup_{n\to\infty}\|X_n - \mathit{\Pi}_{D^*}(X_n)\| > 2\varepsilon$. On the other hand, (9.28) yields

$$\begin{aligned} 2\sum_{i=0}^{n-1}&\gamma_{i+1}(X_i - \mathit{\Pi}_{D^*}(X_i))^T\nabla f_i(X_i) \\ &\le \|X_0 - \mathit{\Pi}_{D^*}(X_0)\|^2 - \|X_n - \mathit{\Pi}_{D^*}(X_n)\|^2 + \sum_{i=1}^{n}\varepsilon_i \qquad (9.33)\\ &\le 4K_\rho^2 + \sum_{i=1}^{n}\varepsilon_i, \quad n \ge 1. \end{aligned}$$

Since $\liminf_{n\to\infty}(X_n - \mathit{\Pi}_{D^*}(X_n))^T\nabla f_n(X_n) > 0$ results from $\liminf_{n\to\infty}\|X_n - \mathit{\Pi}_{D^*}(X_n)\| > 0$ (due to Lemma 3), B1 and (9.33) imply that $\liminf_{n\to\infty}\|X_n - \mathit{\Pi}_{D^*}(X_n)\| = 0$. Otherwise, there would exist $\delta \in (0,\infty)$ and $j_0 \ge 1$ (both depending on ω) such that $(X_n - \mathit{\Pi}_{D^*}(X_n))^T\nabla f_n(X_n) \ge \delta$, $n \ge j_0$, which combined with B1 would yield

$$\begin{aligned} &\sum_{n=0}^{\infty}\gamma_{n+1}(X_n - \mathit{\Pi}_{D^*}(X_n))^T\nabla f_n(X_n) \\ &\ge \sum_{n=0}^{j_0}\gamma_{n+1}(X_n - \mathit{\Pi}_{D^*}(X_n))^T\nabla f_n(X_n) + \delta\sum_{n=j_0+1}^{\infty}\gamma_n = \infty. \end{aligned}$$

Let $l_0 = \inf\{n \ge 0 : \|X_n - \mathit{\Pi}_{D^*}(X_n)\| \le \varepsilon\}$, and define for $k \ge 0$,

$$\begin{aligned} n_k &= \inf\{n \ge l_k : \|X_n - \mathit{\Pi}_{D^*}(X_n)\| \ge 2\varepsilon\}, \\ m_k &= \sup\{n \le n_k : \|X_n - \mathit{\Pi}_{D^*}(X_n)\| \le \varepsilon\}, \\ l_{k+1} &= \inf\{n \ge n_k : \|X_n - \mathit{\Pi}_{D^*}(X_n)\| \le \varepsilon\}. \end{aligned}$$

Obviously, $\{l_k\}_{k\ge0}$, $\{m_k\}_{k\ge0}$, $\{n_k\}_{k\ge0}$ are well-defined, finite and satisfy $l_k \le m_k < n_k < l_{k+1}$ for $k \ge 0$. Moreover,

$$\|X_{m_k} - \mathit{\Pi}_{D^*}(X_{m_k})\| \le \varepsilon, \quad \|X_{n_k} - \mathit{\Pi}_{D^*}(X_{n_k})\| \ge 2\varepsilon, \quad k \ge 0, \qquad (9.34)$$

and

$$\|X_n - \mathit{\Pi}_{D^*}(X_n)\| \ge \varepsilon \quad \text{for } m_k < n \le n_k,\ k \ge 0. \qquad (9.35)$$

Due to (9.34), (9.35) and the fact that $\mathit{\Pi}_{D^*}(\cdot)$ is non-expansive (see, *e.g.*, [42]),

$$\begin{aligned} \varepsilon \le \|X_{m_k+1} - \mathit{\Pi}_{D^*}(X_{m_k+1})\| \le& \varepsilon + \|X_{m_k+1} - \mathit{\Pi}_{D^*}(X_{m_k+1})\| \\ &- \|X_{m_k} - \mathit{\Pi}_{D^*}(X_{m_k})\| \\ \le& \varepsilon + \|(X_{m_k+1} - X_{m_k}) \\ &- (\mathit{\Pi}_{D^*}(X_{m_k+1}) - \mathit{\Pi}_{D^*}(X_{m_k}))\| \\ \le& \varepsilon + 2\|X_{m_k+1} - X_{m_k}\|, \quad k \ge 0. \end{aligned}$$

Therefore, $\lim_{k\to\infty} \|X_{m_k+1} - \mathit{\Pi}_{D^*}(X_{m_k+1})\| = \varepsilon$. Then, (9.35) yields

$$\limsup_{k\to\infty}(\|X_{n_k} - \mathit{\Pi}_{D^*}(X_{n_k})\|^2 - \|X_{m_k+1} - \mathit{\Pi}_{D^*}(X_{m_k+1})\|^2) \geq \varepsilon^2. \qquad (9.36)$$

On the other hand, Lemma 3 and (9.35) imply

$$\liminf_{k\to\infty} \min_{m_k<n\leq n_k} (X_n - \mathit{\Pi}_{D^*}(X_n))^T \nabla f_n(X_n) > 0.$$

Consequently, there exists $k_0 \geq 0$ (depending on ω) such that

$$\sum_{i=m_k+1}^{n_k-1} \gamma_{i+1}(X_i - \mathit{\Pi}_{D^*}(X_i))^T \nabla f_i(X_i) \geq 0 \qquad (9.37)$$

for $k \geq k_0$. Owing to (9.28), (9.37),

$$\begin{aligned}&\|X_{n_k} - \mathit{\Pi}_{D^*}(X_{n_k})\|^2 - \|X_{m_k+1} - \mathit{\Pi}_{D^*}(X_{m_k+1})\|^2\\ &\leq \sum_{m_k+1}^{n_k-1} \gamma_{i+1}(X_i - \mathit{\Pi}_{D^*}(X_i))^T \nabla f_i(X_i) + \sum_{i=m_k+2}^{n_k} \varepsilon_i \leq \sum_{i=m_k+2}^{n_k} \varepsilon_i\end{aligned}$$

for $k \geq k_0$. However, this is not possible, since (9.36) and the limit process $k \to \infty$ applied to (9.37) yield $\varepsilon^2 \leq 0$. Hence, $\lim_{n\to\infty} \|X_n - \mathit{\Pi}_{D^*}(X_n)\| = 0$. This completes the proof. □

Acknowledgement. Research conducted by Sean P. Meyn is supported by the National Science Foundation under Award Nos. ECS 02 17836 and DMI 00 85165. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Part of this research was performed while Roberto Tempo was visiting the Coordinated Science Laboratory of the University of Illinois as a CSL Visiting Research Professor, the support is gratefully acknowledged.

Part III

Probabilistic Methods in Identification and Control

10

A Learning Theory Approach to System Identification and Stochastic Adaptive Control

Mathukumalli Vidyasagar[1] and Rajeeva L. Karandikar[2]

[1] Tata Consultancy Services
No. 1, Software Units Layout
Madhapur, Hyderabad 500 081, India
sagar@atc.tcs.co.in

[2] Indian Statistical Institute
S. J. S. Sansawal Marg
New Delhi 110 016, India
rlk@isid.ac.in

Summary. In this chapter, we present an approach to system identification based on viewing identification as a problem in statistical learning theory. Apparently, this approach was first mooted in [396]. The main motivation for initiating such a program is that traditionally system identification theory provide *asymptotic* results. In contrast, statistical learning theory is devoted to the derivation of *finite time estimates*. If system identification is to be combined with robust control theory to develop a sound theory of indirect adaptive control, it is essential to have finite time estimates of the sort provided by statistical learning theory. As an illustration of the approach, a result is derived showing that in the case of systems with fading memory, it is possible to combine standard results in statistical learning theory (suitably modified to the present situation) with some fading memory arguments to obtain finite time estimates of the desired kind. It is also shown that the time series generated by a large class of BIBO stable nonlinear systems has a property known as β-mixing. As a result, earlier results of [394] can be applied to many more situations than shown in that paper.

10.1 Introduction

10.1.1 The System Identification Problem

The aim of system identification is to fit given data, usually supplied in the form of a time series, with models from within a given model class. One can divide the main challenges of system identification into three successively stronger questions, as follows. As more and more data is provided to the identification algorithm:

1. Does the estimation error between the outputs of the identified model and the actual time series approach the minimum possible estimation error

achievable by any model within the given model class? In other words, is the performance of the identification algorithm asymptotically optimal?

2. Does the identified model converge to the best possible model within the given model class? In other words, assuming that the minimum possible estimation error is achievable by one or more 'best possible models,' does the output of the identification algorithm approach one of these best possible models?
3. Assuming that the data is generated by a 'true' system whose output is corrupted by measurement noise, does the identified model converge to the 'true' system? In other words, if both the true system and the family of models are parameterized by a vector of parameters, does the estimated parameter vector converge to the true parameter vector?

From a technical standpoint, Questions 2 and 3 are easier to answer than Question 1. Following the notational conventions of system identification, let $\{h(\theta), \theta \in \Theta\}$ denote the family of models, where θ denotes a parameter that characterizes the model, and Θ is a topological space (usually a subset of $\mathbb{R}^\ell$ for some ℓ), and let $J(\theta)$ denote the estimation error when the model $h(\theta)$ is used to predict the next measurement. Since identification is carried out recursively, the output of the identification algorithm is a sequence of estimates $\{\theta_t\}_{t \geq 1}$, or what is the same thing, a sequence of estimated models $\{h(\theta_t)\}_{t \geq 1}$. Suppose that we are able to show that Question 1 has an affirmative answer, *i.e.*, that $J(\theta_t) \to J^*$, where J^* denotes the minimum possible estimation error. In such a case, with very few additional assumptions it is possible to answer both Questions 2 and 3 in the affirmative. Traditionally a positive answer to Question 2 is assured by assuming that Θ is a *compact* set, which in turn ensures that the sequence $\{\theta_t\}$ contains at least one convergent subsequence. For convenience, let us relabel this subsequence again as $\{\theta_t\}$. If the answer to Question 1 is 'yes,' if θ^* is any limit point of the sequence, and if $J(\theta)$ is continuous (or at worst, lower semi-continuous) with respect to θ, then it readily follows that $J(\theta_t) \to J^*$. In other words, the model $h(\theta^*)$ is an 'optimal' fit to the data among the family $\{h(\theta), \theta \in \Theta\}$. Coming now to Question 3, suppose θ_{true} is the parameter of the 'true' system, and let f_{true} denote the 'true' system. In order for Question 3 to have an affirmative answer, the true system f_{true} must belong to the model family $\{h(\theta), \theta \in \Theta\}$; otherwise we cannot hope that $h(\theta_t)$ will converge to f_{true}. In such a case, the minimum achievable estimation error is zero, *i.e.*, $J^* = 0$. Now suppose θ^* is a limit point of the sequence $\{\theta_t\}$. The traditional way to ensure that $\theta_{\text{true}} = \theta^*$ is to assume that the input to the true system is 'persistingly exciting' or 'sufficiently rich,' so that the only way for $h(\theta^*)$ to match the performance of f_{true} is to have $\theta^* = \theta_{\text{true}}$.

None of what has been said above is new. Indeed, because of the arguments presented above, the main emphasis in system identification theory has been to study conditions to ensure that Question 1 has an affirmative answer, *i.e.*, that the identification algorithm is asymptotically optimal. In a seminal paper [208], Lennart Ljung has shown that indeed Question 1 can be answered in the affirmative provided empirical estimates of the performance of each model $h(\theta)$ converge *uniformly* to the corresponding true performance, where the uniformity is with respect to $\theta \in \Theta$. In earlier work [68], Caines had established an affirmative answer to Question 1 using ergodic theory. However, so far as the present authors are able to determine, Ljung was the first to address Question 1 *using the notion of uniform convergence of empirical means* (though he did not use that terminology); see [208, Lemma 3.1.], Ljung also showed that this particular uniform convergence property does hold, provided three assumptions are satisfied, namely:

- The parameter set Θ is compact.
- The model class consists of uniformly exponentially stable systems.
- The parameter θ enters the description of the model $h(\theta)$ in a 'differentiable' manner. (Coupled with the assumption that Θ is a compact set, this assumption implies that various quantities have bounded gradients with respect to θ.)

10.1.2 The Need for a Quantitative Identification Theory

By tradition, identification theory is *asymptotic* in nature. In other words, the theory analyzes what happens as the number of samples approaches infinity. In contrast, the objective of the present contribution is to derive *finite time*, or nonasymptotic, estimates of the rate at which the output of the identification process converges to the best possible model. We now give a brief justification as to why such a theory would be useful.

In so-called 'indirect' adaptive control, one first carries out an identification of an unknown system, and after a finite amount of time has elapsed, designs a controller based on the current model of the unknown system. The philosophy behind indirect adaptive control is that, after sufficient time has passed, the output of the identification algorithm (the current model) is 'sufficiently close' to the true system that a controller designed on the basis of the current model will also perform satisfactorily for the true system. In order for the above argument to be made precise, we need to be able to give *quantitative* answers to two questions:

(i) What is the distance (in some reasonable metric) between the currently identified model and the true system?
(ii) How far can the identified model be from the true system (in some reasonable metric) in order that a controller designed on the basis of the current model also performs satisfactorily for the true system?

Of these, the second question falls in the realm of robust control, and there are many satisfactory theories to address this question. But to date very little attention has been paid to the first question. The objective of this chapter is to address this first question, and to derive some preliminary results.

In order to address this first question meaningfully, we must decide what is meant by a ‘reasonable metric’ between the identified model and the true system. If both the true system and the identified model are unstable, then one can use either the so-called ‘gap’ metric [136, 408] or the graph metric [379]. However, indirect adaptive control is rarely used when the true system is unstable, because while identification is taking place the system is not under any control. It is more common to use indirect adaptive control when the true system is stable. In this case, the purpose of applying control is not to stabilize the true system, but to improve its performance in some other way. The metric used to measure the distance of the true system from the identified model must, in some sense, take into account the performance criterion. For the purposes of this presentation, we use a very simple measure, namely the mean-squared error of the system response. This measure is by far the most commonly used error measure. Moreover, the least-squares error measure is very amenable to the kind of analysis carried out here. This is the measure used in previous work on finite-time estimates, such as [78, 394, 395, 397] for example. On the other hand, from the standpoint of robustness analysis, the ℓ_1-error measure would be much more natural. In general, the two error measures are not directly related, unless one imposes some restrictions on the McMillan degree of the various models. It is our hope that future researchers will be able to apply the present methods to more meaningful distance measures.

10.1.3 Review of Previous Work on Finite-Time Estimates

As stated earlier, the paper of Ljung [208] is apparently the first to study the asymptotic optimality of system identification algorithms using the idea that empirical means must converge uniformly to their true values. In that paper, Ljung also establishes the desired property under some assumptions. *In principle*, the arguments in [208] can be used to provide finite-time estimates of the rate at which the identification algorithm converges. However, by tradition the system identification community has not focused on finite-time estimates. So far as the authors are able to determine, the first papers to state the derivation of finite-time estimates as a desirable property in itself are by Weyer *et al.* [396, 397]. In those papers, it is assumed that the time series to which the system identification algorithm is applied consists of so-called ‘finitely-dependent’ data; in other words, it is assumed that the time series can be divided into blocks that are independent. In a later paper [394], Weyer replaced the assumption of finite dependence by the assumption that the time series is β-mixing.[3] However, he did not derive conditions under which a time series is β-mixing. In a still later paper by Campi and Weyer [78], the authors restrict themselves to the case where the data is generated by an ARMA model driven by i.i.d. Gaussian noise, and the model family also consists of

[3] These notions are defined in subsequent sections.

ARMA models driven by Gaussian noise. In this paper, the authors say that they are motivated by the observation that 'signals generated by dynamical systems are not β-mixing in general'. This is why their analysis is limited to an extremely restricted class. However, their statement that dynamical systems do not generate β-mixing sequences is simply incorrect. Specifically, all the systems studied in [78] are themselves β-mixing! (See Theorems 6 and 7.) In another paper [180], the present authors have shown that *practically any exponentially stable system driven by i.i.d. noise with bounded variance* is β-mixing. The systems under study can be linear or nonlinear. This result is reproduced in later sections. Hence, far from being restrictive, the results of Weyer [394] have broad applicability, though he did not show it at the time.

10.1.4 Contributions of the Present Chapter

The present chapter essentially has three contributions.

1. In Section 10.3, we derive a very general result relating the uniform convergence properties of the cost function to the finite-time estimates of the rate of convergence of an identification algorithm. Though previous papers allude to such arguments indirectly, in the present chapter we make the connection quite explicit.
2. In Section 10.4, we study the case where the time series is generated by a 'true but unknown' system, and show that the desired uniform convergence property holds whenever the family of 'error models' (*i.e.*, the difference between the true system and the model family) satisfies two properties: each error model (i) has exponentially decaying memory, and (ii) is exponentially stable. Note that the true system and/or models can be nonlinear and/or infinite-dimensional. So far as we are aware, such a general situation has not been studied by previous papers in the literature.
3. In Section 10.5, we derive bounds on the P-dimension of *nonlinear* ARMA models. There are not too many such bounds in the literature.
4. From Section 10.6 onwards, we reproduce a result from [180] which gives conditions under which a time series is β-mixing, and give its proof. As a consequence of this result, it follows that the time series studied *in all previous papers in this subject* are β-mixing. This had not been recognized until now. As a consequence, the results of [394] have far broader applicability than shown in that paper.

10.2 Problem Formulation

The problem of system identification studied in this chapter can be stated as follows: let $U \subseteq \mathbb{R}^l$ denote the input set, and $Y \subseteq \mathbb{R}^k$ denote the output set, where k, l are appropriate integers. To avoid technical difficulties, it is assumed that both U and Y are *bounded.* This ensures that any random variable assuming values in U, Y or $U \times Y$ has bounded moments of all orders. One is given a time series $\{(u_t, y_t)\}$, where u_t denotes the input to the unknown system at time t, and y_t denotes the output at time t. The time series, as the name implies, is measured one time step at a time. One is also given a family of models $\{h(\theta), \theta \in \Theta\}$, parameterized by a parameter vector θ belonging to a set Θ. Usually Θ is a subset of a finite-dimensional Euclidean space. At time t, the data available to the modeller consists of all the past measurements until time $t-1$. Based on these measurements, the modeller chooses a parameter θ_t, with the objective of making the best possible prediction of the next measured output y_t. The method of choosing θ_t is called the identification algorithm. Traditionally, the aim of identification theory has been to study the behavior of the model $h(\theta_t)$ as $t \to \infty$.

To make this formulation a little more precise, let us introduce some notation. Define $\mathcal{U} \doteq \prod_{-\infty}^{\infty} U$, and define $\mathcal{Y}$ analogously. Note that the input sequence $\{u_t\}$ belongs to the set $\mathcal{U}$, while the output sequence belongs to $\mathcal{Y}$. Let $U_{-\infty}^{0}$ denote the one-sided infinite cartesian product $U_{-\infty}^{0} \doteq \prod_{-\infty}^{0} U$, and for a given two-sided infinite sequence $\mathbf{u} \in \mathcal{U}$, define

$$\mathbf{u}_t \doteq (u_{t-1}, u_{t-2}, u_{t-3}, \ldots) \in U_{-\infty}^{0}.$$

Thus the symbol $\mathbf{u}_t$ denotes the infinite past of the input signal $\mathbf{u}$ at time t. The family of models $\{h(\theta), \theta \in \Theta\}$ consists of a collection of maps $h(\theta), \theta \in \Theta$, where each $h(\theta)$ maps $U_{-\infty}^{0}$ into Y. Thus, at time t, the quantity $h(\theta) \cdot \mathbf{u}_t \doteq \hat{y}_t(\theta)$ is the 'predicted' output if the model parameter is chosen as θ. Note that the above notation automatically builds in the requirement that each model is time-invariant. The quality of this prediction is measured by a 'loss function' $\ell : Y \times Y \to [0,1]$. Thus $\ell(y_t, h(\theta) \cdot \mathbf{u}_t)$ is the loss we incur if we use the model $h(\theta)$ to predict the output at time t, and the actual output is y_t.

To illustrate this notation, consider the most common case where the model family consists of LTI systems described by their unit pulse responses. Hence each $h(\theta)$ is described by a sequence $\{h_i(\theta)\}_{i \geq 1}$. The output of this model at time t is

$$\hat{y}_t(\theta) = \sum_{i=1}^{\infty} h_i(\theta) u_{t-i}.$$

Choose constants μ_U, μ_Y such that $\| u \| \leq \mu_U \ \forall u \in U$, and similarly for Y. Suppose that, for each $\theta \in \Theta$, the unit impulse response sequence $\{h_i(\theta)\}_{i \geq 1}$ is absolutely summable, and that

$$\sum_{i=1}^{\infty} |h_i(\theta)| \leq \mu_Y / \mu_U.$$

Then it is easy to see that each model $h(\theta)$ maps every input sequence assuming values in U to an output sequence assuming values in Y. By far the most commonly used loss function is $\ell(y, z) \doteq \| y - z \|^2$. Hence

$$\ell(y_t, h(\theta) \cdot \mathbf{u}_t) = \| y_t - \hat{y}_t(\theta) \|^2$$

is the mean-squared error between the actual and predicted output.

Since we are dealing with a time series, all quantities are *random*. Hence, to assess the quality of the prediction made using the model $h(\theta)$, we should take the *expected value* of the loss function $\ell(y_t, \hat{y}_t(\theta))$. For this purpose, let $\tilde{P}_{\mathbf{u},\mathbf{y}}$ denote the law of the time series $\{(u_t, y_t)\}$. Observe that $\tilde{P}_{\mathbf{u},\mathbf{y}}$ is a probability measure on the infinite cartesian product set $\mathcal{U} \times \mathcal{Y}$, and describes the statistics of the time series. Given a parameter vector θ, the quality of the prediction made using this choice of model is defined as

$$J(\theta) \doteq \mathbb{E}[\ell(y_t, h(\theta) \cdot \mathbf{u}_t), \tilde{P}_{\mathbf{u},\mathbf{y}}]. \tag{10.1}$$

The quantity $J(\theta)$ is referred to hereafter as the **objective function**. Note that $J(\theta)$ depends solely on θ and nothing else. Also, since the time series is assumed to be stationary, the probability measure $\tilde{P}_{\mathbf{u},\mathbf{y}}$ is shift-invariant, which in turn implies that the quantity $J(\theta)$ is independent of t.

A key observation at this stage is that the probability measure $\tilde{P}_{\mathbf{u},\mathbf{y}}$ is *unknown*. This is because, if the statistics of the time series are known ahead of time, then there is nothing to identify! To make this point more forcefully, let us consider the common situation where y_t is the output of a 'true' system corrupted by measurement noise, as in (10.2). Given the laws of u_t and η_t, *and given* f_{true}, we can, at least in some abstract sense, derive the joint law of the process $\{(u_t, y_t)\}$. Hence assuming that the time series has a known law is tantamount to assuming that the true system is known.

Now at last we come to a precise formulation of the system identification problem.

Definition 1 (System Identification Problem). *Given the time series* $\{(u_t, y_t)\}$ *with* unknown *law* $\tilde{P}_{\mathbf{u},\mathbf{y}}$*, construct if possible an iterative algorithm for choosing* θ_t *as a function of* t*, in such a way that*

$$J(\theta_t) \rightarrow \inf_{\theta \in \Theta} J(\theta).$$

While the above problem definition appears to be rather abstract, actually in many cases the problem can be interpreted as one of approximating an unknown system using a model from a specified family of models. Suppose the measured output y_t corresponds to a noise-corrupted output of a 'true' system f_{true}, and that ℓ is the squared error, as above. Note that it is *not* assumed that the true system f_{true} belongs to the model family $\{h(\theta), \theta \in \Theta\}$. In such a case, the problem formulation becomes the following: suppose the input sequence $\{u_t\}_{-\infty}^{\infty}$ is distributed according to some joint law Q, and that $\{\eta_t\}_{-\infty}^{\infty}$ is a zero-mean i.i.d. measurement noise sequence with one-dimensional law P. Suppose in addition that u_i, η_j are independent for each i, j. Now suppose that

$$y_t = f_{\text{true}} \cdot \mathbf{u}_t + \eta_t, \ \forall t. \tag{10.2}$$

In such a case, the expected value in (10.1) can be expressed in terms of the probability measure $Q \times P^\infty$, and becomes

$$\begin{aligned} J(\theta) &= \mathbb{E}[\| \, (f_{\text{true}} - h(\theta)) \cdot \mathbf{u}_t + \eta_t \, \|^2, Q \times P^\infty] \\ &= \mathbb{E}[\| \, \tilde{h}(\theta) \cdot \mathbf{u}_t \, \|^2, Q] + \mathbb{E}[\| \, \eta \, \|^2, P^\infty], \end{aligned}$$

where $\tilde{h}(\theta) \doteq h(\theta) - f_{\text{true}}$. Since the second term is independent of θ, we effectively minimize only the first term. In other words, by minimizing $J(\theta)$ with respect to θ, we will find the best approximation to the true system f_{true} in the model family $\{h(\theta), \theta \in \Theta\}$. Recall that it is *not* assumed the true system f_{true} belongs to $\{h(\theta), \theta \in \Theta\}$. In case there is a 'true' value of θ, call it θ_{true} such that $f_{\text{true}} = h(\theta_{\text{true}})$, then *an* optimal choice of θ is θ_{true}. If in addition we impose some assumptions to the effect that the input sequence $\{u_t\}$ is sufficiently exciting, then $\theta = \theta_{\text{true}}$ becomes the *only* minimizer of $J(\cdot)$.

10.3 A General Result

As stated in Section 10.2, the system identification problem is to choose a parameter vector θ so as to minimize the objective function $J(\theta)$. As stated just after (10.1), the probability measure $\tilde{P}_{\mathbf{u},\mathbf{y}}$ is *unknown.* In other words, the objective function $J(\theta)$ *cannot be computed* on the basis of the available data.

Thus the key to the system identification problem is that the objective function to be minimzed cannot be computed exactly. To circumvent this difficulty, one replaces the 'true' objective function $J(\cdot)$ by an 'empirical approximation,' as defined next. For each $t \geq 1$ and each $\theta \in \Theta$, define the empirical error

$$\hat{J}_t(\theta) \doteq \frac{1}{t} \sum_{i=1}^{t} \ell[y_i, h(\theta) \cdot \mathbf{u}_i].$$

For example, if $\ell(y, z) =\| \, y - z \, \|^2$, then

$$\hat{J}_t(\theta) = \frac{1}{t} \sum_{i=1}^{t} \| \, y_i - \hat{y}_i(\theta) \, \|^2$$

is the average cumulative mean-squared error between the actual output y_i and the predicted error $\hat{y}_i(\theta)$, from time 1 to time t. Note that, unlike the quantity $J(\theta)$, the function $\hat{J}(\theta)$ *can* be computed on the basis of the available data. Hence, in principle at least, it is possible to choose θ_t so as to minimize the ‘approximate’ (but computable) objective function $\hat{J}(\theta)$ in the hope that, by doing so, we will somehow minimize the ‘true’ (but uncomputable) objective function $J(\theta)$. The next theorem gives some sufficient conditions for this approach to work. Specifically, if a particular property known as UCEM (uniform convergence of empirical means) holds, then the natural approach of choosing θ_t to minimize the *empirical* (or cumulated average) error will lead to a solution of the system identification problem.

Theorem 1. *At time t, choose θ_t^* so as to minimize $\hat{J}_t(\theta)$; that is,*

$$\theta_t^* = \mathrm{argmin}_{\theta \in \Theta}\ \hat{J}_t(\theta).$$

Let

$$J^* \doteq \inf_{\theta \in \Theta} J(\theta).$$

Define the quantity

$$q(t,\epsilon) \doteq \tilde{P}_{\mathbf{u},\mathbf{y}}\{\sup_{\theta \in \Theta} |\hat{J}_t(\theta) - J(\theta)| > \epsilon\}. \tag{10.3}$$

Suppose it is the case that $q(t,\epsilon) \to 0$ as $t \to \infty$, $\forall \epsilon > 0$. Then

$$\tilde{P}_{\mathbf{u},\mathbf{y}}\{J(\theta_t^*) > J^* + \epsilon\} \to 0 \text{ as } t \to \infty,\ \forall \epsilon > 0. \tag{10.4}$$

In other words, the quantity $J(\theta_t^*)$ converges to the optimal value J^* in probability, with respect to the measure $\tilde{P}_{\mathbf{u},\mathbf{y}}$.

Corollary 1. *Suppose that $q(t,\epsilon) \to 0$ as $t \to \infty$, $\forall \epsilon > 0$. Given $\epsilon, \delta > 0$, choose $t_0 = t_0(\epsilon,\delta)$ such that*

$$q(t,\epsilon) < \delta\ \forall t \geq t_0(\epsilon,\delta).$$

Then

$$\tilde{P}_{\mathbf{u},\mathbf{y}}\{J(\theta_t^*) > J^* + \epsilon\} < \delta\ \forall t \geq t_0(\epsilon/3,\delta). \tag{10.5}$$

Proof of Theorem 1. Suppose $q(t,\epsilon) \to 0$ as $t \to \infty$, $\forall \epsilon > 0$. To establish the desired conclusion (10.4), we need to establish the following: given arbitrarily small numbers $\epsilon, \delta > 0$, there exists a $t_0 = t_0(\epsilon,\delta)$ such that

$$\tilde{P}_{\mathbf{u},\mathbf{y}}\{J(\theta_t^*) > J^* + \epsilon\} < \delta\ \forall t \geq t_0.$$

For this purpose, we proceed as follows. Given $\epsilon, \delta > 0$, choose t_0 large enough such that

$$\tilde{P}_{\mathbf{u},\mathbf{y}}\{\sup_{\theta \in \Theta} |\hat{J}_t(\theta) - J(\theta)| > \epsilon/3\} < \delta\ \forall t \geq t_0. \tag{10.6}$$

Select a $\theta_\epsilon \in \Theta$ such that $J(\theta_\epsilon) \leq J^* + \epsilon/3$. Such a θ_ϵ exists in view of the definition of J^*. Then, in view of (10.6), whenever $t \geq t_0$ we can say with confidence $1 - \delta$ that

$$\hat{J}(\theta_t) \geq J(\theta_t) - \epsilon/3, \text{ and } \hat{J}(\theta_\epsilon) \leq J(\theta_\epsilon) + \epsilon/3.$$

By definition,

$$\hat{J}(\theta_t) \leq \hat{J}(\theta_\epsilon).$$

Combining these two inequalities shows that

$$J(\theta_t) \leq \hat{J}(\theta_t) + \epsilon/3 \leq \hat{J}(\theta_\epsilon) + \epsilon/3 \leq J(\theta_\epsilon) + 2\epsilon/3 \leq J^* + 2\epsilon/3 + \epsilon/3 = J^* + \epsilon.$$

This statement holds with confidence $1 - \delta$, that is,

$$\tilde{P}_{\mathbf{u},\mathbf{y}}\{J(\theta_t) > J^* + \epsilon\} < \delta.$$

Since this argument can be repeated for every $\epsilon, \delta > 0$, it follows that

$$\tilde{P}_{\mathbf{u},\mathbf{y}}\{J(\theta_t) > J^* + \epsilon\} \to 0 \text{ as } t \to \infty \; \forall \epsilon > 0,$$

which is the desired conclusion. □
A proof of Corollary 1 is contained in the proof of Theorem 1.

Several points are noteworthy about this theorem.

1. The theorem states that, under appropriate conditions, the quantity $J(\theta_t)$ approaches the infimum J^*, even though we cannot compute either of these quantities. In particular, the performance of the estimated model $h(\theta_t)$ is *asymptotically optimal*, if the conditions of the theorem are satisfied.
2. The condition that $q(t, \epsilon) \to 0$ as $t \to \infty$ is usually referred to in the statistical learning theory as the property of *uniform convergence of empirical means (UCEM).* Thus the theorem states that if the family of error measures $\{J(\theta), \theta \in \Theta\}$ has the UCEM property, then the natural algorithm of choosing θ_t so as to minimize the empirical estimate $\hat{J}(\theta)$ at time t is 'asymptotically optimal.'
3. The corollary turns this asymptotic result into a finite-time result. Specifically, as shown in (10.5), if we can compute a number t_0 such that (10.6) holds, then it can be stated with confidence $1 - \delta$ that the currently identified model θ_t is within ϵ of the optimal performance. This quantification of the finite time performance of the identification algorithm is the additional feature of using statistical learning theory.
4. The number $t_0(\epsilon, \delta)$ such that $q(t, \epsilon) < \delta \; \forall t \geq t_0$ is called the *sample complexity* corresponding to the 'accuracy' ϵ and 'confidence' δ. Thus Corollary 1 states that the sample complexity of achieving ϵ-optimality with confidence δ is no worse than the sample complexity of achieving $\epsilon/3$-accuracy with confidence δ.
5. Note that such an approach is already adopted in the paper of Ljung; see [208, Lemma 3.1]. Thus he was among the first to recognize the importance of the UCEM property in establishing the asymptotic optimality of identification algorithms. Moreover, he was also able to *establish* that

the UCEM property holds under appropriate conditions. In contrast, in the statistical learning theory literature, very general conditions for the UCEM property to hold for real-valued functions were derived by Vapnik and Chervonenkis only in 1981; see [377]. (Similar results for binary-valued, as opposed to real-valued, functions were published by the same authors ten years earlier; see [376].)

6. Note that the result given here is not by any means the most general possible. In particular, it is possible to show that if θ_t is chosen so as to 'nearly' minimize the empirical error 'with high probability,' then the resulting algorithm will still be asymptotically optimal. In this more general version of the theorem, θ_t need not *always* minimize the empirical error $\hat{J}_t$. Rather, the quantity $\tilde{P}_{\mathbf{u},\mathbf{y}}\{\hat{J}(\theta) > \hat{J}(\theta_t)\}$ should approach zero as $t \to \infty$. For an exposition of this approach to the standard PAC learning problem, see [382, Section 3.2]. However, while such an approach makes sense in the context of PAC learning theory (where the underlying probability measure is known), this approach would be meaningless in the context of system identification, where the underlying probability measure $\tilde{P}_{\mathbf{u},\mathbf{y}}$ is unknown.

10.4 A Result on the Uniform Convergence of Empirical Means

In this section, it is shown that the UCEM property of Theorem 1 does indeed hold in the commonly studied case where y_t is the output of a 'true' system corrupted by additive noise, and the loss function ℓ is the squared error. By Theorem 1, this implies that by choosing the estimated model $h(\theta_t)$ so as to minimize the cumulated least squares error, we will eventually obtain the best possible fit to the given time series. Note that no particular attempt is made here to state or prove the 'best possible' result. Rather, the objective is to give a flavour of the the statistical learning theory approach by deriving a result whose proof is free from technicalities.

We begin by listing below the assumptions regarding the family of models employed in identification, and on the time series. Recall that the symbol $\tilde{h}(\theta)\cdot\mathbf{u}_t$ denotes the function $(f_{\text{true}} - h(\theta))\cdot\mathbf{u}_t$. Define the collection of functions $\mathcal{G}$ mapping $\mathcal{U}$ into $\mathbb{R}$ as follows:

$$g(\theta) \doteq \mathbf{u} \mapsto \| \, (f - h(\theta)) \cdot \mathbf{u}_0 \, \|^2 \colon \mathcal{U} \to \mathbb{R},$$

$$\mathcal{G} \doteq \{g(\theta) : \theta \in \Theta\}.$$

Now the various assumptions are listed.

A1. There exists a constant M such that

$$|g(\theta) \cdot \mathbf{u}_0| \leq M, \; \forall \theta \in \Theta, \mathbf{u} \in \mathcal{U}.$$

This assumption can be satisfied, for example, by assuming that the true system and each system in the family $\{h(\theta), \theta \in \Theta\}$ is BIBO stable (with an upper bound on the gain, independent of θ), and that the set U is bounded (so that $\{u_t\}$ is a bounded stochastic process).

A2. For each integer $k \geq 1$, define

$$g_k(\theta) \cdot \mathbf{u}_t \doteq g(\theta) \cdot (u_{t-1}, u_{t-2}, \ldots, u_{t-k}, 0, 0, \ldots). \tag{10.7}$$

With this notation, define

$$\mu_k \doteq \sup_{\mathbf{u} \in \mathcal{U}} \sup_{\theta \in \Theta} |(g(\theta) - g_k(\theta)) \cdot u_0|.$$

Then the assumption is that μ_k is finite for each k and approaches zero as $k \to \infty$. This assumption essentially means that each of the systems in the model family has decaying memory (in the sense that the effect of the values of the input at the distant past on the current output becomes negligibly small). This assumption is satisfied, for example, if

- Each of the models $h(\theta)$ is a linear ARMA model of the form

$$y_t = \sum_{i=1}^{l} a_i(\theta) u_{t-i} + b_i(\theta) y_{t-i},$$

- The characteristic polynomials

$$\phi(\theta, z) \doteq z^{l+1} - \sum_{i=1}^{l} b_i(\theta) z^{l-i}$$

  all have their zeros inside a circle of radius $\rho < 1$, where ρ is independent of θ.
- The numbers $a_i(\theta)$ are uniformly bounded with respect to θ.

The extension of the above condition to MIMO systems is straight-forward and is left to the reader.

A3. Consider the collection of maps $\mathcal{G}_k = \{g_k(\theta) : \theta \in \Theta\}$, viewed as maps from U^k into $\mathbb{R}$. For each k, this family $\mathcal{G}_k$ has finite P-dimension, denoted by $d(k)$. (See [382, Chapter 4] for a definition of the P-dimension.)

Now we can state the main theorem.

Theorem 2. *Define the quantity $q(t, \epsilon)$ as in (10.3) and suppose Assumptions A1 through A3 are satisfied. Given an $\epsilon > 0$, choose $k(\epsilon)$ large enough that $\mu_k \leq \epsilon/4$ for all $k \geq k(\epsilon)$. Then for all $t \geq k(\epsilon)$ we have*

$$q(t, \epsilon) \leq 8k(\epsilon) \left(\frac{32e}{\epsilon} \ln \frac{32e}{\epsilon} \right)^{d(k(\epsilon))} \cdot \exp(-\lfloor t/k(\epsilon) \rfloor \epsilon^2 / 512 M^2), \tag{10.8}$$

where $\lfloor t/k(\epsilon) \rfloor$ denotes the largest integer part of $t/k(\epsilon)$.

Theorem 3. *Let all notation be as in Theorem 2. Then, in order to ensure that the current estimate θ_t satisfies the inequality $J(\theta_t) \leq J^* + \epsilon$ (i.e., is ϵ-optimal) with confidence $1-\delta$, it is enough to choose the number of samples t large enough that*

$$\lfloor t/k(\epsilon) \rfloor \geq \frac{512M^2}{\epsilon^2} \left[\ln\left(\frac{24k(\epsilon)}{\delta}\right) + d(k(\epsilon)) \ln\left(\frac{32e}{\epsilon}\right) + d(k(\epsilon)) \ln\ln\left(\frac{32e}{\epsilon}\right) \right]. \tag{10.9}$$

Proof of Theorem 2. Write $g(\theta) = g_k(\theta) + (g(\theta) - g_k(\theta))$, and define

$$q_1^k(t,\epsilon) \doteq \Pr\{\sup_{\theta \in \Theta} \left| \frac{1}{t} \sum_{i=1}^{t} g_k(\theta) \cdot \mathbf{u}_i - \mathbb{E}[g_k(\theta) \cdot \mathbf{u}_i, \tilde{P}] \right| > \epsilon\},$$

$$q_2^k(t,\epsilon) \doteq \Pr\{\sup_{\theta \in \Theta} \left| \frac{1}{t} \sum_{i=1}^{t} (g(\theta) - g_k(\theta)) \cdot \mathbf{u}_i - \mathbb{E}[(g(\theta) - g_k(\theta)) \cdot \mathbf{u}_i, \tilde{P}] \right| > \epsilon\},$$

Then it is easy to see that

$$q(t,\epsilon) \leq q_1^k(t,\epsilon/2) + q_2^k(t,\epsilon/2).$$

Now observe that if k is sufficiently large that $\mu_k \leq \epsilon/4$, then $q_2^k(t,\epsilon/2) = 0$. This is because, if $|(g(\theta) - g_k(\theta)) \cdot \mathbf{u}_i|$ is always smaller than $\epsilon/4$, then its expected value is also smaller than $\epsilon/4$, so that their difference can be at most equal to $\epsilon/2$. Since this is true for all $\mathbf{u}$ and all θ, the above observation follows. Thus it follows that if $k(\epsilon)$ is chosen large enough that $\mu_k \leq \epsilon/4$ for all $k \geq k(\epsilon)$, then

$$q(t,\epsilon) \leq q_1^{k(\epsilon)}(t,\epsilon/2) \ \forall t \geq k(\epsilon), \ \forall \epsilon. \tag{10.10}$$

Hence the rest of the proof consists of estimating $q_1^{k(\epsilon)}(t,\epsilon)$ when $t \geq k(\epsilon)$.

From here onwards, let us replace $k(\epsilon)$ by k in the interests of notational clarity. When $t \geq k$, define $l \doteq \lfloor t/k \rfloor$, and $r = t - kl$. Partition $\{1,\ldots,t\}$ into k intervals, as follows:

$$I_j \doteq \{i, i+k, \ldots, i+lk\} \text{ for } 1 \leq j \leq r, \text{ and}$$

$$I_j \doteq \{i, i+k, \ldots, i+(l-1)k\} \text{ for } r+1 \leq j \leq k.$$

Then we can write

$$\frac{1}{t} \sum_{i=1}^{t} g_k(\theta) \cdot \mathbf{u}_i = \frac{1}{t} \sum_{j=1}^{k} \sum_{i \in I_j} g_k(\theta) \cdot \mathbf{u}_i.$$

Now define

$$\alpha_j \doteq \frac{1}{l+1} \left| \sum_{i \in I_j} \left(g_k(\theta) \cdot \mathbf{u}_i - \mathbb{E}[g_k(\theta) \cdot \mathbf{u}_i, \tilde{P}] \right) \right|, \ 1 \leq j \leq r, \text{ and}$$

$$\alpha_j \doteq \frac{1}{l} \left| \sum_{i \in I_j} \Big(g_k(\theta) \cdot \mathbf{u}_i - \mathbb{E}[g_k(\theta) \cdot \mathbf{u}_i, \tilde{P}] \Big) \right|, \; r+1 \leq j \leq k.$$

Then, noting that $\mathbb{E}[g_k(\theta) \cdot \mathbf{u}_i, \tilde{P}]$ is independent of i due to the stationarity assumption, we get

$$\left| \frac{1}{t} \sum_{i=1}^{t} g_k(\theta) \cdot \mathbf{u}_i - \mathbb{E}[g_k(\theta) \cdot \mathbf{u}_i, \tilde{P}] \right| \leq \left| \sum_{j=1}^{r} \frac{l+1}{t} \alpha_j + \sum_{j=r+1}^{k} \frac{l}{t} \alpha_j \right|.$$

It follows that if $\alpha_j \leq \epsilon$ for each j, then the left side of the equality is also less than ϵ. So the following containment of events holds:

$$\left\{ \sup_{\theta \in \Theta} \left| \frac{1}{t} (g_k \cdot \mathbf{u}_i - \mathbb{E}[g_k \cdot \mathbf{u}_i, \tilde{P}] \right| > \epsilon \right\} \subseteq \bigcup_{j=1}^{k} \{\alpha_j > \epsilon\}.$$

Hence

$$q_1^k(t, \epsilon) \leq \sum_{j=1}^{k} \Pr\{\alpha_j > \epsilon\}. \tag{10.11}$$

Now note that each $g_k \cdot \mathbf{u}_i$ depends on only u_{i-1} through u_{i-k}. Hence, in the summation defining each of the α_j, the various quantities being summed are independent. Since it is assumed that the family $\{g_k(\theta), \theta \in \Theta\}$ has finite P-dimension $d(k)$, standard results from statistical learning theory can be used to bound each of the probabilities on the right side of (10.11). A small adjustment is necessary, however. The results stated in [382] for example assume that all the functions under study assume values in the interval $[0, 1]$, whereas in the present instance the functions $h(\theta) \cdot \mathbf{u}_i$ all assume values in the interval $[-M, M]$. Thus the range of values now has width $2M$ instead on one. With this adjustment, Equation (7.1) of [382] implies that

$$\Pr\{\alpha_j > \epsilon\} \leq 8 \left(\frac{16e}{\epsilon} \ln \frac{16e}{\epsilon} \right)^{d(k)} \exp(-(l{+}1)^2 \epsilon^2 / 128 M^2), \text{ for } 1 \leq j \leq r, \text{ and}$$

$$\Pr\{\alpha_j > \epsilon\} \leq 8 \left(\frac{16e}{\epsilon} \ln \frac{16e}{\epsilon} \right)^{d(k)} \exp(-l^2 \epsilon^2 / 128 M^2), \text{ for } r+1 \leq j \leq k.$$

Since $\exp(-(l+1)^2) < \exp(-l^2)$, the $l+1$ term can be replaced by l in the first inequality as well. Substituting these estimates into (10.11) yields the desired estimate

$$q_1^k(t, \epsilon) \leq 8k \left(\frac{16e}{\epsilon} \ln \frac{16e}{\epsilon} \right)^{d(k)} \exp(-l^2 \epsilon^2 / 128 M^2).$$

Finally, the conclusion (10.8) is obtained by replacing ϵ by $\epsilon/2$ in the above expression, and then applying (10.10). □

Proof of Theorem 3. By Corollary 1, we can conclude that $J(\theta_t) \leq J^* + \epsilon$ with confidence $1-\delta$ provided $q(t, \epsilon/3) \leq \delta/3$. This can be achieved by setting the right side of (10.9) less than or equal to $\delta/3$ and solving for t. Thus we wish to have

$$8k(\epsilon)\left(\frac{32e}{\epsilon}\ln\frac{32e}{\epsilon}\right)^{d(k(\epsilon))} \cdot \exp(-\lfloor t/k(\epsilon)\rfloor \epsilon^2/512M^2) \leq \delta/3,$$

or

$$\exp(\lfloor t/k(\epsilon)\rfloor \epsilon^2/512M^2) \geq \frac{24k(\epsilon)}{\delta}\left(\frac{32e}{\epsilon}\ln\frac{32e}{\epsilon}\right)^{d(k(\epsilon))},$$

or

$$\lfloor t/k(\epsilon)\rfloor \geq \frac{512M^2}{\epsilon^2}\left[\ln\left(\frac{24k(\epsilon)}{\delta}\right) + d(k(\epsilon))\ln\left(\frac{32e}{\epsilon}\right) + d(k(\epsilon))\ln\ln\left(\frac{32e}{\epsilon}\right)\right].$$

This completes the proof. □

10.5 Bounds on the P-Dimension

In order for the estimate in Theorem 2 to be useful, it is necessary for us to derive an estimate for the P-dimension of the family of functions defined by

$$\mathcal{G}_k \doteq \{g_k(\theta) : \theta \in \Theta\},$$

where $g_k(\theta) : U^k \to \mathbb{R}$ is defined by

$$g_k(\theta)(\mathbf{u}) \doteq \| (f - h(\theta)) \cdot \mathbf{u}_k \|^2,$$

where

$$\mathbf{u}_k \doteq (\ldots, 0, u_k, u_{k-1}, \ldots, u_1, 0, 0, \ldots).$$

Note that, in the interests of convenience, we have denoted the infinite sequence with only k non-zero elements as $u_k, \ldots, u_1$ rather than $u_0, \ldots, u_{1-k}$ as done earlier. Clearly this makes no difference. In this section, we state and prove such an estimate for the commonly occuring case where each system model $h(\theta)$ is an ARMA model where the parameter θ enters linearly. Specifically, it is supposed that the model $h(\theta)$ is described by

$$x_{t+1} = \sum_{i=1}^{l} \theta_i \; \phi_i(x_t, u_t), \; y_t = x_t, \tag{10.12}$$

where $\theta = (\theta_1, \ldots, \theta_l) \in \Theta \subseteq \mathbb{R}^l$, and each $\phi_i(\cdot,\cdot)$ is a polynomial of degree no larger than r in the components of x_t, u_t.

Theorem 4. *With the above assumptions, we have that*

$$\begin{aligned}\text{P-dim}(\mathcal{G}_k) &\leq 9l + 2l\log_2[2(r^{k+1}-1)/(r-1)] \\ &\approx 9l + 2lk\log_2(2r) \ \textit{if}\ r > 1.\end{aligned}$$

In case $r = 1$ so that each system is linear, the above bound can be simplified to

$$\text{P-dim}(\mathcal{G}_k) \leq 9l + 2l\log_2(2k).$$

Remark 1. It is interesting to note that the above estimate is *linear* in both the number of parameters l and the duration k of the input sequence $\mathbf{u}$, but is only logarithmic in the degree of the polynomials ϕ_i. In the practically important case of linear ARMA models, even k appears inside the logarithm.

Proof. For each function $g_k(\theta) : U^k \to \mathbb{R}$ defined as in (10.7), define an associated function $g'_k : U^k \times [0,1] \to \{0,1\}$ as follows:

$$g'_k(\theta)(\mathbf{u}, c) \doteq \eta[g_k(\theta)(\mathbf{u}) - c],$$

where $\eta(\cdot)$ is the Heaviside or 'step' function. Then it follows from [382, Lemma 10.1] that

$$\text{P-dim}(\mathcal{G}_k) = \text{VC-dim}(\mathcal{G}'_k).$$

Next, to estimate VC-dim$(\mathcal{G}'_k)$, we use [382, Corollary 10.2] which states that, if the condition $\eta[g_k(\theta)\mathbf{u} - c] = 1$ can be stated as a Boolean formula involving s polynomial inequalities, each of degree no larger than d, then

$$\text{VC-dim}(\mathcal{G}'_k) \leq 2l\log_2(4eds). \tag{10.13}$$

Thus the proof consists of showing that the conditions needed to apply this bound hold, and of estimating the constants d and s.

Towards this end, let us back-substitute repeatedly into the ARMA model (10.12) to express the inequality

$$\| \, (f - h(\theta))\mathbf{u}_k \, \|^2 - c < 0$$

as a polynomial inequality in $\mathbf{u}$ and the θ-parameters. To begin with, we have

$$\begin{aligned} x_{k+1} &= \sum_{i=1}^{l} \theta_i \ \phi_i(x_k, u_k) \\ &= \sum_{i=1}^{l} \theta_i \phi_i \left(\sum_{j=1}^{l} \theta_j \ \phi_j(x_{k-1}, u_{k-1}) \right) \\ &= \ldots \end{aligned}$$

Thus each time one of the functions ϕ_i is applied to its argument, the degree with respect to any of the θ_j goes up by a factor of r. In other words, the total degree of x_{k+1} with respect to each of the θ_j is no larger than $1 + r + r^2 + \ldots + r^k = (r^{k+1} - 1)/(r - 1)$. If $r = 1$, then the degree is simply k. Next, we can write

$$\| x_{k+1} \|^2 - c < 0 \iff x'_{k+1} x_{k+1} - c < 0.$$

This is a single polynomial inequality. Moreover, the degree of this polynomial in the components of θ is at most $2(r^{k+1} - 1)/(r - 1)$ if $r > 1$, and $2k$ if $r = 1$. Thus we can apply the bound (10.13) with and $s = 1$, and

$$d = \begin{cases} \frac{2(r^{k+1}-1)}{r-1} & \text{if } r > 1, \\ 2k & \text{if } r = 1. \end{cases}$$

The desired estimate now follows on noting that $\log_2 e < 1.5$, so that $\log_2(8e) < 4.5$. □

10.6 Definition of Beta-Mixing and Significance

The main result of the remainder of the chapter shows that an exponentially stable *nonlinear* system driven by i.i.d. noise with bounded variance, and satisfying a few additional technical assumptions, generates a β-mixing sequence. This is the first time that such a general result is available in the literature. Thus the results of [394] have much wider applicability than is shown in that paper.

10.6.1 Mixing Coefficients of Stochastic Processes

Given a stationary stochastic process $\{\mathcal{X}_t\}$, it is desirable to have a notion of how dependent $\{\mathcal{X}_{t+k}, \mathcal{X}_{t+k+1}, \ldots\}$ are on $\{\mathcal{X}_t, \mathcal{X}_{t-1}, \ldots\}$. There are several different notions of mixing used in the literature, but only three are introduced here, namely α-mixing, β-mixing and ϕ-mixing. Actually, β-mixing is the notion with which we are most concerned. However, the other two definitions are widely used in the literature. Moreover, an earlier paper [77] uses ϕ-mixing processes. Thus they are introduced for the purposes of completeness.

A little bit of notation is introduced first to facilitate the definitions. For each index k, let $\varSigma_{-\infty}^{k}$ denote the σ-algebra generated by the coordinate random variables $\mathcal{X}_i, i \leq k$, and similarly let $\varSigma_k^{\infty}$ denote the σ-algebra generated by the coordinate random variables $\mathcal{X}_i, i \geq k$. Next, suppose we are given the probability measure $\tilde{P}$ on the doubly infinite Cartesian product space $\varXi$, and note that $\varXi$ is itself the product of the singly infinite product spaces $\mathcal{X}_- \doteq \prod_{i=-\infty}^{0} X$ and $\mathcal{X}^+ \doteq \prod_{i=1}^{\infty} X$. Let $\tilde{P}_{-\infty}^{0}$ denote the marginal probability of $\tilde{P}$ on X_-, and similarly, let $\tilde{P}_1^{\infty}$ denote the marginal probability of $\tilde{P}$ on X^+. Finally define $\tau_0(\tilde{P}) \doteq \tilde{P}_{-\infty}^{0} \times \tilde{P}_1^{\infty}$. Then it is clear that $\tau_0(\tilde{P})$ is the unique probability measure on $(\varXi, \S^{\infty})$ such that:

1. The laws of $\{\mathcal{X}_i, i \leq 0\}$ under $\tilde{P}$ and under $\tau_0(\tilde{P})$ are the same.

2. The laws of $\{\mathcal{X}_j, j \geq 1\}$ under $\tilde{P}$ and under $\tau_0(\tilde{P})$ are the same.
3. Under the measure $\tau_0(\tilde{P})$, the variables $\{\mathcal{X}_i, i \leq 0\}$ are independent of $\{\mathcal{X}_j, j \geq 1\}$. This means that each $\mathcal{X}_i, i \leq 0$ is independent of each $\mathcal{X}_j, j \geq 1$.

Let $\bar{\Sigma}_1^{k-1}$ denote the σ-algebra generated by the random variables $\mathcal{X}_i, i \leq 0$ as well as $\mathcal{X}_j, j \geq k$. Thus the bar over the Σ serves to remind us that the random variables between 1 and $k-1$ are missing from the list of variables that generate Σ.

Now we are ready to state the definitions.

Definition 2. *The* α-mixing coefficient *of the stochastic process* $\{\mathcal{X}_t\}$ *is defined as*

$$\alpha(k) \doteq \sup_{A \in \Sigma_{-\infty}^0, B \in \Sigma_k^\infty} |\tilde{P}(A \cap B) - \tilde{P}(A) \cdot \tilde{P}(B)|.$$

The β-mixing coefficient *of the stochastic process is defined as*

$$\beta(k) \doteq \sup_{C \in \bar{\Sigma}_1^{k-1}} |\tilde{P}(C) - (\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty)(C)|.$$

The ϕ-mixing coefficient *of the stochastic process is defined as*

$$\phi(k) \doteq \sup_{A \in \Sigma_{-\infty}^0, B \in \Sigma_k^\infty} |\tilde{P}(B|A) - \tilde{P}(B)|.$$

In the definition of the α-mixing coefficient, A is an event that depends only on the 'past' random variables $\{\mathcal{X}_i, i \leq 0\}$ while B is an event that depends only on the 'future' random variables $\{\mathcal{X}_i, i \geq k\}$. If the future event B were to be truly independent of the past event A, then the probability $\tilde{P}(A \cap B)$ would exactly equal $\tilde{P}(A)\tilde{P}(B)$. Thus the α-mixing coefficient measures how near to independence future events are of past events, by taking the supremum of the difference between the two quantities $\tilde{P}(A \cap B)$ and $\tilde{P}(A)\tilde{P}(B)$. Similarly, if the future event B were to be truly independent of the past event A, then the conditional probability $\tilde{P}(B|A)$ would exactly equal the unconditional probability $\tilde{P}(B)$. The ϕ-mixing coefficient measures how near to independence future events are of past events, by taking the supremum of the difference between the two quantities $\tilde{P}(B|A)$ and $\tilde{P}(B)$. The β-mixing coefficient has a somewhat more involved interpretation. If the future events beyond time k were to be truly independent of the past events before time 0, then the probability measure $\tilde{P}$ would exactly equal the product measure $\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty$ when restricted to the σ-algebra $\bar{\Sigma}_1^{k-1}$. The β-mixing coefficient of the stochastic process equals the total variation metric between the true probability measure $\tilde{P}$ and the product $\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty$ when restricted to the σ-algebra $\bar{\Sigma}_1^{k-1}$. Thus the mixing coefficient $\beta(k)$ measures how nearly the product measure $\tilde{P}_{-\infty}^0 \times \tilde{P}_1^\infty$ approximates the actual measure $\tilde{P}$ on $\bar{\Sigma}_1^{k-1}$.

Since $\Sigma_{k+1}^{\infty} \subseteq \Sigma_k^{\infty}$, it is obvious that

$$\alpha(k+1) \leq \alpha(k), \ \beta(k+1) \leq \beta(k), \ \phi(k+1) \leq \phi(k).$$

Moreover, it can also be shown that

$$\alpha(k) \leq \beta(k) \leq \phi(k) \ \forall k.$$

Definition 3. *The stochastic process $\{\mathcal{X}_t\}$ is said to be* α-mixing *if $\alpha(k) \to 0$ as $k \to \infty$,* β-mixing *if $\beta(k) \to 0$ as $k \to \infty$, and* ϕ-mixing *if $\phi(k) \to 0$ as $k \to \infty$.*

It is ironic that some authors refer to α-mixing as 'strong' mixing, even though it is the weakest notion of mixing. In some papers, especially the Russian literature, α-mixing is also referred to as 'strong regularity,' β-mixing as 'complete regularity,' and ϕ-mixing as 'uniform regularity.'

In case any of these mixing coefficients decays at an exponential rate, we say that the mixing is 'geometric.' Thus, for example, if $\beta(k) = O(r^k)$ for some $r < 1$, then the stochastic process is said to be 'geometrically β-mixing.'

10.6.2 Significance of Beta-Mixing Sequences

The significance of β-mixing arises from a result proved in [394], which shows that it is possible to derive finite time bounds for system identification in the case where the time series to be identified is β-mixing, and the model family consists of ARMA models of bounded degree. Specifically, the problem studied in [394] is as follows:

W1. The time series to be identified is denoted by $\{(u_t, y_t)\}$, and is assumed to be β-mixing with a geometrically decaying β-mixing coefficient.

W2. The family of models consists of systems of the form

$$A(\mathrm{d})y = B(\mathrm{d})u,$$

where d denotes a one time-step delay, and

$$A(\mathrm{d}) = 1 + \sum_{i=1}^{r} a_i \mathrm{d}^i, \ B(\mathrm{d}) = \sum_{i=1}^{s} b_i \mathrm{d}^i.$$

If we define $\theta \doteq (a_1, \ldots, a_r, b_1, \ldots, b_s)$, then the system can also be written as

$$y_t = \theta \cdot \phi_t,$$

where ϕ_t is the regression vector

$$\phi_t \doteq (-y_{t-1}, \ldots, -y_{t-r}, u_{t-1}, \ldots, u_{t-s}).$$

In [394] it is assumed that the system in SISO (single-input, single-output) but this assumption is not necessary. In particular, Lemma A.11 of [394] can be readily modified to the case where the system is MIMO. Let q, p denote respectively the dimensions of the vectors u and y.

W3. The loss function $l(y, \hat{y})$ is taken as *any increasing function* of the ℓ_∞ norm $\| y - \hat{y} \|_\infty$. Note that in [394] he uses $|y - \hat{y}|$. In order to generalize his arguments to the MIMO case, we replace the absolute value by the ℓ_∞ norm. It is further assumed that the loss function takes values in an interval $[0, M]$.

With these assumptions, the following result is shown in [394, Lemma 7].

Theorem 5. *With the above notation and assumptions, for every time t and every integer $k \leq t$, define $l \doteq \lfloor t/k \rfloor$, the integer part of t/k. Define the 'dimension'*

$$d \doteq 2p(pr + qs) + 1.$$

Then, whenever $\epsilon > 4M/l$, the quantity $q(t, \epsilon)$ defined in (10.3) is bounded by

$$q(t, \epsilon) \leq 4ed \left(\frac{32eM}{\epsilon} \right)^d \exp(-l\epsilon^2/128M^2) + 2l\beta(k).$$

A couple of points are noteworthy about this version of Weyer's theorem.

- In [394], the system is assumed to be SISO. As a result, the regression vector has dimension $r + s$. In the present instance, the regression vector has dimension $pr + qs$. Hence the integer k in [394, Lemma A.10] now becomes $pr + qs$. Similarly, since y and $\hat{y}$ are now p-dimensional vectors, the inequality $\| y - \hat{y} \|_\infty > c$ can be written as a set of $2p$ inequalities

$$(y - \hat{y})_i > c \text{ or } (y - \hat{y})_i < -c, i = 1, \ldots, p.$$

  With these adjustments, the P-dimension estimate $2k + 1$ in Lemma A.10 of [394] now becomes the quantity d defined above.
- The multiplier 16 on the right side of Lemma 7 appears to us a simple error in arithmetic and should be 4 instead.

In [394], the author did not give conditions under which any time series is β-mixing. Indeed, in [78], the authors say that they are motivated by the observation that 'signals generated by dynamical systems are not β-mixing in general'. As shown below in Theorem 7, actually this statement is *quite false – practically every* time series encountered in system identification is β-mixing. Thus the results of [394] have very wide applicability, though this was not shown in that paper.

10.7 Statement of Main Results on Beta-Mixing of Markov Chains

Since β-mixing plays such a central role in the present chapter, we begin with an extremely general result that shows that a large class of nonlinear recursions are β-mixing. There appears to be some confusion in the literature about β-mixing as a property. Indeed, in [78], the authors make the statement that 'signals generated by dynamical systems are not β-mixing in general'. Actually, *exactly the opposite* is true: Theorem 6 below shows that a very wide class of Markov chains naturally occuring in control and system theory *are* β-mixing.

We state the main result at once so that the reader can see where we are going. The proof itself is spread over the next two sections.

Throughout the remainder of the chapter, $| \cdot |$ denotes the Euclidean, or ℓ_2-norm on $\mathbb{R}^k$ and on $\mathbb{R}^m$. Where convenient, we also use the same symbol $| \cdot |$ to denote the *matrix* norm induced by the ℓ_2-norm, that is, the largest singular value of a matrix. Thus, in this notation, if $A \in \mathbb{R}^{k \times k}$ and $v \in \mathbb{R}^k$, we have $|Av| \leq |A| \cdot |v|$.

Throughout, we consider Markov chains described by the recursion relation

$$\mathcal{X}_{t+1} = f(\mathcal{X}_t, \mathbf{e}_t), \tag{10.14}$$

where $x_t \in \mathbb{R}^k, \mathbf{e}_t \in \mathbb{R}^m$ for some integers k, m, and $\{\mathbf{e}_t\}$ is a stationary noise sequence. It is assumed that the following assumptions are satisfied:

A1. The function $f : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^k$ is 'smooth,' *i.e.*, is C^∞, and in addition, f is globally Lipschitz continuous. Thus there exist constants L and K such that

$$|f(x, u) - f(y, v)| \leq L|x - y| + K|u - v|.$$

A2. The noise sequence $\{\mathbf{e}_t\}$ is i.i.d., has finite variance, and has a continuous multivariate density function $\phi(\cdot)$ that is positive in some neighbourhood Ω of the origin in $\mathbb{R}^m$.

A3. When $\mathbf{e}_t = 0 \; \forall t$, the 'unforced' system

$$x_{t+1} = f(x_t, 0)$$

is globally exponentially stable with the origin as the unique globally attractive equilibrium. This means that there exist constants M' and $l < 1$ such that

$$|x_t| \leq M'|x_0| l^t, \; \forall t \geq 1, \; \forall x_0.$$

By taking $M \doteq \max\{M', 1\}$, one can write the above inequality as

$$|x_t| \leq M|x_0| l^t, \; \forall t \geq 0, \; \forall x_0.$$

A4. The associated deterministic control system

$$x_{t+1} = f(x_t, u_t) \tag{10.15}$$

is 'globally forward accessible' from the origin with the control set Ω. In other words, for every $y \in \mathbb{R}^k$, there exist a time N and a control sequence $\{u_0, \ldots, u_{N-1}\} \subseteq \Omega$ such that, with $x_0 = 0$ we have $x_N = y$.

A5. The associated deterministic control system (10.15) is 'locally controllable' to the origin with the control set Ω. This means that there exists a neighbourhood $\mathcal{B}$ of the origin in $\mathbb{R}^k$ such that, for every $y \in \mathcal{B}$ there exist a time N and a control sequence $\{u_0, \ldots, u_{N-1}\} \subseteq \Omega$ such that, with $x_0 = y$ we have $x_N = 0$.

Now we can state the main result.

Theorem 6. *Suppose assumptions A1 through A5 hold. Then the state sequence $\{\mathcal{X}_t\}$ is geometrically β-mixing.*

Theorem 7. *Suppose assumptions A1 through A5 hold. Then the sequence $\{\mathcal{Y}_t = (\mathcal{X}_t, \epsilon_t)\}$ is geometrically β-mixing.*

As a concrete illustration of the above theorems, consider the linear system

$$\mathcal{X}_{t+1} = A\mathcal{X}_t + Be_t,\ \mathcal{Y}_t = C\mathcal{X}_t,$$

where the matrix A has all of its eigenvalues inside the unit circle, and the pair (A, B) is controllable; note that it is *not* assumed that the pair (C, A) is observable. Under these conditions, if $\{e_t\}$ is an i.i.d. sequence with bounded variance (*e.g.*, Gaussian noise), then both the state sequence $\{\mathcal{X}_t\}$ and the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ are β-mixing. This is in sharp contrast to ϕ-mixing. The following result due to Athreya and Pantula shows that ϕ-mixing is an extremely restrictive concept.

Lemma 1. ([19], Theorem 2) *Consider the first-order recursion*

$$\mathcal{X}_{t+1} = l\mathcal{X}_t + e_t,$$

where $l \in [0, 1)$ is some constant, and $\{e_t\}$ is an i.i.d. sequence independent of $\mathcal{X}_0$. Suppose:

1. *$\mathbb{E}[\{\log(e_1)\}_+] < \infty$, where $(\cdot)_+$ denotes the positive part.*
2. *For some $n \geq 1$, the random variable $\sum_{i=1}^{n} l^i e_i$ has a nontrivial absolutely continuous component. (This assumption is satisfied if $l > 0$ and e_1 has a nontrivial absolutely continuous component.)*

Then $\{\mathcal{X}_t\}$ is ϕ-mixing if and only if the noise sequence $\{e_t\}$ is essentially bounded, that is, there exists a constant M such that

$$|e_t| \leq M a.s.$$

The interesting part of the above lemma is the 'only if' part. This lemma implies that even the simple situation of a stable recursion driven by Gaussian noise is not ϕ-mixing, since Gaussian noise is unbounded. In contrast, such a sequence is indeed β-mixing. Thus it appears that β-mixing is a more natural and useful notion than ϕ-mixing.

10.8 Beta Mixing Properties of Markov Chains

In this section we present some sufficient conditions to ensure that a Markov chain is β-mixing. These conditions are stated in terms of a property known as V-geometric ergodicity. Throughout this section, the principal reference is [227]. Then we show that, if a stationary stochastic process is β-mixing, then so is any 'measurement' process obtained from it. This result is relevant to so-called 'hidden Markov models.'

10.8.1 Background Material on Markov Chains

We begin by introducing some background material on Markov chains and hidden Markov models (HMMs). Suppose $(X, \S)$ is a measurable space. For the purposes of the present discussion, a *Markov chain* is a sequence of random variables $\{\mathcal{X}_m\}_{m \geq 0}$ together with a set of probability measures $P^n(x, A), x \in X, A \in \S$ denoting the 'transition probabilities.' It is assumed that

$$\Pr\{\mathcal{X}_{n+m} \in A | \mathcal{X}_j, j \leq m\} = P^n(\mathcal{X}_m, A).$$

Thus $P^n(x, A)$ denotes the probability that the state $\mathcal{X}$ will belong to the set A after n time steps, starting from the initial state x at time m. It is common to denote the 'one-step' transition probability by $P(x, A)$, so that $P^1(x, A) = P(x, A)$. The fact that the transition probability does not depend on the values of $\mathcal{X}$ prior to time m is the Markov property, and the fact that the transition probability does not depend on the 'initial time' m means that the Markov chain is stationary.

Suppose the Markov chain is set in motion with the initial state at time $t = 0$ distributed according to the probability measure Q_0. Then the definition of $P(\cdot, \cdot)$ implies that

$$Q_1(A) \doteq \Pr\{\mathcal{X}_1 \in A\} = \int_X P(x, A)\, Q_0(dx).$$

Under suitable conditions (see [227] for a detailed treatment), a stationary Markov chain has an *invariant measure* or a *stationary distribution* π on $(X, \S)$ with the property that

$$\pi(A) = \int_X P(x, A)\, \pi(dx).$$

Thus, if the Markov chain is started off with the initial state distributed according to the the stationary distribution π, then at all subsequent times the state continues to be distributed according to π.

10.8.2 An Expression for the Beta-Mixing Coefficient of a Markov Chain

The main result of this subsection gives a characterization of the β-mixing coefficient in terms of an abstract integral. Note that the formula (10.16) below is actually due to Davydov [97], but a complete proof is given here for the convenience of the reader.

Theorem 8. *Suppose a Markov chain has m-step transition probability $P^m(\cdot,\cdot)$ and a stationary distribution π. Then its β-mixing coefficient is given by*

$$\beta(m) = \mathbb{E}\{\rho[P^m(x,\cdot),\pi],\pi\} = \int_X \rho[P^m(x,\cdot),\pi]\ \pi(dx). \tag{10.16}$$

Note that the third expression is just a restatement of the second expression. Thus the importance of the theorem is in relating the β-mixing coefficient to the expected value of the difference between the m-step transition probability $P^m(x,\cdot)$ and the invariant probability π.

In order to prove the theorem we require two preliminary lemmas. The first is on decomposition of measures, also known as existence of regular conditional probabilities. The second lemma shows that two distinct ways of defining the β-mixing coefficient are in fact equivalent. This lemma is important since both definitions are widely used in the literature, but so far as can be ascertained, the equivalence of the two formulas is not explicitly stated anywhere.

Let us begin with a little notation. Suppose Z_1, Z_2 are complete separable metric spaces, and that $\S_1, \S_2$ are the corresponding Borel σ-algebras of subsets of Z_1 and Z_2 respectively. Define $Z = Z_1 \times Z_2$ and let $\S = \S_1 \times \S_2$ be the corresponding product algebra. Define $\mathcal{G}_1 = \S_1 \times \{\emptyset, Z_2\}$, and similarly $\mathcal{G}_2 = \{\emptyset, Z_1\} \times \S_2$. Suppose P is a probability measure on $(Z, \S)$, and let P_1, P_2 denote the marginal probability measures of P on Z_1 and Z_2 respectively. Thus for $A_1 \in \S_1, A_2 \in \S_2$ we have

$$P_1(A_1) = P(A_1 \times Z_2),$$

and similarly for P_2. Now we are ready to state the lemma on existence of regular conditional probabilities.

Lemma 2. *With the above notation, there exists a probability transition function $Q : Z_1 \times \S_2 \to [0,1]$, that is, $Q(z_1,\cdot)$ is a probability measure on $(Z_2, \S_2)$ for all $z_1 \in Z_1$, and $Q(\cdot, A_2) \in \S_1$ for all $A_2 \in \S_2$, such that for all $A \in \S$ we have*

$$P(A) = \int_{Z_1} Q(z_1, A(z_1))\ P_1(dz_1),$$

where $A(z_1)$: the z_1-section of A of is given by

$$A(z_1) \doteq \{z_2 : (z_1, z_2) \in A\}.$$

Further,

$$E_P(I_A|\mathcal{G}_1) = Q(\cdot, A(\cdot)),$$

where $E_P(I_A|\mathcal{G}_1)$ *denotes the best approximation to the indicator function* $I_A(\cdot)$ *among functions measurable with respect to* $\mathcal{G}_1$*, and the error measure is the* L_2*-norm with respect to the measure* P*. In other words,* $f(z_1) = Q(z_1, A(z_1))$ *satisfies*

$$E_P(I_A|\mathcal{G}_1) = f(\cdot).$$

The proof can be found in, for example, [64].

Next, it is shown that two distinct-looking definitions of the β-mixing coefficient that are widely used in the literature are in fact equivalent.

Lemma 3. *With the notation as above, let* $\mathcal{H}_2 \subseteq \S_2$ *be a sub-*σ*-algebra on* Z_2 *such that* $(Z_1, \S_1)$*,* $(Z_2, \mathcal{H}_2)$ *are standard-Borel. Let*

$$\beta \doteq \sup_{A \in \S_1 \times \mathcal{H}_2} |P(A) - (P_1 \times P_2)(A)|,$$

$$\begin{aligned}\theta &\doteq \mathbb{E}[\sup_{A_2 \in \mathcal{H}_2} |Q(z_1, A_2) - P_2(A_2)|, P_1] \\ &= \int_{Z_1} \sup_{A_2 \in \mathcal{H}_2} |Q(z_1, A_2) - P_2(A_2)| \; P_1(dz_1).\end{aligned}$$

Then $\beta = \theta$*.*

Remark 2. We will show in the course of the proof that the expression appearing in the previous line is measurable.

Proof. To prove that $\theta \leq \beta$, we proceed as follows. Let

$$R(z_1, A_2) = Q(z_1, A_2) + P(A_2), \; A_2 \in \mathcal{H}_2.$$

Then for all z_1, $Q(z_1, \cdot)$ is absolutely continuous with respect to $R(z_1, \cdot)$. Let us denote the Radon-Nikodym derivative by $f(z_1, \cdot)$ and let $g(z_1, \cdot)$ be the Radon-Nikodym derivative of P_2 with respect to $R(z_1, \cdot)$. It is well known in the probability literature that $f(z_1, z_2)$ and $g(z_1, z_2)$ can be chosen to be jointly measurable with respect to $\S_1 \times \mathcal{H}_2$ and hence

$$A \doteq \{(z_1, z_2) : f(z_1, z_2) \geq g(z_1, z_2)\}$$

belongs to $\S_1 \times \mathcal{H}_2$. Let

$$A(z_1) \doteq \{z_2 : f(z_1, z_2) \geq g(z_1, z_2)\}.$$

Then $z_1 \longrightarrow (Q(z_1, A(z_1)) - P_2(A(z_1)))$ is measurable and it can be checked that

$$Q(z_1, A(z_1)) - P_2(A(z_1)) = \sup_{C_2 \in \mathcal{H}_2} |Q(z_1, C_2) - P_2(C_2)|.$$

Hence

$$\begin{aligned} P(A) - (P_1 \times P_2)(A) &= \int_{Z_1} [Q(z_1, A(z_1)) - P_2(A(z_1))] \, P_1(dz_1) \\ &= \theta. \end{aligned}$$

Therefore

$$\beta = \sup_{A \in \S_1 \times \mathcal{H}_2} |P(A) - (P_1 \times P_2)(A)| \geq \theta.$$

To show that $\beta \leq \theta$, suppose that $A \in \S_1 \times \mathcal{H}_2$. Then

$$P(A) = \int_{Z_1} Q(z_1, A(z_1)) \, P_1(dz_1), \; (P_1 \times P_2)(A) = \int_{Z_1} P_2(A(z_1)) \, P_1(dz_1).$$

Hence

$$\begin{aligned} |P(A) - (P_1 \times P_2)(A)| &\leq \int_{Z_1} |Q(z_1, A(z_1)) - P_2(A(z_1))| \, P_1(dz_1) \\ &\leq \int_{Z_1} \sup_{A_2 \in \mathcal{H}_2} |Q(z_1, A_2) - P_2(A_2)| \, P_1(dz_1) = \theta. \end{aligned}$$

Here we use the fact that $A(z_1) \in \mathcal{H}_2$ since $A \in \S_1 \times \mathcal{H}_2$. Since the above argument holds for every A, it follows that $\beta \leq \theta$. This completes the proof. □

Proof of Theorem 8. Let $\Xi, \S^\infty$ and $\{\mathcal{X}_t\}$ be as before. Let $Y = \prod_{i=1}^\infty X$ and let $\mathcal{T}$ be the corresponding product σ-algebra generated by $\S$. (The difference between Ξ and Y is that Ξ is a doubly infinite Cartesian product whereas Y is a singly infinite Cartesian product; similarly for $\S^\infty$ *vs* $\mathcal{T}$.) Let $\tilde{P}$ be a probability measure on $(\Xi, \S^\infty)$ such that $\{X_n\}$ is a stationary Markov chain with the transition probability $P(x, A)$ and stationary distribution π. Similarly, let Q_x be the probability measure on $(Y, \mathcal{T})$ such that $\{\mathcal{Y}_n\}$ is a stationary Markov chain with transition probability $P(x, A)$ and initial distribution δ_x, where $\{\mathcal{Y}_n\}$ denotes the 'co-ordinate random variables' on $(Y, \mathcal{T})$. To apply Lemma 3, we identify

$$Z_1 = \prod_{i=-\infty}^{0} X, \; Z_2 = \prod_{i=1}^{\infty} X.$$

Let $\tilde{P}_1, \tilde{P}_2$ be the marginal measures of $\tilde{P}$ on Z_1 and Z_2 respectively. Here we make use of the Markov property which implies that the conditional probability

$$\tilde{P}\{(\mathcal{X}_1, \mathcal{X}_2, \ldots) \in A_2 | \mathcal{X}_i : i \leq 0\}$$

depends only on $\mathcal{X}_0$ and equals $Q_{\mathcal{X}_0}(A_2)$. Hence we identify $Q(z_1, A_2)$ with $Q_{x_0}(A_2)$ where $z_1 = \{\ldots, x_{-1}, x_0\}$. For $D \in \mathcal{T}$, we have

$$P\{(\mathcal{X}_{m+1}, \mathcal{X}_{m+2}, \ldots) \in D | \sigma\{\mathcal{X}_i, i \leq m\}\} = Q_{\mathcal{X}_m}(D).$$

Now define $\mathcal{H}_2 \doteq \sigma\{\mathcal{X}_i, i \geq m\}$. Then the β-mixing coefficient $\beta(m)$ is given, using the result of Lemma 3, by

$$\int_{\Xi} \sup_D |P\{(\mathcal{X}_{m+1}, \mathcal{X}_{m+2}, \ldots) \in D|\sigma(\mathcal{X}_i, i \leq 0)\} - P\{(\mathcal{X}_{m+1}, \mathcal{X}_{m+2}, \ldots) \in D\}| \; dP.$$

Note that

$$\begin{aligned} P\{(\mathcal{X}_{m+1}, x_{m+2}, \ldots) \in D\} &= \int_{\Xi} P\{(\mathcal{X}_{m+1}, x_{m+2}, \ldots) \in D|\sigma(X_m)\} dP \\ &= \int_{\Xi} Q_{\mathcal{X}_m}(D) dP = \int_X Q_y(D) \; \pi(dy), \end{aligned}$$

since the only random variable under the integral sign is $\mathcal{X}_m$. Similarly

$$\begin{aligned} &P\{(\mathcal{X}_{m+1}, \mathcal{X}_{m+2}, \ldots) \in D|\mathcal{X}_i, i \leq 0\} \\ &= \mathbb{E}[P\{(\mathcal{X}_{m+1}, \mathcal{X}_{m+2}, \ldots) \in D|\sigma\{X_i, i \leq m\}|\sigma\{X_i, i \leq 0\}] \\ &= \mathbb{E}[Q_{\mathcal{X}_m}(D)|\sigma\{X_i, i \leq 0\}] \\ &= \int_X Q_y(D) \; P^m(\mathcal{X}_0, dy) \end{aligned}$$

Thus

$$\beta(m) = \int_X \sup_D \left| \int_X Q_y(D) \; P^m(x_0, dy) - \int_X Q_y(D) \; \pi(dy) \right| \; \pi(dx_0). \quad (10.17)$$

Since $Q_y(D) \leq 1$, it is clear that

$$\sup_D \left| \int_X Q_y(D) \; P^m(x_0, dy) - \int_X Q_y(D) \; \pi(dy) \right| \leq \rho[P^m(x_0, \cdot), \pi(\cdot)],$$

where ρ is the total variation metric. If in (10.17) we take D to be of the form $B \times X \times X \times \ldots$ where $B \in \S$, it follows that the left side is in fact no smaller than the right side. Therefore we finally have

$$\beta(m) = \int_X \rho[P^m(x_0, \cdot), \pi(\cdot)] \; \pi(dx_0),$$

which is the same as (10.16). □

10.8.3 Characterization of Beta-Mixing in Terms of *V*-Geometric Ergodicity

In this subsection, we begin by recalling a notion called V-geometric ergodicity from [227]. Then it is shown that V-geometric ergodicity implies geometric β-mixing.

A (stationary) Markov chain is said to be *geometrically ergodic* if there exist constants μ and $l < 1$ such that

$$\rho[P^n(x,\cdot),\pi] \leq \mu l^n,\ \forall x \in X.$$

Note that here ρ denotes the total variation metric between two probability measures. Thus in a geometrically ergodic Markov chain, the total variation metric distance between the n-step transition probability $P^n(x,\cdot)$ and the stationary distribution π decays to zero at a geometric rate; moreover, this rate is *independent of the initial state* x. If the state space X is not compact, it is not reasonable to expect such a strong type of convergence to hold. To cater to the general situation, a more liberal notion called 'V-geometric ergodicity' is introduced. A stationary Markov chain is said to be *V-geometrically ergodic* with respect to the measurable function $V : X \to [1,\infty)$ if there exist constants μ and $l < 1$ such that

$$\rho[P^n(x,\cdot),\pi] \leq \mu l^n V(x),\ \forall x \in X,$$

and in addition,

$$\mathbb{E}[V,\pi] = \int_X V(x)\ \pi(dx) < \infty.$$

Actually, the notion of V-geometric ergodicity as defined in [227] is more restrictive than the above. Specifically, in [227] the total variation metric $\rho[P^n(x,\cdot),\pi]$ is replaced by a larger quantity that can be thought of as the total variation *with respect to all functions bounded by* V. Since V is bounded below by one, this latter quantity is no smaller than $\rho[P^n(x,\cdot),\pi]$. Consequently V-geometric ergodicity in the sense of [227] implies the above inequality.

Thus a Markov chain is V-geometrically ergodic if two conditions hold. First, there is a non-negative-valued function V such that the total variation distance between the n-step transition probability $P^n(x,\cdot)$ and the invariant measure π approaches zero at a geometric rate *multiplied by* $V(x)$. Thus the *rate* of geometric convergence is independent of x, but the *multiplicative constant* is allowed to depend on x. To ensure that the property is meaningful, the second condition is imposed, namely that the 'growth function' $V(\cdot)$ has finite expectation with respect to the invariant measure π. Thus 'on average' the total variation metric distance between the n-step transition probability and the stationary distribution decays to zero at a geometric rate.

Theorem 9. *Suppose a Markov chain is V-geometrically ergodic. Then it is geometrically β-mixing, i.e., there exist constants B and $l < 1$ such that $\beta(m) \leq Bl^m$ for all m.*

Proof. Since the Markov chain is V-geometrically ergodic, it follows that

$$\rho[P^m(x_0,\cdot),\pi(\cdot)] \leq V(x_0)\mu l^m$$

for some function $V : X \to [1,\infty)$ such that $\mathbb{E}[V,\pi] < \infty$, and some constants μ and $l < 1$. Consequently, it follows from (10.16) that

$$\beta(m) \leq \mu\mathbb{E}[V,\pi]l^m,\ \forall m,$$

which shows that the Markov chain is geometrically β-mixing. □

10.8.4 Hidden Markov Models

Next we introduce the notion of hidden Markov models (HMMs) as used in this chapter. Suppose $\{\mathcal{X}_m\}_{m\geq 0}$ is a stationary Markov chain assuming values in a set X with associated σ-algebra §, and that Y is a complete separable metric space, called the 'output space.'[4] Let $\mathcal{B}(Y)$ denote the Borel σ-algebra on Y. Suppose $\mu : X \times \mathcal{B}(Y) \rightarrow [0,1]$ is a 'transition probability' function. This means that, for each $x \in X$, $\mu(x,\cdot)$ is a probability measure on Y, and for each $A \in \mathcal{B}(Y)$, $\mu(\cdot, A)$ is a measurable function on $(X, §)$. Then a stochastic process $\{\mathcal{Y}_m\}_{m\geq 0}$ is called a *hidden Markov model (HMM)* if

$$\Pr\{\mathcal{Y}_m \in A | \mathcal{Y}_i, i \leq m-1, \mathcal{X}_j, j \leq m\} = \mu(\mathcal{X}_m, A),\ \forall A \in \mathcal{B}(Y).$$

In other words, the Markov process $\{\mathcal{X}_m\}$ generates a probability $\mu(\mathcal{X}_m,\cdot)$ on Y, and this is the conditional law of $\mathcal{Y}_m$ given $\{\mathcal{Y}_i, i \leq m-1\}$ and $\{\mathcal{X}_j, j \leq m\}$.

The next result shows that if a Markov chain is (geometrically) β-mixing, so is any hidden Markov model generated from the Markov chain. Actually, the result is more general than that.

Theorem 10. *Suppose $\{\mathcal{X}_t\}_{t\geq 0}$ is a stationary stochastic process assuming values in a set X with associated σ-algebra §. Suppose Y is a complete separable metric space, and let $\mathcal{B}(Y)$ denote the Borel σ-algebra on Y. Suppose $\mu : X \times \mathcal{B}(Y) \rightarrow [0,1]$ is a transition probability function. Thus for each $x \in X$, $\mu(x,\cdot)$ is a probability measure on Y, and for each $A \in \mathcal{B}(Y)$, $\mu(\cdot, A)$ is a measurable function on $(X, §)$. Finally, suppose $\{\mathcal{Y}_t\}_{t\geq 0}$ is a Y-valued stochastic process such that*

$$\Pr\{\mathcal{Y}_t \in A | \mathcal{Y}_i, i \leq t-1, \mathcal{X}_j, j \leq t\} = \mu(\mathcal{X}_t, A).$$

Under these assumptions, if $\{\mathcal{X}_t\}$ is β-mixing, so is $\{\mathcal{Y}_t\}$.

The proof of this theorem is given at the end of this section.

Next we give a proof of Theorem 10. The proof is based on a couple of preliminary lemmas.

Lemma 4. *Suppose a real-valued stochastic process $\{\mathcal{X}_t\}$ is α-, β-, or ϕ-mixing, and that $\mathcal{Y}_t = f(\mathcal{X}_t)$ where $f : X \rightarrow \mathbb{R}$. Then $\{\mathcal{Y}_t\}$ is also α-, β-, or ϕ-mixing, as appropriate.*

Proof. Note that mixing is really a property of the σ-algebras generated by the stochastic process. Since $\mathcal{Y}_t$ is a measurable function of $\mathcal{X}_t$, we see that the σ-algebra generated by any collection of the $\mathcal{Y}_t$ is a subset of (and perhaps equal to) the σ-algebra generated by the corresponding collection of $\mathcal{X}_t$. Hence the $\mathcal{Y}_t$ stochastic process inherits the mixing properties of the $\{\mathcal{X}_t\}$ sequence. □

[4] The assumption that the output space Y is a complete separable metric space is made to facilitate some of the measure-theoretic arguments in the sequel.

Lemma 5. *Suppose $\{\mathcal{X}_t\}$ is β-mixing, and that $\{\mathcal{U}_t\}$ is i.i.d. and also independent of $\{\mathcal{X}_t\}$. Suppose $\mathcal{Y}_t = f(\mathcal{X}_t, \mathcal{U}_t)$, where f is a fixed measurable function. Then $\{\mathcal{Y}_t\}$ is also β-mixing.*

Proof. Note that under the hypotheses, it follows that the joint process $\{(\mathcal{X}_t, \mathcal{U}_t)\}$ is β-mixing. Now the desired conclusion follows from Lemma 4. □

Proof of Theorem 10. The theorem is proved by constructing a representation of $\mathcal{Y}_t$ as a deterministic function of $\mathcal{X}_t$ and another random variable $\mathcal{U}_t$ that is i.i.d. and also independent of $\mathcal{X}_t$. The conclusion then follows from Lemma 5. Specifically, it is shown that there exists a measurable mapping $\psi : X \times [0,1] \to Y$ such that the process $\{\mathcal{Z}_t\}_{t \geq 0}$ defined by

$$\mathcal{Z}_t = \psi(\mathcal{X}_t, \mathcal{U}_t)$$

has the same distribution as $\{\mathcal{Y}_t\}$, where $\{\mathcal{U}_t\}_{t \geq 0}$ is a sequence of i.i.d. random variables whose common distribution is the uniform distribution on $[0,1]$.

Recall (see, *e.g.*, [263]) that if Y is a complete separable metric space, then there exists a Borel subset E of $[0,1]$ and a one-to-one onto mapping ϕ from Y into E such that both ϕ and ϕ^{-1} are measurable. With ϕ as above, define the transition function $\nu : X \times \mathcal{B}(E) \to [0,1]$ as follows:

$$\nu(x, B) \doteq \mu(x, \phi^{-1}(B)), \ \forall x \in X, B \in \mathcal{B}(E).$$

Here $\mathcal{B}(E)$ denotes the σ-algebra of Borel subsets of E. Now define the map $\psi_0 : X \times [0,1] \to [0,1]$ as follows:

$$\psi_0(x, s) \doteq \lim_{m \to \infty} \frac{1}{2^m} \inf\{k \geq 0 : \nu(x, (-\infty, k/2^m)) \geq s\}.$$

It readily follows from the above definition that the function ψ_0 is jointly measurable. Moreover, it is easy to see that

$$\psi_0(x, s) = \inf\{u \geq 0 : \nu(x, (-\infty, u)) \geq s\}.$$

However, the above equation is not used as a definition of ψ_0 since it involves an infimum over an uncountable set, and it is therefore not clear that the resulting function is jointly measurable.

From the above equation it can be seen that

$$\psi_0(x, s) \leq u \text{ if and only if } \nu(x, (-\infty, u)) \geq s.$$

Hence, if l denotes the Lebesgue measure on $[0,1]$, it follows that

$$l\{s : \psi_0(x, s) \leq u\} = \nu(x, (-\infty, u)).$$

Now define

$$\psi(x, s) \doteq \phi^{-1}(\psi_0(x, s)).$$

Then for each $A \in \mathcal{B}(Y)$ we have

$$l\{s : \psi(x, s) \in A\} = \mu(x, A).$$

Therefore, if $\{\mathcal{U}_t\}_{t \geq 0}$ is a sequence of i.i.d. random variables whose common distribution is the uniform distribution on $[0, 1]$, then the process $\{\mathcal{Z}_t\}_{t \geq 0}$ defined by

$$\mathcal{Z}_t = \psi(\mathcal{X}_t, \mathcal{U}_t)$$

has the same distribution as $\{\mathcal{Y}_t\}$.

Finally, by Lemma 5, if $\{\mathcal{X}_t\}$ is β-mixing, so is $\{\mathcal{Y}_t\}$. □

Similar results can be proven for α- and ϕ-mixing. Moreover, notice that the above proof does not require the process $\{\mathcal{X}_t\}$ to be Markovian. However, in the absence of a result like Theorem 6 that guarantees the mixing properties of the sequence $\{\mathcal{X}_t\}$, Theorem 10 might not be very useful by itself.

10.9 Proofs of Main Results

In this section we give proofs of Theorems 6 and 7. Actually, all the effort goes into proving Theorem 6. The proof of this is very long and requires a great deal of build up, to address lots of technical issues. The proof is given in Section 10.9.3.

10.9.1 'Petiteness' of Compact Sets

Consider the Markov Chain defined by (10.14). Note that the probability transition function $P(x, A)$ for this chain is given by

$$P(x, A) = \int_{\mathbb{R}^m} I_A(f(x, e))\phi(e)de$$

(Recall that ϕ denotes the (common) density of the noise sequence $\mathbf{e}_t$). In this subsection, it is shown that when Conditions A1 through A5 hold, every compact set in $\mathbb{R}^k$ is 'petite' in the sense defined in [227]. This is a technical condition needed to apply the main theorem of [227] to deduce that the Markov chain (10.14) is V-geometrically ergodic. It is possible that Lemma 6 holds under weaker conditions than A1 through A5, but this requires further investigation.

Lemma 6. *Suppose assumptions A1 through A5 hold. Then every compact set in $\mathbb{R}^k$ is 'petite' in the sense of [227].*

Proof. To prove this claim we proceed as follows:
Claim 1: *The system (10.14) is a T-chain.*

To establish this claim, we invoke [227, Proposition 7.1.5], which states that the system (10.14) is a T-chain if the associated control system (10.15) is 'forward accessible' in the sense of [227], *i.e.*, if it is possible to start from any initial state $y \in \mathbb{R}^k$ and reach any final state $z \in \mathbb{R}^k$ in a finite number of time steps, using only controls from Ω. But this kind of forward accessibility is immediate. If $y \in \mathcal{B}$ (see Condition A5), then it is possible to steer the state y to the origin in a finite number of time steps using only controls from Ω. Then, from condition A4, it is possible to steer the state from the origin to z in another finite number of steps, again using only controls from Ω. Now suppose y lies outside $\mathcal{B}$. Then by applying the control $u_t = 0$, it is possible to ensure that the resulting state x_t enters $\mathcal{B}$ in a finite number of time steps, since the unforced system is globally exponentially stable (see Condition A3). Now apply the previous argument.

Claim 2: The origin in $\mathbb{R}^k$ is a globally attractive state in the sense of [227, p. 160].[5] Moreover, the system (10.14) is a ψ-irreducible aperiodic T-chain.

To establish the first part of the claim, it is necessary to show that, for every $y \in \mathbb{R}^k$, there exist an integer N and a control sequence in Ω such that the resulting state x_N comes arbitrarily close to the origin (the claimed globally attractive state). In fact we can do better than that, by showing that x_N actually *equals* 0. Let $\mathcal{B}$ be the neighbourhood of 0 in Condition A5. Since the unforced system is globally asymptotically stable, if we simply apply no control (and recall that the origin in the control space $\mathbb{R}^m$ belongs to the control set Ω), then within a finite number of time steps the state trajectory starting from y enters the set $\mathcal{B}$. Then in another finite number of steps the state can be steered to the origin using only controls from Ω.

Now let $\overline{A_+(0)}$ denote the closure of the set of all states reachable from 0 (see [227, Equation 7.10]). By condition A4, this set in fact equals $\mathbb{R}^k$. By [227, Proposition 7.2.5], $\mathbb{R}^k$ is the unique minimal set of this Markov chain. Clearly $\mathbb{R}^k$ is connected, so the Markov chain is aperiodic; see [227, Proposition 7.3.4]. Next, by [227, Proposition 7.3.5], the Markov chain is an aperiodic ψ-irreducible T-chain.

Finally, by [227, Theorem 6.2.5], all compact sets are petite. □

10.9.2 A Converse Lyapunov Theorem for Discrete-Time Systems

Converse Lyapunov theory refers to results which state that, if a system has a particular type of stability property, then there exists a corresponding Lyapunov function satisfying an associated set of conditions. Converse Lyapunov theory for continuous-time systems is well-studied but nowadays it is rare to find this theory mentioned in textbooks. Most of the books that discuss converse Lyapunov theory are now out of print, and [383] is among the few extant books that discusses this theory. However, even [383] does not explicitly discuss converse Lyapunov theory for *discrete-time* systems, which is needed in the present context.

[5] Note that the phrase 'globally attractive state' is used in [227] in a different sense than in stability theory.

Though Theorem 11 is used here as an intermediate step in the proof of the main result, namely Theorem 6, it is of independent interest and is thus stated and proved separately.

Theorem 11. *Consider the system*

$$g(x) \doteq f(x,0) : \mathbb{R}^k \to \mathbb{R}^k.$$

Suppose the following conditions hold:[6]

B1. g is C^1, and there exists a constant L such that

$$|\nabla g(x)| \leq L, \; \forall x \in \mathbb{R}^k.$$

B2. Define the functions $g^n : \mathbb{R}^k \to \mathbb{R}^k$ recursively as follows:

$$g^0(x) \doteq x, \; g^n(x) \doteq g[g^{n-1}(x)].$$

Suppose there exist constants $M < \infty, l < 1$ such that

$$|g^n(x)| \leq M|x|l^n, \; \forall n \geq 0. \tag{10.18}$$

Under these conditions, there exists a C^1 function $S : \mathbb{R}^k \to [0,\infty)$ satisfying the following properties, for suitable constants $c_1 \in [1,\infty)$, $c_2, c_3 > 0$:

L1. $|x|^2 \leq S(x)$, $\forall x$.
L2. $S(x) \leq c_1|x|^2$, $\forall x$.
L3. $S[g(x)] - S(x) \leq -c_2 S(x) \leq -c_2|x|^2$, $\forall x$.
L4. $|\nabla S(x)| \leq c_3(x)$, $\forall x$.
L5. $S(x+y) - S(x) \leq c_3|x| \cdot |y| + (c_3/2)|y|^2$, $\forall x, y$.

Remark 3. Note that Properties L1 through L4 are quite standard in converse Lyapunov theory. However, Property L5 is not usually included as a part of the theorem, but is needed here. Hence we give an *ab initio* proof of the theorem, though the various steps in the proof are by now well-known in the Lyapunov theory literature.

Proof. We begin by defining an intermediate function $W : \mathbb{R}^k \to [0,\infty)$ as follows: Choose an integer p large enough that $Ll^{2p-1} < 1$; this is possible since $l < 1$. Define

$$W(x) \doteq \sum_{n=0}^{\infty} |g^n(x)|^{2p} = \sum_{n=0}^{\infty} \{[g^n(x)]^t g^n(x)\}^p.$$

It is now shown that the function W has the following properties:

W1. $W(x) \geq |x|^{2p}$, $\forall x$.

[6] Note that Condition B1 is weaker than Condition A1, in that g is required only to be C^1, not C^∞.

W2. We have

$$W(x) \le \frac{M^{2p}}{1-l^{2p}} |x|^{2p}, \ \forall x.$$

W3. $W[g(x)] - W(x) = -|x|^{2p}, \ \forall x.$

W4. We have

$$|\nabla W(x)| \le 2p \frac{M^{2p-1}}{1 - Ll^{2p-1}} |x|^{2p-1}, \ \forall x.$$

Property W1 is obvious, since $|x|^2$ is the first term in the infinite series that defines $W(x)$.

To prove Property W2, use (10.18). Thus

$$W(x) \le \sum_{n=0}^{\infty} M^{2p} l^{2p} |x|^{2p} = \frac{M^{2p}}{1-l^{2p}} |x|^{2p}.$$

To prove Property W3, note that from the definition of $W(\cdot)$ we have

$$W[g(x)] = \sum_{n=0}^{\infty} |g^{n+1}(x)|^{2p} = \sum_{n=1}^{\infty} |g^n(x)|^{2p} = W(x) - |x|^{2p},$$

since $g^0(x) = x$.

Finally, to prove Property W4, let us compute $\nabla W(x)$ by differentiating the infinite series for $W(x)$ inside the summation (which can be easily justified). This gives

$$\begin{aligned}\nabla W(x) &= \sum_{n=0}^{\infty} p[|g^n(x)|^2]^{p-1} \cdot \nabla[|g^n(x)|^2] \\ &= 2p \sum_{n=0}^{\infty} |g^n(x)|^{2p-2} \cdot \nabla g^n(x) \cdot g^n(x).\end{aligned}$$

Hence

$$|\nabla W(x)| \le 2p \sum_{n=0}^{\infty} |g^n(x)|^{2p-1} |\nabla g^n(x)|.$$

However, since $|\nabla g(x)| \le L \ \forall x$, it follows by induction that $|\nabla g^n(x)| \le L^n \ \forall x$. Therefore, after invoking (10.18), we get

$$|\nabla W(x)| \le 2p \sum_{n=0}^{\infty} M^{2p-1} l^{(2p-1)n} L^n |x|^{2p-1} = 2p \frac{M^{2p-1}}{1 - Ll^{2p-1}} |x|^{2p-1}.$$

To complete the proof, define the function $S : \mathbb{R}^k \to [0, \infty)$ by

$$S(x) \doteq [W(x)]^{1/p}.$$

It is now shown that the function S has each of the claimed properties L1 through L5.

L1 follows readily from W1. L2 follows readily from W2, where the definition of the constant c_1 is obvious.

To prove L3, note that from W3 we have

$$W[g(x)] = W(x) - |x|^{2p} \leq (1-d)W(x),$$

where

$$d \doteq \frac{1 - l^{2p}}{M^{2p}} < 1.$$

Taking the $1/p$-th power of both sides leads to

$$S[g(x)] \leq (1-d)^{1/p} S(x) \doteq (1-c_2)S(x),$$

where $c_2 \doteq 1 - (1-d)^{1/p} > 0$. This is the same as

$$S[g(x)] - S(x) \leq -c_2 S(x) \leq -c_2 |x|^2,$$

where the last inequality follows from L1 which is already established.

To prove L4, note that

$$\nabla S(x) = \frac{1}{p}[W(x)]^{-(p-1)/p} \nabla W(x).$$

Hence it follows from W4 that

$$|\nabla S(x)| \leq \text{const.}|x|^{-2(p-1)} \cdot |x|^{2p-1} = \text{const.}|x|.$$

If let c_3 denote the constant in the above we get L4.

Finally, to prove L5, define $h(\alpha) \doteq S(x+\alpha y)$, and note that $h(1) = S(x+y)$ and $h(0) = S(x)$. Therefore

$$\begin{aligned} S(x+y) - S(y) = h(1) - h(0) &= \int_0^1 h'(\alpha) d\alpha \\ &= \int_0^1 \nabla S(x+\alpha y) \cdot y d\alpha \\ &\leq \int_0^1 c_3 |x+\alpha y| \cdot |y| d\alpha \\ &\leq \int_0^1 c_3 [|x| + \alpha |y|] \cdot |y| d\alpha \\ &= c_3 |x| \cdot |y| + \frac{c_3}{2} |y|^2. \end{aligned}$$

This completes the proof. □

10.9.3 Proof of Theorem 6

At last we are in a position to give a proof of Theorem 6.

Proof of Theorem 6. Define the function $S : \mathbb{R}^k \to [0, \infty)$ as in Theorem 11, and define the function $V : \mathbb{R}^k \to [1, \infty)$ by

$$V(x) \doteq 1 + S(x), \ \forall x.$$

Recall ([227, p. 174]) that the 'drift' of the stochastic Lyapunov function V for the system (10.14) is defined as

$$\mathrm{d}V(x) = \int_X V(y)\ P(x, dy) - V(x) = \int_{\mathbb{R}^m} \{V(f(x, e)) - V(x)\}\phi(e)de.$$

By Lemma 6, we already know that every compact set in $\mathbb{R}^k$ is petite. Thus, by a theorem in [227, p. 354], the Markov chain is V-geometrically ergodic if there exist constants γ, ν such that

$$\mathrm{d}V(x) \leq -\gamma V(x) + I_{B(\nu)}\ \forall x, \tag{10.19}$$

where $B(\nu)$ denotes the closed ball of radius ν centered at the origin, and $I.$ denotes the indicator function. Thus the proof consists of obtaining an upper bound for $\mathrm{d}V(x)$ and showing that (10.19) is satisfied. This will establish V-geometric ergodicity of the Markov chain, which in turn implies geometric β-mixing by Theorem 9.

Computing directly, we have (using property L3 of S)

$$\begin{aligned} V[f(x, e)] - V(x) &= V[f(x, 0)] - V(x) + (V[f(x, e)] - V[f(x, 0)]) \\ &\leq -c_2|x|^2 + V[f(x, e)] - V[f(x, 0)]. \end{aligned}$$

Now write

$$f(x, e) = f(x, 0) + z,$$

where

$$z = f(x, e) - f(x, 0).$$

By Condition A1 we have $|z| \leq K|e|$. By Property L5 of $S(\cdot)$, we have

$$\begin{aligned} V[f(x, e)] - V[f(x, 0)] &= S[f(x, 0) + z] - S[f(x, 0)] \\ &\leq c_3|f(x, 0)| \cdot |z| + \frac{c_3}{2}|z|^2 \\ &\leq c_3 L|x| \cdot |z| + \frac{c_3}{2}|z|^2 \\ &\leq c_3 LK|x| \cdot |e| + \frac{c_3 K^2}{2}|e|^2 \\ &\doteq c_4|x| \cdot |e| + c_5|e|^2. \end{aligned}$$

Now the drift term can be estimated as follows:

$$\begin{aligned}
\mathrm{d}V(x) &= \int_{\mathbb{R}^m} [V(f(x,e)) - V(x)]\ \phi(e)de \\
&\leq \int_X [-c_2|x|^2 + c_4|x| \cdot |e| + c_5|e|^2]\phi(e)de \\
&= -c_2|x|^2 + c_4|x| \cdot \parallel \mathbf{e}_1 \parallel_1 + c_5 \parallel \mathbf{e}_1 \parallel_2^2 \\
&\leq -c_2|x|^2 + c_4|x| \cdot \parallel \mathbf{e}_1 \parallel_2 + c_5 \parallel \mathbf{e}_1 \parallel_2^2,
\end{aligned}$$

Here, $\parallel \mathbf{e}_1 \parallel_1$ and $\parallel \mathbf{e}_1 \parallel_2$ are the L^1 and L^2 norms of the random variable $\mathbf{e}_1$ which are finite as $\mathbf{e}_1$ is assumed to have finite variance. Let ν' denote the positive root of the quadratic equation

$$-\frac{c_2}{2} r^2 + c_4 r \parallel \mathbf{e}_1 \parallel_2 + c_5 \parallel \mathbf{e}_1 \parallel_2^2 = 0,$$

and let $\nu \doteq \max\{\nu', 1\}$. Then, whenever $|x| \geq \nu$, we have

$$-\frac{c_2}{2}|x|^2 + c_4|x| \cdot \parallel \mathbf{e}_1 \parallel_2 + c_5 \parallel \mathbf{e}_1 \parallel_2^2 \leq 0.$$

As a result,

$$\begin{aligned}
\mathrm{d}V(x) &\leq -\frac{c_2}{2}|x|^2 \leq -\frac{c_2}{2}\left[\frac{1}{1+c_1}|x|^2 + \frac{c_1}{1+c_1}|x|^2\right] \\
&\leq -\frac{c_2}{2}\left[\frac{1}{1+c_1} + \frac{1}{1+c_1}S(x)\right] \\
&\leq -\frac{c_2}{2(1+c_1)}V(x).
\end{aligned}$$

Hence (10.19) holds with ν as above and $\gamma \doteq c_2/[2(1+c_1)]$. This shows that the Markov chain is V-geometrically ergodic and hence geometrically β-mixing.

□

Proof of Theorem 7. We have shown that the Markov Chain $\{\mathcal{X}_t\}$ defined by (10.14) is V-geometrically ergodic and hence is β-mixing. We now show that the chain consisting of the state $\mathcal{X}_t$ and the noise e_t together also inherit these properties. Let $\mathcal{Y}_t = (\mathcal{X}_t, e_t)$ denote the augumented chain. Note that for Borel sets A, B

$$\begin{aligned}
P\{\mathcal{Y}_{n+m} \in A \times B | \mathcal{Y}_i,\ i \leq m\} &= P\{(\mathcal{X}_{n+m}, e_{n+m}) \in A \times B | \mathcal{X}_i, e_i,\ i \leq m\} \\
&= P\{\mathcal{X}_{n+m} \in A | \mathcal{X}_i, e_i,\ i \leq m\} P\{e_{n+m} \in B\} \\
&= P\{\mathcal{X}_{n+m} \in A | \mathcal{X}_m, e_m\} \mu_\phi(B) \\
&= P\{\mathcal{X}_{n+m} \in A | f(\mathcal{X}_m, e_m)\} \mu_\phi(B) \\
&= P^{n-1}(f(\mathcal{X}_m, e_m)) \mu_\phi(B)
\end{aligned}$$

where μ_ϕ is the (common) distribution of $_t$. Thus it is clear that $\{\mathcal{Y}_t\}$ is a Markov chain whose n- step transition probability function $Q^n((x,e), A)$ is given by

$$Q^n((x,e),\cdot) = P^{(n-1)}(f(x,e),\cdot) \otimes \mu_\phi$$

and $P^{(n-1)}(f(x,e),\cdot)\otimes\mu_\phi$ denotes the product of the measure $P^{(n-1)}(f(x,e),\cdot)$ (on $\mathbb{R}^k$) and and $\pi\mu_\phi$ (on $\mathbb{R}^m$).

Thus if π denotes the stationary distribution for the Markov chain $\mathcal{X}_t$, it follows that the stationary distribution of $\mathcal{Y}_t$ is given by $\widetilde{\pi} = \pi \otimes \mu_\phi$. Since the second component of the two product measures, $Q^n((x,e),\cdot)$ and $\widetilde{\pi}$ are the same, it follows that

$$\rho(Q^n((x,e),\cdot),\widetilde{\pi}) = \rho(P^{(n-1)}(f(x,e),\cdot),\pi)$$

Since $\{\mathcal{X}_t\}$ is V-geometrically ergodic, it follows that there exists $\lambda < 1$, μ such that

$$\rho(P^n(x,\cdot),\pi) \le \mu\lambda^n V(x)$$

with

$$\int V(x)\pi(dx) < \infty.$$

Defining $\widetilde{V}(x,e) = V(x), \widetilde{\mu} = \frac{\mu}{\lambda}$, using (10.20) we conclude that

$$\rho(Q^n((x,e),\cdot),\widetilde{\pi}) \le \widetilde{\mu}\lambda^n\widetilde{V}(x,e)$$

Further,

$$\int \widetilde{V}(x,e)\delta\widetilde{\pi}(x,e) = \int V(x)\pi(dx) < \infty.$$

Thus we have shown that the chain $\{\mathcal{Y}_t\}$ is $\widetilde{V}$-geomoetrically ergodic and hence is also beta mixing. □

10.10 Conclusions

In this chapter, a beginning has been made towards showing that it is possible to use the methods of statistical learning theory to derive *finite time* estimates for use in system identification theory. Obviously there is a great deal of room for improvement in the *specific results* presented here. For instance, in Sections 10.4 and 10.5, it would be desirable to combine the fading memory argument and the ARMA model into a single step. This would require new results in statistical learning theory, whereby one would have to compute the VC-dimension of mappings whose range is an infinite-dimensional space. This has not been the practice thus far.

In summary, the message of this contribution is that both system identification theory and statistical learning theory can enrich each other. Much work remains to be done to take advantage of this potential.

11

Probabilistic Design of a Robust Controller Using a Parameter-Dependent Lyapunov Function

Yasuaki Oishi

Department of Mathematical Informatics
Graduate School of Information Science and Technology
The University of Tokyo
Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
oishi@mist.i.u-tokyo.ac.jp

Summary. An extension of a randomized algorithm is considered for the use of a parameter-dependent Lyapunov function. The proposed algorithm is considered to be useful for a less conservative design of a robust state-feedback controller against nonlinear parametric uncertainty. Indeed, one can avoid two sources of conservatism with the proposed algorithm: covering the uncertainty by a polytope and using a common Lyapunov function for all parameter values. After a bounded number of iterations, the proposed algorithm either gives a probabilistic solution to the provided control problem with high confidence or detects infeasibility of the problem in an approximated sense. Convergence to a non-strict deterministic solution is discussed. Usefulness of the proposed algorithm is illustrated by a numerical example.

11.1 Introduction

Although a robust-controller design is a classical control problem, it still remains difficult against nonlinear parametric uncertainty. One existing approach to this problem consists of two simplifications: (i) to cover the uncertainty by a polytope and consider the plants corresponding to the vertices of the polytope, and (ii) to assume a common Lyapunov function for all of those plants. These two simplifications enable us to formulate and solve the above problem in terms of a linear matrix inequality (LMI). On the other hand, these simplifications produce conservatism, that is, even if the original problem is solvable, the simplified one may not.

In order to reduce conservatism, we propose in this chapter a randomized algorithm that allows the use of a parameter-dependent Lyapunov function. This is based on a randomized algorithm, which was proposed in [250] as an extension of [76, 176, 271, 273], and also on a technique to use a parameter-dependent Lyapunov function, which was developed by de Oliveira *et al.* [100]

with further extensions by [11,13,101,116,328,334]. In particular, the proposed algorithm does not need to cover the uncertainty and allows a Lyapunov function having any parameter dependence.

A problem to be considered is formulated in terms of a parameter-dependent LMI having two types of variables: parameter-independent variables, which describe a controller, and parameter-dependent variables, which correspond to a Lyapunov function. The proposed algorithm iterates the following three steps: (i) random selection of uncertain parameter values; (ii) optimization of the parameter-dependent variables; (iii) update of the parameter-independent variables based on the sensitivity of the preceding optimization problem. This algorithm terminates after a bounded number of iterations. At its termination, it gives a probabilistic solution of the provided problem with high confidence or detects that the problem has no deterministic solution in an approximated sense. The above sensitivity is efficiently evaluated based on a theory of semidefinite programming. In [76, 176, 271, 273], it is shown that their algorithms find a non-strict deterministic solution instead of a probabilistic one after a finite number of iterations with probability one. We can prove a corresponding result on our algorithm. However, this result is not emphasized here because it has some practical difficulties as is discussed in [250] for a special case.

In Section 11.2, the problem to be considered is presented with an example. An algorithm is proposed in Section 11.3 with a theoretical guarantee on its performance. Section 11.4 describes how one can compute an infimum and a subgradient, which are required in the algorithm. In Section 11.5, convergence to a non-strict deterministic solution is discussed. Section 11.6 provides a numerical example and Section 11.7 concludes this chapter.

Throughout this chapter, ln stands for the natural logarithm. For a real number a, the symbol $\lceil a \rceil$ denotes the minimum integer that is larger than or equal to a. The symbol $\mathbb{R}^n$ designates the n-dimensional Euclidean space and vol the volume in $\mathbb{R}^n$. The symbol $\|\cdot\|$ denotes the Euclidean norm. The transpose of a matrix or a vector is expressed by $^{\mathrm{T}}$. The maximum (most positive) eigenvalue of a symmetric matrix A is written as $\overline{\lambda}[A]$. Negative definiteness and negative semidefiniteness of a symmetric matrix A are denoted by $A \prec 0$ and $A \preceq 0$, respectively.

11.2 Problem

The problem to be considered is stated in a general form. Let $V(x, y, \theta)$ be a symmetric-matrix-valued function in $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, and $\theta \in \Theta \subseteq \mathbb{R}^p$. At this moment, the parameter set Θ can be any set. The function $V(x, y, \theta)$ is affine in x and y and is written as

$$V(x, y, \theta) = V_0(\theta) + \sum_{i=1}^{n} x_i V_i(\theta) + \sum_{i=1}^{m} y_i V_{n+i}(\theta),$$

where x_i and y_i are the ith elements of x and y, respectively.

Problem 1. Find $x \in \mathbb{R}^n$ such that for any $\theta \in \Theta$ there exists $y \in \mathbb{R}^m$ satisfying $V(x, y, \theta) \prec 0$. □

The choice of y can depend on the parameter θ. The inequality $V(x, y, \theta) \prec 0$ is called a *parameter-dependent LMI.* When Θ consists of infinitely many components, our problem is equivalent to solving infinitely many LMIs with infinitely many variables. Note also that our problem reduces to the one considered in the existing papers [76,176,250,271,273] in the special case that V does not depend on y.

Here is an example that can be formulated in the above general form.

Example 1. Consider a discrete-time plant $\xi[k+1] = A(\theta)\xi[k] + B(\theta)u[k]$, which depends nonlinearly on a time-invariant but uncertain parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Suppose that we want a state-feedback controller $u[k] = K\xi[k]$ that robustly stabilizes this plant for all $\theta \in \Theta$. This problem can be stated in terms of a parameter-dependent LMI based on the result of de Oliveira *et al.* [100]. Namely, consider to find two matrices G and L such that for any $\theta \in \Theta$ there exists a symmetric matrix Y satisfying

$$-\begin{bmatrix} Y & A(\theta)G + B(\theta)L \\ G^{\mathrm{T}}A(\theta)^{\mathrm{T}} + L^{\mathrm{T}}B(\theta)^{\mathrm{T}} & G + G^{\mathrm{T}} - Y \end{bmatrix} \prec 0. \tag{11.1}$$

Then $K := LG^{-1}$ gives a desired controller. Here, $\eta[k]^{\mathrm{T}}Y\eta[k]$ works as a Lyapunov function for the transposed version of the closed-loop system $\eta[k+1] = [A(\theta) + B(\theta)K]^{\mathrm{T}}\eta[k]$. Since Y can be chosen depending on θ, the Lyapunov function is allowed to be parameter-dependent. This means that a less conservative result can be expected than in the case that a common Lyapunov function is used for all θ. In order to express this problem in our general form, construct x by the components of G and L, construct y by the independent components of Y, and write the left-hand side of (11.1) as V. □

The idea in this example has been extended to robust performance problems [11,101] and to problems on continuous-time systems [13,116,328,334]. These can be stated in our general form, too.

The x desired in Problem 1 is called a *deterministic solution.* The set of all deterministic solutions is referred to as the *solution set* $S \subseteq \mathbb{R}^n$. Rather than a deterministic solution itself, we consider to obtain an approximate solution in this chapter. For a given probability measure $\mathbb{P}$ on Θ and a given $0 < \epsilon < 1$, we mean by a *probabilistic solution* an $x \in \mathbb{R}^n$ that satisfies

$$\mathbb{P}\big\{\theta \in \Theta : \exists y \in \mathbb{R}^m \text{ s.t. } V(x, y, \theta) \prec 0\big\} > 1 - \epsilon.$$

By a *non-strict deterministic solution*, we mean an $x \in \mathbb{R}^n$ such that for any $\theta \in \Theta$ there exists $y \in \mathbb{R}^m$ satisfying $V(x, y, \theta) \preceq 0$. Note that the strict inequality in Problem 1 is replaced by the non-strict one.

11.3 Algorithm

This is the main section of this chapter and proposes a randomized algorithm for finding a probabilistic solution of Problem 1. The basic idea is as follows. Note that $V(x, y, \theta) \prec 0$ is equivalent to $\overline{\lambda}[V(x, y, \theta)] < 0$. Hence, for provided x and θ, there exists $y \in \mathbb{R}^m$ satisfying $V(x, y, \theta) \prec O$ if and only if $\inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x, y, \theta)] < 0$. This $\inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x, y, \theta)]$ is convex in x because $\overline{\lambda}[V(x, y, \theta)]$ is convex in x and y. By these facts, we can extend the randomized algorithms in [250], which are for the special case that V does not depend on y. Although the extended algorithm requires computation of the infimum value, $\inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x, y, \theta)]$, and its subgradient with respect to x, their computation is efficiently carried out as we will see in the next section.

In [250], two randomized algorithms are presented: the gradient-based algorithm, which was originally proposed by Polyak and Tempo [273] and Calafiore and Polyak [76], and the ellipsoid-based algorithm, which was by Kanev *et al.* [176]. We present only an extension of the ellipsoid-based algorithm because an extension of the gradient-based one is similar.

The algorithm to be proposed iteratively updates an ellipsoid, which is expected to contain a deterministic solution. The ellipsoid at the kth iteration, $E^{(k)}$, is specified by its center $x^{(k)}$ and a positive definite matrix $Q^{(k)}$ in the form of $\{x \in \mathbb{R}^n : (x - x^{(k)})^{\mathrm{T}} (Q^{(k)})^{-1} (x - x^{(k)}) < 1\}$.

Before executing the algorithm, we choose an initial ellipsoid $E^{(0)} = (x^{(0)}, Q^{(0)})$ and three numbers $0 < \mu$, $0 < \epsilon < 1$, and $0 < \delta < 1$. We usually choose these three numbers close to zero. We use two counters in the algorithm: k counts the number of iterations and ℓ the number of updates. We define an integer

$$\overline{\ell} := \left\lceil 2(n+1) \ln \frac{\operatorname{vol} E^{(0)}}{\mu} \right\rceil$$

and a function

$$\kappa(\ell) := \left\lceil \Bigl(\ln \frac{\pi^2 (\ell+1)^2}{6\delta} \Bigr) \Big/ \ln \frac{1}{1-\epsilon} \right\rceil .$$

Algorithm 11.1
Initialization: Set $k := 0$ and $\ell := 0$.

Step 1. If ℓ reaches $\overline{\ell}$, stop with no output.
Step 2. If $\lambda^{(k-1)}, \lambda^{(k-2)}, \ldots, \lambda^{(k-\kappa(\ell))}$ are all well-defined and negative, stop and give $x^{(k)}$ as an output.
Step 3. Randomly sample $\theta^{(k)} \in \Theta$ according to the given probability measure $\mathbb{P}$.
Step 4. Check $\lambda^{(k)} := \inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^{(k)}, y, \theta^{(k)})]$ for negativity.
Case 1 The value $\lambda^{(k)}$ is non-negative.
Compute a subgradient, say $d^{(k)}$, of $\inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^{(k)}, y, \theta^{(k)})]$ as a convex function in x. Update the current ellipsoid by setting

$$x^{(k+1)} := x^{(k)} - \frac{Q^{(k)} d^{(k)}}{(n+1)\sqrt{(d^{(k)})^{\mathrm{T}} Q^{(k)} d^{(k)}}},$$

$$Q^{(k+1)} := \frac{n^2}{n^2-1}\left[Q^{(k)} - \frac{2Q^{(k)} d^{(k)} (d^{(k)})^{\mathrm{T}} Q^{(k)}}{(n+1)(d^{(k)})^{\mathrm{T}} Q^{(k)} d^{(k)}}\right].$$

Case 2 The value $\lambda^{(k)}$ is negative.

Keep the current ellipsoid by setting $x^{(k+1)} := x^{(k)}$ and $Q^{(k+1)} := Q^{(k)}$.

Step 5. If $\lambda^{(k)} \geq 0$, set $\ell := \ell + 1$.

Step 6. Set $k := k + 1$. Go back to Step 1.

The performance of this algorithm is guaranteed by the following theorem, which is a direct generalization of Theorem 21 in [250], a result on the special case that V does not depend on y. In particular, the statement (a) evaluates the computational complexity and the statements (b) and (c) give properties of the output. Although the proof is almost the same as in the special case, it is presented in Appendix 11.8 for convenience of the readers. Notice that the probabilistic behavior of the algorithm is determined by the sequence $\theta^{(0)}, \theta^{(1)}, \ldots$ We hence analyze the behavior of the algorithm according to the probability measure on such sequences, which can be derived from $\mathbb{P}$ and is denoted by $\mathbb{P}^\infty$. See for example [335, Section II.3.4].

Theorem 1. *The following statements hold on Algorithm 11.1.*

(a) The number of iterations k is bounded as

$$k \leq \overline{\ell}\kappa(\overline{\ell} - 1) = \overline{\ell}\left\lceil \left(\ln \frac{\pi^2 \overline{\ell}^2}{6\delta}\right) \Big/ \ln \frac{1}{1-\epsilon} \right\rceil =: \overline{k}.$$

(b) If the algorithm terminates at Step 1, the volume of the set $E^{(0)} \cap S$ is less than μ.

(c) The probability that the algorithm stops at Step 2 but still the corresponding output $x^{(k)}$ fails to satisfy $\mathbb{P}\{\theta \in \Theta : \exists y \in \mathbb{R}^m \text{ s.t. } V(x^{(k)}, y, \theta) \prec 0\} > 1-\epsilon$ is less than or equal to δ, where the probability is measured with respect to $\mathbb{P}^\infty$.

The statement (a) of Theorem 1 means that our algorithm stops after at most $\overline{k}$ iterations. The number $\overline{k}$ is of order $\mathrm{O}((\overline{\ell}/\epsilon)\ln(\overline{\ell}^2/\delta))$, which is a polynomial in n, $\ln(\mathrm{vol}\, E^{(0)}/\mu)$, $1/\epsilon$, and $\ln(1/\delta)$. Note that this number does not depend on the dimension of the parameter p. This forms a sharp contrast with deterministic algorithms whose complexity is usually of exponential order in p.

When the algorithm stops at Step 1, we have $\mathrm{vol}\,(E^{(0)} \cap S) < \mu$ by the statement (b). This means that our choice of an initial ellipsoid $E^{(0)}$ is not good or the solution set S itself is too small. On the other hand, termination at Step 2 implies that the associated output is a probabilistic solution with high confidence. Note that the algorithm always stops at Step 2 if $\mathrm{vol}\,(E^{(0)} \cap S) \geq \mu$.

As is considered in Section 6 of [250] for the special case, we can adapt the algorithm for an optimization problem. Indeed, since the algorithm gives a negative or a positive result, its bisectional use leads us to finding a solution that approximately optimizes a given objective function. For the explicit algorithm and its properties, see Section 6 of [250].

11.4 Computation of the Infima and the Subgradients

In order to carry out Algorithm 11.1 in the previous section, we need to compute the infimum value, $\lambda^{(k)} = \inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^{(k)}, y, \theta^{(k)})]$, and its subgradient $d^{(k)}$. This section provides a method for this task.

Let us notice that the infimum value $\lambda^{(k)}$ is the optimal value of the following semidefinite programming problem:

$$\begin{aligned} &\text{minimize} && \lambda \\ &\text{subject to} && V(x^{(k)}, y, \theta^{(k)}) - \lambda I \preceq 0, \end{aligned} \tag{11.2}$$

where the optimization variables are y and λ. By following a general theory on semidefinite programming (see *e.g.* [331]), its dual problem is

$$\begin{aligned} &\text{maximize} && \operatorname{tr}[UV(x^{(k)}, 0, \theta^{(k)})] \\ &\text{subject to} && U \succeq 0, \quad \operatorname{tr} U = 1, \\ & && \operatorname{tr}[UV_{n+i}(\theta^{(k)})] = 0 \quad \text{for } i = 1, \dots, m, \end{aligned} \tag{11.3}$$

with the optimization variable being U, where tr designates the trace of a matrix. The following properties hold on this dual problem.

Theorem 2. *If the optimal value of the primal problem (11.2) is finite, the dual problem (11.3) has an optimal solution and the optimal value is equal to the infimum value $\lambda^{(k)}$. In this case, the set of all subgradients of the infimum value is equal to the set of vectors $\left[\operatorname{tr}[UV_1(\theta^{(k)})] \ \cdots \ \operatorname{tr}[UV_n(\theta^{(k)})]\right]^{\mathrm{T}}$, where U is an optimal solution of (11.3).*

Proof. The first statement follows from the duality theorem on semidefinite programming [331, Theorem 4.1.3]. Notice that the Slater condition is satisfied in our primal problem (11.2), that is, there exist y and λ that satisfy $V(x^{(k)}, y, \theta^{(k)}) - \lambda I \prec 0$. The second statement is derived from sensitivity analysis on semidefinite programming [331, Theorem 4.1.2]. □

Theorem 2 reduces computation of the infimum value $\lambda^{(k)}$ and the subgradient $d^{(k)}$ to solving the dual semidefinite programming problem (11.3). Since one can efficiently solve a semidefinite programming problem, computation of $\lambda^{(k)}$ and $d^{(k)}$ is now possible.

Remark 1. In some situations, one can check $\lambda^{(k)}$ for negativity without computing its value explicitly. In the case of Example 1, for example, $\lambda^{(k)}$ is negative if and only if the controller corresponding to $x^{(k)}$ stabilizes the plant at $\theta = \theta^{(k)}$. Stability of the system can be checked by location of its poles. Using such an alternative is preferable when it is faster and/or numerically more robust than semidefinite programming. □

Remark 2. In the algorithms of [76, 176, 271, 273], the norm $\|V(x,y,\theta)^+\|_{\mathrm{F}}$ is used in place of our $\overline{\lambda}[V(x,y,\theta)]$, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm and $V(x,y,\theta)^+$ is the projection of $V(x,y,\theta)$ onto the cone of positive semidefinite matrices. There are two reasons why we adopt the maximum eigenvalue. First, the infimum value and its subgradient can be efficiently computed as we have seen. Second, $V(x,y,\theta) \prec 0$ can be differentiated from $V(x,y,\theta) \preceq 0$ with $\overline{\lambda}[V(x,y,\theta)]$, while this is not possible with $\|V(x,y,\theta)^+\|_{\mathrm{F}}$. The latter point is important because a strict inequality is often required in control applications. The use of $\overline{\lambda}[V(x,y,\theta)]$ is due to [205]. □

11.5 Convergence to a Non-Strict Deterministic Solution

In the papers [76, 176, 271, 273], they considered a non-strict deterministic solution in our terminology and showed that their randomized algorithms obtain this solution after a finite number of iterations with probability one. We prove in this section a corresponding result for our algorithm and investigate the expected number of necessary iterations. For this purpose, we slightly modify Algorithm 11.1 in Section 11.3 by removing Steps 1 and 2. Similarly to [76,176,271,273], the following assumptions are made.

Assumption 11.1. *For $x^* \in \mathbb{R}^n$, the probability*

$$\mathbb{P}\{\theta \in \Theta : \inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^*,y,\theta)] > 0\}$$

is positive whenever $\inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(x^*,y,\theta)] > 0$ holds for some $\theta \in \Theta$.

Assumption 11.2. *The intersection between the initial ellipsoid $E^{(0)}$ and the solution set S has a positive volume.*

Theorem 3. *Under Assumptions 11.1 and 11.2, there exists with probability one a finite k such that $x^{(k)}$ produced by the modified algorithm is a non-strict deterministic solution, where the probability is measured with $\mathbb{P}^\infty$.*

Proof. The proof is basically the same as that of Lemma 2 in [176]. Since $E^{(k)} \supseteq E^{(0)} \cap S$ as in the proof of Theorem 1 (b), the volume of the ellipsoid $E^{(k)}$ has to be greater than or equal to that of $E^{(0)} \cap S$, which is positive due to Assumption 11.2. This means that the number of updates is finite by the same reasoning as in the proof of Theorem 1 (b). In the modified algorithm, it is allowed that no update is made for an infinite number of consecutive iterations. However, the probability that this occurs is zero by Assumption 11.1. □

For each sequence $\{\theta^{(k)}\}$, there may exist multiple k's having the property of the theorem. We write the minimum such k as k_{N}, which stands for the number of iterations necessary to find a non-strict deterministic solution. Note that k_{N} is a random number.

Though the above result is theoretically attractive, it is less practical than the search for a probabilistic solution proposed in the preceding sections. The reason is as follows. First, it is difficult to choose an initial ellipsoid $E^{(0)}$ so that Assumption 11.2 holds because we do not know the location of the solution set S in many cases. If Assumption 11.2 fails to hold, the finite-time convergence is not guaranteed. Second, it is difficult to detect when $x^{(k)}$ becomes a non-strict deterministic solution. Indeed, in order to detect it, we need to check $\inf_{y\in\mathbb{R}^m} V(x^{(k)}, y, \theta) \preceq 0$ for all $\theta \in \Theta$, which is impossible in a typical case that Θ consists of infinitely many parameter values. Finally, the number of necessary iterations k_{N} is distributed with a heavy tail. In fact, the expectation of this number is infinite as shown in the following theorem. The corresponding result has been reported in the special case that V does not depend on y [250, 252].

Theorem 4. *Suppose that an initial ellipsoid $E^{(0)}$ is provided. Under Assumptions 11.3–11.6 below, the number of necessary iterations k_{N} in the modified algorithm has infinite expectation, where the expectation is taken according to $\mathbb{P}^\infty$.*

The proof is similar to that in the special case. See Appendix 11.9.

Due to the dependence on y, the required assumptions are stronger than in the special case.

Assumption 11.3. *The parameter set Θ is a bounded closed set having a non-empty interior and a finite-measure boundary. There exists an open set that includes Θ and has $V_0(\theta), \ldots, V_{n+m}(\theta)$ continuously differentiable there. The probability measure $\mathbb{P}$ has a density function possessing a finite upper bound and a positive lower bound.*

Assumption 11.4. *There exists a finite sequence of parameter values $\{\hat{\theta}^{(k)}\}_{k=0}^{k^*}$ having the following properties, where $\hat{\theta}^{(0)}, \ldots, \hat{\theta}^{(k^*-1)}$ are in the interior of Θ and $\hat{\theta}^{(k^*)}$ is in Θ. If we choose $\theta^{(0)} = \hat{\theta}^{(0)}$, …, $\theta^{(k^*)} = \hat{\theta}^{(k^*)}$ in the modified algorithm, we obtain $x^{(0)} = \hat{x}^{(0)}$, …, $x^{(k^*)} = \hat{x}^{(k^*)}$, with which the following statements hold:*

(a) For each of $k = 0, \ldots, k^* - 1$, there holds $\inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(\hat{x}^{(k)}, y, \hat{\theta}^{(k)})] > 0$;
(b) $\inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(\hat{x}^{(k^*)}, y, \hat{\theta}^{(k^*)})] = \max_{\theta\in\Theta} \inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(\hat{x}^{(k^*)}, y, \theta)] = 0$;
(c) For each of $k = 0, \ldots, k^*$, there exists $\hat{y}^{(k)} \in \mathbb{R}^m$ that attains the infimum value $\inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(\hat{x}^{(k)}, y, \hat{\theta}^{(k)})]$;

(d) *For each of* $k = 0, \ldots, k^*$, *the maximum eigenvalue* $\overline{\lambda}[V(\hat{x}^{(k)}, \hat{y}^{(k)}, \hat{\theta}^{(k)})]$ *is simple;*
(e) *For each of* $k = 0, \ldots, k^*$, *the Hessian of* $\overline{\lambda}[V(x, y, \theta)]$ *with respect to* y *is positive definite at* $(\hat{x}^{(k)}, \hat{y}^{(k)}, \hat{\theta}^{(k)})$*;*
(f) *The gradient of* $\overline{\lambda}[V(x, y, \theta)]$ *with respect to* θ *is non-zero at* $(\hat{x}^{(k^*)}, \hat{y}^{(k^*)}, \hat{\theta}^{(k^*)})$.

The assumption (d) guarantees the existence of the Hessian and the gradient considered in the assumptions (e) and (f). By the assumptions (c), (d), and (e), the infimum value, $\inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x, y, \theta)]$, is continuously differentiable in neighborhoods of $(\hat{x}^{(k)}, \hat{\theta}^{(k)})$, $k = 0, \ldots, k^*$. In particular, the gradient of $\inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x, y, \theta)]$ with respect to θ is non-zero at $(\hat{x}^{(k^*)}, \hat{\theta}^{(k^*)})$ by the assumption (f). Due to this fact and the assumption (b), the point $\hat{\theta}^{(k^*)}$ has to lie on the boundary of Θ.

Since $x^{(k^*)}$ depends on the choice of $\theta^{(0)}, \ldots, \theta^{(k^*-1)}$, it is possible to regard $\inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^{(k^*)}, y, \theta^{(k^*)})]$ as a function of $\theta^{(0)}, \ldots, \theta^{(k^*)}$. This function is continuously differentiable in a neighborhood of $(\hat{\theta}^{(0)}, \ldots, \hat{\theta}^{(k^*)})$ by the above discussion and the update rule of the algorithm. It is also possible to regard $\max_{\theta^{(k^*)} \in \Theta} \inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^{(k^*)}, y, \theta^{(k^*)})]$ as a function of $\theta^{(0)}, \ldots, \theta^{(k^*-1)}$. The next assumption is made on this function.

Assumption 11.5. *The maximum* $\max_{\theta^{(k^*)} \in \Theta} \inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^{(k^*)}, y, \theta^{(k^*)})]$ *is attained at a unique* $\theta^{(k^*)}$ *for each* $(\theta^{(0)}, \ldots, \theta^{(k^*-1)})$ *in a neighborhood of* $(\hat{\theta}^{(0)}, \ldots, \hat{\theta}^{(k^*-1)})$.

It follows from this assumption that the maximum value, $\max_{\theta^{(k^*)} \in \Theta} \inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^{(k^*)}, y, \theta^{(k^*)})]$, is continuously differentiable in a neighborhood of $(\hat{\theta}^{(0)}, \ldots, \hat{\theta}^{(k^*-1)})$ and that the unique maximizing $\theta^{(k^*)}$ is continuous there. The final assumption is now made.

Assumption 11.6. *There exists at least one* $k = 0, \ldots, k^* - 1$ *such that the gradient of* $\max_{\theta^{(k^*)} \in \Theta} \inf_{y \in \mathbb{R}^m} \overline{\lambda}[V(x^{(k^*)}, y, \theta^{(k^*)})]$ *with respect to this* $\theta^{(k)}$ *is non-zero at* $(\hat{\theta}^{(0)}, \ldots, \hat{\theta}^{(k^*-1)})$.

11.6 Example

In order to illustrate our approach, we consider robust stabilization of a tower crane model, which is taken from [356] with some simplification. Application of Algorithm 11.1 successfully gave a probabilistic solution as we will see below. On the other hand, we were not able to find a solution due to the conservatism with the existing approach based on a polytopic cover and a common Lyapunov function. These results show practical usefulness of our approach.

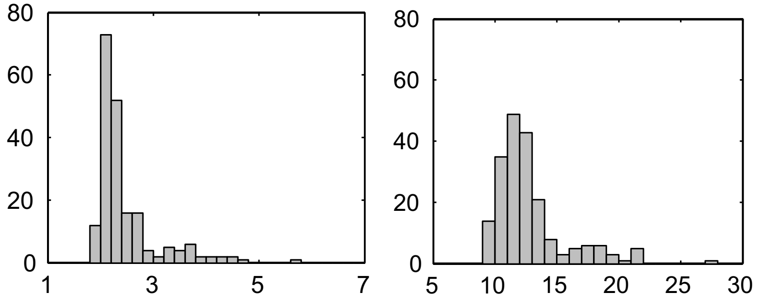


Figure 11.1. Histograms for 200 executions of the algorithm. *Left:* number of iterations ($\times 10^4$); *Right:* running time (s).

By discretizing the model with the sampling period 0.01 s, we obtain the plant $\xi[k+1] = A\xi[k] + Bu[k]$, where the dimension of $\xi[k]$ is four. The coefficients A and B nonlinearly depend on the three-dimensional parameter θ, which expresses the rope length, the boom angle at the equilibrium point, and the load weight. More details of the plant is found in [251]. It is supposed that each component of θ can take any value in some interval, which means that our parameter set Θ is a hyper rectangle. We consider to design a robustly stabilizing controller by using the formulation of Example 1 and applying Algorithm 11.1, which results in $n = 20$ and $m = 10$. We use the uniform distribution as $\mathbb{P}$ and choose the initial ellipsoid $E^{(0)}$ by letting $x^{(0)}$ be the zero vector and $Q^{(0)}$ be the identity. The parameters are set as $\mu = 10^{-9}\mathrm{vol}\, E^{(0)}$, $\epsilon = 0.001$, and $\delta = 0.0001$.

We executed the algorithm 200 times with Pentium 4 of 2.4 GHz and memory of 2.0 GByte. We used SDPA-M[1] of the version 2.00 to solve the semidefinite programming problem (11.3). In every execution, the algorithm stopped at Step 2 and gave an output. Figure 11.1 (a) shows the histogram of the number of iterations and (b) shows that of the running time. We can see from Figure 11.1 (b) that the running time is less than 20 seconds in many cases, which is practical enough. By Theorem 1 (c), on the other hand, the probability that the algorithm stops at Step 2 but the obtained output is not a probabilistic solution with $\epsilon = 0.001$ is less than or equal to $\delta = 0.0001$. This means that the obtained output is a probabilistic solution with high confidence.

11.7 Conclusion

A randomized algorithm that allows the use of a parameter-dependent Lyapunov function is proposed. It is considered to be useful for a less conservative

[1] http://grid.r.dendai.ac.jp/sdpa/index.html

robust-controller design against nonlinear parametric uncertainty. This algorithm provides a probabilistic solution after a bounded number of iterations, which is of polynomial order in the problem size. Convergence to a non-strict deterministic solution is considered and its practical difficulties are pointed out. Since our problem is described in a general form, the application of the proposed algorithm is not limited to robust-controller design.

11.8 Appendix: Proof of Theorem 1

(a) By the construction of the algorithm, at most $\kappa(\ell_0)$ iterations are made during the period that the counter ℓ has the value ℓ_0. Since the algorithm stops when ℓ reaches $\overline{\ell}$, the number of iterations k has to be less than or equal to $\kappa(0)+\cdots+\kappa(\overline{\ell}-1) \leq \overline{\ell}\kappa(\overline{\ell}-1)$.

(b) Suppose that the ellipsoid $E^{(k)}$ is updated. Then the ellipsoid $E^{(k+1)}$ is in fact the minimum-volume ellipsoid that contains the intersection between the original ellipsoid $E^{(k)}$ and the half space $\{x \in \mathbb{R}^n : (d^{(k)})^{\mathrm{T}}(x-x^{(k)}) < 0\}$ [50]. As we will see below, this half space includes the solution set S. Therefore $E^{(k+1)}$ includes the set $E^{(k)} \cap S$. It can also be shown that $\operatorname{vol} E^{(k+1)} < (\operatorname{vol} E^{(k)})\mathrm{e}^{-1/2(n+1)}$ [50]. Hence, if the ellipsoid $E^{(k)}$ has experienced $\overline{\ell}$ updates, it satisfies $E^{(k)} \supseteq E^{(0)} \cap S$ and $\operatorname{vol} E^{(k)} < (\operatorname{vol} E^{(0)})\mathrm{e}^{-\overline{\ell}/2(n+1)} \leq \mu$, which shows the claim.

We show that the half space $\{x \in \mathbb{R}^n : (d^{(k)})^{\mathrm{T}}(x-x^{(k)}) < 0\}$ includes the solution set S. To this end, note that $\inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(x,y,\theta^{(k)})] \geq \inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(x^{(k)},y,\theta^{(k)})] + (d^{(k)})^{\mathrm{T}}(x-x^{(k)}) \geq (d^{(k)})^{\mathrm{T}}(x-x^{(k)})$ because $d^{(k)}$ is the subgradient of $\inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(x^{(k)},y,\theta^{(k)})]$ and the value $\inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(x^{(k)}, y,\theta^{(k)})]$ is non-negative. By this inequality, if x belongs to S, it satisfies $\inf_{y\in\mathbb{R}^m} \overline{\lambda}[V(x,y,\theta^{(k)})] < 0$ and thus $(d^{(k)})^{\mathrm{T}}(x-x^{(k)}) < 0$.

(c) This is an extension of a result of [182] and [358]. Write the solution candidate after ℓ updates as $x^{[\ell]}$ noting that the candidate remains unchanged until the next update. We define two events for each of $\ell = 0, 1, \ldots$:

M_ℓ : Update is made at least ℓ times and the candidate $x^{[\ell]}$ is not updated for consecutive $\kappa(\ell)$ iterations;

B_ℓ : Update is made at least ℓ times and $x^{[\ell]}$ satisfies $\mathbb{P}\{\theta \in \Theta : \exists y \in \mathbb{R}^m \text{ s.t. } V(x^{[\ell]},y,\theta) \prec 0\} \leq 1-\epsilon$.

In the following, we will show

$$\mathbb{P}^\infty\big[(M_0 \cap B_0) \cup (M_1 \cap B_1) \cup \cdots\big] \leq \delta, \tag{11.4}$$

which establishes the claim.

We have, for any ℓ,

$$\mathbb{P}^\infty(M_\ell \cap B_\ell) = \mathbb{P}^\infty(M_\ell|B_\ell)\mathbb{P}^\infty(B_\ell) \leq \mathbb{P}^\infty(M_\ell|B_\ell) \leq (1-\epsilon)^{\kappa(\ell)},$$

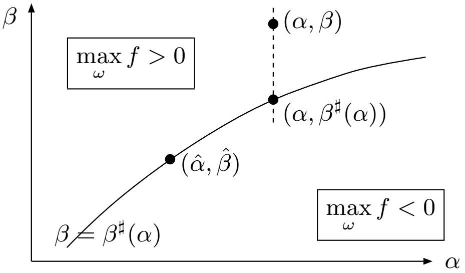


Figure 11.2. The function $\beta^\sharp(\alpha)$

where $\mathbb{P}^\infty(M_\ell|B_\ell)$ expresses the conditional probability of M_ℓ with B_ℓ being assumed. Therefore, we have

$$\begin{aligned}\mathbb{P}^\infty\big[(M_0\cap B_0)\cup(M_1\cap B_1)\cup\cdots\big] &\le (1-\epsilon)^{\kappa(0)}+(1-\epsilon)^{\kappa(1)}+\cdots\\ &\le \frac{6\delta}{\pi^2}\Big(1+\frac{1}{2^2}+\cdots\Big)=\delta,\end{aligned}$$

which shows (11.4).

11.9 Appendix: Proof of Theorem 4

We prepare some notation. As was noticed after Assumption 11.4, the infimum value, $\inf_{y\in\mathbb{R}^m}\overline{\lambda}[V(x^{(k^*)},y,\theta^{(k^*)})]$, is regarded as a function of $(\theta^{(0)},\ldots,\theta^{(k^*)})$, which is continuously differentiable in a neighborhood of $(\hat\theta^{(0)},\ldots,\hat\theta^{(k^*)})$. Let us write this function as $f(\alpha,\beta,\omega)$, where $[\alpha^{\mathrm{T}}\ \ \beta]^{\mathrm{T}}$ stands for $[(\theta^{(0)})^{\mathrm{T}}\ \ \ldots\ \ (\theta^{(k^*-1)})^{\mathrm{T}}]^{\mathrm{T}}$ with β being the last element of the vector and ω stands for $\theta^{(k^*)}$. We also use the notation $[\hat\alpha^{\mathrm{T}}\ \ \hat\beta]^{\mathrm{T}}=[(\hat\theta^{(0)})^{\mathrm{T}}\ \ \ldots\ \ (\hat\theta^{(k^*-1)})^{\mathrm{T}}]^{\mathrm{T}}$ and $\hat\omega=\hat\theta^{(k^*)}$. By rearranging the order of the elements if necessary, we can assume due to Assumption 11.6 that $(\partial/\partial\beta)\max_{\omega\in\Theta}f(\alpha,\beta,\omega)\neq 0$ at $(\hat\alpha,\hat\beta)$. Using the implicit function theorem, we can define a continuously differentiable function $\beta^\sharp(\alpha)$ so that $\beta^\sharp(\hat\alpha)$ is equal to $\hat\beta$ and $\max_{\omega\in\Theta}f(\alpha,\beta,\omega)=0$ is equivalent to $\beta=\beta^\sharp(\alpha)$ in a neighborhood of $(\hat\alpha,\hat\beta)$. The situation is illustrated in Figure 11.2.

For each pair of (α,β), a point $x^{(k^*)}$ is determined. A pair (α,β) satisfies $\max_{\omega\in\Theta}f(\alpha,\beta,\omega)>0$ if and only if the corresponding $x^{(k^*)}$ is not a non-strict deterministic solution, in which case this $x^{(k^*)}$ needs to be further updated. We will evaluate the probability of this update, which is equal to $\mathbb{P}^\infty\{\omega\in\Theta : f(\alpha,\beta,\omega)\ge 0\}$. Indeed, the two lemmas below show that this probability is bounded by a linear function of $|\beta-\beta^\sharp(\alpha)|$. Once such evaluation is obtained, the theorem is easily proved.

We need more preparation to present the first lemma. By Assumption 11.5, the maximum value, $\max_{\omega\in\Theta}f(\alpha,\beta^\sharp(\alpha),\omega)$, is attained at a unique ω, which is written as $\omega^\sharp(\alpha)$. This $\omega^\sharp(\alpha)$ is continuous in α as was noted after Assumption 11.5. Distinguishing the first element of ω from the others, we write $\omega = [\omega_1 \;\; \omega_2^{\mathrm{T}}]^{\mathrm{T}}$, $\hat{\omega} = [\hat{\omega}_1 \;\; \hat{\omega}_2^{\mathrm{T}}]^{\mathrm{T}}$, and so on. We can assume without loss of generality that $(\partial/\partial\omega_1)f(\alpha,\beta,\omega)$ is non-zero at $(\hat{\alpha},\hat{\beta},\hat{\omega})$ since the gradient is non-zero as was discussed after Assumption 11.4.

Lemma 1. *There exist a neighborhood of $(\hat{\alpha},\hat{\beta})$ and a positive number c such that, if (α,β) belongs to that neighborhood and $\omega \in \Theta$ satisfies $f(\alpha,\beta,\omega) = 0$, there exists a real number t such that $[\omega_1 + t \;\; \omega_2^{\mathrm{T}}]^{\mathrm{T}}$ is on the boundary of Θ and $|t| \le c|\beta - \beta^\sharp(\alpha)|$.*

Proof. Since $(\partial/\partial\omega_1)f(\alpha,\beta,\omega)$ is non-zero at $(\hat{\alpha},\hat{\beta},\hat{\omega})$, we can apply the implicit function theorem and define a continuously differentiable function $\omega_1(\alpha,\beta,\omega_2)$ so that $\omega_1(\hat{\alpha},\hat{\beta},\hat{\omega}_2)$ is equal to $\hat{\omega}_1$ and $f(\alpha,\beta,[\omega_1,\omega_2^{\mathrm{T}}]^{\mathrm{T}}) = 0$ is equivalent to $\omega_1 = \omega_1(\alpha,\beta,\omega_2)$ in a neighborhood of $(\hat{\alpha},\hat{\beta},\hat{\omega})$. Moreover, it is possible to assume the existence of a positive number c such that

$$\left|\frac{\partial\omega_1(\alpha,\beta,\omega_2)}{\partial\beta}\right| = \left|\frac{(\partial/\partial\beta)f(\alpha,\beta,\omega)}{(\partial/\partial\omega_1)f(\alpha,\beta,\omega)}\right| \le c. \tag{11.5}$$

At $(\hat{\alpha},\hat{\beta})$, the relationship $f(\alpha,\beta,\omega) = 0$ holds only at $\hat{\omega}$; Hence, it is possible to choose a neighborhood of $(\hat{\alpha},\hat{\beta})$, in which $f(\alpha,\beta,\omega) = 0$ holds only at ω obtained as $[\omega_1(\alpha,\beta,\omega_2) \;\; \omega_2^{\mathrm{T}}]^{\mathrm{T}}$. Now, suppose that $f(\alpha,\beta,\omega) = 0$ holds with (α,β) in this neighborhood. By integrating (11.5) from β to $\beta^\sharp(\alpha)$, we have the lemma because no $\omega \in \Theta$ satisfies $f(\alpha,\beta^\sharp(\alpha),\omega) = 0$ except for $\omega^\sharp(\alpha)$. □

Lemma 2. *There exist a neighborhood of $(\hat{\alpha},\hat{\beta})$ and a positive number c' such that all (α,β) in that neighborhood satisfy $\mathbb{P}\{\omega \in \Theta : f(\alpha,\beta,\omega) \ge 0\} \le c'|\beta - \beta^\sharp(\alpha)|$.*

Proof. In Lemma 1, a neighborhood of $(\hat{\alpha},\hat{\beta})$ can be chosen to be convex without loss of generality. Suppose that $f(\alpha,\beta,\omega) \ge 0$ holds for some (α,β) in that neighborhood and $\omega \in \Theta$. Since $f(\alpha,\beta^\sharp(\alpha),\omega) \le 0$, continuity of $f(\alpha,\beta,\omega)$ implies that there exists $\tilde{\beta}$ satisfying $f(\alpha,\tilde{\beta},\omega) = 0$ on the line segment connecting β and $\beta^\sharp(\alpha)$. Lemma 1 guarantees the existence of t such that $[\omega_1 + t \;\; \omega_2^{\mathrm{T}}]^{\mathrm{T}}$ is on the boundary of Θ and $|t| \le c|\tilde{\beta} - \beta^\sharp(\alpha)| \le c|\beta - \beta^\sharp(\alpha)|$. Due to Assumption 11.3, the probability of $\omega \in \Theta$ having this property is bounded by $c'|\beta - \beta^\sharp(\alpha)|$ for some $c' > 0$. Note that we can choose the same c' for all α in a neighborhood of $\hat{\alpha}$. □

We now prove the theorem. We can assume without loss of generality that $(\partial/\partial\beta)\max_{\omega\in\Theta} f(\hat{\alpha},\hat{\beta},\omega)$ is not only non-zero but also positive. We choose positive numbers a and b small enough that the set $A = \{(\alpha,\beta) : \|\alpha - \hat{\alpha}\| \le a$ and $\beta^\sharp(\alpha) < \beta \le \beta^\sharp(\alpha) + b\}$ is contained in the neighborhood of Lemma 2. For a pair $(\alpha,\beta) \in A$, the corresponding $x^{(k^*)}$ is not a non-strict deterministic solution and thus needs to be updated. Because the probability of update is bounded by $c'|\beta - \beta^\sharp(\alpha)|$ due to Lemma 2, the expected number of iterations required to make one update is larger than or equal to $1/c'|\beta - \beta^\sharp(\alpha)|$. If we integrate this number in A with $\mathbb{P}^\infty$, it gives a lower bound of the expected number of iterations necessary to find a non-strict deterministic solution. The result is infinity, which completes the proof.

Acknowledgement. The author is thankful to the following people for their comments: Stephen Boyd, Marco Campi, Fabrizio Dabbene, Shinji Hara, Satoru Iwata, Masakazu Kojima, Kazuo Murota, Maria Prandini, Roberto Tempo, and Takashi Tsuchiya. This work is supported by the 21st Century COE Program on Information Science and Technology Strategic Core and a Grant-in-Aid of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

12

Probabilistic Robust Controller Design: Probable Near Minimax Value and Randomized Algorithms

Yasumasa Fujisaki[1] and Yasuaki Kozawa[2]

[1] Department of Computer and Systems Engineering
Kobe University, Nada, Kobe 657-8501, Japan
fujisaki@cs.kobe-u.ac.jp
[2] Graduate School of Science and Technology
Kobe University, Nada, Kobe 657-8501, Japan

Summary. A probabilistic approach to robust controller design is presented. The design can be recast as a minimax problem with a cost function in general. In order to solve the problem efficiently, the definition of probable near minimax value is introduced. A probable near minimax value of the function can be calculated with a certain accuracy and a certain confidence by using a randomized algorithm, where independent identically distributed samples of optimized parameters are generated according to probability measures. It is shown that the necessary number of samples depends on the accuracy and the confidence, and is independent of the number of parameters. Furthermore, a special case where the cost function has a global saddle point is investigated. The definition of probable near saddle value, which is weaker than that of probable near minimax value, is introduced. Then, it is shown that the necessary number of samples is smaller in this case.

12.1 Introduction

Analysis and synthesis of robust control systems can be formulated as mathematical problems defined by structured singular value μ in general [253]. However, exact computation of μ is NP-hard with respect to the number of uncertain parameters in the model of a plant [62]. That is, analysis and synthesis of robust control systems involves an essential difficulty from the view point of computational complexity.

To cope with this difficulty in a practical way, probabilistic approach has been investigated, *e.g.*, in [24, 74, 76, 129, 176, 182, 252, 273, 347, 358, 371, 381, 382]. In contrast to deterministic approach, where *all* members in the model set of the plant are considered, probabilistic approach which chooses *almost all* members randomly provides a practical method with low computational complexity and low risk.

A probabilistic method for robustness analysis of control systems is investigated in [182, 347, 358]. This method is used for $\mathcal{H}_\infty$ controller synthesis in [182], which presents a randomized algorithm to find a low-order $\mathcal{H}_\infty$ controller that is as efficient as the full-order $\mathcal{H}_\infty$ controller given by analytical method. Furthermore, a probabilistic robust controller design with respect to plant uncertainties is investigated in [381], which proposes a randomized algorithm to find a controller that minimizes an *average* performance of the closed-loop system with respect to plant uncertainties.

In this chapter, we present a probabilistic approach to robust controller design in a general setting. We propose a randomized algorithm to find a controller that minimizes the *worst case* performance under plant uncertainties.

More specifically, robust controller design can be formulated as a minimax problem with a cost function. In order to develop a probabilistic method to solve the minimax problem, we introduce the definition of *probable near minimax value*. Then, we propose a randomized algorithm which calculates a probable near minimax value of the function with a certain accuracy and a certain confidence, where independent identically distributed samples of optimized parameters are generated according to probability measures. The proposed algorithm is based on so-called Monte Carlo simulation. The main result of this chapter is to show that the necessary number of samples depends on the accuracy and the confidence, and is independent of the number of parameters.

Furthermore, we consider a special case of the minimax problem, where the cost function has a global saddle point and its value at the point gives the exact minimax value. In this case, if we introduce a weaker notion, that is, *probable near saddle value*, then we can reduce the necessary number of samples corresponding to the plant uncertain parameters. This fact means that the necessary number of samples depends on problem structure, and we may invent more efficient randomized algorithms utilizing a particular structure of control problems.

The chapter is organized as follows. In Section 12.2, we summarize basic materials in a probabilistic approach [381], that is, the definitions of probable near minimum and maximum, and randomized algorithms to find them. In Section 12.3, we revisit robust controller design and recall that the design can be recast as a minimax problem with a cost function. Then, in Section 12.4, we present the main result of this chapter. That is, we define probable near minimax value and propose a randomized algorithm to find it with low computational complexity. Subsequently, in Section 12.5, we derive a similar result for the case that the function has a global saddle point. Here we compare the necessary numbers of samples given by these results. In Section 12.6, we demonstrate effectiveness of the proposed randomized algorithm through numerical examples. Finally, in Section 12.7, we give concluding remarks and make some comments on a possible extension of these results.

12.2 Preliminaries

Let us consider a measurable function $g : Y \to \mathbb{R}$, where Y is a measurable subset of some finite-dimensional Euclidean space. It is well-known that finding the exact minimum value of the function $g(\cdot)$

$$g^* = \inf_{y \in Y} g(y)$$

is in general NP-hard with respect to the dimension of the vector $y \in Y$. To cope with this difficulty, the following notion of an approximation of the exact minimum is introduced in [381].

Definition 1. *Suppose that $g : Y \to \mathbb{R}$, that P_Y is a given probability measure on Y, and that $\alpha \in (0,1)$ is a given number. A number $g_0 \in \mathbb{R}$ is said to be a probable near minimum of $g(\cdot)$ to level α if*

$$\begin{aligned} &g^* \leq g_0 \\ &P_Y\{\tilde{y} \in Y : g(\tilde{y}) < g_0\} \leq \alpha. \end{aligned} \tag{12.1}$$

If $g_0 = g^*$ then (12.1) holds for any α. This means that the exact minimum is also a probable near minimum to any level. On the other hand, if $g_0 > g^*$ then small α implies that the vector $\tilde{y}$ such that $g(\tilde{y}) < g_0$ hardly occurs when $\tilde{y}$ is chosen randomly. In this sense, g_0 is near g^*, although $g_0 - g^*$ may not be small actually.

There exists an efficient randomized algorithm to find a probable near minimum [381].

Lemma 1. *Suppose that a probability measure P_Y on Y, a measurable function $g : Y \to \mathbb{R}$, a level parameter $\alpha \in (0,1)$, and a confidence parameter $\delta_\alpha \in (0,1)$ are given. Choose an integer N such that*

$$N \geq \frac{\ln(1/\delta_\alpha)}{\ln[1/(1-\alpha)]}. \tag{12.2}$$

Generate independent identically distributed (i.i.d.) samples $y_1, y_2, \ldots, y_N \in Y$ distributed according to P_Y. Define

$$\begin{aligned} y &= [y_1 \ y_2 \ \ldots \ y_N]^T \ \in Y^N \\ \hat{g}(y) &= \min_{1 \leq i \leq N} g(y_i). \end{aligned}$$

Then, it can be said with confidence at least $1 - \delta_\alpha$ that $\hat{g}(y)$ is a probable near minimum of $g(\cdot)$ to level α.

The number N of i.i.d. samples required for estimation of a probable near minimum depends only on parameters α and δ_α, and does not depend on the dimension of the vector y. Thus, if i.i.d. samples can be generated with a polynomial time, then finding a probable near minimum is not NP-hard with respect to the dimension. Note that such an efficient algorithm to generate random samples is derived for robustness analysis of control systems in [74].

Now, let us consider a measurable function $h : X \to \mathbb{R}$, where X is a measurable subset of some finite-dimensional Euclidean space. In the same way, for finding the exact maximum of $h(\cdot)$

$$h^* = \sup_{x \in X} h(x)$$

we can state the corresponding notion and result as follows.

Definition 2. *Suppose that $h : X \to \mathbb{R}$, that P_X is a given probability measure on X, and that $\beta \in (0,1)$ is a given number. A number $h_0 \in \mathbb{R}$ is said to be a probable near maximum of $h(\cdot)$ to level β if*

$$\begin{aligned} & h^* \geq h_0 \\ & P_X\{\tilde{x} \in X : h(\tilde{x}) > h_0\} \leq \beta. \end{aligned}$$

Lemma 2. *Suppose that a probability measure P_X on X, a measurable function $h : X \to \mathbb{R}$, an level parameter $\beta \in (0,1)$, and a confidence parameter $\delta_\beta \in (0,1)$ are given. Choose an integer M such that*

$$M \geq \frac{\ln(1/\delta_\beta)}{\ln[1/(1-\beta)]}.$$

Generate i.i.d. samples $x_1, x_2, \dots, x_M \in X$ distributed according to P_X. Define

$$\begin{aligned} x &= [x_1 \ x_2 \ \dots \ x_M]^T \ \in X^M \\ \hat{h}(x) &= \max_{1 \leq i \leq M} h(x_i). \end{aligned}$$

Then, it can be said with confidence at least $1-\delta_\beta$ that $\hat{h}(x)$ is a probable near maximum of $h(\cdot)$ to level β.

12.3 Problem Statement

Let us consider a family of plants $\{G(x), x \in X\}$, where x is the vector of all the uncertain elements of the plant description and X is the set of all possible plant parameter vectors x. We assume that x is a finite-dimensional vector and X is a measurable subset of some finite-dimensional Euclidean space. This plant description $\{G(x), x \in X\}$ can represent any plant family with structured uncertainty.

We also consider a family of controllers $\{K(y), y \in Y\}$, where y is the vector of all adjustable design parameters of the controller description and Y is the set of all possible controller parameter y. Similarly, we assume that

y is a finite-dimensional vector and Y is a measurable subset of some finite-dimensional Euclidean space. It should be noted that, although the function $K(y)$ is assumed to be given in advance, this formulation can treat not only a fixed structure of the controllers but also a fixed design procedure of the controllers. This detail can be found in Example 1 of Section 12.6.

In general, the objective of robust controller design is to find a single fixed controller $K(y_0), y_0 \in Y$ that achieves a given performance for all plants $G(x), x \in X$. Thus, we introduce a cost function $f : X \times Y \to \mathbb{R}$, where we assume that f is a measurable function and lower value of f means better performance of the control system. Then, the robust controller design becomes to find a controller parameter that minimizes the worst case value of the cost function with respect to all possible plant parameters. That is, the design can be recast as a minimax problem: find $y_0 \in Y$ that achieves

$$f^* = \inf_{y \in Y} \sup_{x \in X} f(x, y).$$

In the rest of this chapter, we refer to f^* as the exact minimax value.

Note that even if we fix a controller and assume that X is a hyperbox, in principle, we must check at least 2^m extreme values of X in order to find the worst case value of the cost function, where m is the dimension of the vector x. In other words, this minimax problem is at least NP-hard with respect to the dimension, thus heavy computational effort could be required when the problem has large dimensions [62].

12.4 Probable Near Minimax Value and Randomized Algorithms

In this section, we consider the minimax problem from the view point of probabilistic approximation. We first introduce a notion of an approximation of the exact minimax value f^* as follows.

Definition 3. *Suppose that $f : X \times Y \to \mathbb{R}$, that P_X, P_Y are given probability measures on X, Y respectively, and that $\alpha, \beta \in (0, 1)$ are given numbers. A number $f_0 \in \mathbb{R}$ is said to be a probable near minimax value of $f(\cdot, \cdot)$ to minimum level α and maximum level β if there exist a number $f_U \in \mathbb{R}$ and a measurable function $f_L : Y \to \mathbb{R}$ such that*

$$\inf_{y \in Y} f_L(y) \le f_0 \le f_U$$

$$P_Y\{\tilde{y} \in Y : \sup_{x \in X} f(x, \tilde{y}) < f_U\} \le \alpha \tag{12.3}$$

$$P_X\{\tilde{x} \in X : f(\tilde{x}, y) > f_L(y)\} \le \beta, \forall y \in Y. \tag{12.4}$$

Here, let us confirm that the exact minimax value is a probable near minimax value to any levels, that is,

$$\inf_{y \in Y} f_L(y) \le f^* \le f_U \tag{12.5}$$

together with (12.3) and (12.4) holds for any α and any β. Furthermore, we will see that small α implies that f_U is near f^* in the sense of probable near minimum, and small β implies that $\inf_{y \in Y} f_L(y)$ is near f^* in the sense of probable near maximum. As a result, this ensures that f_0 is near f^* in these senses.

First, we consider f_U. Note that there always exists f_U satisfying

$$f^* \le f_U \tag{12.6}$$

and (12.3) for any α because the selection $f_U = f^*$ ensures this fact. Then, comparing (12.6) and (12.3) with Definition 1, we see that every value f_U is a probable near minimum of the function $\sup_{x \in X} f(x,y)$. Hence, small α implies that f_U is near f^* in the sense of Definition 1. That is, if α is small, probability such that $f_0 > f^*$ becomes small.

Next, we consider $f_L(y)$. Note that there always exists $f_L(y)$ satisfying

$$\sup_{x \in X} f(x,y) \ge f_L(y), \ \forall y \in Y \tag{12.7}$$

and (12.4) for any β because the selection $f_L(y) = \sup_{x \in X} f(x,y)$ ensures this fact. Choosing such $f_L(y)$, we see that

$$\inf_{y \in Y} f_L(y) \le f^*$$

holds. Thus, when $f_0 = f^*$, we can say that there exists $f_L(y)$ satisfying the conditions of Definition 3 for any β. Now, for a fixed y, comparing (12.7) and (12.4) with Definition 2, we see that $f_L(y)$ is a probable near maximum of the function $f(x,y)$. Hence, small β implies that $f_L(y)$ is near $\sup_{x \in X} f(x,y)$ in the sense of Definition 2. That is, if β is small, probability of $f_L(y)$ such that $\sup_{x \in X} f(x,y) < f_L(y)$ becomes small for each y, and therefore probability of f_0 such that $f_0 < f^*$ becomes small.

As a result, we can say that the exact minimax value f^* is also a probable near minimax value f_0 to any levels. Furthermore, if we consider each probability such that f_0 is greater or less than f^*, in the sense of Definitions 1 and 2, we can reduce each probability with small α or small β respectively.

The main result of this contribution is the following theorem.

Theorem 1. *Suppose that probability measures P_X, P_Y on X, Y, a measurable function $f : X \times Y \to \mathbb{R}$, level parameters $\alpha, \beta \in (0,1)$, and confidence parameters $\delta_\alpha, \delta_\beta \in (0,1)$ are given. Choose integers M, N such that*

$$M \ge \frac{\ln(N/\delta_\beta)}{\ln[1/(1-\beta)]}, \qquad N \ge \frac{\ln(1/\delta_\alpha)}{\ln[1/(1-\alpha)]}. \tag{12.8}$$

Generate i.i.d. samples $x_1, x_2, \ldots, x_M \in X$ and $y_1, y_2, \ldots, y_N \in Y$ distributed according to P_X and P_Y respectively. Define

$$\begin{aligned} x &= [x_1 \ x_2 \ \dots \ x_M] \\ y &= [y_1 \ y_2 \ \cdots \ y_N] \\ \hat{f}(x,y) &= \min_{1 \le i \le N} \max_{1 \le j \le M} f(x_j, y_i). \end{aligned}$$

Then, it can be said with confidence at least $1 - (\delta_\alpha + \delta_\beta)$ that $\hat{f}(x,y)$ is a probable near minimax value of $f(\cdot,\cdot)$ to minimum level α and maximum level β.

Proof. We first consider $f_L(y)$. Here we fix the subscription i as any number within $1 \le i \le N$. As we see, for any β, there exists $f_L(y)$ satisfying (12.7), (12.4), and

$$\inf_{y \in Y} f_L(y) \le f_L(y_i).$$

Now, let us define

$$\tilde{\delta}_\beta = \frac{\delta_\beta}{N}.$$

Then, M of (12.8) is rewritten as

$$M \ge \frac{\ln(1/\tilde{\delta}_\beta)}{\ln[1/(1-\beta)]}.$$

Thus, from Lemma 2, we see that $\max_{1 \le j \le M} f(x_j, y_i)$ is a probable near maximum of the function $f(x, y_i)$ to level β with confidence at least $1 - \tilde{\delta}_\beta$. We therefore see that there exists $f_L(y)$ such that

$$f_L(y_i) \le \max_{1 \le j \le M} f(x_j, y_i).$$

It turns out that there exists $f_L(y)$ such that

$$\inf_{y \in Y} f_L(y) \le \max_{1 \le j \le M} f(x_j, y_i)$$

together with (12.4) with confidence at least $1 - \tilde{\delta}_\beta$, for each i. Computing the confidence such that this holds for all N samples, we conclude that

$$\inf_{y \in Y} f_L(y) \le \hat{f}(x,y) \tag{12.9}$$

and (12.4) hold with confidence at least $1 - N\tilde{\delta}_\beta$, that is, $1 - \delta_\beta$.

We next consider f_U. If we define

$$f_U = \min_{1 \le i \le N} \sup_{x \in X} f(x, y_i),$$

then

$$\hat{f}(x,y) \le f_U \tag{12.10}$$

holds for any M. Noting here that N satisfies (12.8), from Lemma 1, we see that f_U is a probable near minimum of the function $\sup_{x\in X} f(x,y)$ to level α with confidence at least $1-\delta_\alpha$. That is, (12.10) and (12.3) hold with confidence at least $1-\delta_\alpha$.

Finally, notice that (12.9) or (12.10) always holds because (12.5) holds. We therefore see that all of the conditions

$$\inf_{y\in Y} f_L(y) \leq \hat{f}(x,y) \leq f_U,$$

(12.3), and (12.4) hold with confidence at least

$$(1-\delta_\alpha)+(1-\delta_\beta)-1 = 1-(\delta_\alpha+\delta_\beta).$$

This completes the proof of the theorem. □

As we see, smaller level parameters α and β imply that $\hat{f}(x,y)$ becomes better approximation of f^*, and smaller confidence parameters δ_α and δ_β imply that $\hat{f}(x,y)$ is a probable near minimax value with higher probability. These parameters α, β, δ_α, and δ_β can be designed by the user, and the necessary numbers N and M of samples for computing $\hat{f}(x,y)$ are determined from (12.8). Although N and M increase if we make the parameters small, they are still independent of the dimensions of x and y.

12.5 Probable Near Saddle Value and Randomized Algorithms

In this section, we investigate a special case of the problem we have considered. Here we assume that

$$\sup_{x\in X}\inf_{y\in Y} f(x,y) = \inf_{y\in Y}\sup_{x\in X} f(x,y).$$

That is, the function $f(\cdot,\cdot)$ has a global saddle point in the space $X\times Y$. However, we do not make any further assumption, for example, convexity of the function. Thus, finding the exact saddle point is still NP-hard and intractable in general.

Here we introduce a definition of a probably approximate value of the function $f(\cdot,\cdot)$ at the saddle point.

Definition 4. *Suppose that $f : X\times Y \to \mathbb{R}$, that P_X, P_Y are given probability measures on X, Y respectively, and that $\alpha,\beta\in(0,1)$ are given numbers. A number $f_0\in\mathbb{R}$ is said to be a probable near saddle value of $f(\cdot,\cdot)$ to minimum level α and maximum level β if there exist $f_U\in\mathbb{R}$ and $f_L\in\mathbb{R}$ such that*

$$f_L \leq f_0 \leq f_U$$

$$P_Y\{\tilde{y}\in Y : \sup_{x\in X} f(x,\tilde{y}) < f_U\} \leq \alpha \tag{12.11}$$

$$P_X\{\tilde{x}\in X : \inf_{y\in Y} f(\tilde{x},y) > f_L\} \leq \beta. \tag{12.12}$$

This notion is weaker than that of probable near minimax value. That is, a probable near minimax value is always a probable near saddle value, and the converse does not hold in general. Although this is immediately true from the definitions, this can be also confirmed by the fact that, for any function $f(\cdot,\cdot)$ which may not have a saddle point,

$$\sup_{x\in X}\inf_{y\in Y} f(x,y) \leq \inf_{y\in Y}\sup_{x\in X} f(x,y) \tag{12.13}$$

holds and the equality is *not* attained in general. Then, referring the discussion of the previous section, we can regard f_U as an approximation of an upper bound of $\inf_{y\in Y}\sup_{x\in X} f(x,y)$, while from the similarity we can also regard f_L as an approximation of a lower bound of $\sup_{x\in X}\inf_{y\in Y} f(x,y)$. The existence of the gap in (12.13) suggests that the notion above is weaker than the previous one.

Now, the necessary number of samples for computing a probable near saddle value is smaller than that for computing a probable near minimax value. That is, the following theorem holds.

Theorem 2. *Suppose that probability measures P_X, P_Y on X, Y, a measurable function $f : X \times Y \to \mathbb{R}$, level parameters $\alpha, \beta \in (0,1)$, and confidence parameters $\delta_\alpha, \delta_\beta \in (0,1)$ are given. Choose integers M, N such that*

$$M \geq \frac{\ln(1/\delta_\beta)}{\ln[1/(1-\beta)]}, \qquad N \geq \frac{\ln(1/\delta_\alpha)}{\ln[1/(1-\alpha)]}. \tag{12.14}$$

Generate i.i.d. samples $x_1, x_2, \ldots, x_M \in X$ and $y_1, y_2, \ldots, y_N \in Y$ distributed according to P_X and P_Y respectively. Define

$$\begin{aligned} x &= [x_1\ x_2\ \ldots\ x_M] \\ y &= [y_1\ y_2\ \ldots\ y_N] \\ \hat{f}(x,y) &= \min_{1\leq i\leq N}\ \max_{1\leq j\leq M} f(x_j, y_i). \end{aligned}$$

Then, it can be said with confidence at least $1-(\delta_\alpha+\delta_\beta)$ that $\hat{f}(x,y)$ is a probable near saddle value of $f(\cdot,\cdot)$ to minimum level α and maximum level β.

The proof of this theorem is similar to the theorem in the previous section. The difference appears in the lower bound f_L, but its proof is parallel to the upper bound f_U. Thus, the proof is omitted here.

We immediately see that the necessary samples of (12.14) is smaller than that of (12.8). For example, if we choose $\alpha = \beta = 0.05$ and $\delta_\alpha = \delta_\beta = 0.025$, then the smallest M or N satisfying (12.14) is 72. On the other hand, the smallest M satisfying (12.8) becomes 156. If we choose $\alpha = \beta = 0.01$ and $\delta_\alpha = \delta_\beta = 0.005$, then the smallest M or N satisfying (12.14) is 528. On the other hand, the smallest M satisfying (12.8) becomes 1151.

12.6 Numerical Examples

In this section, we demonstrate the proposed randomized algorithms through numerical examples.

Example 1. We assume that $X = Y = [-5, 5]$ and consider the cost function

$$f(x, y) = y^2 - x^2 + 5x - 4y + xy + 5. \tag{12.15}$$

We set both probability measures P_X, P_Y on X, Y as uniform distribution. Furthermore, we choose $\alpha = \beta = 0.05$, $\delta_\alpha = \delta_\beta = 0.025$, and $N = M = 72$ which satisfies (12.14).

The function (12.15) has a global saddle point, which can be computed analytically, and we see that $f^* = 10.8$ at $x = 2.8$, $y = 0.6$. On the other hand, f_U and f_L can be selected arbitrarily under the condition that they satisfy (12.11) and (12.12). Here we define $\bar{f}_U$ as the largest f_U satisfying (12.11) and $\bar{f}_L$ as the largest f_L satisfying (12.12). Computing $\bar{f}_L$ and $\bar{f}_U$ using grids of x and y with interval 0.005, we obtained $\bar{f}_L \approx 10.7363$ and $\bar{f}_U \approx 10.8781$.

Based on Theorem 2, we generated i.i.d. samples $x_1, x_2, \ldots, x_M \in X$ and $y_1, y_2, \ldots, y_N \in Y$ according to P_X and P_Y. Then, we obtained $\hat{f}(x, y) = 10.8012$, which is a probable near saddle value. Furthermore, we executed 100 trials of calculating $\hat{f}(x, y)$. The result is shown in Figure12.1, where black bar denotes the case that the corresponding $\hat{f}(x, y)$ does not satisfy $\bar{f}_L \leq \hat{f}(x, y) \leq \bar{f}_U$. As you see, the condition $\bar{f}_L \leq \hat{f}(x, y) \leq \bar{f}_U$ was satisfied 96 times of 100. Thus, we confirm that the estimated confidence is greater than the theoretical value $\{1 - (\delta_\alpha + \delta_\beta)\} \times 100 = 95$ given by Theorem 2.

Example 2. We next consider the example in [273], which is the state feedback design for the lateral motion of an aircraft. The state space equation of the model is given by

$$\dot{z}(t) = A(\xi)z(t) + Bv(t),$$

$$A(\xi) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & \xi_1 & \xi_2 & \xi_3 \\ \xi_4 & 0 & \xi_5 & -1 \\ \xi_4\xi_6 & \xi_7 & \xi_8 + \xi_5\xi_6 & \xi_9 - \xi_6 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & -3.91 \\ 0.035 & 0 \\ -2.53 & 0.31 \end{bmatrix}.$$

where the state variables z_1, z_2, z_3, z_4 are the bank angle, its derivative, the side-slip angle, and the yaw rate, while the inputs v_1 and v_2 represent the rudder and aileron deflections respectively. The vector of uncertain parameter ξ is allowed to vary in a set of 15% of its nominal value $\bar{\xi}$, *i.e.*,

$$\xi \in \varXi = \left\{\xi \in \mathbb{R}^9 \; : \; \xi_i \in [0.85\,\bar{\xi}_i, 1.15\,\bar{\delta}_i], i = 1, 2, \ldots, 9\right\}$$

where

$$\bar{\xi} = \begin{bmatrix} -2.93 \; -4.75 \; 0.78 \; 0.086 \; -0.11 \; 0.1 \; -0.042 \; 2.601 \; -0.29 \end{bmatrix}^T.$$

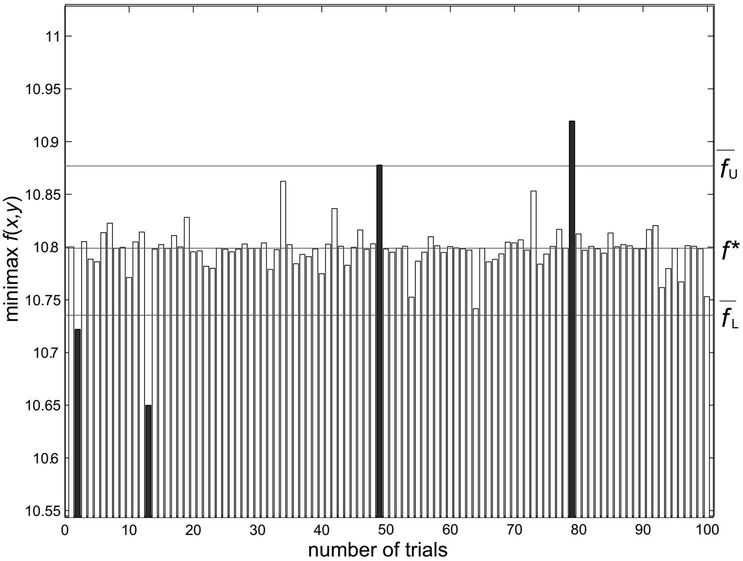


Figure 12.1. Probable near saddle values (100 trials)

We assume that the p.d.f. of ξ is uniform on Ξ, and define x and X by ξ and Ξ.

For this plant, we introduce a quadratic performance index

$$J = \int_0^\infty \{z^T(t)Qz(t) + v^T(t)Rv(t)\}\,\mathrm{dt}$$

where we choose $Q = 0.01I$ and $R = \nu I$. Here the parameter $\nu \in [0.1,\ 100]$ is used for the design of state feedback gain. That is, we assume that the p.d.f. of ν is uniform on this interval, and define y and Y by ν and $[0.1,\ 100]$. Then, the state feedback gain $K(\nu)$ for a fixed ν is selected as the optimal regulator with respect to the nominal system, *i.e.*,

$$v(t) = K(\nu)z(t), \qquad K(\nu) = -R^{-1}B^TP$$

where P is the symmetric and positive definite solution of

$$PA(\bar{\xi}) + A^T(\bar{\xi})P - PBR^{-1}B^TP + Q = 0.$$

We employ the following cost function:

$$f(x,y) = \begin{cases} 1 & \text{if the closed-loop system is unstable} \\ \dfrac{\mathrm{tr}\,\bar{P}}{1+\mathrm{tr}\,\bar{P}} & \text{if the closed-loop system is stable} \end{cases}$$

where $\bar{P}$ is the symmetric and positive definite solution of

$$\bar{P}\{A(\xi)+BK(\nu)\}+\{A(\xi)+BK(\nu)\}^T\bar{P}+Q+100K^T(\nu)K(\nu)=0.$$

Since $z^T(0)\bar{P}z(0)$ gives the value of J with $R = 100I$ for the gain $K(\nu)$, the above definition of the cost function enables us to expect a smaller input induced by this J.

We now describe the results of the simulations. We chose $\alpha = \beta = 0.05$ and $\delta_\alpha = \delta_\beta = 0.025$. Following (12.8), we set $M = 156$ and $N = 72$. We randomly generated x (*i.e.*, ξ) and y (*i.e.*, ν), and computed a probable near minimax value. Then, we obtained $\hat{f}(x, y) = 0.6523$. This point is marked as 'o' in Figure12.2, while the solid line indicates $\max_{1\le j\le M} f(x_j, y)$, which was minimized on y_i, $1 \le i \le N$. The corresponding state feedback gain was

$$K(33.42)=\begin{bmatrix}0.0164\ 0.0055\ -0.0003\ 0.0120\\0.0120\ 0.0041\ -0.0002\ 0.0070\end{bmatrix}.$$

On the other hand, when the extreme points of X are selected and a grid on Y is introduced, the minimax value can be computed as 0.7074 approximately. This point is marked as 'x' in Figure12.2, while the dashed line indicates the function which has been maximized on the extreme points of X. This figure shows that the obtained feedback gain can stabilize the systems at all the extreme points robustly. It also shows that the lower bound of the probable near minimax value is not so tight as the upper bound, which is consistent with the definition of probable near minimax value.

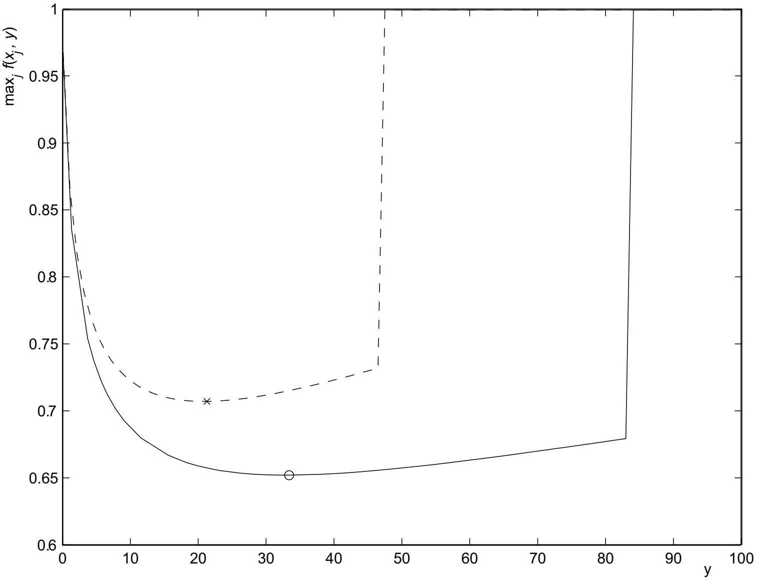


Figure 12.2. Cost function of robust LQR design

12.7 Concluding Remarks

In this chapter, we have proposed a probabilistic approach to robust controller design, which can be recast as a minimax problem with a cost function.

We have introduced probability measures on the optimized parameters and have defined *probable near minimax value*, which is an approximation of the true minimax value. Clarifying the relation to probable near minimum or maximum, we have investigated the meaning of the approximation, that is, in what sense the defined probable near minimax value is *near* the true minimax value. Then, we have proposed an efficient randomized algorithm which finds a probable near minimax value to given levels with given confidences.

Furthermore, we have investigated a special case such that a global saddle point exists. For this case, we have introduced a weaker notion of the minimax value, that is, *probable near saddle value.* It is shown that the number of samples required by the randomized algorithm is smaller than that of the general case. This fact suggests that we may invent more efficient randomized algorithms utilizing a particular structure of control problems. This insight may be also useful for improving a sophisticated version of randomized algorithms which interests many researchers recently [76,129,176,252,273,371].

Acknowledgement. We are thankful to Yasuaki Oishi for helpful discussions on a technical point in the main theorem.

13

Sampling Random Transfer Functions

Constantino M. Lagoa, Xiang Li, Maria Cecilia Mazzaro, and Mario Sznaier

EE Dept., The Pennsylvania State University.
University Park, PA 16802, USA,
lagoa@engr.psu.edu, xiangli@psu.edu, cmazzaro@gandalf.ee.psu.edu,
msznaier@frodo.ee.psu.edu

Summary. Recently, considerable attention has been paid to the use of probabilistic algorithms for analysis and design of robust control systems. However, since these algorithms require the generation of random samples of the uncertain parameters, their application has been mostly limited to the case of parametric uncertainty. In this chapter, we provide the means for further extending the use of probabilistic algorithms for the case of dynamic causal uncertain parameters. More precisely, we exploit both time and frequency domain characterizations to develop efficient algorithms for generation of random samples of causal, linear time-invariant uncertain transfer functions. The usefulness of these tools will be illustrated by developing algorithms that address the problem of risk-adjusted model invalidation. Furthermore, procedures are also provided for solving some multi-disk problems arising in the context of synthesizing robust controllers for systems subject to structured dynamic uncertainty.

13.1 Introduction

A large number of control problems of practical importance can be reduced to the robust performance analysis framework illustrated in Figure 13.1. The family of systems under consideration consists of the interconnection of a known stable LTI plant with some bounded uncertainty $\Delta \subset \mathbf{\Delta}$, and the goal is to compute the worst-case, with respect to $\mathbf{\Delta}$, of the norm of the output to some class of exogenous disturbances.

Depending on the choice of models for the input signals and on the criteria used to assess performance, this prototype problem leads to different mathematical formulations such as $\mathcal{H}_\infty$, ℓ^1, $\mathcal{H}_2$ and ℓ^∞ control. A common feature to all these problems is that, with the notable exception of the $\mathcal{H}_\infty$ case, no tight performance bounds are available for systems with uncertainty Δ being a causal bounded LTI operator[1]. Moreover, even in the $\mathcal{H}_\infty$ case, the problem

[1] Recently some tight bounds have been proposed for the $\mathcal{H}_2$ case, but these bounds do not take causality into account; see [254].

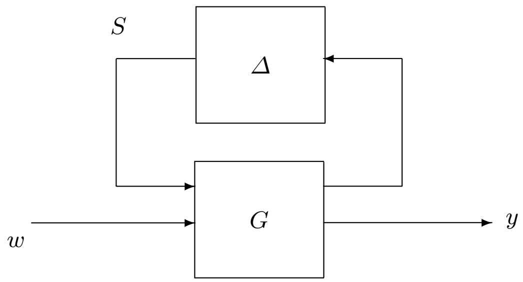


Figure 13.1. The robust performance analysis problem

of computing a tight performance bound is known to be NP-hard in the case of structured uncertainty, with more than two uncertainty blocks [62].

Given the difficulty of computing these bounds, over the past few years, considerable attention has been devoted to the use of probabilistic methods. This approach furnishes, rather than worst case bounds, risk-adjusted bounds; *i.e.*, bounds for which the probability of performance violation is no larger than a prescribed risk level ϵ. An appealing feature of this approach is that, contrary to the worst-case approach case, here, the computational burden grows moderately with the size of the problem. Moreover, in many cases, worst-case bounds can be too conservative, in the sense that performance can be substantially improved by allowing for a small level of performance violation. The application of Monte Carlo methods to the analysis of control systems was recently proposed in the work by Stengel and coworkers [216,290, 347] and was followed, among others, by [24,27,74,84,182,358,404,411]. The design of controllers under risk specifications is also considered in some of the work above as well as in [25,85,199,390,405].

At the present time the domain of applicability of Monte Carlo techniques is largely restricted to the finite-dimensional parametric uncertainty case. The main reason for this limitation resides in the fact that up to now, the problem of sampling causal bounded operators (rather than vectors or matrices) has not appeared in the systems literature. A notable exceptions to this limitation is the algorithm for generating random fixed order state space representations in [72]. In this chapter, we provide two algorithms aimed at removing this limitation when the set $\mathbf{\Delta}$ consists of balls in $\mathcal{H}_\infty$. We use results on interpolation theory to develop three new procedures for random transfer function generation. The first algorithm generates random samples of the first n Markov parameters of transfer functions whose $\mathcal{H}_\infty$ norm is less than or equal to one. This algorithm is particularly useful for problems like time-domain based model invalidation where only the first few Markov parameters of the systems involved are used. The second algorithm generates random transfer functions having the property that, for a given frequency, the frequency response is uniformly distributed over the interior of the unit circle. This algorithm is useful for problems such as time-domain based model (in)validation, where the uncertainty that validates the model description is not necessarily on the boundary of the uncertainty set $\mathbf{\Delta}$. Finally, the third algorithm provides samples uniformly distributed over the unit circle, and is useful for cases such as some robust performance analysis/synthesis problems where the worst-case uncertainty is known to be on the boundary of $\mathbf{\Delta}$.

The usefulness of these tools is illustrated by developing algorithms for model invalidation. Moreover, we also provide an algorithm aimed at solving some multi-disk problems arising in the context of synthesizing robust controllers for systems subject to structured dynamic uncertainty. More precisely, we provide a modification of the algorithm in [201] that when used together with the sampling schemes mentioned above, enables one to solve the problem of designing a controller that robustly stabilizes the system for a 'large' set of uncertainties while guaranteeing a given performance level on a 'smaller' uncertainty subset.

13.2 Preliminaries

13.2.1 Notation

Below we summarize the notation used in this chapter:

$\lfloor x \rfloor$	largest integer smaller than or equal to $x \in \mathbb{R}$.
$\overline{\sigma}\,(A)$	maximum singular value of the matrix A.
$\mathcal{BX}(\gamma)$	open γ-ball in a normed space $\mathcal{X}$: $\mathcal{BX}(\gamma) = \{x \in \mathcal{X} \colon \|x\|_{\mathcal{X}} < \gamma\}$.
$\overline{\mathcal{BX}}(\gamma)$	closure of $\mathcal{BX}(\gamma)$.
$\mathcal{BX}$ $(\overline{\mathcal{BX}})$	open (closed) unit ball in $\mathcal{X}$.
$\mathcal{X} \subset \mathbb{R}^k$.	
$\mathrm{Proj}^l(\mathcal{C})$	projection operator. Given a set $\mathcal{C} \subset \mathbb{R}^m$ and $l < m$: $\mathrm{Proj}^l(\mathcal{C}) \doteq \left\{x \in \mathbb{R}^l \colon \left[x^T y^T\right]^T \in \mathcal{C} \text{ for some } y \in \mathbb{R}^{k-l}\right\}.$
$\mathcal{S}^k_{\mathcal{C}}$	k-th section of a set $\mathcal{C} \subset \mathbb{R}^n$. Given $y \in \mathbb{R}^{n-k}$: $\mathcal{S}^k_{\mathcal{C}}(y) \doteq \left\{x \in \mathbb{R}^k \colon \left[y^T x^T\right]^T \in \mathcal{C}\right\}.$
$\mathbb{E}[X\|Y]$	conditional expected value of X given Y.
$\mathcal{F}_l(M, \Delta)$	lower linear fractional transformation: $\mathcal{F}_l(M, \Delta) = M_{11} + M_{12}\Delta(I - M_{22}\Delta)^{-1}M_{21}.$
$\mathcal{F}_u(M, \Delta)$	upper linear fractional transformation: $\mathcal{F}_u(M, \Delta) = M_{21}\Delta(I - M_{11}\Delta)^{-1}M_{12} + M_{22}.$
ℓ^m_p	extended Banach space of vector valued real sequences equipped with the norm: $\|x\|_p \doteq \left(\sum_{i=0}^{\infty} \|x_i\|_p^p\right)^{\frac{1}{p}} \quad p \in [1, \infty),\ \|x\|_\infty \doteq \sup_i \|x_i\|_\infty.$
$\mathcal{L}_\infty$	Lebesgue space of complex-valued matrix functions essentially bounded on the unit circle, equipped with the norm: $\|G\|_\infty \doteq ess\sup_{\|z\|=1} \overline{\sigma}\,(G(z))$.

$\mathcal{H}_\infty$ subspace of transfer matrices in $\mathcal{L}_\infty$ with bounded analytic continuation inside the unit disk, equipped with the norm: $\|G\|_\infty \doteq ess\sup_{|z|<1} \overline{\sigma}\,(G(z))$.

$\mathcal{H}_{\infty,\rho}$ space of transfer matrices in $\mathcal{H}_\infty$ with analytic continuation inside the disk of radius $\rho \geq 1$, equipped with the norm $\|G\|_{\infty,\rho} \doteq \sup_{|z|<\rho} \overline{\sigma}\,(G(z))$.

$\mathcal{BH}^n_\infty$ set of $(n-1)^{\text{th}}$ order FIR transfer matrices that can be completed to belong to $\mathcal{BH}_\infty$, *i.e.* $\mathcal{BH}^n_\infty \doteq \big\{H(z) = H_0 + H_1 z + \ldots + H_{n-1}z^{n-1} \colon H(z) + z^n G(z) \in \mathcal{BH}_\infty, \text{ for some } G(z) \in \mathcal{H}_\infty\big\}$.

$\mathcal{RX}$ subspace of $\mathcal{X} \subseteq \mathcal{L}_\infty$ composed of real rational transfer matrices.

$\mathcal{H}_2$ Hilbert space of complex matrix valued functions analytic in the set $\{z \in \mathbb{C} : |z| \geq 1\}$, equipped with the inner product

$$\langle H, T\rangle = \frac{1}{2\pi}\int_0^{2\pi} \text{Re}\{\text{tr}[H(e^{j\theta})^* T(e^{j\theta})]\} d\theta.$$

and norm

$$\|T\|_2 = \left(\frac{1}{2\pi}\int_0^{2\pi} \text{Re}\{\text{tr}[T(e^{j\theta})^* T(e^{j\theta})]\} d\theta\right)^{\frac{1}{2}}.$$

$\mathcal{RH}_2$ subspace of all rational functions in $\mathcal{H}_2$.

$X(z)$ z transform of a right-sided real sequence $\{x\}$, evaluated at $\frac{1}{z}$: $X(z) = \sum\limits_{i=0}^{\infty} x_i z^i$.

13.2.2 Space of Proper Rational Transfer Functions

Define the space $\mathcal{G}$ as the space of rational functions $G : \mathbb{C} \rightarrow \mathbb{C}^{n \times m}$ that can be represented as

$$G(z) = G_s(z) + G_u(z).$$

where $G_s(z)$ analytic in the set $\{z \in \mathbb{C} : |z| \geq \alpha\}$ and $G_u(z)$ is strictly proper, analytic in the set $\{z \in \mathbb{C} : |z| < \alpha\}$ and $0 < \beta < \alpha < 1$. Now, given two functions $G, H \in \mathcal{G}$ define the distance function d as

$$d(G,H) \doteq \left(\|G_s(z) - H_s(z)\|_2^2 + \|G_u(\beta/z) - H_u(\beta/z)\|_2^2)\right)^{\frac{1}{2}}.$$

The results later in the chapter that make use of this distance function hold for arbitrary $0 < \beta < \alpha < 1$. However, α and β are usually taken to be very close to one since one is usually interested in distinguishing between the stable and anti-stable parts of a transfer function. Finally, define the projection $\pi_s : \mathcal{G} \rightarrow \mathcal{H}_2$ as

$$\pi_s(G) \doteq G_s.$$

13.2.3 Convex Functions and Subgradients

Consider a convex function $g : \mathcal{H}_2 \rightarrow \mathbb{R}$. Then, given any $G_0 \in \mathcal{H}_2$, there exists a $\partial_G g(G_0) \in \mathcal{H}_2$ such that

$$g(G) - g(G_0) \geq \langle \partial g_G(G_0), G - G_0 \rangle.$$

for all $G \in \mathcal{H}_2$. The quantity $\partial_G g(G_0)$ is said to be a subgradient of g at the point G_0. For example if $g(G) = \|G\|_2$ and G is a scalar; *i.e.*,

$$g(G) = \left(\frac{1}{2\pi} \int_0^{2\pi} |G(e^{j\theta})|^2 d\theta \right)^{1/2}$$

then [57] indicates that

$$\partial_G g(G) = \frac{1}{2\pi \|G\|_2} G.$$

13.2.4 Closed-Loop Transfer Function Parametrization

Central to the results on design presented in this chapter is the parameterization of all achievable closed-loop transfer functions. Consider the closed-loop plant in Figure 13.2 with uncertain parameters $\Delta \in \mathbf{\Delta}$. The Youla parameterization (see, *e.g.*, [318]) indicates that, given $\Delta \in \mathbf{\Delta}$ and a stabilizing controller $C \in \mathcal{G}$, the closed-loop transfer function can be represented as

$$T_{CL}(z, \Delta, C) = T^1_\Delta(z) + T^2_\Delta(z) Q_{\Delta,C}(z) T^3_\Delta(z),$$

where $T^1_\Delta, T^2_\Delta, T^3_\Delta \in \mathcal{RH}_2$ are determined by the plant $G(z, \Delta)$ (and, hence, they also depend on the uncertainty Δ) and $Q_{\Delta,C} \in \mathcal{RH}_2$ depends on both the open loop plant $G(z, \Delta)$ and the controller $C(z)$. Also, given any $Q_{\Delta,C}(s) \in \mathcal{RH}_2$, there exists a controller $C \in \mathcal{G}$ such that the equality above is satisfied.

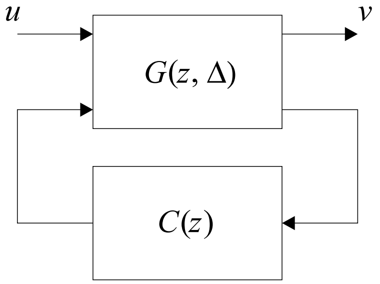


Figure 13.2. Closed loop system

This parameterization also holds for all closed-loop transfer functions, stable and unstable. Using a frequency scaling reasoning one can prove the following result: given $\Delta \in \mathbf{\Delta}$ and a controller $C \in \mathcal{G}$, the closed-loop transfer function can be represented as

$$T_{CL}(z, \Delta, C) = T^1_\Delta(z) + T^2_\Delta(z) Q_{\Delta,C}(z) T^3_\Delta(z),$$

where $T^1_\Delta, T^2_\Delta, T^3_\Delta \in \mathcal{RH}_2$ are the same as above and $Q_{\Delta,C}(s) \in \mathcal{G}$. Furthermore, given any $Q_{\Delta,C}(s) \in \mathcal{G}$ there exists a controller $C \in \mathcal{G}$ such that the equality above is satisfied. See [201] for a discussion on this extension of the Youla parameterization.

Note that the mapping from Δ to $T^1_\Delta, T^2_\Delta, T^3_\Delta$ is not unique. In what follows, we assume that a unique mapping has been selected. The results to follow do not depend on how this mapping is chosen.

13.3 Sampling the Class $\mathcal{BH}^n_\infty$

The use of Monte Carlo methods for risk assessment and volume estimation has been widely studied in the probabilistic literature (see, *e.g.*, [126] and references therein). However, a key issue that needs to be addressed before these methods can be applied is the generation of samples of a random variable with the appropriate distribution. In particular, as we will show in the sequel, using a risk-adjusted approach to perform time-domain model (in)validation and to assess finite horizon robust performance[2] requires solving the following problem.

Problem 1. Given n, generate uniformly distributed samples from a suitable finite dimensional representation of the convex set $\mathcal{BH}^n_\infty$.

In the problem above, n is given by the specific application under consideration: for model invalidation problems, n is given by the number of experimental data points; for performance analysis, n corresponds to the horizon length of interest.

In principle sampling general convex sets is a hard problem, even in the finite-dimensional case. However, as we will show in the sequel, in the case under consideration here, the special structure of the problem can be exploited to obtain a computationally efficient algorithm.

13.3.1 Reducing the Problem to Sampling Finite Dimensional Sets

We begin by showing how Problem 1 can be reduced to the problem of sampling a *finite-dimensional* convex set. From Carathéodory-Fejér Theorem (see

[2] We will also show that this approach allows for assessing *infinite-horizon* robust performance by resorting to an iterative process.

Section 13.8.1) it follows that given the first n Markov parameters $H_i \in \mathbb{R}^{s\times m}$, $i = 0, 1, \ldots, n-1$ of a matrix operator $H(z) \in \mathcal{H}_\infty$, the corresponding $H(z) \in \mathcal{BH}^n_\infty$ if and only if

$$\overline{\sigma}\left(T^n_H\right) \leq 1,$$

where

$$T^n_H(H_0, H_1, \ldots, H_{n-1}) \doteq \begin{bmatrix} H_{n-1} & \cdots & H_1 & H_0 \\ H_{n-2} & \cdots & H_0 & 0 \\ \vdots & & & \vdots \\ H_0 & 0 & \cdots & 0 \end{bmatrix}.$$

Thus, a natural representation for $\mathcal{BH}^n_\infty$ in Problem 1 is the set

$$\mathcal{C}_{\mathcal{H}_n} \doteq \left\{ \{H_i\}_{i=0}^{n-1} : \overline{\sigma}\left(T^n_H\right) \leq 1 \right\}.$$

This leads to the problem below.

Problem 2. Given $n > 0$, generate uniform samples over the convex set $\mathcal{C}_{\mathcal{H}_n}$.

In the sequel, we present an algorithm for generating uniform samples over arbitrary finite dimensional convex sets and we solve Problem 2 as a special case.

13.3.2 Generating Uniform Samples over Convex Sets

Let $\mathcal{C} \subset \mathbb{R}^n$ denote an arbitrary convex set. Given $x \in \mathcal{C}$, partition the vector conformably to some given structure in the following form

$$x = \left[x_1^T \; x_2^T \; \cdots \; x_m^T\right]^T$$

where $x_i \in \mathbb{R}^{n_i}$ and $\sum_{i=1}^m n_i = n$.

Consider now the following algorithm:

Algorithm 13.1

Step 1. Let $k = 0$. Generate N samples, x_1^l, $l = 1, 2, \ldots, N$, uniformly distributed over the set $\mathcal{I}_o \doteq Proj^{n_1}(\mathcal{C})$.

Step 2. Let $k := k + 1$. For every generated sample $(x_1^l, x_2^l, \ldots, x_{k-1}^l)$, let

$$\mathcal{C}_k(x_1^l, x_2^l, \ldots, x_{k-1}^l) \doteq \mathcal{S}^{n^*}_{\mathcal{C}}\left(\left[(x_1^l)^T \; (x_2^l)^T \; \cdots \; (x_{k-1}^l)^T\right]^T\right)$$
$$\mathcal{I}_k(x_1^l, x_2^l, \ldots, x_{k-1}^l) \doteq \mathrm{Proj}^{n_k}(\mathcal{C}_k),$$

with $n^* \doteq \sum_{i=k}^m n_i$. Generate

$$N_k \doteq \lfloor \alpha_k N \mathrm{vol}\left(\mathcal{I}_k\right) \rfloor$$

samples uniformly over the set $\mathcal{I}_k$, where α_k is an arbitrary positive constant.

Step 3. If $k < m$ go to step 2. Else Stop.

Next we show that the probability distribution of the samples generated by this algorithm converges, with probability one, to a uniform distribution as $N \to \infty$.

Theorem 1. *Consider any set $\mathcal{A} \subseteq \mathcal{C}$. For a given N, denote by $N_t(N)$ and $N_{\mathcal{A}}(N)$*[3] *the total number of samples generated by Algorithm 13.1 and the number of those samples that belong to $\mathcal{A}$, respectively. Then*

$$\frac{N_{\mathcal{A}}(N)}{N_t(N)} \overset{w.p.1}{\longrightarrow} \frac{\mathrm{vol}(\mathcal{A})}{\mathrm{vol}(\mathcal{C})}.$$

Proof. See Appendix 13.9.

Remark 1. The main reason that prevents the estimate of probability produced by the samples generated by Algorithm 1 from being unbiased is the fact that, in general, at step s,

$$\frac{\lfloor N\alpha_s v_s(X_1^k, X_2^m, \ldots, X_{s-1}^n)\rfloor}{N} \neq \alpha_s v_s(X_1^k, X_2^m, \ldots, X_{s-1}^n).$$

due to the rounding. Indeed, for any union of hyper-rectangles $\mathcal{A} \subseteq \mathcal{C}$ satisfying

$$\frac{\lfloor N\alpha_s v_s(X_1^k, X_2^m, \ldots, X_{s-1}^n)\rfloor}{N} = \alpha_s v_s(X_1^k, X_2^m, \ldots, X_{s-1}^n).$$

it can be shown that, for any value of N,

$$\frac{\mathbb{E}[N_{\mathcal{A}}]}{\mathbb{E}[N_t]} = \frac{\mathrm{vol}(\mathcal{A})}{\mathrm{vol}(\mathcal{C})}$$

Unfortunately, in the general case this equality is not true. However, as we show next, the difference between these values can be made very small even for relatively small values of N.

Theorem 2. *Consider a set $\mathcal{A} \subseteq \mathcal{C}$. Then, there exist constants k_1, k_2 and k_3 such that, for any N,*

$$\left|\frac{\mathbb{E}[N_{\mathcal{A}}]}{\mathbb{E}[N_t]} - \frac{\mathrm{vol}(\mathcal{A})}{\mathrm{vol}(\mathcal{C})}\right| \leq \frac{1}{N}\frac{k_1}{k_2 + \frac{k_3}{N}}.$$

Proof. see Appendix 13.10.

Remark 2. The main difference between 'traditional' Monte Carlo simulation for risk assessment and risk assessment using Monte Carlo methods together with the sample generation algorithm above is the fact that here one has to determine the volume of several sets in order to compute the samples.

[3] Note that both N_t and $N_{\mathcal{A}}$ are random variables.

However, as we will see in the next section, for the problem at hand we do not need to estimate these volumes. The structure is such that one can determine them up to a multiplicative constant. Therefore, the number of samples needed to compute reliable estimates of risk is similar to the ones in 'traditional' Monte Carlo simulations. For bounds on the number of samples required for reliable estimation of risk, see [182, 358].

13.3.3 $\mathcal{BH}_\infty^n$ as a Simpler Case

In the case of a general convex set $\mathcal{C}$, Algorithm 13.1 requires knowledge of the volume of the projection sets up to a multiplying constant. However, as we show in the sequel, for sets of the form $\mathcal{C}_{\mathcal{H}_n} \doteq \left\{\{H_i\}_{i=0}^{n-1} : \overline{\sigma}\left(T_H^n\right) \leq 1\right\}$ it is possible to find these quantities *analytically.* Since these are precisely the sets arising in the context of Problem 1, and since the linear spaces $\mathbb{R}^{s\times m}$ and $\mathbb{R}^{sm}$ are isomorphic, it follows that this problem can be efficiently solved by applying Algorithm 13.1.

Specifically, given $\{H_0, H_1, \ldots, H_{k-1}\}$, $1 \leq k \leq n$, consider the problem of determining the set

$$\begin{aligned}\mathrm{Proj}^{n_k}&\left(\mathcal{C}_k(H_0, H_1, \ldots, H_{k-1})\right)\\ &\doteq \big\{H_k \colon (H_0, \ldots, H_{k-1}, H_k, H_{k+1}, \ldots, H_{n-1}) \in \mathcal{C}_{\mathcal{H}_n},\\ &\qquad\qquad \mathrm{forsome}(H_{k+1}, \ldots, H_{n-1})\big\}. \quad (13.1)\end{aligned}$$

From Parrott's Theorem (Appendix 13.8) it follows that the set (13.1) is given by:

$$\left\{H_k \colon \overline{\sigma}\left(T_H^{k+1}(H_0, H_1, \ldots, H_k)\right) \leq 1\right\}.$$

Moreover, an explicit parameterization of this set can be obtained as follows. Consider the partition

$$T_H^{k+1}(H_0, H_1, \ldots, H_k) = \begin{bmatrix} H_k & B \\ C & A \end{bmatrix}$$

and let the matrices Y and Z be a solution of the linear equations

$$\begin{aligned} B &= Y(I - A^T A)^{\frac{1}{2}} \\ C &= (I - AA^T)^{\frac{1}{2}} Z \\ \overline{\sigma}\,(Y) &\leq 1, \overline{\sigma}\,(Z) \leq 1. \end{aligned}$$

Then

$$\begin{aligned}&\left\{H_k \colon \overline{\sigma}\left(T_H^{k+1}\right) \leq 1\right\}\\ &\qquad = \left\{H_k \colon H_k = -YA^T Z + (I - YY^T)^{\frac{1}{2}} W (I - Z^T Z)^{\frac{1}{2}}, \overline{\sigma}\,(W) \leq 1\right\}.\end{aligned}$$

Hence, generating uniform samples over the set (13.1) reduces to the problem of uniformly sampling the set $\{W \colon \overline{\sigma}\,(W) \leq 1\}$. Algorithms to do sampling over such sets are readily available (see for instance [74]). In addition, this parameterization allows for easily computing, up to a multiplying constant, the volume of the set $\mathrm{Proj}^{n_k}(\mathcal{C}_k)$, required in step 2 of Algorithm 13.1. This follows from the fact that $\mathrm{Proj}^{n_k}\big(\mathcal{C}_k(H_0, H_1, \ldots, H_{k-1})\big)$ is a linear transformation of the set $\mathcal{M}\big(\{W \colon \overline{\sigma}(W) \leq 1\}\big)$ and thus

$$J(H_0, H_1, \ldots, H_{k-1}) = \frac{\mathrm{vol}\left(\mathrm{Proj}^{n_k}\left(\mathcal{C}_k(H_0, H_1, \ldots, H_{k-1})\right)\right)}{\mathrm{vol}\left(\mathcal{M}\left(\{W \colon \overline{\sigma}(W) \leq 1\}\right)\right)}.$$

where

$$J(H_0, H_1, \ldots, H_{k-1}) = |(I - YY^T)^{\frac{1}{2}}|^m |(I - Z^T Z)^{\frac{1}{2}}|^s$$

is the Jacobian of the transformation above (see [74], Appendix F). Combining these observations leads to the following algorithm for solving Problem 1.

Algorithm 13.2

Step 1. Let $k = 0$. Generate N samples uniformly distributed over the set

$$\{H_0 \colon \overline{\sigma}(H_0) \leq 1\}.$$

Step 2. Let $k := k + 1$. For every generated sample $(H_0^l, H_1^l, \ldots, H_{k-1}^l)$, consider the partition

$$T_H^{k+1}(H_0^l, H_1^l, \ldots, H_k) = \begin{bmatrix} H_k & B \\ C & A \end{bmatrix}$$

and let the matrices Y and Z be a solution of the linear equations

$$B = Y(I - A^T A)^{\frac{1}{2}}$$
$$C = (I - AA^T)^{\frac{1}{2}} Z$$
$$\overline{\sigma}(Y) \leq 1, \overline{\sigma}(Z) \leq 1.$$

Generate $\lfloor NJ(H_0, H_1, \ldots, H_{k-1})\rfloor$ samples uniformly over the set $\{W \colon \overline{\sigma}(W) \leq 1\}$ and for each of those samples W^i, take

$$H_k^i = -YA^T Z + (I - YY^T)^{\frac{1}{2}} W^i (I - Z^T Z)^{\frac{1}{2}}.$$

Step 3. If $k < m$ go to Step 2. Else Stop.

13.3.4 Extension to the Infinite Horizon Case

In this section we show that the results above can be extended to assess *infinite horizon* robust performance. Due to space constraints, we provide only an outline of the ideas involved. Begin by noting that Carathéodory-Fejér only specifies the values of the function and its first $n - 1$ derivatives at $z = 0$.

However, these conditions do not impose any constraints on the smoothness of the function over the unit disk and can lead to transfer functions which do not represent a physical uncertainty. For example, $h = \{0, 0, \ldots, 0, \frac{\gamma}{(1+\gamma)^2}\}$ has all the h_i, $i \leq n-1$ arbitrarily small and satisfies the Carathéodory-Fejér theorem. Moreover, it can be easily shown that a suitable interpolant is given by

$$H(z) = \frac{\gamma}{1 + \gamma - z^n}.$$

Clearly, $H(z) \in \mathcal{BH}_\infty$. However, $\|\frac{d}{dz}H(z)\|_\infty = \frac{n}{\gamma} \to \infty$. Since these functions are arguably not a good abstraction of physical uncertainty, estimating worst-case performance bounds using samples from the set $\mathcal{F}_n$ can lead to conservative results. This effect can be avoided by working with the ball $\mathcal{BH}_{\infty,\rho}$, instead of $\mathcal{BH}_\infty$, since restricting all the poles of the system to the exterior of the disk $|z| \geq \rho$ induces a smoothness constraint. This leads to the following modified version of Problem 1.

Problem 3. Given $n > 0, \rho > 1, \rho \sim 1$, generate uniformly distributed samples over an appropriate finite-dimensional representation of the set

$$\mathcal{F}_{n,\rho} \doteq \big\{H(z) = H_0 + H_1 z + \ldots + H_{n-1} z^{n-1} \colon H(z) + z^n G(z) \in \mathcal{BH}_{\infty,\rho}, \\ \text{for some } G(z) \in \mathcal{BH}_{\infty,\rho}\big\}.$$

As we show next, this problem readily reduces to Problem 2 and thus can be solved using Algorithm 13.1. To this end, note that $F(z) = H(z) + z^n G(z) \in \mathcal{BH}_{\infty,\rho}$ is equivalent to $F(\rho z) \in \mathcal{BH}_\infty$. Combining this observation with Carathéodory-Fejér Theorem, it follows that, given $\{H_0, H_1, \ldots, H_{n-1}\}$, then there exists $G(z) \in \mathcal{BH}_{\infty,\rho}$ such that $\sum_{i=0}^{n-1} H_i z^i + z^n G(z) \in \mathcal{BH}_{\infty,\rho}$ if and only if

$$\overline{\sigma}\left(T^n_{\hat{H}}\right)(\hat{H}_0, \hat{H}_1, \ldots, \hat{H}_{n-1}) \leq 1,$$

where $\hat{H}_i = \rho^i H_i$. It follows that Problem 3 reduces to Problem 2 simply with the change of variables $H_k \to \rho^k H_k$.

Next, we show that the norm of the tail $\|z^n G(z)\|_\infty \to 0$ as $n \to \infty$. Thus, sampling the set $\mathcal{F}_{n,\rho}$ indeed approximates sampling the ball $\mathcal{BH}_{\infty,\rho}$. To establish this result note that if $F \in \mathcal{BH}_{\infty,\rho}$, then its Markov parameters satisfy

$$F_k = \frac{1}{2\pi} \oint_{\partial D_\rho} F(z) \frac{dz}{z^{k+1}} \Rightarrow \overline{\sigma}\left(F_k\right) \leq \frac{1}{\rho^k}$$

where D_ρ denotes the disk centered at the origin with radius ρ. Thus

$$\|z^n G(z)\|_\infty = \|\sum_{i=n}^{\infty} F_i z^i\|_\infty \leq \sum_{i=n}^{\infty} \frac{1}{\rho^i} = \frac{1}{\rho^{n-1}} \frac{1}{\rho - 1}.$$

From this inequality it follows that $\|F(z) - H(z)\|_\infty \leq \epsilon$ for $n \geq n_o(\epsilon)$, for some $n_o(\epsilon)$ that can be precomputed *a-priori.* Recall (see for instance

Corollary B.5 in [254]) that robust stability of the LFT interconnection shown in Figure 13.1 implies that $(I - M_{11}\Delta)^{-1}$ is uniformly bounded over $\mathcal{BH}_\infty$. In turn, this implies that there exists some finite β such that $\|\mathcal{F}_u(M,\Delta)\|_\infty \leq \beta$ for all $\Delta \in \mathcal{BH}_\infty$. Thus, given some $\epsilon_1 > 0$, one can find ϵ and $n_o(\epsilon)$ such that $\|\mathcal{F}_u(M,F) - \mathcal{F}_u(M,H)\|_* \leq \epsilon_1$, where $\|\cdot\|_*$ denotes a norm relevant to the performance specifications.

Finally, we conclude this section by showing that the proposed algorithm can also be used to assess performance against uncertainty in $\overline{\mathcal{RBH}_\infty}$. Consider a sequence $\rho_i \downarrow 1$ and let Δ_i be the corresponding worst-case uncertainty. Since $\mathcal{BH}_{\infty,\rho} \subset \mathcal{BH}_\infty$ and $\|\mathcal{F}_u(M,\Delta)\|_\infty \leq \beta$ it follows that both Δ_i and $\mathcal{F}_u(M,\Delta_i)$ are normal families (see Appendix 13.8). Thus, they contain a normally convergent subsequence $\Delta_i \to \tilde{\Delta}$ and $\mathcal{F}_l(M,\Delta_i) \to \mathcal{F}_l(M,\tilde{\Delta}_i)$. It can be easily shown that $\tilde{\Delta}$ is indeed the worst case uncertainty over $\overline{\mathcal{RBH}_\infty}$. Thus, robust performance can be assessed by applying the proposed algorithm to a sequence of problems with decreasing values of ρ.

13.4 Sampling $\mathcal{BH}_\infty$ - A Frequency Domain Approach

We now present two algorithms, based on Nevanlinna-Pick interpolation, for generating random transfer functions in $\mathcal{BH}_\infty$.

13.4.1 Sampling the 'Inner' $\mathcal{BH}_\infty$

The first one, based on 'ordinary' Nevanlinna-Pick interpolation, provides transfer functions with $\mathcal{H}_\infty$ norm less or equal than 1 and whose frequency response, at given frequency grid points, is uniformly distributed over the complex plane unit circle.

Algorithm 13.3

Step 1. Given an integer N, pick N frequencies λ_i such that $|\lambda_i| = 1$, $i = 1, 2, \ldots, N$.

Step 2. Generate N independent samples w_i uniformly distributed over the set $\{w \in \mathbb{C} : |w| < 1\}$.

Step 3. Find $0 < r < 1$ such that the matrix Λ with entries

$$\Lambda_{i,j} = \frac{1 - w_i w_j^*}{1 - r^2 \lambda_i \lambda_j^*}$$

is positive definite.

Step 4. Find a rational function $h_r(\lambda)$ analytic inside the unit circle satisfying

$$h_r(r\lambda_i) = w_i; \quad i = 1, 2, \ldots, N$$

$$\|h_r\|_\infty \leq 1$$

by solving a 'traditional' Nevanlinna-Pick interpolation problem.

Step 5. The random transfer function is given by

$$h(z) = h_r(rz^{-1}).$$

We refer the reader to Appendix 13.8 for a brief review of results on Nevanlinna-Pick interpolation and state space descriptions of the interpolating transfer function $h(z)$.

Remark 3. Note that, there always exists an $0 < r < 1$ that will make the matrix Λ positive definite. This is a consequence of the fact that the diagonal entries are positive real numbers and that, as one increases $r < 1$, the matrix will eventually be diagonally dominant.

13.4.2 Sampling the Boundary of $\mathcal{BH}_\infty$

We now present a second algorithm for random generation of rational functions. The algorithm below generates random transfer functions whose frequency response, at given frequency grid points, is uniformly distributed over the boundary of the unit circle. Recall that the rational for generating these samples is that in many problems it is known that the worst case uncertainty is located in the boundary of the uncertainty set $\cdot$, and thus there is no point in generating and testing elements with $\|\Delta\|_\infty < 1$.

Algorithm 13.4

Step 1. Given an integer N, pick N frequencies λ_i such that $|\lambda_i| = 1$, $i = 1, 2, \ldots, N$.

Step 2. Generate N independent samples w_i uniformly distributed over the set $\{w \in \mathbb{C} : |w| = 1\}$.

Step 3. Find the smallest possible $\rho \geq 0$ such that the matrix Λ with entries

$$\Lambda_{i,j} = \begin{cases} \frac{1-w_i^* w_j}{1-\lambda_i^* \lambda_j} & i \neq j \\ \rho & i = j \end{cases}$$

is positive definite.

Step 4. Let

$$\theta(\lambda) = \begin{bmatrix} \theta_{11}(\lambda) & \theta_{12}(\lambda) \\ \theta_{21}(\lambda) & \theta_{22}(\lambda) \end{bmatrix}$$

be a 2×2 transfer function matrix given by

$$\theta(\lambda) = I + (\lambda - \lambda_0) C_0 (\lambda I - A_0)^{-1} \Lambda^{-1} (\lambda I - A_0^*)^{-1} C_0^* J$$

where

$$C_0 = \begin{bmatrix} w_1 \ldots w_N \\ 1 \ \ldots \ 1 \end{bmatrix}; \quad A_0 = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix}; \quad J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

and λ_0 *is a complex number of magnitude 1 and not equal to any of the numbers* $\lambda_1, \lambda_2, \ldots, \lambda_N$*.*

Step 5. The random transfer function is given by

$$h(z) = \frac{\theta_{12}(z^{-1})}{\theta_{22}(z^{-1})}.$$

The algorithm above provides a solution of the boundary Nevanlinna-Pick interpolation problem

$$h(\lambda_i) = w_i; \quad i = 1, 2, \ldots, N$$

$$h'(\lambda_i) = \rho \lambda_i^* w_i; \quad i = 1, 2, \ldots, N$$

$$\|h\|_\infty = 1.$$

A proof of this result can be found in [23]. A more complete description of the results on boundary Nevanlinna-Pick interpolation used in this chapter is given in Appendix 13.9.

Remark 4. The search for the lowest ρ that results in a positive definite matrix Λ is equivalent to finding the interpolant with the lowest derivative.

13.5 Application 1: Risk-Adjusted Time-Domain Model (In)validation

In this section we exploit the sampling framework introduced in Section 13.3.3 to solve the problem of model (in)validation in the presence of structured LTI uncertainty using time-domain data. Consider the lower LFT interconnection, shown in Figure 13.3, of a known model M and structured dynamic uncertainty Δ.

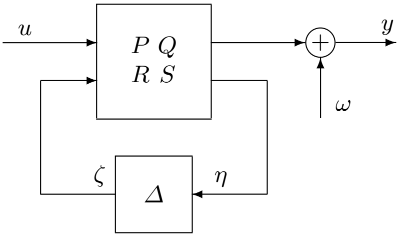


Figure 13.3. The model (in)validation set-up

The block M

$$M \doteq \begin{bmatrix} P & Q \\ R & S \end{bmatrix}$$

consists of a nominal model P of the actual system and a description, given by the blocks Q, R and S^4 of how uncertainty enters the model. The block $\varDelta$ is known to belong to a given set $\boldsymbol{\varDelta}_{st}$:

$$\boldsymbol{\varDelta}_{st}(\gamma) \doteq \{\varDelta\colon \varDelta = \mathrm{diag}(\varDelta_1, \ldots, \varDelta_l),\ \varDelta_i \in \overline{\mathcal{BH}_\infty}(\gamma), \forall i = 1, \ldots, l\}$$

Finally, the signals u and y represent a known test input and its corresponding output respectively, corrupted by measurement noise $\omega \in \mathcal{N} \doteq \mathcal{B}\ell_p^m[0,n](\epsilon_t)$. The goal is to determine whether the measured values of the pair (u, y) are consistent with the assumed nominal model and uncertainty description, as formalized in the following problem.

Problem 4. Given the time-domain experiments:

$$u \doteq \{u_0, u_1, \ldots, u_n\},\ y \doteq \{y_0, y_1, \ldots, y_n\}$$

determine if they are consistent with the assumed *a-priori* information $(M, \mathcal{N}, \boldsymbol{\varDelta}_{st})$, *i.e.* whether the consistency set

$$\mathcal{T}(y) = \{(\varDelta, \omega)\colon \varDelta \in \boldsymbol{\varDelta}_{st},\ \omega \in \mathcal{N} \text{ and } y_k = \big(\mathcal{F}_l(M, \varDelta) * u + \omega\big)_k, k = 0, \ldots, n\}$$

is non-empty.

Model (in)validation of Linear Time Invariant (LTI) systems has been extensively studied in the past decade (see for instance [83, 274, 337] and references therein). The main result shows that in the case of unstructured uncertainty, it reduces to a convex optimization problem that can be efficiently solved. In the case of structured uncertainty, the problem leads to bilinear matrix inequalities, and has been shown to be NP-hard in the number of uncertainty blocks [362]. However, (weaker) necessary conditions in the form of LMIs are available, by reducing the problem to a scaled unstructured (in)validation one ([83, 362]).

13.5.1 Reducing the Problem to Finite-Dimensional Sampling

In the sequel we show that the computational complexity of the model (in)validation problem can be overcome by pursuing a *risk-adjusted* approach. The basic idea is to sample the set $\boldsymbol{\varDelta}_{st}$ in an attempt to find an element that, together with an admissible noise, explains the observed experimental data. If no such uncertainty can be found, then we can conclude that, with a certain probability, the model is invalid. Note that, given a *finite* set of n input/output

[4] We will assume that $\|S\|_\infty < \gamma^{-1}$, so that the interconnection $\mathcal{F}_l(M, \varDelta)$ is well-posed.

measurements, since Δ is causal, *only the first n Markov parameters* affect the output y. Thus, in order to approach this problem from a risk-adjusted perspective, we only need to generate uniform samples of the first n Markov parameters of elements of the set $\boldsymbol{\Delta}_{st}$. Combining this observation with Algorithm 13.2, leads to the following model (in)validation algorithm:

Algorithm 13.5 *Given γ_{st}, take N_s samples of $\boldsymbol{\Delta}_{st}(\gamma_{st})$, $\{\Delta^i(z)\}_{i=1}^{N_s}$, according to the procedure described in Section 13.3.3.*

1. *At step s, let*
$$\omega^s \doteq \{(y - \mathcal{F}_l(M, \Delta^s) * u)_k\}_{k=0}^n. \tag{13.2}$$
2. *Find whether $\omega^s \in \mathcal{N}$. If so, stop. Otherwise, consider next sample $\Delta^{s+1}(z)$ and go back to Step 2.*

Clearly, the existence of at least one $\omega^s \in \mathcal{N}$ is equivalent to $\mathcal{T}(y) \neq \emptyset$. The algorithm finishes, either by finding one admissible uncertainty $\Delta^s(z)$ that makes the model not invalidated by the data or after N_s steps, in which case the model is deemed to be invalid. The following Lemma gives a bound on the probability of the algorithm terminating without finding an admissible uncertainty even though the model is valid, *e.g.* the probability of discarding a valid model.

Lemma 1. *Let* (ϵ, δ) *be two positive constants in* $(0,1)$. *If N in Algorithm 13.2 is chosen such that*
$$N \geq \frac{\ln(1/\delta)}{\ln(1/(1-\epsilon))},$$
then, with probability greater than $1 - \delta$, the probability of rejecting a model which is not invalidated by the data is smaller than ϵ.

Proof. Define the function $f(\Delta^s(z)) \doteq \epsilon_t - \|\omega^s\|_{p[0,N]}$, with ω^s given by (13.2). Note that the model is not invalidated by the data whenever one finds at least one $\Delta^s(z)$ so that $f(\Delta^s(z)) > 0$. Equivalently, if $\forall \Delta^s$, $f(\Delta^s(z)) \leq 0$, we might be rejecting a model which is indeed not invalidated by the data. Following [358], since the number of independent samples is at least N, then
$$\mathbb{P}^N\big\{\mathbb{P}\{\exists \Delta(z) \colon f(\Delta(z)) > 0 | \{f(\Delta^i(z))\}_{i=1}^{N_s} \leq 0\} \leq \epsilon\big\} \geq (1-\delta),$$
which yields the desired result. □

Thus, by introducing an (arbitrarily small) risk of rejecting a possibly *good* candidate model, we can substantially alleviate the computational complexity entailed in validating models subject to structured uncertainty.

In addition, as pointed out in [410] the worst-case approach to model invalidation is optimistic since a candidate model will be accepted even if there exists only a very small set (or even a single) of pairs (uncertainty,noise) that validate the experimental record. On the other hand, both the approach in [410] and the one proposed here will reject (with probability close to 1) such models. The main difference between these approaches is related to the experimental data and the *a-priori* assumptions. The approach in [410] uses frequency domain data and relies heavily on the whiteness of the noise process and independence between samples at different frequencies – as a consequence of Nevannlina-Pick boundary interpolation theory – to obtain mathematically tractable, frequency-by-frequency estimates of the probability of the model not being invalidated by the data. On the other hand, the approach pursued in this chapter is based on time-domain data, and while $\varDelta$ is treated as a random variable, the risk estimates are independent of the specific density function [358].

13.5.2 A Simple (In)Validation Example

In order to illustrate the proposed method, consider the following system:

$$\hat{G}(z) = \mathcal{F}_l(M, \hat{\varDelta}),$$

with:

$$P(z) = \frac{0.2(z+1)^2}{18.6z^2 - 48.8z + 32.6} \qquad Q(z) = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}$$

$$R(z) = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \qquad S(z) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$\hat{\varDelta}(z) = \begin{bmatrix} \frac{0.125(5.1-4.9z)}{(6.375-3.6250z)} & 0 & 0 \\ 0 & \frac{0.1(5.001-4.9990z)}{(6.15-3.85z)} & 0 \\ 0 & 0 & \frac{0.05(5.15-4.85z)}{(6.95-3.05z)} \end{bmatrix}.$$

Our experimental data consists of a set of $n = 20$ samples of the impulse response of $\hat{G}(z) = \mathcal{F}_\ell(P, \hat{\varDelta})$, corrupted by additive noise in $\mathcal{N} \doteq \mathcal{B}\ell_\infty[0, n](0.0041)$. The noise bound ϵ_t represents a 10% of the peak value of the impulse response. Our goal is to find the minimum size of the uncertainty, γ_{st}, so that the model is not invalidated by the data. A coarse lower bound on γ_{st} can be obtained by performing an invalidation test using *unstructured* uncertainty $\varDelta(s) \in \boldsymbol{\varDelta}_u$, which reduces to an LMI feasibility problem [83]. In our case, this approach led to the lower bound $0.0158 \leq \gamma_{st}$.

Direct application of Lemma 1 indicates that using $N_s = 6000$ samples guarantees a probability of at least 0.9975 that $\text{prob}\{f(\varDelta) > 0\} \leq 0.001$. Thus, starting from $\gamma_{st} = 0.0158$, we generated 3 sets of $N_s = 6000$ samples over $\mathcal{BH}_\infty(\gamma_{st})$, one for each of the scalar blocks $\varDelta_i(z)$, $i = 1, 2, 3$, which yields one single set of samples $\{\varDelta^n(z)\}_{n=1}^{N_s}$ over $\boldsymbol{\varDelta}_{st}$[5]. Following Section 13.3, at each given value of γ_{st}, we evaluated the function

[5] The corresponding samples over the set $\boldsymbol{\varDelta}(\gamma_{st})$ were obtained by appropriate scaling of the impulse response of each given sample by γ_{st}.

$$f(\Delta^s) = \epsilon_t - \| \{\mathcal{F}_l(M, \Delta^s) * u - y\}_{k=0}^n \|_{\infty[0,n]}$$

for all $\Delta^s \in \boldsymbol{\Delta}_{st}(\gamma_{st})$. If $\forall \Delta^s$, $f(\Delta^s) < 0$, then the model is invalidated by the data with high probability. It is then necessary to increase the value of γ_{st} and continue the (in)validation test. In this particular example, the test was repeated over a grid of 1000 points of the interval $\mathcal{I}$ until we obtained the value γ_{st} of 0.0775, the minimum value of γ_{st} for which the model was not invalidated by the given experimental evidence.

The proposed approach differs from the one in [83] in that here the invalidation test is performed by searching over $\boldsymbol{\Delta}_{st}$ with the hope of finding one admissible $\Delta \in \boldsymbol{\Delta}_{st}$ that makes the model not invalid; while there it is done by searching over the class of unstructured uncertainties $\boldsymbol{\Delta}_u$ and by introducing, at each step, diagonal similarity scaling matrices with the aim of invalidating the model. More precisely, if at step k the model subject to unstructured uncertainty remains *not* invalidated (which is equivalent to the existence of at least one feasible pair (ζ, D_k) so that a given matrix $M(\zeta, D_k) \leq 0$), one possible strategy is to select the scaling D_{k+1} so as to maximize the trace of M. See [82, Chapter 9, pp. 301–306] for details. However, for this particular example $D_k = \mathrm{diag}(d_{1k}, d_{2k}, d_{3k})$ and this last condition becomes

$$\sup_{d_{1k}, d_{2k}, d_{3k}} -n\big(1 + \frac{1}{\gamma^2}\big)(d_{2k} + d_{3k}) + n\big(1 - \frac{1}{\gamma^2}\big)d_{1k}, \quad d_{1k}, d_{2k}, d_{3k} \geq 0.$$

For $0 < \gamma < 1$, clearly the supremum is achieved at $d_{1k} = 0$, $d_{2k} = 0$ and $d_{3k} = 0$. As an alternative searching strategy, one may attempt to randomly check condition $M(\zeta, D_k) \leq 0$ by sampling appropriately the scaling matrices, following [362]. Using 6000 samples led to a value of γ_{st} of 0.03105 for which the model was invalidated by the data. For larger values of γ_{st} in $[0.03105, 0.125]$ nothing can be concluded regarding the validity of the model.

Combination of these bounds with the risk-adjusted ones obtained earlier shows that the model is definitely invalid for $\gamma_{st} \leq 0.03105$, invalid with probability 0.999 in $0.03105 < \gamma_{st} < 0.0755$ and it is not invalidated by the experimental data available thus far for $0.0755 \leq \gamma_{st} \leq 0.125$. Thus these approaches, rather than competing, can be combined to obtain sharper conditions for rejecting candidate models.

As a final remark, note that it seems possible to reduce the number of samples required by the proposed method, at the expense of requiring additional *a-priori* information on the actual system. This situation may arise for example when it is known that the uncertainty affecting the candidate model is exponentially stable or even real, if the system has uncertain parameters. The former case amounts to sampling $\mathcal{BH}_{\infty,\rho} \subset \mathcal{BH}_\infty$, $\rho > 1$, while the latter involves samples of constant matrices.

13.6 Application 2: Multi-Disk Design Problem

In this section we discuss a second application of the sampling algorithms developed in this chapter. More precisely, we introduce a stochastic gradient based algorithm to solve the so-called multi-disk design problem. We aim at solving the problem of designing a robustly stabilizing controller that results in guaranteed performance in a subset of the uncertainty support set. The algorithm presented is an extension of the algorithms developed in [201]. Before providing the controller design algorithm, we first provide a precise definition of the problem to be solved and the assumptions that are made.

13.6.1 Problem Statement

Consider the closed-loop system in Figure 13.2 and a convex objective function $g_1 : \mathcal{H}_2 \to \mathbb{R}$. Given a performance value γ_1 and uncertainty radii $r_2 > r_1 > 0$, we aim at designing a controller $C^*(s)$ such that the closed-loop system $T_{CL}(z, \Delta, C^*)$ is stable for all $\|\Delta\|_\infty \leq r_2$ and satisfies

$$g\left[T_{CL}(z, \Delta, C^*)\right] \leq \gamma_1$$

for all $\|\Delta\|_\infty \leq r_1$. Throughout this chapter, we will assume that the problem above is feasible. More precisely, the following assumption is made.

Assumption 13.1. *There exists a controller C^* and an $\varepsilon > 0$ such that*

$$d(Q_{\Delta,C^*}, Q) < \varepsilon \Rightarrow g_1\left[T_\Delta^1(z) + T_\Delta^2(z)Q(z)T_\Delta^3(z)\right] \leq \gamma_1$$

for all $\|\Delta\|_\infty \leq r_1$ and there exists a γ_2 (sufficiently large) such that

$$\begin{aligned} &d(Q_{\Delta,C^*}, Q) < \varepsilon \\ &\quad \Rightarrow g_2\left[T_\Delta^1(z) + T_\Delta^2(z)Q(z)T_\Delta^3(z)\right] \doteq \left\|T_\Delta^1(z) + T_\Delta^2(z)Q(z)T_\Delta^3(z)\right\|_2 \leq \gamma_2 \end{aligned}$$

for all $\|\Delta\|_\infty \leq r_2$.

Remark 5. Even though it is a slightly stronger requirement than robust stability, the existence of a large constant γ_2 satisfying the second condition above can be considered to be, from a practical point of view, equivalent to robust stability.

13.6.2 Controller Design Algorithm

We now state the proposed robust controller design algorithm. This algorithm has a free parameter η that has to be specified. This parameter can be arbitrarily chosen from the interval $(0, 2)$.

Algorithm 13.6

Step 1. Let $k = 0$. Pick a controller $C_0(z)$.

Step 2. Generate sample i^k with equal probability or being 1 or 2.

Step 3. Draw sample Δ^k over $\mathcal{BH}_\infty(r_{i^k})$. Given $G(z, \Delta^K)$, compute $T^1_{\Delta^k}(z)$, $T^2_{\Delta^k}(z)$ and $T^3_{\Delta^k}(z)$ as described in [318].

Step 4. Let $Q_k(z)$ be such that the closed-loop transfer function using controller $C_k(s)$ is

$$T_{CL}(z, \Delta^k,\ C_k) = T^1_{\Delta^k}(z) + T^2_{\Delta^k}(z) Q_k(z) T^3_{\Delta^k}(z).$$

Step 5. Do the stabilizing projection[6]

$$Q_{k,s}(z) \ = \ \pi_s(Q_k(z)).$$

Step 6. Perform update

$$Q_{k+1}(z) = Q_{k,s}(z) - \alpha_k(Q_{k,z}, \Delta^k)(z) \partial_Q g_{i^k}(T_{CL}(z, \Delta^k, Q))|_{Q_{k,s}}$$

where

$$\alpha_k(Q_k, \Delta) = \begin{cases} \eta \frac{g_{i^k}(T_{CL}(z,\Delta,Q_k)) - \gamma_{i^k} + \epsilon \, \|\partial_Q g_{i^k}(T_{CL}(z,\Delta,Q))|_{Q_k}\|_2}{\|\partial_Q g_{i^k}(T_{CL}(z,\Delta,Q))|_{Q_k}\|_2^2} & \\ \qquad \textit{if } g_{i^k}(T_{CL}(z, \Delta, Q_k)) > \gamma_{i^k} & \\ 0 \qquad \text{otherwise,} & \end{cases} .$$

Step 7. Determine the controller $C_{k+1}(z)$ so that

$$Q_{\Delta^k, C_{k+1}} = Q_{k+1}.$$

Step 8. Let $k = k + 1$. Go to Step 2.

Conjecture. *Let $g_1 : \mathcal{H}_2 \to \mathbb{R}$ be a convex function with subgradient $\partial g_1 \in \mathcal{RH}_2$ and let $\gamma_1 > 0$ be given. Also let $g_2(H) = \|H\|_2$. Define*

$$P_{k,1} \ \doteq \mathbb{P}\{g_1(T_{CL}(z, \Delta, C_k)) > \gamma_1\}$$

with Δ having the distribution over $\mathcal{BH}_\infty(r_1)$ used in the algorithm. Similarly take

$$P_{k,2} \ \doteq \mathbb{P}\{g_2(T_{CL}(z, \Delta, C_k)) > \gamma_2\}$$

with Δ having the distribution over $\mathcal{BH}_\infty(r_2)$ used in the algorithm. Given this, define

$$P_k \ \doteq \ \frac{1}{2} P_{k,1} + \frac{1}{2} P_{k,2}.$$

Then, consistent numerical experience indicates that the algorithm described above generates a sequence of controllers C_k for which the risk of performance violation satisfies the equality

$$\lim_{k \to \infty} P_k \ = \ 0.$$

Hence, risk tends to zero as $k \to \infty$.

[6] Note that, since C_k is not guaranteed to be a robustly stabilizing controller, Q_k might not be stable.

13.6.3 A Simple Numerical Example

Consider the uncertain system

$$P(z, \Delta) = P_0(z) + \Delta(z),$$

with nominal plant

$$P_0(z) = \frac{0.006135z^2 + 0.01227z + 0.006135}{z^2 - 1.497z + 0.5706}$$

and stable causal dynamic uncertainty Δ. The objective is to find a controller $C(z)$ such that, for all $\|\Delta\|_\infty \leq r_1 = 1$,

$$\|W(z)(1 + C(z)P(z, \Delta))^{-1}\|_2 \leq \gamma_1 = 0.089$$

where

$$W(z) = \frac{0.0582z^2 + 0.06349z + 0.005291}{z^2 + 0.2381z - 0.6032}.$$

and the closed-loop system is stable for all $\|\Delta\|_\infty \leq r_2 = 2$. Since the plant $P(z, \Delta)$ is stable in spite of the uncertainty, according to the Youla parameterization, all stabilizing controllers are of the form

$$C = \frac{Q(z)}{1 - Q(z)P(z, \Delta)}$$

where $Q(z)$ is a stable rational transfer function. To solve this problem using the algorithm presented in the previous section, we take $\gamma_2 = 10^9$ (which is in practice equivalent to requiring robust stability for $\|\Delta\| \leq r_2$) and generate the random uncertainty samples using Algorithm 13.4 by taking $z_i = e^{j2\pi i/11}$, $i = 1, 2, \ldots, 10$.

We first consider a design using only the nominal plant. Using `Matlab`'s function `dh2lqg()`, we obtain the nominal $\mathcal{H}_2$ optimal controller

$$C_{nom}(s) = \frac{138.2z^3 - 93.78z^2 - 90.4z + 64.5}{z^4 + 2.238z^3 + 0.8729z^2 - 0.9682z - 0.6031}$$

and a nominal performance $\|T_{cl}(z)\|_2 = 0.0583$. However, this controller does not robustly stabilize the closed-loop plant for $\|\Delta\|_\infty \leq 2$. We next apply Algorithm 13.6 to design a risk-adjusted controller and, after 1500 iterations, we obtain

$$\begin{aligned} C_1(s) = {} & \frac{-0.003808z^{14}}{z^{14} - 0.1778z^{13} + 0.6376z^{12} + 0.09269z^{11}} \\ & \frac{-0.01977z^{13}}{+0.2469z^{10} + 0.06291z^9 + 0.08426z^8 + 0.0433z^7} \\ & \frac{-0.002939z^{12}}{+0.07403z^6 + 0.0004446z^5 - 0.1107z^4 - 0.07454z^3} \\ & \frac{+0.04627z^{11}}{-0.08156z^2 - 0.05994z + 0.01213}. \end{aligned}$$

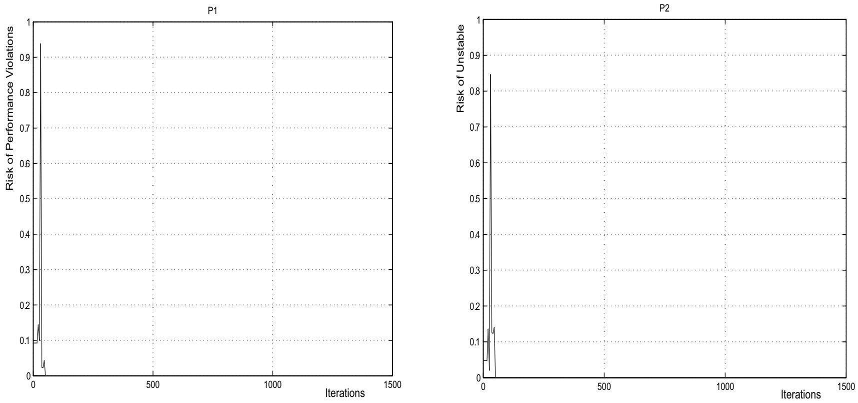
As in last section, define the probability of violating the performance specification

$$P_{k,1} \doteq \text{Prob}\{\|W(z) - W(z)P(z,\Delta)Q_k(z)\|_2 > \ \gamma_1 = 0.089\}; \quad \|\Delta\|_\infty \leq 1$$

and the approximation of probability of instability

$$P_{k,2} \doteq \text{Prob}\{\|W(z) - W(z)P(z,\Delta)Q_k(z)\|_2 > \ \gamma_2 = 10^9\}; \quad \|\Delta\|_\infty \leq 2.$$

Monte Carlo simulations were performed to estimate $P_{k,1}$ and $P_{k,2}$ for each controller $C_k(z)$ and the results are shown in Figures 13.4 (a) and (b) and Figures 13.5 (a) and (b). From these figures, one can see that both the probability of performance violation for $\|\Delta\|_\infty \leq 1$ and the probability of instability for $\|\Delta\|_\infty \leq 2$ quickly converge to zero, being negligible after iteration 200.

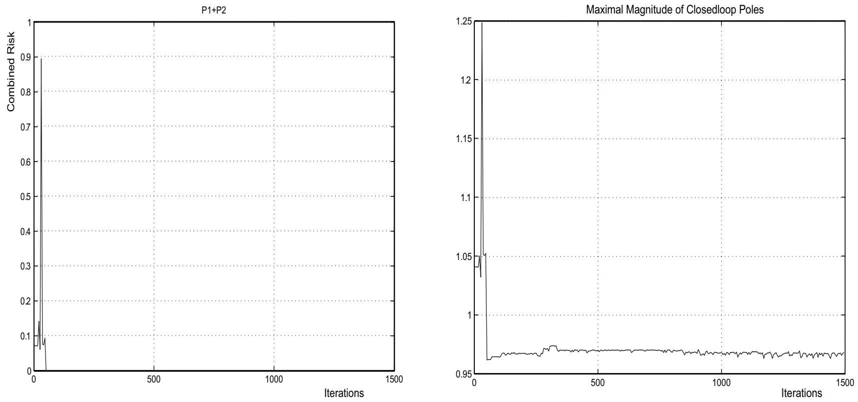


(a) (b)

Figure 13.4. Estimated (a) $P_{k,1}$ and (b) $P_{k,2}$ as a function of iteration k

13.7 Concluding Remarks and Directions for Further Research

In this chapter, we provide efficient algorithms for generation of random samples of causal, linear time-invariant uncertain transfer functions. First, results on matrix dilation are used to develop algorithms for generating random samples of the first n Markov parameters of transfer functions in $\mathcal{BH}_\infty$. Then, results on Nevanlinna-Pick and boundary Nevanlinna-Pick interpolation are exploited to develop two more algorithms. The first one generates samples inside the unit $\mathcal{H}_\infty$ ball and the second one generates random transfer function



(a) (b)

Figure 13.5. (a) Estimated P_k (b) Estimated maximum magnitude of closed-loop poles

on the boundary of the unit $\mathcal{H}_\infty$ ball. The usefulness of these tools is illustrated by developing algorithms for model invalidation and for solving some multi-disk problems arising in the context of synthesizing robust controllers for systems subject to structured dynamic uncertainty.

The results presented suggest several directions for further research. First, we believe that effort should be put in the development of efficient numerical implementations of the algorithms put forth in this contribution. Also, note that the algorithm for controller design proposed in this chapter only guarantees that one obtains a robustly stabilizing controller if one performs and infinite number of iterations (although our experiments have revealed that, in most cases, one quickly converges to a robustly stabilizing controller). Therefore, a possible direction for further research is the development of stochastic gradient algorithms for controller design which would guarantee that one would obtain a robustly stabilizing controller after a finite number of steps.

13.8 Appendix: Background Results

In this appendix we recall, for ease of reference, some results on matrix norm optimization, interpolation theory and complex analysis. These results are used only in the technical proofs and can therefore be skipped in a first reading.

13.8.1 Matrix Dilations

Theorem 3 (Parrott's Theorem). *([409], page 40). Let A, B and C be given matrices of compatible dimensions. Then*

$$\min_X \left\| \begin{bmatrix} X & B \\ C & A \end{bmatrix} \right\| \doteq \gamma_o = \max \left\{ \| [C \ A] \| , \left\| \begin{bmatrix} B \\ A \end{bmatrix} \right\| \right\},$$

where $\|\cdot\|$ stands for $\overline{\sigma}(\cdot)$. Moreover, all the solutions X to the above problem are parameterized by

$$X = -YA^TZ + \gamma_o(I - YY^T)^{\frac{1}{2}}W(I - Z^TZ)^{\frac{1}{2}}$$

where the free parameter W is an arbitrary contraction and the matrices Y and Z solve the linear equations

$$\begin{aligned} B &= Y(\gamma_o^2 I - A^TA)^{\frac{1}{2}} \\ C &= (\gamma_o^2 I - AA^T)^{\frac{1}{2}}Z \\ \overline{\sigma}(Y) &\leq 1, \overline{\sigma}(Z) \leq 1. \end{aligned}$$

Theorem 4 (Carathéodory-Fejér, [23]). *Given a matrix-valued sequence $\{\mathbf{L}_i\}_{i=0}^{n-1}$, there exists a causal, LTI operator $L(z) \in \mathcal{BH}_\infty$ such that*

$$L(z) = \mathbf{L}_0 + \mathbf{L}_1 z + \mathbf{L}_2 z^2 + \ldots + \mathbf{L}_{n-1} z^{n-1} + \ldots$$

if and only if $M_c \doteq I - T_L^n (T_L^n)^T \geq 0$ where

$$T_L^n = \begin{bmatrix} \mathbf{L}_0 & \mathbf{L}_1 & \cdots & \mathbf{L}_{n-1} \\ \mathbf{0} & \mathbf{L}_0 & \cdots & \mathbf{L}_{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{L}_0 \end{bmatrix}.$$

13.8.2 Complex Analysis

Let f_n denote a sequence of complex-valued functions, each of whose domain contains an open subset U of the complex plane. The sequence f_n *converges normally* in U to f if f_n is pointwise convergent to f in U and this convergence is uniform on each compact subset of U. A family $\mathcal{F}$ of functions analytic in U is said to be *normal* if each sequence f_n from $\mathcal{F}$ contains at least one normally convergent subsequence. Given a sequence of functions f_n, each of whose terms is analytic in an open set U, it is of interest to know whether f_n is normal, *i.e.*, if it is possible to extract a normally convergent subsequence. An answer to this question is given by Montel's theorem, which requires a certain equi-boundedness assumption. A family $\mathcal{F}$ is said to be *locally bounded in U* if its members are uniformly bounded on each compact set in U.

Theorem 5 (Montel's Theorem, [255]). *Let $\mathcal{F}$ be a family of functions that are analytic in an open set U. Suppose that $\mathcal{F}$ is locally bounded in U. Then $\mathcal{F}$ is a normal family in this set.*

In particular, if $\mathcal{F} \subset \mathcal{H}_\infty$ is such that $f \in \mathcal{F} \Rightarrow \|f\|_\infty \leq M$, then the theorem implies that $\mathcal{F}$ is normal inside the unit disk. Thus, every sequence $\{f_i\} \in \mathcal{F}$ contains a normally convergent subsequence.

13.8.3 Nevanlinna-Pick Interpolation

We start by focusing our attention on a more general result in interpolation theory. Let $\mathcal{T}$ and $\mathcal{BT}$ denote the space of complex valued rational functions continuous in $|\lambda| = 1$ and analytic in $|\lambda| < 1$, equipped with the $\|.\|_{\mathcal{L}_\infty}$ norm, and the (open) unit ball in this space, respectively (*i.e.* $f(\lambda) \in \mathcal{BT} \iff f(\frac{1}{z}) \in \mathcal{BH}_\infty$).

Theorem 6. *There exists a transfer function $f(\lambda) \in \mathcal{BT}$ ($\overline{\mathcal{BT}}$) such that*

$$\sum_{\lambda_o \in \mathcal{D}} \operatorname{Res}\left\{f(\lambda)C_-(\lambda I - A)^{-1}\right\}_{\lambda=\lambda_o} = C_+ \tag{13.3}$$

if and only if the following discrete time Lyapunov equation has a unique positive (semi) definite solution

$$M = A^*MA + C_-^*C_- - C_+^*C_+ \tag{13.4}$$

where A, C_- and C_+ are constant complex matrices of appropriate dimensions and $\mathcal{D}$ denotes the open unit circle. If $M > 0$ then the solution $f(\lambda)$ is non-unique and the set of solutions can be parameterized in terms of $q(\lambda)$, an arbitrary element of $\overline{\mathcal{BT}}$, as follows:

$$f(\lambda) = \frac{T_{11}(\lambda)q(\lambda) + T_{12}(\lambda)}{T_{21}(\lambda)q(\lambda) + T_{22}(\lambda)} \tag{13.5}$$

$$T(\lambda) = \begin{bmatrix} T_{11}(\lambda) & T_{12}(\lambda) \\ T_{21}(\lambda) & T_{22}(\lambda) \end{bmatrix} \tag{13.6}$$

where $T(\lambda)$ is the J-lossless[7] *matrix:*

$$\begin{aligned}
T(\lambda) &\equiv \left[\begin{array}{c|c} A_T & B_T \\ \hline C_T & D_T \end{array}\right] \\
A_T &= A \\
B_T &= M^{-1}\left(A^* - I\right)^{-1}\left[-C_+^* \; C_-^*\right] \\
C_T &= \begin{bmatrix} C_+ \\ C_- \end{bmatrix}(A - I) \\
D_T &= I + \begin{bmatrix} C_+ \\ C_- \end{bmatrix} M^{-1}\left(A^* - I\right)^{-1}\left[-C_+^* \; C_-^*\right]
\end{aligned}$$

Proof. See [23, 305].

Note that the matrices A and C_- provide the structure of the interpolation problem while C_+ provides the interpolation values. The following corollaries show that both the Nevanlinna-Pick and the Carathéodory-Fejér problems are special cases of this theorem, corresponding to an appropriate choice of the matrices A and C_-

[7] A transfer function $H(\lambda)$ is said to be J-lossless if $H^T(1/\lambda)JH(\lambda) = J$ when $|\lambda| = 1$, and $H^T(1/\lambda)JH(\lambda) < J$ when $|\lambda| < 1$. Here $J = \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix}$.

Corollary 1 (Nevanlinna-Pick). *Let* $\Gamma = \mathrm{diag}\{\lambda_i\} \in \mathbb{C}^{r\times r}$ *and take*

$$\begin{aligned} A &= \Gamma \\ C_- &= \begin{bmatrix} 1\ 1 \ldots 1 \end{bmatrix} \in \mathbb{R}^r \\ C_+ &= \begin{bmatrix} w_1\ w_2 \ldots w_r \end{bmatrix} \end{aligned}$$

then (13.3) is equivalent to

$$f(\lambda_i) = w_i \qquad i = 1, \ldots, r$$

and the solution to (13.4) is the standard Pick matrix:

$$P = \left[\frac{1 - \bar{w}_i w_j}{1 - \bar{\lambda}_i \lambda_j} \right]_{ij} .$$

Proof. Replace A, C_-, C_+ in (13.3). See [305] for details. □

13.8.4 Using these Results for Boundary Interpolation

In the case of boundary interpolation $|\lambda_i| = 1$, $|w_i| < 1$, these results can be used as follows:

1. Find a scalar $r < 1$ such that the equation

$$M = r^2 A^* M A + C_-^* C_- - C_+^* C_+$$

   has a positive definite solution $M > 0$.
2. Find the modified interpolant using the formulas (13.5)–(13.6) with $A = r\Gamma = r\mathrm{diag}\{\lambda_i\}$.
3. The desired interpolant is given by $G(\lambda) = G_r(r\lambda)$.

13.8.5 Boundary Nevanlinna-Pick Interpolation

We now elaborate on the results on boundary Nevanlinna-Pick interpolation used in this chapter. For a extensive treatment of this problem see [23]. Let $\mathcal{D}$ denote the unit circle in the complex plane with boundary $\partial\mathcal{D}$ and consider the following interpolation problem.

Problem 5. Given N distinct points λ_1, λ_2, …, λ_N in $\partial\mathcal{D}$, N complex numbers $w_1, w_2, \ldots, w_N$ of unit magnitude and N positive real numbers $\rho_1, \rho_2, \ldots, \rho_N$, find all rational functions $f(\lambda)$ mapping $\mathcal{D}$ into $\mathcal{D}$ such that

$$\begin{aligned} f(\lambda_i) &= w_i \\ f'(\lambda_i) &= \lambda_i^* w_i \rho_i \end{aligned}$$

for all $i = 1, 2, \ldots, N$.

The following theorem provides a solution for the problem above.

Theorem 7. *Let* $\lambda_1, \lambda_2, \ldots, \lambda_N$*,* $w_1, w_2, \ldots, w_N$ *and* $\rho_1, \rho_2, \ldots, \rho_N$ *be as in the statement of Problem 5 and define the matrix* $\Lambda = [\Lambda_{ij}]_{1 \le i,j \le N}$ *by*

$$\Lambda_{i,j} = \begin{cases} \frac{1-w_i^* w_j}{1-\lambda_i^* \lambda_j} & i \neq j \\ \rho_i & i = j \end{cases}.$$

Then a necessary condition for Problem 5 to have a solution is that Λ *be positive semidefinite and a sufficient condition is that* Λ *be positive definite. In the latter case, the set of all solution is given by*

$$f(\lambda) = \frac{\theta_{11}(\lambda)g(\lambda) + \theta_{12}(\lambda)}{\theta_{21}(\lambda)g(\lambda) + \theta_{22}(\lambda)}$$

where $g(\lambda)$ *is an arbitrary scalar rational function analytic on* $\mathcal{D}$ *with* $\sup\{|g(\lambda)| : z \in \mathcal{D}\} \le 1$ *such that* $\theta_{21}(\lambda)g(\lambda) + \theta_{22}(\lambda)$ *has a simple pole at the points* $\lambda_1, \lambda_2, \ldots, \lambda_N$*. Here*

$$\theta(\lambda) = \begin{bmatrix} \theta_{11}(\lambda) & \theta_{12}(\lambda) \\ \theta_{21}(\lambda) & \theta_{22}(\lambda) \end{bmatrix}$$

is given by

$$\theta(\lambda) = I + (\lambda - \lambda_0) C_0 (\lambda I - A_0)^{-1} \Lambda^{-1} (\lambda I - A_0^*)^{-1} C_0^* J$$

where

$$C_0 = \begin{bmatrix} w_1 & \ldots & w_N \\ 1 & \ldots & 1 \end{bmatrix}; \quad A_0 = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix}; \quad J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

and λ_0 *is a complex number of magnitude 1 and not equal to any of the numbers* $\lambda_1, \lambda_2, \ldots, \lambda_N$*.*

Proof. See [23].

Note that if only the values $w_1, w_2, \ldots, w_N$ of magnitude one are specified at the boundary points $\lambda_1, \lambda_2, \ldots, \lambda_N$, then the matrix Λ in the theorem above can always be made positive definite by choosing the unspecified quantities $\rho_1, \rho_2, \ldots, \rho_N$ sufficiently large. This leads to the following corollary.

Corollary 2. *Let 2N complex numbers of magnitude one* $\lambda_1, \lambda_2, \ldots, \lambda_N$ *and* $w_1, w_2, \ldots, w_N$ *be given, where* $\lambda_1, \lambda_2, \ldots, \lambda_N$ *are distinct. Then, there always exist scalar rational functions* $f(\lambda)$ *analytic in* $\mathcal{D}$ *with*

$$\sup\{|f(\lambda)| : \lambda \in \mathcal{D}\} \le 1$$

which satisfy the set of interpolation conditions

$$f(\lambda_i) = w_i; \quad i = 1, 2, \ldots, N.$$

13.9 Proof of Theorem 1

For the sake of notational simplicity we will prove the result for the case where the number of partitions of the vector x is $m = 4$, but the same reasoning applies to arbitrary dimensions.

Consider a rectangle

$$R \doteq R_1 \times R_2 \times R_3 \times R_4 \subseteq \mathcal{C}$$

where $R_i \subset \mathbb{R}^{k_i}$, $i = 1, 2, 3, 4$. Let N_t and N_R be the total number of samples generated and the number of hits of R respectively. We will show that

$$\frac{N_R}{N_t} \xrightarrow{w.p.1} \frac{\mathrm{vol}(R_1)\mathrm{vol}(R_2)\mathrm{vol}(R_3)\mathrm{vol}(R_4)}{\mathrm{vol}(\mathcal{C})}.$$

In other words, the ratio converges with probability one to a constant which is equal to the probability of the rectangle R under a uniform distribution over the set $\mathcal{C}$. Henceforth, the symbol $\rightarrow$ denotes convergence with probability one.

Let X_1^k be the k-th sample of the first component of the vector. Similarly, denote by X_2^{mk} and X_3^{nmk} the m-th sample of the second component of the vector when the first component is X_1^k, and the n-th sample of the third component of the vector when the first two components are X_1^k and X_2^{mk}, respectively. Finally, denote by $v_k \doteq \mathrm{vol}\,[I_k(\mathbf{X_1}, \ldots, X_{k-1})]$. Consider now

$$\frac{N_t}{N^4} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{N}\sum_{m=1}^{\lfloor \alpha_2 N v_2(X_1^k)\rfloor}\frac{1}{N}\sum_{n=1}^{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}\frac{1}{N}\lfloor N\alpha_4 v_4(X_1^k, X_2^{mk}, X_3^{nmk})\rfloor$$

$$= \frac{1}{N}\sum_{k=1}^{N}\alpha_2 v_2(X_1^k)\frac{\lfloor N\alpha_2 v_2(X_1^k)\rfloor}{N\alpha_2 v_2(X_1^k)}$$

$$\frac{1}{\lfloor N\alpha_2 v_2(X_1^k)\rfloor}\sum_{m=1}^{\lfloor N\alpha_2 v_2(X_1^k)\rfloor}\alpha_3 v_3(X_1^k, X_2^{mk})\frac{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}{N\alpha_3 v_3(X_1^k, X_2^{mk})}$$

$$\frac{1}{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}\sum_{n=1}^{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}\frac{1}{N}\lfloor N\alpha_4 v_4(X_1^k, X_2^{mk}, X_3^{nmk})\rfloor$$

The Strong Law of Large Numbers (see, *e.g.*, [147]) indicates that, as $N \rightarrow \infty$,

$$\alpha_3 v_3(X_1^k, X_2^{mk}) \tfrac{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}{N\alpha_3 v_3(X_1^k, X_2^{mk})}$$

$$\tfrac{1}{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor} \textstyle\sum_{n=1}^{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor} \tfrac{1}{N} \lfloor N\alpha_4 v_4(X_1^k, X_2^{mk}, X_3^{nmk})\rfloor$$

$$\to \alpha_3 v_3(X_1^k, X_2^{mk}) \mathbb{E}[\alpha_4 v_4(X_1^k, X_2^{mk}, X_3) | X_1^k, X_2^{mk}]$$

$$= \alpha_3 \alpha_4 \mathrm{vol}[\mathcal{S}_2(X_1^k, X_2^{mk})].$$

Furthermore,

$$\alpha_2 v_2(X_1^k) \frac{\lfloor N\alpha_2 v_2(X_1^k)\rfloor}{N\alpha_2 v_2(X_1^k)} \frac{1}{\lfloor N\alpha_2 v_2(X_1^k)\rfloor} \sum_{m=1}^{\lfloor N\alpha_2 v_2(X_1^k)\rfloor} \alpha_3 \alpha_4 \mathrm{vol}[\mathcal{S}_2(X_1^k, X_2^{mk})]$$

$$\to \alpha_2 v_2(X_1^k) \mathbb{E}\left[\alpha_3 \alpha_4 \mathrm{vol}[\mathcal{S}_2(X_1^k, X_2)] | X_1^k\right] = \alpha_2 \alpha_3 \alpha_4 \mathrm{vol}[\mathcal{S}_3(X_1^k)]$$

and

$$\frac{1}{N} \sum_{k=1}^{N} \alpha_2 \alpha_3 \alpha_4 \mathrm{vol}[\mathcal{S}_3(X_1^k)] \to \mathbb{E}\left[\alpha_2 \alpha_3 \alpha_4 \mathrm{vol}[\mathcal{S}(X_1)]\right] = \frac{\alpha_2 \alpha_3 \alpha_4}{v_1} \mathrm{vol}(\mathcal{C}).$$

Hence

$$\frac{N_t}{N^4} \to \frac{\alpha_2 \alpha_3 \alpha_4}{v_1} \mathrm{vol}(\mathcal{C})$$

as $N \to \infty$.

Next, consider the number of hits of the rectangle R, which we denote by N_R. The Strong Law of Large Numbers implies that

$$\left.\frac{N_R}{N}\right| X_1^k \in R_1, X_2^{mk} \in R_2, X_3^{nmk} \in R_3$$

$$\to \alpha_4 v_4(X_1^k, X_2^{mk}, X_3^{nmk}) \frac{\mathrm{vol}(R_4)}{v_4(X_1^k, X_2^{mk}, X_3^{nmk})} = \alpha_4 \mathrm{vol}(R_4)$$

which is independent of the values of X_1^k, X_2^{mk} and X_3^{nmk}. Using the same reasoning, we have

$$\left.\frac{N_R}{N^2}\right| X_1^k \in R_1, X_2^{mk} \in R_2 \to \alpha_3 v_3(X_1^k, X_2^{mk}) \frac{\mathrm{vol}(R_3) \alpha_4 \mathrm{vol}(R_4)}{v_3(X_1^k, X_2^{mk})}$$

$$= \alpha_3 \mathrm{vol}(R_3) \alpha_4 \mathrm{vol}(R_4)$$

$$\left.\frac{N_R}{N^3}\right| X_1^k \in R_1 \to \alpha_2 v_2(X_1^k) \frac{\mathrm{vol}(R_2) \alpha_3 \mathrm{vol}(R_3) \alpha_4 \mathrm{vol}(R_4)}{v_2(X_1^k)}$$

$$= \alpha_2 \mathrm{vol}(R_2) \alpha_3 \mathrm{vol}(R_3) \alpha_4 \mathrm{vol}(R_4).$$

Finally, this implies that

$$\frac{N_R}{N^4} \rightarrow \frac{1}{v_1}\alpha_2\alpha_3\alpha_4 \mathrm{vol}(R_1)\mathrm{vol}(R_2)\mathrm{vol}(R_3)\mathrm{vol}(R_4).$$

Hence,

$$\frac{N_R}{N_t} \rightarrow \frac{\mathrm{vol}(R_1)\mathrm{vol}(R_2)\mathrm{vol}(R_3)\mathrm{vol}(R_4)}{\mathrm{vol}(\mathcal{C})}$$

as $N \rightarrow \infty$. This completes the proof. □

13.10 Proof of Theorem 2

As in the proof of Theorem 1, only $m = 4$ is considered and it is assumed that

$$\mathcal{A} \doteq R_1 \times R_2 \times R_3 \times R_4 \subseteq \mathcal{C}$$

where $R_i \subset \mathbb{R}^{k_i}$, $i = 1, 2, 3, 4$, satisfy

$$\mathrm{vol}(R_i) = dx_i.$$

The proof can be easily generalized for other values of m and other sets $\mathcal{A}$.

Using the notation in the proof of Theorem 1, we first consider $\mathbb{E}[N_t]$. The reasoning to follow relies on the fact that, given two random variables, X and Y, $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$. Indeed,

$$\mathbb{E}\left[\frac{N_t}{N^4}\right] = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\frac{1}{N}\sum_{m=1}^{\lfloor \alpha_2 N v_2(X_1^k)\rfloor} \right.$$
$$\left. \frac{1}{N}\sum_{n=1}^{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}\frac{1}{N}\lfloor N\alpha_4 v_4(X_1^k, X_2^{mk}, X_3^{nmk})\rfloor\right].$$

Moreover,

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}\frac{1}{N}\lfloor N\alpha_4 v_4(X_1^k, X_2^{mk}, X_3^{nmk})\rfloor\,\middle|\, X_1^k, X_2^{mk}\right]$$
$$= \frac{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}{N}\left(\alpha_4\frac{\mathrm{vol}(S_2(X_1^k, X_2^{mk}))}{v_3(X_1^k, X_2^{mk})} - \frac{\varepsilon_4}{N}\right)$$

where $\varepsilon_4 \in [0, 1]$. Also,

$$\mathbb{E}\left[\frac{1}{N}\sum_{m=1}^{\lfloor \alpha_2 N v_2(X_1^k)\rfloor}\frac{\lfloor N\alpha_3 v_3(X_1^k, X_2^{mk})\rfloor}{N}\alpha_4\frac{\mathrm{vol}(S_2(X_1^k, X_2^{mk}))}{v_3(X_1^k, X_2^{mk})}\,\middle|\, X_1^k\right]$$
$$= \frac{\lfloor \alpha_2 N v_2(X_1^k)\rfloor}{N}\left(\alpha_3\alpha_4\frac{\mathrm{vol}(S_1(X_1^k))}{v_2(X_1^k)} - \frac{\varepsilon_3}{N}\mathbb{E}\left[\alpha_4\frac{\mathrm{vol}(S_2(X_1^k, X_2^{mk}))}{v_3(X_1^k, X_2^{mk})}\,\middle|\, X_1^k\right]\right)$$

where $\varepsilon_3 \in [0,1]$. Finally,

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\frac{\lfloor \alpha_2 N v_2(X_1^k)\rfloor}{N}\alpha_3\alpha_4\frac{\mathrm{vol}(S_1(X_1^k))}{v_2(X_1^k)}\right]$$

$$= \frac{\alpha_2\alpha_3\alpha_4}{v1}\mathrm{vol}(\mathcal{C}) - \mathbb{E}\left[\alpha_3\alpha_4\frac{\varepsilon_2}{N}\frac{\mathrm{vol}(S_1(X_1^k))}{v_2(X_1^k)}\right]$$

where $\varepsilon_2 \in [0,1]$. Hence,

$$\mathbb{E}\left[\frac{N_t}{N^4}\right] = \frac{\alpha_2\alpha_3\alpha_4}{v1}\mathrm{vol}(\mathcal{C}) - \frac{\beta_1+\beta_2+\beta_3}{N}$$

where

$$\beta_1 = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\frac{1}{N}\sum_{m=1}^{\lfloor \alpha_2 N v_2(X_1^k)\rfloor}\frac{1}{N}\sum_{n=1}^{\lfloor N\alpha_3 v_3(X_1^k,X_2^{mk})\rfloor}\varepsilon_4\right];$$

$$\beta_2 = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\frac{1}{N}\sum_{m=1}^{\lfloor \alpha_2 N v_2(X_1^k)\rfloor}\mathbb{E}\left[\varepsilon_3\alpha_4\frac{\mathrm{vol}(S_2(X_1^k,X_2^{mk}))}{v_3(X_1^k,X_2^{mk})}\middle| X_1^k\right]\right];$$

$$\beta_3 = \mathbb{E}\left[\varepsilon_2\alpha_3\alpha_4\frac{\mathrm{vol}(S_1(X_1^k))}{v_2(X_1^k)}\right];$$

$$\varepsilon_2 \in [0,1]; \quad \varepsilon_3 \in [0,1]; \quad \varepsilon_4 \in [0,1].$$

Since β_1, β_2 and β_3 above are bounded functions of N, there exists a constant β such that

$$-\frac{\beta}{N} \le \mathbb{E}\left[\frac{N_t}{N^4}\right] - \frac{\alpha_2\alpha_3\alpha_4}{v_1}\mathrm{vol}(\mathcal{C}) \le 0.$$

Next consider

$$\mathbb{E}\left[\frac{N_{\mathcal{A}}}{N^4}\right] = \mathbb{E}\left[\frac{1}{N^4}\sum_{k=1}^{N}\sum_{m=1}^{\lfloor \alpha_2 N v_2(X_1^k)\rfloor}\sum_{n=1}^{\lfloor N\alpha_3 v_3(X_1^k,X_2^{mk})\rfloor}\right.$$

$$\left.\sum_{l=1}^{\lfloor N\alpha_4 v_4(X_1^k,X_2^{mk},X_3^{nmk})\rfloor} I_{X_1^k\in R_1,X_2^{mk}\in R_2,X_3^{nmk}\in R_3,X_4^{nmkl}\in R_4}\right]$$

where I denotes the indicator function; *i.e.*,

$$I_{X_1^k\in R_1,X_2^{mk}\in R_2,X_3^{nmk}\in R_3,X_4^{nmkl}\in R_4}$$

$$= \begin{cases} 1 & \text{if} X_1^k \in R_1, X_2^{mk} \in R_2, X_3^{nmk} \in R_3, X_4^{nmkl} \in R_4; \\ 0 & \text{otherwise.} \end{cases}$$

Indeed, by using conditional expectations again, it follows that

$$\mathbb{E}\left[\frac{1}{N}\sum_{l=1}^{\lfloor N\alpha_4 v_4(X_1^k,X_2^{mk},X_3^{nmk})\rfloor} I_{X_1^k\in R_1,X_2^{mk}\in R_2,X_3^{nmk}\in R_3,X_4^{nmkl}\in R_4}\right| \\ \left. X_1^k\in R_1, X_2^{mk}\in R_2, X_3^{nmk}\in R_3\right]$$

$$= \frac{1}{N}\lfloor N\alpha_4 v_4(X_1^k,X_2^{mk},X_3^{nmk})\rfloor \frac{dx_4}{v_4(X_1^k,X_2^{mk},X_3^{nmk})}$$

$$= \alpha_4 dx_4 - \frac{1}{N}\tilde{\varepsilon}_4 \frac{dx_4}{v_4(X_1^k,X_2^{mk},X_3^{nmk})}$$

where $\tilde{\varepsilon}_4 \in [0,1]$. Repeating this reasoning for the other three coordinates leads to the following result:

$$\mathbb{E}\left[\frac{N_{\mathcal{A}}}{N^4}\right] = dx_1dx_2dx_3dx_4\frac{\alpha_2\alpha_3\alpha_4}{v_1}\mathrm{vol}(\mathcal{C}) - \frac{\gamma_1+\gamma_2+\gamma_3}{N}$$

where

$$\gamma_3 = \mathbb{E}\left[\frac{1}{N^3}\sum_{k=1}^{N}\sum_{m=1}^{\lfloor\alpha_2 N v_2(X_1^k)\rfloor}\sum_{n=1}^{\lfloor N\alpha_3 v_3(X_1^k,X_2^{mk})\rfloor}\tilde{\varepsilon}_4\frac{dx_4}{v_4(X_1^k,X_2^{mk},X_3^{nmk})}\right]$$

$$\gamma_2 = \mathbb{E}\left[\frac{1}{N^2}\sum_{k=1}^{N}\sum_{m=1}^{\lfloor\alpha_2 N v_2(X_1^k)\rfloor}\tilde{\varepsilon}_3\frac{dx_3dx_4}{v_3(X_1^k,X_2^{mk})}\right]$$

$$\gamma_1 = \mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N}\tilde{\varepsilon}_2\frac{dx_2dx_3dx_4}{v_2(X_1^k)}\right]$$

$$\tilde{\varepsilon}_2\in[0,1];\quad \tilde{\varepsilon}_3\in[0,1];\quad \tilde{\varepsilon}_4\in[0,1].$$

Since γ_1, γ_2 and γ_3 above are bounded function of N, there exists a constant γ such that

$$-\frac{\gamma}{N} \le \mathbb{E}\left[\frac{N_{\mathcal{A}}}{N^4}\right] - dx_1dx_2dx_3dx_4\frac{\alpha_2\alpha_3\alpha_4}{v_1}\mathrm{vol}(\mathcal{C}) \le 0.$$

The proof is completed by noting that given the results above, one can determine constants k_1, k_2 and k_3 such that

$$\left|\frac{\mathbb{E}[N_{\mathcal{A}}]}{\mathbb{E}[N_t]} - \frac{\mathrm{vol}(\mathcal{A})}{\mathrm{vol}(\mathcal{C})}\right| = \left|\frac{\mathbb{E}[N_{\mathcal{A}}/N^4]}{\mathbb{E}[N_t/N^4]} - \frac{\mathrm{vol}(\mathcal{A})}{\mathrm{vol}(\mathcal{C})}\right|$$

$$\le \frac{1}{N}\frac{k_1}{k_2+\frac{k_3}{N}}.$$

□

Acknowledgement. Funding for this research was provided by the National Science Foundation under grants ANI-0125653, ECS-9984260, ECS-0115946, IIS-0117387, ECS-0221562 and ITR-0312558.

14

Nonlinear Systems Stability via Random and Quasi-Random Methods

Peter F. Hokayem[1], Silvia Mastellone[1], and Chaouki T. Abdallah[2]

[1] Coordinated Science Laboratory
University of Illinois
1308 W. Main St.
Urbana, IL 61801, USA
{hal,smastel2}@uiuc.edu

[2] Electrical & Computer Engineering Department
MSC01 1100, 1 University of New Mexico
Albuquerque, NM 87131-0001, USA
chaouki@ece.unm.edu

Summary. This chapter addresses the utility of sampling techniques in studying stability of nonlinear systems, and in certain instances expanding the stability region obtained from Lyapunov-like analysis. To this end, we provide an overview of random and quasi-random methods, error bounds and various transformations of general nonlinear systems into their polynomial-like counterparts for which preliminary Lyapunov analysis is feasible.

14.1 Introduction

Many practical control problems are so complex that they defy traditional analysis and design methods. Consequently, sampling methods that provide approximate solutions to such 'difficult' problems have emerged in recent years. Sampling techniques in general fall into three categories: gridding, random [135, 175], and quasi-random [244]. Due to the 'curse of dimensionality' [33], gridding techniques produce a number of samples that grows exponentially with the dimension of the problem. Hence, we revert to random and quasi-random techniques that require a fixed number of samples irrespective of the dimension, to produce approximate answers. In a recent paper by the authors [163], both methods were compared with respect to their convergence ability when the number of samples increases. The quality of the resulting answer in both cases was studied by the authors in the random sense [189, 190] and in the deterministic sense [164].

In this chapter, we review random and quasi-random sampling techniques, and adapt them to the problem of stability analysis for nonlinear systems. We reformulate the nonlinear stability problem through various transformations (approximation, S-procedure, generalized Sum-of-Squares) and project it into the smaller subclass of nonlinear polynomial systems. The approximation and the S-procedure techniques provide a mapping from the original nonlinear system into a polynomial domain and indirectly provides stability guarantees of the original system through the study of the polynomial system and the corresponding mapping. The generalized Sum-of-Squares technique studies the nonlinear system directly by partitioning the system's state-space equations into polynomial and non-polynomial parts. Throughout our proposed work, the main tool for studying the stability of the nonlinear system, or its counter part in the polynomial-like subclass, will be Lyapunov functions and the negativity of their derivatives along the trajectories of the corresponding state equations. All such techniques provide sufficient stability conditions.

Some of the issues we will address: an overview of random and quasi-random methods, error bounds, transformation of general nonlinear systems into their polynomial-like counterparts, stability analysis via Lyapunov methods either of the original nonlinear system or its polynomial-like counterpart, performance verification of the different methods via simulations. Moreover, since Lyapunov techniques are conservative in most cases, we propose to extend the stability region beyond that obtained analytically by sampling outside the guaranteed stability regions within a set ($[-\epsilon, +\epsilon]^n$), where $\epsilon > 0 \in \mathbb{R}$ and n is the dimension of the problem, and increasing ϵ until we hit the first instability point. This method, although not guaranteed to provide an error bound on the quality of our answer under deterministic sampling, is still useful from the practical point of view when analytical results fail to exist for the extended region.

The chapter starts by reviewing the concepts of random and quasi-random sampling methods in Sections 14.2 and 14.3, respectively. In Section 14.4 we present some transformations that allow us to study the stability of non-polynomial nonlinear systems. Section 14.6 provides a brief example that illustrates some of the concepts discussed, and Section 14.7 concludes the chapter. It is important to note that the authors became aware of similar work on Sum-of-Squares transformations in [256, 257].

14.2 Monte Carlo Method

The Monte Carlo method was first published in 1949 by Metropolis and Ulam at Los Alamos National laboratory. Since then it has been used extensively in various areas of science such as statistical and stochastic physical systems [175], derivative estimation [228], and integral evaluation [135].

Loosely defined, *Monte Carlo is a numerical method based upon random sampling of the parameters space.* Given a function $g(x)$, it is required to find $\int_{\mathbb{I}^d} g(x)dx$ ($\mathbb{I}^d$ - the d-dimensional unit hypercube). Usually the dimension 'd' is high, and numerical solutions are computationally expensive. That is when Monte Carlo method comes into the picture, because it overcomes the curse of dimensionality. The first step is to equip the integration region with a d-dimensional probability density Π, usually uniform if no prior knowledge is available. The second step is to integrate with respect to the probabilistic distribution as follows:

$$\phi = \int_{\mathbb{I}^d} g(x)dx = \lambda_d(\mathbb{I}^d) \int_{\mathbb{I}^d} g(\eta)d\eta = \mathbb{E}\{g(\eta)\} \tag{14.1}$$

where λ_d is an d-dimensional Lebesgue measure and $\mathbb{I}^d$ is transformed into a probability space equipped with a probability density $d\eta = \frac{dx}{\lambda_d(\mathbb{I}^d)}$ [135, 244]. Finally, the integral is replaced by a summation over samples drawn according to a specific strategy. As a result, the problem of evaluating the integral has been simply transformed into evaluating the 'empirical' expected value on the probability space, which provides an approximate answer. For an extensive overview on Monte Carlo methods in robust control problems see [95, 189, 190, 359, 360].

The dimension d may be extremely large in some applications, but the probabilistic results obtained using Monte Carlo methods are dimension-independent. Finally, the convergence error in (14.1) between empirical and actual expected values is of order $\mathcal{O}(N^{-1/2})$, where N is the number of samples. The constant by which the order is multiplied is a function of the variance of the samples. That is why different Monte Carlo methods are usually targeted at decreasing the variance of the samples (see [135]). Figure 14.1 illustrates uniform random sampling in the two-dimensional unit plane. It can be easily spotted that there are several clusters in the sample set, and substantial gaps as a result.

14.3 Quasi-Monte Carlo Methods

In this section we review the basic definitions involved in quasi-Monte Carlo (QMC) methods and state the basic inequalities governing the quality of the approximation of integrals using deterministic sampling methods. The main idea in QMC methods is to evaluate an integrand at specific points and approximate the integral by the average of the results obtained at these specific points. While this is exactly the same approach adopted in Monte Carlo sampling, the critical difference between the two approaches resides in the actual choice of sample points, the first being random and the second deterministic.

14.3.1 Discrepancy

The *discrepancy* is a measure of the 'regularity in distribution' of a set of points in the sample space [244]. In order to define it mathematically, we need to define the following counting function:

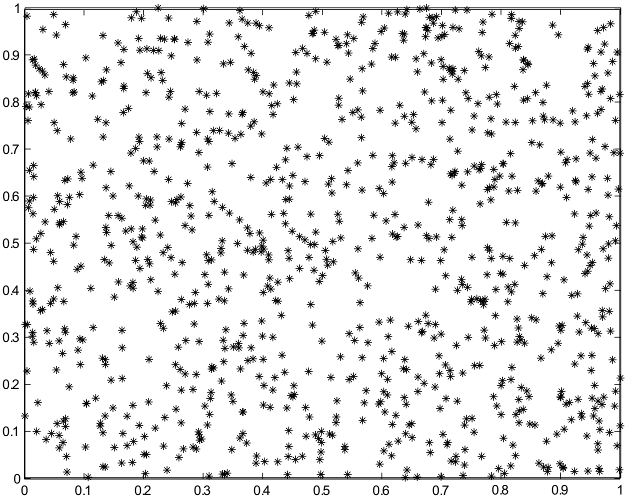


Figure 14.1. Uniform random sampling in 2D for 1000 points

$$A(B,P) = \sum_{i=1}^{N} I_B(X_i)$$

where $B \subset \mathbb{I}^d$ is an arbitrary set, $P = (X_1, \ldots, X_N)$ is a set of points, N is the number of sample points, and I_B is an indicator function. Thus $A(B,P)$ measures the number of points taken from P that happen to land inside the set B.

Definition 1. *The general formula for the evaluation of the discrepancy is given by*

$$\mathcal{D}_N(\mathcal{B}, P) = \sup_{B \in \mathcal{B}} \left| \frac{A(B,P)}{N} - \lambda_d(B) \right| \tag{14.2}$$

where $\lambda_d(B)$ is the d-dimensional Lebesgue measure of the arbitrary set B ($\subset \mathbb{I}^d$) and $\mathcal{B}$ is the family of all lebesgue measurable subsets B of $\mathbb{I}^d$.

Definition 1 can be specialized into the following two cases:

- The *star discrepancy* $D_N^\star(X_1, \ldots, X_N)$ is obtained by restricting $\mathcal{B}$ in (14.2) to be defined as follows:

$$\mathcal{B}^\star = \left\{ B : B = \prod_{i=1}^{d} [0, u_i) \right\}$$

  i.e. the set of all d-dimensional subsets of $\mathbb{I}^d$ that have a vertex at the origin, and u_i's being arbitrary points in the corresponding one-dimensional space.

- The *extreme discrepancy* $D_N(X_1, \ldots, X_N)$ is obtained by letting $\mathcal{B}$ in (14.2) be defined as follows $\mathcal{B} = \left\{B : B = \prod_{i=1}^{d}[v_i, u_i)\right\}$, where v_i's and u_i's are both arbitrary points in the corresponding one-dimensional space.

The star discrepancy and extreme discrepancy are related through the following inequality $D_N^\star(P) \leq D_N(P) \leq 2^d D_N^\star(P)$.

14.3.2 Point Sets Generation

In this section we briefly describe how to generate quasi-Monte Carlo low discrepancy points in an d-dimensional sample space. Since the points result from a deterministic method of generation, they possess a certain regularity property of distribution in the sample space described by their discrepancy.

For brevity, we are not going to present the various methods used in the generation of the sample points. Instead, we refer the reader to [163] for a compact presentation and [244] for a more involved one.

Van der Corput

The van der Corput sequence in base $b \in \mathbb{N}$, where $b \geq 2$, is a one-dimensional sequence of points that possesses the property of having a low discrepancy in the unit interval $\mathbb{I} = [0, 1] \subset \mathbb{R}$. *The main idea is to express every integer* $n \in \mathbb{N}$ *in base* b *and then reflect the expansion into the unit interval* $\mathbb{I}$. This is done as follows:

1. Let $R_b = \{0, 1, \ldots, b-1\}$ be the remainder set modulo b.
2. Any integer $n \geq 0$ can be expanded in base b as, $n = \sum_{k=0}^{\infty} a_k(n) b^k$, where $a_k(n) \in R_b, \forall k$.
3. Finally, we get the sequence $\{X_n\}$ through $X_n = \phi_b(n) = \sum_{k=0}^{\infty} a_k(n) b^{-k-1}$.

As will be seen, the van der Corput sequence will be used to generate higher dimensional vector samples, by varying the expansion base b. Finally, the star discrepancy of the van der Corput sequence is calculated to be: $D_N^\star(X_1, \ldots, X_N) = \mathcal{O}(N^{-1}\log(N))$, with the order constant depending on the base of expansion.

Halton sequence

The Halton sequence is a generalization of the van der Corput sequence to span a d-dimensional sample space. The main idea is to generate d one-dimensional sequences and form the corresponding d-dimensional vector sample points. Let $b_1, b_2, \ldots, b_d$ be the corresponding expansion bases for each dimension, preferably relatively prime.[3] Let $\phi_{b_1}, \phi_{b_2}, \ldots, \phi_{b_d}$ be the corresponding reflected

[3] Choosing the expansion bases relatively prime reduces the discrepancy, hence the error bound.

expansions according to the corresponding bases. Then the d-dimensional sequences $\{X_n^d\}$ are formed as follows:

$$X_n = (\phi_{b_1}, \phi_{b_2}, \dots, \phi_{b_d}) \in \mathbb{I}^d \tag{14.3}$$

Assume that the bases for the expansion are relatively prime, then the star discrepancy is bounded by (see [244])

$$D_N^\star(X_1, \dots, X_N) < \frac{d}{N} + \frac{1}{N} \prod_{i=1}^{d} \left(\frac{b_i - 1}{2 \log b_i} \log N + \frac{b_i + 1}{2} \right).$$

14.3.3 Total Variation

The problem of bounding the error involved in evaluating the integral of a function using QMC methods depends on our ability to obtain the value of total variation of the function under consideration, as will be seen in the next section. Consequently, in this section we will concentrate on defining several notions of variation of a function defined on an interval $[0,1]^d$.

Definition 2. *[65] A finite function $f(x)$ defined on interval $[0,1]$ is said to have* 'bounded variation' *if there exists a number M, such that for any partition p of the interval $[0,1]$*

$$v_p = \sum_{i=1}^{n} |f(X_i) - f(X_{i-1})| < M.$$

Moreover, the 'total variation' *of $f(x)$ on $[0,1]$ is defined as $V(f) = \sup_{p \in \mathcal{P}} (v_p)$, where $\mathcal{P}$ is the set of all partitions on $[0,1]$.*

Notice that Definition 2 pertains to functions of a single variable and does not require that the function be continuous. However, the function has to have a countable number of discontinuities on the interval under study. If it is further assumed that the function $f(x)$ is differentiable on $[0,1]$, then the total variation is found to be

$$V(f) = \int_0^1 \left| \frac{df}{dx} \right| dx \tag{14.4}$$

Remark 1. The total variation of a function can be understood as the sum of all the heights of monotone segments. That is why we integrate over the absolute value of the gradient in (14.4).

The total variation of a function f defined on a one-dimensional unit interval $\mathbb{I} = [0,1]$ is fairly easy to calculate. However, if f is defined on $\mathbb{I}^d$ the problem of calculating $V^{(d)}(f)$ (the d-dimensional total variation) is more involved (see [167, 244]). In what follows we only present the definitions of the total variation for continuous and differentiable functions.

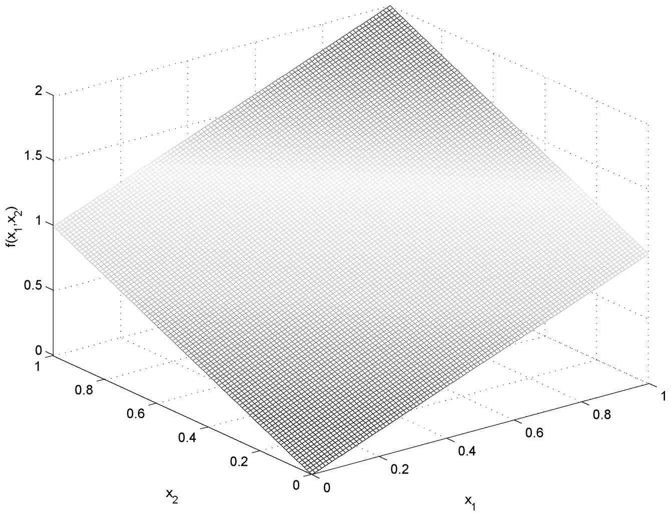


Figure 14.2. Plot of $f(x_1, x_2) = x_1 + x_2$

Definition 3. *The total variation of a function f defined on $\mathbb{I}^d$ in the sense of Vitali is defined as*

$$V^{(d)} = \int_0^1 \cdots \int_0^1 \left| \frac{\partial^{(d)} f}{\partial \eta_1 \partial \eta_2 \dots \partial \eta_d} \right| d\eta_1 d\eta_2 \dots d\eta_d \tag{14.5}$$

whenever the indicated partial derivative is continuous on $\mathbb{I}^d$. If $V^{(d)} < +\infty$, then the function f is said to have a 'bounded total variation in the sense of Vitali'.

Note that the Definition 3 only measures the variation of f over all the variables at once. However, the indicated partial derivative in (14.5) might be zero, but still the variation over the domain is not equal to zero as illustrated in the following example.

Example 1. Let $f(x_1, x_2) = x_1 + x_2 \Rightarrow \frac{\partial^{(2)} f}{\partial x_1 \partial x_2} = 0$ and the total variation as defined in *(14.5)* is equal to zero. However, when we plot the function $f(x_1, x_2)$, it is varying over the interval $[0, 1]^2$ as seen in Figure 14.2.

The problem encountered in the Definition 3 can be remedied via the following enhanced definition of the total variation.

Definition 4. *[194, 244] Let f be a function defined on $\mathbb{I}^d$ with bounded variation in the sense of Vitali. Suppose that the restriction of f to each face F of $\mathbb{I}^d$ of dimension $k = 1, 2, \dots, d-1$ is also of bounded variation on F in the sense of Vitali. Then the function f is said to be of 'bounded variation in the sense of Hardy and Krause'.*

Remark 2. The restriction of the function f to the face F in Definition 4 is achieved through setting the $d-k$ variables equal to 1.

Definition 4 overcomes the difficulties we encountered with Definition 3 as seen in the following example.

Example 2. Let us revisit the same function in Example 1. Using Definition 4 we get the following formula for the total variation of this second order function

$$\begin{aligned} V^{(2)}(f) = & \int_0^1 \int_0^1 \left| \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \right| dx_1 dx_2 \\ & + \int_0^1 \left| \frac{\partial f(x_1,1)}{\partial x_1} \right| dx_1 + \int_0^1 \left| \frac{\partial f(1,x_2)}{\partial x_2} \right| dx_2. \end{aligned} \tag{14.6}$$

Substituting and performing the necessary partial differentiation and integration we get $V^{(2)}(f) = 2$.

The second order total variation has been used in [353,354], and the following intuitive bound on the variation on (14.6) was suggested in [353]:

$$\begin{aligned} V^{(2)}(f) \leq & \max_{x_1,x_2} \left| \frac{\partial^2 f(x_1,x_2)}{\partial x_1 \partial x_2} \right| \\ & + \max_{x_1} \left| \frac{\partial f(x_1,1)}{\partial x_1} \right| + \max_{x_2} \left| \frac{\partial f(1,x_2)}{\partial x_2} \right|. \end{aligned}$$

14.3.4 Error in Quasi-Monte Carlo

The error in quasi-Monte Carlo methods integration over the unit hypercube for N samples is defined as follows:

$$e = \int_{\mathbb{I}^d} f(\eta) d\eta - \frac{1}{N} \sum_{n=1}^{N} f(X_n). \tag{14.7}$$

The following two theorems provide bounds on the error (14.7), for the cases of one-dimensional and d-dimensional integration, respectively.

Theorem 1. *Koksma's Inequality [244]*
Let $f(\cdot)$ be a function defined on $\mathbb{I} = [0,1]$ of bounded total variation $V(f)$. Then

$$\left| \int_{\mathbb{I}^d} f(\eta) d\eta - \frac{1}{N} \sum_{i=1}^{N} f(X_n) \right| \leq V(f) D_N^{\star}(X_1, \ldots, X_N)$$

Theorem 2. *Koksma-Hlawka Inequality [244]*
Let $f(\cdot)$ be a function defined on $\mathbb{I}^d$ of bounded variation in the sense of Hardy and Krause. Then

$$\left| \int_{\mathbb{I}^d} f(\eta) d\eta - \frac{1}{N} \sum_{i=1}^{N} f(X_n) \right| \leq V^{(d)}(f) D_N^{\star}(X_1, \ldots, X_N).$$

In fact, Theorems 1 and 2 state that the magnitude of the error depends on the total variation (defined in Section 14.3.3) of the function and the star discrepancy of the point set chosen. That is why we always seek low star discrepancy point sets in quasi-Monte Carlo methods. It is also worth mentioning that the error bounds are conservative, *i.e.* if the variation of the function is large, we get a large bound on the error, although the actual error might be small.

In what follows we review a class of nonlinear systems for which sampling techniques may be used to provide stability results, and state how QMC methods may be used to do so.

14.4 Stability of Non-Polynomial Systems

Studying the stability of general nonlinear systems is usually reduced to the difficult task of identifying Lyapunov function candidates. Such a task is slightly simplified when the class of nonlinear systems at hand is restricted. We proceed by giving a formal definition of the class of systems we deal with in the remainder of this chapter.

Definition 5. *Consider the class S of a nonlinear multivariate function defined as follows*

$$\begin{aligned} S = \ & \{f : f(x) = \sum_{i=1}^{l} p_i(x) g_i(x),\, p_i : \mathbb{R}^n \to \mathbb{R},\, monomials \\ & g_i : D_1^n \to D_2,\, nonlinear\ functions,\, D_1^n \subseteq \mathbb{R}^n, \\ & D_2 \subseteq \mathbb{R},\, l \in \mathbb{N}\}. \end{aligned}$$

In particular, the elements of class S are sums of polynomial functions, non-polynomial functions, and product of both.

Definition 6. *Recall again the class S of multivariate functions defined in Definition 5 as the functions composed by a sum of terms in which there are polynomial and non-polynomial elements. Consider a subset $S_1 \subset S$ in which only part of the variables appear in the polynomial functions $p(x)$, i.e.*

$$\begin{aligned} f(x) = \ & \sum_{i} p_i(x) g_i(X_g),\, x \in \mathbb{R}^n,\, X_g \in \mathbb{R}^{(n-k)} \\ & X_g(j) = x(j),\, j = k+1, ..., n,\, k \leq n \end{aligned}$$

where p_i are multivariate polynomial functions and g_i are multivariate non polynomial functions. Observe that the first k components of x only appear in the polynomial part and form a so-called polynomial vector $X_p(i) = x(i)$, $i = 1, \dots, k$ where $X_p \in \mathbb{R}^k$. We refer to these variables as 'polynomial variables' and to the remaining variables in the state vector as 'global variables', and they form a 'global vector' $X_g(j) = x(j)$, $j = k+1, \dots, n$ and $X_g \in \mathbb{R}^{(n-k)}$, we have $x = [X_p \ X_g]^T$. The meaning of such notation will be made clearer next.

Definition 7. *Consider the class of systems that produce a derivative of the quadratic Lyapunov function along the trajectory that belongs to the class S_1 defined in Definition 6, we refer to this class of system as 'decoupled state systems', in which the state vector can be split into two parts, the first part of the state vector only contribute to the dynamic of the system through polynomial functions, i.e.*

$$\begin{aligned} \dot{x} &= P(x)G(X_g), \ x \in \mathbb{R}^n, \, X_g \in \mathbb{R}^{(n-k)} \\ X_g(j) &= x(j) \ \ j = k+1, ..., n, \ \ k \leq n \end{aligned} \tag{14.8}$$

where P and G are respectively a vector polynomial function and a vector non-polynomial function. Consider the state vector x and split it into two parts $x = [\xi \ \varphi]^T$, in which ξ is the vector of polynomial variables and φ is the vector of global variables as defined in Definition 6, we can then rewrite (14.8) as follows:

$$\begin{aligned} \begin{bmatrix} \dot{\xi} \\ \dot{\varphi} \end{bmatrix} &= P(\xi, \varphi)G(\varphi) \\ \xi(i) &= x_i \quad i = 1, \dots, k \\ \varphi(j) &= x_j \quad j = k+1, \dots n \ \ k \leq n. \end{aligned}$$

14.4.1 S-Procedure Approach

Our goal in this section is to simplify the structure of non-polynomial functions through a transformation that allows us to rewrite the function as a multivariate polynomial whose variables are subject to some inequality constraints. The positivity of the original function can then be investigated, by studying the new set of inequalities, of the transformed function and the constraints. In [261] a technique to test the polynomial non-negativity over a finite set described by polynomial equalities and inequalities is proposed. We propose here an alternative approach. The S-procedure will allow us to obtain sufficient conditions for the positivity of the system of inequalities. We start by stating the problem of determining the positivity of a multivariate nonlinear function over $\mathbb{R}^n$.

Problem 1. Consider a multivariate nonlinear function composed of sum of an arbitrary number of nonlinear functions, $f = \sum_1^l f_i$, $f_i : D_1^n \rightarrow D_2$ where

$D_1^n \subseteq \mathbb{R}^n$ and $D_2 \subseteq \mathbb{R}$. Our objective is to determine if $f(x)$ is non-negative for all $x \in D_1^n$.

Next we consider the problem of deciding positivity of a multivariate polynomial function, whose variables are subject to inequality constraints.

Problem 2. Consider a multivariate polynomial function $p : \mathbb{R}^n \to \mathbb{R}$ where D is an n-dimensional domain. We aim to determine if $p(x)$ is non-negative for all $x \in \mathbb{R}^n$ subject to inequality constraints *i.e.* $p(x) \geq 0$, $\forall x \in \mathbb{R}^n$, $\underline{x}_i < x_i < \overline{x}_i$, $i = 1, \ldots, n$.

Using a special transformation as in [222], Problem 1 can be reformulated as Problem 2. Then the S-procedure [59] can be used to solve Problem 2. The S-procedure is briefly stated next.

S-procedure for quadratic functions

Let $F_0 \ldots, F_k$ be quadratic functions of the variables $z \in \mathbb{R}^n$:

$$F_i(z) = z^T T_i z + 2u_i^T z + v_i,\ i = 0, \ldots, k$$

where $T_i = T_i^T$, are $n \times n$, u_i are $n \times 1$ vectors and v_i are scalars. Then a sufficient condition for the following statement

$$\forall z \ \ such \ \ that \ \ F_i(z) \geq 0,\ i = 1, \ldots k \Rightarrow F_0 \geq 0$$

is that there exists $\tau_1, \ldots, \tau_k \geq 0$ such that

$$F_0 \geq \tau_1 F_1 + \cdots + \tau_k F_k.$$

14.4.2 Generalized Sum of Squares Decomposition

In [260], it was shown how SOS programming can be applied to analyze the stability of nonlinear systems described by polynomial functions. The tool has also been extended to several applications other than stability analysis [170, 262]. We aim in this section to extend this approach to systems that are not characterized by polynomial functions. The main advantage of the proposed approach is the computational tractability of the SOS decomposition for multivariate polynomials. See [222] for a more detailed exposure to this section.

As stated repeatedly in this chapter, many problems in nonlinear systems can be reduced to the basic problem of checking the global non-negativity of a function of several variables [56]. First, we will show that using semi-definite programming (SDP) it is possible to test if a given polynomial admits an SOS decomposition [260].

Theorem 3. *Given a multivariate polynomial* $p : x \in \mathbb{R}^n \to \mathbb{R}$ *of degree* $2m$*, a sufficient condition for the existence of SOS representation* $p(x) = p(z) = z^T Q z$ *is* $Q \succeq 0$ *where* z *is a vector of monomials in* x *of degree* m*.*

So the test for SOS of a polynomial function has been reduced to a linear matrix inequality (LMI) [59]. Then for a symmetric matrix Q we obtain the following eigenvalue factorization [9]: $Q = L^T T L$, from which follows the decomposition $p(x) = \sum_i (Lz)_i^2$. In general we have that the SOS representation might not be unique, depending on the choice of the components of the z vector. In particular, different choices of the vector z correspond to different matrices Q that satisfy the SOS representation. It could be that only some of those matrices are PSD, so the existence of SOS decomposition for a polynomial may depend on the representation. If at least one of the matrices of the linear subspace is positive semidefinite (*i.e.* the intersection of the linear subspace of matrices satisfying the SOS representation with the positive semidefinite matrix cone is non-empty), then $p(x)$ is SOS and therefore PSD. In general we will choose the components of z to be linearly independent, and we will say that the corresponding representation is minimal.

14.4.3 SOS Generalization: A Partial State Vector Approach

We will show how, under certain assumptions, it is possible to apply the SOS procedure to a nonlinear, non-polynomial function. The main idea is based on the use of SOS procedure, considering the generic nonlinear function as a polynomial function, in which the non-polynomial parts are treated as coefficients of the function. Rewriting the function as a quadratic form we get $f(x) = z(X_p)^T Q(X_g) z(X_p)$ where z is a vector of monomials formed from the polynomial variables $x_1, \ldots, x_k$ which are elements of the polynomial vector X_p (defined in Section 14.4), and Q is a matrix of appropriate dimension, which depends on the global variables $x_{k+1}, \ldots, x_n$ which form the global vector X_g (also defined earlier in Section 14.4). From SOS theory, a sufficient condition for $f(x) > 0$ is that $Q(X_g)$ is positive definite. In order to apply the SOS procedure to a generic nonlinear non-polynomial function, we need to restrict the class of systems we deal with, in particular we will consider the class of systems defined in (7). The state vector x is divided into two parts, φ and ξ. In choosing a quadratic Lyapunov function $V = x^T x$ and applying the SOS procedure to determine the sign of $-\dot{V}$ we aim to find conditions on φ that guarantee V is decreasing along the trajectory of the system for all ξ *i.e.*

$$-\dot{V} = z(\xi)^T Q(\varphi) z(\xi) > 0 \quad \forall \xi \in \mathbb{R}^k$$

14.5 Nonlinear Stability Analysis via Sampling

In the previous sections we have reviewed several concepts pertaining to stability of non-polynomial systems and the QMC technique for deterministic sampling. In what follows we propose linking these concepts in order to analyze the stability of non-polynomial systems. Starting with a positive definite Lyapunov function and taking its derivative along the state trajectories, we propose two approaches to study the stability of the underlying non-polynomial nonlinear systems:

- The deterministic sampling (QMC) method which determines with certain guaranties, depending on the total variation, the sign definiteness of the derivative of the Lyapunov function along the state trajectories.
- The transformation methods discussed earlier (Approximation, S-procedure and Generalized SOS) in order to conservatively identify, using always a Lyapunov framework, a subregion of the state-space where the system is asymptotically stable.

Moreover, in the second scenario it is also possible to use the approximation approach, to further explore beyond the stability region obtained above via sampling. Such a process may be iterated in order to reduce with certain error guarantees the conservativeness in the Lyapunov approach. It is important however to realize that the approximation method gives a sure answer, while the extension beyond this certain stability region involves certain error due to our inability to test the continuum of the initial conditions in the state-space.

1. Consider a non-polynomial system $\dot{x} = f(x)$, $x \in \mathbb{R}^n$ for which we determined using one of the transformation techniques provided in Section 14.4 a subregion of stability, say $[-c, c]^n$.
2. Augment the region with an $\epsilon > 0$ obtaining a new subregion, $[-c-\epsilon, c+\epsilon]^n \backslash [-c, c]^n$.
3. Using random or quasi-random sampling, generate points in the above subregion and test the stability at each point by either:
   a. Simulating the response of the initial conditions using the state equations, in which case we cannot get a definite bound on the performance of the method, or
   b. In the case of extending the region of stability obtained via the approximation technique, we can obtain a deterministic bound on the error through the total variation of the derivative of the Lyapunov function.

14.6 Example

We aim to apply the approaches described above to study the stability of a mobile robot model described in [94].Consider the dynamic model of a vehicle

$$\begin{aligned} m\ddot{x} &= -\eta\dot{x} + (F_s + F_p)\cos\theta \\ m\ddot{y} &= -\eta\dot{y} + (F_s + F_p)\sin\theta \\ J\ddot{\theta} &= -\psi\dot{\theta} + (F_s - F_p)r \end{aligned}$$

where (x, y) are the position coordinates, θ is the orientation angle. As for the physical parameters, m is the mass of the vehicle, J is the rotational inertia,

F_s and F_p are respectively the starboard and the port fan forces, r is the moment arm of the forces and η and ψ are the coefficients of viscous friction and rotational friction, respectively.

In order to achieve controllability [94], consider the error dynamics around a constant velocity $\dot{x}_{nom}$, $\dot{y}_{nom}$ and heading $\dot{\theta}$ are given by

$$\begin{aligned} m\ddot{x}_e &= -\eta(\dot{x}_e + \dot{x}_{nom}) + (F_s + F_p)\cos(\theta_e + \theta_{nom}) \\ m\ddot{y}_e &= -\eta(\dot{y}_e + \dot{y}_{nom}) + (F_s + F_p)\sin(\theta_e + \theta_{nom}) \\ J\ddot{\theta}_e &= -\psi\dot{\theta}_e + (F_s - F_p)r \end{aligned}$$

with nominal input $F_s = F_p = \frac{(\eta\dot{x}_{nom})}{2\cos\theta_{nom}}$. We use the following nominal values for linear velocity and angular position: $\dot{x}_{nom} = \dot{y}_{nom} = 10$, $\dot{\theta}_{nom} = \frac{\pi}{4}$. Let us consider the state vector $\xi = [x\ y\ \theta\ \dot{x}\ \dot{y}\ \dot{\theta}]^T$ we get the following system:

$$\begin{aligned} \dot{\xi}_1 &= \xi_4 \\ \dot{\xi}_2 &= \xi_5 \\ \dot{\xi}_3 &= \xi_6 \\ \dot{\xi}_4 &= -\frac{\eta}{m}(\xi_4 + \dot{x}_{nom}) + \frac{F_s + F_p}{m}\cos(\xi_3 + \theta_{nom}) \\ \dot{\xi}_5 &= -\frac{\eta}{m}(\xi_5 + \dot{y}_{nom}) + \frac{F_s + F_p}{m}\sin(\xi_3 + \theta_{nom}) \\ \dot{\xi}_6 &= -\frac{\psi}{J}\xi_6 + \frac{F_s - F_p}{J}m. \end{aligned}$$

We aim to study how the stabilizing controller designed for a linearized system perform on the real system; in other words if designing a controller for stabilizing the linearized system we can guarantee the stability of the original system at least in a neighborhood of the point around which we did the approximation.

The first step is to linearize the original system around an equilibrium point that chosen for simplicity as the origin. Using the numerical values $J = .05$, $\psi = .084$, $\eta = 5.5$, and $m = 5.15$ we observe that the linearized system is unstable. Since the system is controllable, we design a controller in order to achieve a stable closed-loop system

$$\begin{aligned} \dot{\xi} &= A\xi + Bu \\ u &= -K\xi \end{aligned}$$

where $u = [F_s\ \ F_p]^T$, and the resulting controller gain matrix K is

$$K = \begin{bmatrix} 3.2122 & 3.3217 & 1.8706 & 3.0078 & 3.1102 & 0.7258 \\ 3.3217 & 3.2122 & -1.8706 & 3.1102 & 3.0078 & -0.7258 \end{bmatrix}.$$

We can now proceed to studying the stability of the closed loop nonlinear system. In particular we want to compare the three results we get by applying the described procedures.

At first, since we know the closed loop system is stable at the origin we can think of exploring the neighborhood in order to determinate the region of attraction. By using the QMC approach and picking samples in the region of the hypercube $[-5, 5]^6$ we get that the system is stable in that region as it is shown in the graphic of Figure 14.3. In particular the dynamics of the six state variables, the errors on the linear and angular position and velocity, are plotted and for initial condition inside the region $[-5, 5]^6$ they converge.

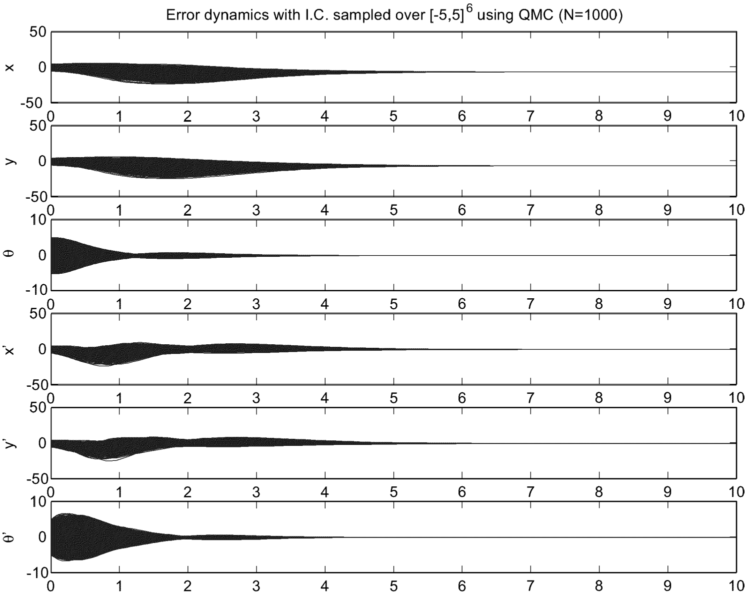


Figure 14.3. Error dynamics

Then using a Lyapunov based analysis and employing the techniques described in Section 14.4, we obtain a stability region $[-0.5, 0.5]^6$, which is much more conservative than the answer obtained earlier, yet more precise.

14.7 Conclusion

In this chapter we have reviewed the main ideas pertaining to the sampling methods used in control systems analysis and design, in particular quasi-Monte Carlo method. We also presented various transformation techniques that allow us to change the problem of stability analysis of general nonlinear systems into that of studying the stability of a polynomial or partial polynomial systems.

15

Probabilistic Control of Nonlinear Uncertain Systems

Qian Wang[1] and Robert F. Stengel[2]

[1] Mechanical Engineering, Penn State University
University Park, PA 16802
quw6@psu.edu

[2] Mechanical and Aerospace Engineering, Princeton University
Princeton, NJ 08544
stengel@princeton.edu

Summary. Robust controllers for nonlinear systems with uncertain parameters can be reliably designed using probabilistic methods. In this chapter, a design approach based on the combination of stochastic robustness and dynamic inversion is presented for general systems that have a feedback-linearizable nominal system. The efficacy of this control approach is illustrated through the design of flight control systems for a hypersonic aircraft and a highly nonlinear, complex aircraft model. The proposed stochastic robust nonlinear control explores the direct design of nonlinear flight control logic; therefore the final design accounts for all significant nonlinearities in the aircraft's high-fidelity simulation model. Monte Carlo simulation is used to estimate the likelihood of closed-loop system instability and violation of performance requirements subject to variations of the probabilistic system parameters. The stochastic robustness cost function is defined in terms of the probabilities that design criteria will not be satisfied. We use randomized algorithms, in particular genetic algorithms, to search the design parameters of the parameterized controller with feedback linearization structure. The design approach is an extension of earlier methods for probabilistic robust control of linear systems. Prior results are reviewed, and the nonlinear approach is presented.

15.1 Introduction

Control systems should be designed to run satisfactorily not only with assumed plant parameters but with possible variations in operating conditions. Perfect models of systems to be controlled are rarely available when controllers are being designed, parameters of similar plants are likely to vary from one example to the next, and dynamic characteristics may change as parts wear or operating points shift. Control system designs must be tolerant of these differences for practical control to take place, that is, they must be robust. For parametric uncertainty, guaranteed stability-bound estimates often are unduly conservative, and the resulting controller usually needs very high control effort. With respect to computational complexity, many worst-case deterministic robust control problems are proved to be NP hard. If instead of worst-case guaranteed conclusions, probabilistic robustness is acceptable, computational complexity can be reduced significantly. In probabilistic robust control design, randomized algorithms with polynomial complexity are used to characterize system robustness and to identify satisfactory controllers.

Many problems in system synthesis can be formulated as the minimization of an objective function with respect to the parameters of a parameterized controller. The probabilistic robust control problem is transformed to a stochastic optimization problem. Combinations of a variety of pre-existing control methodologies and the probabilistic approach to robustness have been applied to control designs such as Linear-Quadratic-Gaussian regulators [217, 273], transfer function sweep designs [391], quadratic stabilization for linear systems [25], robust Linear Matrix Inequality (LMI) or Quadratic Matrix Inequality (QMI) [76], Linear-Parameter-Varying control [129], robust $\mathcal{H}_2$ control [201] and Model Predictive control [177].

The probabilistic approach is readily applied to nonlinear designs as well as to linear designs. We present a framework for nonlinear robust control that merges the stochastic approach with feedback linearization. There has been intensive research in deterministic nonlinear robust control using, for example, Lyapunov redesign, backstepping, sliding-mode control, and neural network based adaptive robust control [165]. The probabilistic approach to control design could reduce design conservativeness significantly, and it provides a viable treatment for system robustness with respect to uncertain parameters that may enter the system in an arbitrary way. In this chapter, the proposed stochastic feedback linearization approach is illustrated through two flight control applications. The first application is to the control of the longitudinal motion of a NASA Langley hypersonic aircraft [333] cruising at a Mach number of 15 and at an altitude of $110,000$ ft. There are 28 uncertain parameters in characterizing the aircraft's inertial and aerodynamic model. Robustness metrics include system stability and 38 performance specifications for velocity and altitude command responses in the presence of uncertain parameter variations. The probabilistic robust control design is formulated as a stochastic optimization of a cost function that is a weighted quadratic sum of these probabilities of violation of design specifications. Due to the non-convex and non-deterministic nature of this stochastic optimization problem, genetic algorithms are used here to search the controller parameters. We apply a similar approach to the probabilistic robust control of the six-degree-of-freedom motion of a High Incidence Research Model (HIRM) [210] whose highly nonlinear aerodynamic model is described by a combination of analytic equations and look-up tables. Due to the complexity of the system model, a two-time-scale decomposition is used in the design of controller structure. The resulting nonlinear control design is evaluated and compared against existing designs with respect to handling qualities for a wide range of flight envelopes and in the presence of system parametric uncertainties.

This chapter is organized as follows. Section 15.2 summarizes prior results in the probabilistic design of constant-coefficient controllers for linear systems. In Section 15.3, we present a general approach for probabilistic robust control of nonlinear systems. In Section 15.4, the proposed approach is applied to the flight control design for a NASA Langley hypersonic aircraft model and the design for the High Incidence Research Model is presented in Section 15.5; simulation results are presented for stability and performance robustness of the closed-loop system.

15.2 Stochastic Analysis and Design for Linear, Time-Invariant Systems

15.2.1 Stochastic Robustness Analysis (SRA)

Stochastic stability theory provides a logical starting point, as satisfactory stability is often a necessary condition for satisfactory performance. A typical problem is to determine bounds on the parameter vector p of an unforced, continuous-time system [192, 195],

$$\dot{x}(t) = f[p(t), x(t)],\ x \in \mathbb{R}^n,\ x(0) = x_0,\ f \in \mathbb{R}^n,\ p \in \mathbb{R}^l$$

where x is the dynamic state and $p(t)$ is a random process, such that stability can be expected with a probability of one (or arbitrarily close to one). A corresponding linear control problem is to find a satisfactory control gain matrix C for the linear plant and control law,

$$\dot{x}(t) = F[p(t), t]x(t) + G[p(t), t]u(t), u \in \mathbb{R}^m, F \in \mathbb{R}^{n\times n}, G \in \mathbb{R}^{n\times m} \quad (15.1)$$

$$u(t) = -Cx(t), C \in \mathbb{R}^{m\times n}.$$

The system dynamics vector $f(\cdot)$ becomes

$$f[p(t), x(t)] = \{F[p(t), t] - G[p(t), t]C\}x(t)$$

and the uncertainty is contained in the varying values of $[F(\cdot), G(\cdot)]$. Probabilistic stability criteria have been developed using expectations of Lyapunov functions, and they require consideration of stochastic integrals and transformations [169, 402]. Analogous discrete-time problems are discussed in [197]. Given infinite (*e.g.*, Gaussian) parameter distributions, the probability of instability is finite, and the escape (or exit) time may be of interest [93].

The principal focus of current robustness research is on ensembles of linear systems for which p is a random constant rather than a random process. For a particular parameter value p_k, F_{p_k} is uncertain but fixed. Deterministic stability criteria apply to each member of the ensemble. Because each dynamic system is linear and time-invariant, its stability is entirely determined by its eigenvalues, that is, the solutions λ_j to the equation

$$|\lambda_j I - F^T(p_k)| = 0, j = 1, \cdots, n. \tag{15.2}$$

Given a vector of the probability density functions of p, $f_p(p)$, equation (15.2) provides an implicit transformation for computing the probability density functions, $f_p(\lambda_j)$, of the corresponding ensemble of eigenvalues $\lambda_j, j = 1, \cdots, n$. An evaluation of the cumulative probability of (in)stability induced by $f_p(p)$ requires integration of the $f_p(\lambda_j)$ over the (right) left-half complex plane. Linear eigenvalue sensitivities, $\partial\lambda_j/\partial p$, can be derived and applied for analytic evaluation of the integral [207,268], and additional studies of eigenvalue and eigenvector sensitivities can be found in [120,143,166,231,345,346].

Analytical solutions to this integral have limited utility for evaluating the probability of (in)stability. The most practical approach for evaluating the probability of (in)stability in the general case is to use numerical computation, as expanded below. Numerical evaluation of probabilities involves sampling of parameter probability distributions [229,259] and computation of their consequences using either exhaustive sampling or Monte Carlo methods [66]. In the first case, all possible parameter combinations in a finite set are sampled, and the exact probability of hypothesis $\mathcal{H}$ (in the current discussion, the stability or instability of the controlled system) is computed as

$$\mathbb{P}(\mathcal{H}) = N_{\mathcal{H}}/N_{\text{Total}} \tag{15.3}$$

where $N_{\mathcal{H}}$ is the number of instances of $\mathcal{H}$, and N_{Total} is the total number of trials. For the second method, each scalar parameter is represented by a random number generator, whose characteristics are shaped by the parameter's statistical description. There is no restriction on the shapes or correlations of probability distributions (*i.e.*, they may be bounded, non-Gaussian, *etc.*), and parameters may have different distribution types. For a single trial, each element of p_k is generated, and the related hypothesis is computed. The probability of a hypothesis is computed as before in (15.3), but there is uncertainty in the estimate, as discussed below.

For linear, time-invariant (LTI) systems, the *probability of instability* P can be estimated from repeated eigenvalue calculation [347]. Given a system with l parameters, each of which takes w values with equal probability, P can be calculated exactly from w^l evaluations using exhaustive sampling (equation (15.3)), with $N_{\mathcal{H}}$ equal to the number of unstable cases, and N_{Total} equal to w^l. For Monte Carlo evaluation, the closed-loop eigenvalues, λ_j, are evaluated N_{Total} times with each element of p_k, $k = 1, \cdots, N_{\text{Total}}$, specified by a random number generator whose individual outputs are shaped by $f_p(p)$. The probability-of-(in)stability estimate becomes increasingly precise as N_{Total} becomes large:

$$\mathbb{P}(stable) = P = \lim_{N_{\text{Total}}\to\infty} \frac{N(\sigma_{\max} \le 0)}{N_{\text{Total}}}$$

$$\mathbb{P}(unstable) = P = 1 - \mathbb{P}(stable)$$

$N(\cdot)$ is the number of cases for which all elements of $\underline{\sigma}$, the vector of the real parts of the closed-loop eigenvalues ($\lambda = \sigma + j\omega$), are less than or equal to zero, that is, for which $\sigma_{\max} \leq 0$, where $\sigma_{\max}$ is the maximum real eigenvalue component in $\underline{\sigma}$. For $N_{\text{Total}} < \infty$, the Monte Carlo evaluation is an estimate, $\hat{P}$, whose uncertainty is characterized by a confidence interval.

Because P is a binomial variable (*i.e.*, the outcome of each trial takes one of two values: stable or unstable), confidence intervals are calculated using the binomial test, where lower ($\mathcal{L}$) and upper ($\mathcal{U}$) intervals satisfy the following [92]:

$$\mathbb{P}(N_U \leq n-1) = \sum_{k=0}^{n-1} (N_{\text{Total}}, k)\mathcal{L}^k (1-\mathcal{L})^{N_{\text{Total}}-k} = 1 - \frac{\alpha}{2} \qquad (15.4)$$

$$\mathbb{P}(N_U \leq n) = \sum_{k=0}^{n} (N_{\text{Total}}, k)\mathcal{U}^k (1-\mathcal{U})^{N_{\text{Total}}-k} = \frac{\alpha}{2} \qquad (15.5)$$

N_U is the actual number of unstable cases after N_{Total} evaluations ($N_U = N_{\text{Total}}\hat{P}$), (N_{Total}, k) is the binomial coefficient, $\frac{N_{\text{Total}}!}{k!(N_{\text{Total}}-k)!}$, and $(1-\alpha)$ is the confidence coefficient. Explicit approximations of the binomial test [7, 8] avoid an iterative solution of (15.4) and (15.5) for $(\mathcal{L}, \mathcal{U})$, and they are accurate to within 0.1% [347].

The number of evaluations required to estimate a binomial probability distribution for specified interval widths and a 95% confidence coefficient varies with the true P (Figure 15.1), see [347]. For narrow intervals and small P, large numbers of evaluations are required; however, large percentage interval widths may be acceptable if P is small.

The number of Monte Carlo evaluations needed to yield $\hat{P}$ with a given confidence level is independent of the number of uncertain parameters or their probability distributions. This result has broad implications for the robustness evaluation of complex systems. While exact or approximate exhaustive sampling may be useful when there are few parameters, Monte Carlo simulation has broad application for systems with large numbers of uncertain parameters. Chernoff bounds and related analysis have been used to derive the number of required Monte Carlo evaluations to estimate the probability, see [74]. When the distribution of the underlying uncertain parameters is unknown, the effect of sampling distribution on the stochastic robustness analysis has been investigated in [24], where a uniform parameter distribution is found to be the worst-case unimodal distribution.

15.2.2 Stochastic Robustness Design (SRD)

Design for stochastic robustness follows analysis by incorporating search. The simplest approach is to choose the best from an ensemble of controllers, without regard to the design algorithms employed for each controller. For example,

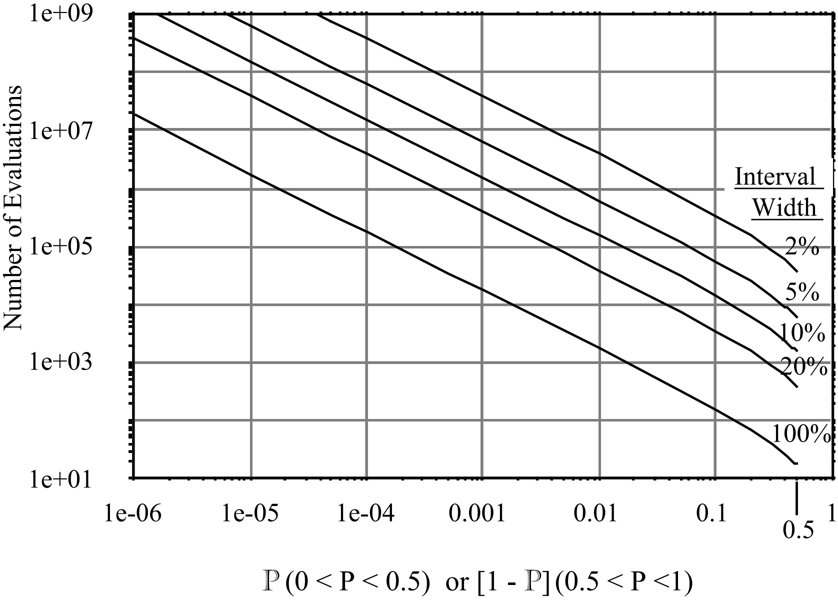


Figure 15.1. Number of evaluations required to estimate a binomial probability distribution for given confidence interval widths and 95% confidence coefficient; interval width is given as percent of P or $(1 - P)$ (from [347])

given the Benchmark Control Problem [398], we could compare the probabilities of instability, P_i, excess control usage, P_u, and excess settling time, P_{T_s}, for the ten design solutions, selecting the one that appears most suitable. The relative importance of the three criteria must be known to make the selection, and the probability distributions of the uncertain parameters that induce them should be well motivated. Guidelines for comparing controller pairs are contained in [288].

Probabilistic *synthesis* of control systems is a natural adjunct to probabilistic analysis; the random or randomized *search* is a dual to Monte Carlo *evaluation*. Building on [342], random-search methods of finding control system gains are explored in [20,368,406]. There are similarities to directed searches that minimize multi-objective cost functions [325], to parameter-space methods [2,336], and to fine-tuning of control gains by search [6]. A genetic algorithm – which performs randomized reproduction, crossover, and mutation on candidate control-gain strings – has been used to design controllers [193], while the stochastic robustness analysis is extended to control design using sequential line searches in [286–290,347]. Statistical learning theory has been applied to control design in [381]. By exploring the convex structure of certain control design problems, sequential stochastic gradient algorithms are used in the minimization of a convex stochastic cost function, see [76,129,177,252] and [359] with references therein.

A typical design procedure has four steps: 1) define cost function of objective probabilities, 2) define controller structure, 3) perform stochastic robustness analysis of the closed-loop system, and 4) conduct numerical search to minimize the cost function.

As an example for Step 1, the quadratic cost function

$$J = \alpha P_i^2 + \beta P_u^2 + \gamma P_{T_s}^2 \tag{15.6}$$

weights the squares of the probabilities to emphasize large values and deemphasize small values. α, β, and γ are scalar weights on the relative importance of instability, excess control usage, and excess settling time over the range of parameter uncertainty. P_i, P_u, and P_{T_s} are in (0, 1). With an LQG controller, the control law and associated estimator for Step 2 are

$$u(t) = -C\hat{x} \tag{15.7}$$

$$\dot{\hat{x}} = F\hat{x} + Gu + K(z - H\hat{x}) \tag{15.8}$$

and the weighting matrices for the LQG problem are chosen as the control design parameters. For a single-input/single-output compensator, the controller structure may simply be a transfer function whose numerator and denominator coefficients are the design parameters. In Step 3, an ensemble of trials is evaluated to compute the probability (15.3) using the dynamic system of (15.1) with randomly generated parameter vectors, p, and closed-loop control specified by (15.7) and (15.8). This Monte Carlo evaluation forms an ‘inner loop’ for the minimization algorithm in Step 4. A genetic algorithm is used to minimize (15.6) through the choice of control design parameters. As an alternative, simulated annealing could be used for the optimization [233]. Execution time for this computationally intensive process can be decreased greatly through the use of parallel computation [321].

15.3 Stochastic Robust Control of Nonlinear Systems

The nonlinear control design is an extension of the probabilistic robust control of linear systems in Section 15.2. A combination of probabilistic robustness with feedback linearization is presented. First, we design a feedback linearization control law for the nominal system, then introduce parametric uncertainty and reformulate the problem in a probabilistic format. The control design parameters are searched to minimize a stochastic robustness cost function that is a weighted quadratic sum of probabilities of violating design specifications.

Consider a nonlinear system that has a nominal system as follows:

$$\begin{gathered} \dot{x} = f(x) + G(x)u, G(x) = \begin{bmatrix} g_1(x) \; g_2(x) \cdots \; g_m(x) \end{bmatrix} \\ y = h(x) \end{gathered} \tag{15.9}$$

where f and g_j $(j = 1, 2, \cdots, m)$ are smooth vector fields on $\mathbb{R}^n$, and h is a smooth function mapping $\mathbb{R}^n \to \mathbb{R}^m$. If this nominal system is feedback linearizable, there exists a nonlinear coordinate transformation $\varsigma = T(x)$

$$\begin{cases} \varsigma_1^i = h_i \\ \varsigma_2^i = \frac{dh_i}{dt} = L_f h_i \\ \vdots \\ \varsigma_{\lambda_i}^i = \frac{d^{(\lambda_i - 1)} h_i}{dt} = L_f^{\lambda_i - 1} h_i \end{cases} \qquad i = 1, 2, \cdots, m$$

such that the nominal system is transformed to a set of decoupled linear systems,

$$\begin{cases} \dot{\varsigma}_1^i = \varsigma_2^i \\ \dot{\varsigma}_2^i = \varsigma_3^i \\ \vdots \\ \dot{\varsigma}_{\lambda_i}^i = L_f^{\lambda_i} h_i + \sum_{j=1}^m L_{g_j}(L_f^{\lambda_i - 1} h_i) u_j = v_i \end{cases} \qquad i = 1, 2, \cdots, m \tag{15.10}$$

where the Lie derivatives are defined as $L_f h_i = \frac{\partial h_i(x)}{\partial x} f(x)$, $L_f^k h_i = L_f(L_f^{k-1} h_i)$, and $L_{g_j} h_i = \frac{\partial h_i(x)}{\partial x} g_j(x)$.

For the decoupled linear systems (15.10), the control law $v = \begin{bmatrix} v_1 \; v_2 \; \cdots \; v_m \end{bmatrix}^T$ could be designed using any existing technique, such as, a linear quadratic control that is parameterized in terms of weighting matrices Q and R. By (15.10), the nonlinear control u is calculated through the new control input v as

$$u = -[G^*(x)]^{-1} f^*(x) + [G^*(x)]^{-1} v \tag{15.11}$$

where

$$f^*(x) = \begin{bmatrix} L_f^{\lambda_1} h_1 \\ L_f^{\lambda_2} h_2 \\ \vdots \\ L_f^{\lambda_m} h_m \end{bmatrix}$$

$$G^*(x) = \begin{bmatrix} L_{g_1} L_f^{\lambda_1 - 1} h_1 & L_{g_2} L_f^{\lambda_1 - 1} h_1 & \cdots & L_{g_m} L_f^{\lambda_1 - 1} h_1 \\ L_{g_1} L_f^{\lambda_2 - 1} h_2 & L_{g_2} L_f^{\lambda_2 - 1} h_2 & \cdots & L_{g_m} L_f^{\lambda_2 - 1} h_2 \\ \cdots & \cdots & \cdots & \cdots \\ L_{g_1} L_f^{\lambda_m - 1} h_m & L_{g_2} L_f^{\lambda_m - 1} h_m & \cdots & L_{g_m} L_f^{\lambda_m - 1} h_m \end{bmatrix}.$$

After the control design is derived for the nominal system, we consider the uncertain nonlinear vector fields $(f(x, q), G(x, q))$ subject to parametric uncertainty $q \in Q$. According to system design requirements, a set of robustness metrics is defined and a stochastic robustness cost function is formulated as a weighted quadratic sum of the probabilities of violating these robustness metrics. We use the parameterized control law of the nominal system as the controller structure for the system with uncertainties, and tune the control design parameters to minimize the stochastic robustness cost function. The input-to-state stability of the nominal closed-loop system is guaranteed, and the stability and other performance metrics of the uncertain system are evaluated by Monte Carlo simulation. As addressed in Section 15.2, the discrepancy between the Monte Carlo estimate and the true value results in apparent 'noise' in the evaluation of the cost function. Furthermore, the cost function may be non-convex, having large plateaus and corners, so traditional gradient-based search algorithms can get stuck in local minima and not escape from large plateau areas. A series of randomized algorithms such as stochastic gradient methods, sequential line search, clustering algorithms, genetic algorithms and simulated annealing has been investigated [215, 233]. In this chapter, genetic algorithms are used to minimize the stochastic robustness cost function.

In the following two sections, we illustrate the application of the above stochastic robust nonlinear control to two flight control examples: a NASA Langley hypersonic aircraft and the High Incidence Research Model. One of the major challenges in the design of flight control systems is model uncertainties and parameter variations in characterizing an aircraft and its operating environment. While many gains have been made in robust control theory over the past several decades, the gap between the new methods and conventional flight control design approaches has precluded their widespread use. The proposed stochastic robust control framework takes into account the engineering design requirements during the design phase, and it gives a direct answer to the likelihood that the design metrics are not satisfied.

15.4 Stochastic Robust Control Design For a Hypersonic Aircraft Model

15.4.1 System Model and Design Specifications

Consider the control of the longitudinal motion of a hypersonic aircraft cruising at a Mach number of 15 and at an altitude of 110, 000 ft [333]. The dynamic equations are

$$\dot{V} = \frac{T\cos\alpha - D}{m} - \frac{\mu\sin\gamma}{r^2}$$

$$\dot{\gamma} = \frac{L + T\sin\alpha}{mV} - \frac{(\mu - V^2 r)\cos\gamma}{Vr^2}$$

$$\dot{h} = V\sin\gamma$$

$$\dot{\alpha} = q - \dot{\gamma}$$

$$\dot{q} = M_{yy}/I_{yy}$$

where

$$L = \frac{1}{2}\rho V^2 S C_L(\alpha)$$

$$D = \frac{1}{2}\rho V^2 S C_D(\alpha)$$

$$T = \frac{1}{2}\rho V^2 S C_T(\delta T, \alpha)$$

$$M_{yy} = \frac{1}{2}\rho V^2 S\bar{c}[C_M(\alpha) + C_M(\delta E) + C_M(q)]$$

$$r = h + R_E$$

We have used relatively simple functions to fit the aerodynamic coefficients and air data around the nominal cruising condition. Twenty-eight inertial and aerodynamic parameters (identified in [389]) are assumed to be uncertain. Each parameter is multiplied by an element of the uncertainty vector, ν, that is assumed to follow a normal distribution with a mean of 1 and a standard deviation of 0.1. At the trimmed cruise condition ($M = 15$, $V = 15,060$ ft/s, $h = 110,000$ ft, $\alpha = 0.0315$ rad, $\delta T = 0.183$, $\delta E = -0.0066$ rad, and $T = 4.6853 \times 10^4$ lbf), a linearized model of the nominal open-loop dynamics has eigenvalues of -0.8, 0.687, $-0.0001 + 0.0263\jmath$, and 0.0008. The first two eigenvalues represent a statically unstable short-period mode; the complex pair of eigenvalues portrays a lightly damped phugoid mode, and the last real eigenvalue indicates a mildly unstable height mode. Consequently, cruising flight would be subject to attitude and height divergence that would require stabilizing feedback control.

Three aspects of flight control robustness are of concern in this design: stability, performance in velocity command response, and performance in altitude command response. The command responses are initiated at the trimmed condition. State histories of the aircraft's nonlinear response to the velocity and altitude commands are evaluated for stability and performance. Table 15.1 lists 39 stability and performance metrics that characterize the responses to a step velocity command change of 100 ft/s and a step altitude command change of 2000 ft. The indicator functions with subscripts 'V' and 'h' denote the metrics for velocity and altitude command responses.

The cost function chosen to guide the design is a weighted quadratic sum of the 39 probabilities of design requirement violation:

$$J = \sum_{j=1}^{39} w_j P_j^2. \tag{15.12}$$

As indicated in Table 15.1, the stability weight w_1 is chosen as 10, the weight for each more-demanding performance metric is selected as 1, and the weight for each less-demanding performance metric is 0.1.

Controller structure

First we consider the nominal dynamics of the hypersonic aircraft with velocity and altitude commands:

Table 15.1. Stability and performance metrics for a hypersonic aircraft

Metric	Weight in J	Indicator function	Design requirement
1	10	I_i	Stability
2 (3)	0.1 (1.0)	I_{V,T_s25} (I_{V,T_s50})	10% settling time less than 25s (50s)
4 (5)	0.1 (1.0)	$I_{V,R25}$ ($I_{V,R50}$)	90% rise time less than 25s (50s)
6	0.1	$I_{V,Rev}$	No reversal of response in V before peaking
7 (8)	0.1 (1.0)	$I_{V,D5}$ ($I_{V,D10}$)	10% dwell time less than 5s (10s)
9 (10)	0.1 (1.0)	$I_{V,OS10}$ ($I_{V,OS20}$)	Overshoot less than 10% (20%)
11 (12)	0.1 (1.0)	$I_{V,\Delta\alpha 0.5}$ ($I_{V,\Delta\alpha 1}$)	Max change in α less than 0.5° (1°)
13 (14)	0.1 (1.0)	$I_{V,g1}$ ($I_{V,g2}$)	Max load factor less than 1g (2g)
15 (16)	0.1 (1.0)	$I_{V,\Delta h0.25}$ ($I_{V,\Delta h0.5}$)	Max change of h less than 0.25% (0.5%)
17 (18)	0.1 (1.0)	$I_{V,\delta T50}$ ($I_{V,\delta T100}$)	Max change in thrust less than 50% (100%)
19 (20)	0.1 (1.0)	$I_{V,\delta E5}$ ($I_{V,\delta E10}$)	Max change in δE less than 5° (10°)
21 (22)	0.1 (1.0)	I_{h,T_s50} (I_{h,T_s100})	10% settling time less than 50s (100s)
23 (24)	0.1 (1.0)	$I_{h,R50}$ ($I_{h,R100}$)	90% rise time less than 50s (100s)
25	0.1	$I_{h,Rev}$	No reversal of response in h before peaking
26 (27)	0.1 (1.0)	$I_{h,D10}$ ($I_{h,D20}$)	10% dwell time less than 10s (20s)
28 (29)	0.1 (1.0)	$I_{h,OS20}$ ($I_{h,OS40}$)	Overshoot less than 20% (40%)
30 (31)	0.1 (1.0)	$I_{h,\Delta\alpha 0.5}$ ($I_{h,\Delta\alpha 1}$)	Max change in α less than 0.5° (1°)
32 (33)	0.1 (1.0)	$I_{h,g1}$ ($I_{h,g2}$)	Max load factor less than 1g (2g)
34 (35)	0.1 (1.0)	$I_{h,\Delta V0.25}$ ($I_{h,\Delta V0.5}$)	Max change of V less than 0.25% (0.5%)
36 (37)	0.1 (1.0)	$I_{h,\delta T50}$ ($I_{h,\delta T100}$)	Max change in thrust less than 50% (100%)
38 (39)	0.1 (1.0)	$I_{h,\delta E5}$ ($I_{h,\delta E10}$)	Max change in δE less than 5° (10°)

$$y_{com} = \begin{bmatrix} V \\ h \end{bmatrix}.$$

Integral compensation is used to minimize the steady-state error of the command response; hence define

$$V_I = \int_0^t (V(\tau) - V^*)d\tau,\ h_I = \int_0^t (h(\tau) - h^*)d\tau$$

where V^* and h^* are the commanded values.

Dynamic extension is used to ensure that the vector relative degree is well defined; we assume that engine dynamics take a second-order form,

$$\ddot{\delta T} = k_1 \dot{\delta T} + k_2 \delta T + k_3 \delta T_{com}$$

where choosing $k_1 = k_2 = 0$ and $k_3 = 1$ provides a suitable model.

By augmenting the state variables as

$$x_1 = \begin{bmatrix} V_I \\ V \\ \gamma \end{bmatrix},\ x_2 = \begin{bmatrix} \delta T \\ h_I \\ h \\ \alpha \end{bmatrix},\ x_3 = \begin{bmatrix} \dot{\delta T} \\ q \end{bmatrix}$$

and defining the control vector as

$$u = \begin{bmatrix} \delta T_{com} \\ \delta E \end{bmatrix}$$

the state equation can be put into a triangular form, *i.e.* it is feedback linearizable.

Using the notation

$$z^T = \begin{bmatrix} V \; \gamma \; \alpha \; \delta T \; h \end{bmatrix}$$

we have

$$\begin{cases} \dot{V} = \frac{T\cos\alpha - D}{m} - \frac{\mu \sin\gamma}{r^2} \\ \ddot{V} = \frac{1}{m}\omega_1 \dot{z} \\ V^{(3)} = \frac{1}{m}(\omega_1 \ddot{z} + \dot{z}^T \varOmega_2 \dot{z}) \end{cases} \tag{15.13}$$

$$\begin{cases} \dot{h} = V\sin\gamma \\ \ddot{h} = \dot{V}\sin\gamma + V\dot{\gamma}\cos\gamma \\ h^{(3)} = \ddot{V}\sin\gamma + 2\dot{V}\dot{\gamma}\cos\gamma - V\dot{\gamma}^2\sin\gamma + V\ddot{\gamma}\cos\gamma \\ h^{(4)} = V^{(3)}\sin\gamma + 3\ddot{V}\dot{\gamma}\cos\gamma - 3\dot{V}\dot{\gamma}^2\sin\gamma + 3\dot{V}\ddot{\gamma}\cos\gamma \\ \qquad\quad - 3V\dot{\gamma}\ddot{\gamma}\sin\gamma - V\dot{\gamma}^3\cos\gamma + V\gamma^{(3)}\cos\gamma \end{cases} \tag{15.14}$$

where

$$\begin{cases} \ddot{\gamma} = \pi_1 \dot{z} \\ \gamma^{(3)} = \pi_1 \ddot{z} + \dot{z}^T \varXi_2 \dot{z} \end{cases} .$$

The vectors ω_1, π_1 and matrices $\varOmega_2$, $\varXi_2$ are omitted here for brevity; they can be found in [389].

By separating $\ddot{\alpha}$ and $\delta\ddot{T}$ into control-independent and control-dependent parts,

$$\ddot{\alpha} = \ddot{\alpha}_0 + \ddot{\alpha}_{\delta E}\delta E$$

$$\delta\ddot{T} = \delta\ddot{T}_0 + \delta\ddot{T}_{com}\delta T_{com}$$

where $\ddot{\alpha}_{\delta E}$ represents the first derivatives of $\ddot{\alpha}$ with respect to δE, and $\delta\ddot{T}_{com}$ represents the first derivative of $\delta\ddot{T}$ with respect to δT_{com}, $\ddot{z}^T$ can be written as

$$\begin{aligned} \ddot{z}^T &= \begin{bmatrix} \ddot{V} \; \ddot{\gamma} \; \ddot{\alpha} \; \delta\ddot{T} \; \ddot{h} \end{bmatrix} \\ &= \begin{bmatrix} \ddot{V} \; \ddot{\gamma} \; \ddot{\alpha}_0 \; \delta\ddot{T}_0 \; \ddot{h} \end{bmatrix} \\ &\quad + \begin{bmatrix} \delta T_{com} \; \delta E \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 0 & \delta\ddot{T}_{com} & 0 \\ 0 & 0 & \ddot{\alpha}_{\delta E} & 0 & 0 \end{bmatrix} \\ &= \ddot{z}_0^T + u^T \ddot{z}_u^T . \end{aligned}$$

Therefore, by (15.13) and (15.14), we have

$$\begin{bmatrix} V^{(3)} \\ h^{(4)} \end{bmatrix} = f^*(x) + G^*(x)u$$

with f^* and G^* as

$$f^* = \begin{bmatrix} \frac{1}{m}\omega_1 \ddot{z}_0 + \frac{1}{m}\dot{z}^T \Omega_2 \dot{z} \\ 3\ddot{V}\dot{\gamma}\cos\gamma - 3\dot{V}\dot{\gamma}^2 \sin\gamma + 3\dot{V}\ddot{\gamma}\cos\gamma - 3V\dot{\gamma}\ddot{\gamma}\sin\gamma - V\dot{\gamma}^3\cos\gamma \\ + \left(\frac{1}{m}\omega_1 \ddot{z}_0 + \frac{1}{m}\dot{z}^T \Omega_2 \dot{z}\right)\sin\gamma + V\cos\gamma(\pi_1 \ddot{z}_0 + \dot{z}^T \Pi_2 \dot{z}) \end{bmatrix}$$

$$G^* = \begin{bmatrix} \frac{T_{\delta T}\cos\alpha}{m}\delta T_{com} & \frac{T_\alpha \cos\alpha - T\sin\alpha - D_\alpha}{m}\ddot{\alpha}_{\delta E} \\ \frac{T_{\delta T}\sin(\alpha+\gamma)}{m}\delta T_{com} & \frac{T\cos(\alpha+\gamma) + T_\alpha \sin(\alpha+\gamma) + L_\alpha \cos\gamma - D_\alpha \sin\gamma}{m}\ddot{\alpha}_{\delta E} \end{bmatrix}.$$

The determinant of G^* is calculated as

$$\det(G^*) = \frac{T_{\delta T}\ddot{\delta T}_{com}\ddot{\alpha}_{\delta E}}{m^2}\cos\gamma(T + L_\alpha \cos\alpha + D_\alpha \sin\alpha)$$

where L_α, D_α and T_α denote the partial derivatives of L, D, and T with respect to the angle of attach α; $T_{\delta T}$ denotes the partial derivative of T with respect to the throttle setting δT. The nonsingular condition for G^* can be represented as

$$\det(G^*) \neq 0 \Leftrightarrow (T + L_\alpha \cos\alpha + D_\alpha \sin\alpha)\cos\gamma \neq 0.$$

Therefore, G^* is nonsingular unless the flight path is vertical or $(T + L_\alpha \cos\alpha + D_\alpha \sin\alpha) = 0$.

By assuming desired command-rates as zero, and using (15.13) and (15.14), we define a nonlinear coordinate transformation, $\xi = T_1(x, V^*)$ and $\eta = T_2(x, h^*)$, as

$$\begin{cases} \xi_1 = \int_0^t (V(\tau) - V^*)d\tau \\ \xi_2 = V - V^* \\ \xi_3 = \dot{V} \\ \xi_4 = \ddot{V} \end{cases} \quad \begin{cases} \eta_1 = \int_0^t (h(\tau) - h^*)d\tau \\ \eta_2 = h - h^* \\ \eta_3 = \dot{h} \\ \eta_4 = \ddot{h} \\ \eta_5 = h^{(3)} \end{cases}.$$

This results in decoupled subsystems

$$\dot{\xi} = A_1 \xi + b_1 v_1 \tag{15.15}$$

$$\dot{\eta} = A_2 \eta + b_2 v_2 \tag{15.16}$$

where

$$A_1 = \begin{bmatrix} 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 0\,0\,0\,0 \end{bmatrix}, b_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 0\,1\,0\,0\,0 \\ 0\,0\,1\,0\,0 \\ 0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,1 \\ 0\,0\,0\,0\,0 \end{bmatrix}, b_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

For the transformed linear systems, (15.15) and (15.16), we design the new inputs v_1 and v_2 as linear-quadratic control laws. Considering intermediate objective functions

$$J_1 = \int_0^\infty (\xi^T Q_1 \xi + r_1 v_1^2) dt \tag{15.17}$$

and

$$J_2 = \int_0^\infty (\eta^T Q_2 \eta + r_2 v_2^2) dt \tag{15.18}$$

the new input v_1 is derived by minimizing J_1 subject to (15.15):

$$v_1 = -r_1^{-1} b_1^T P_1 \xi$$

where P_1 is the positive-definite solution to the algebraic Riccati equation with design parameters Q_1 and r_1,

$$A_1^T P_1 + P_1 A_1 - r_1^{-1} P_1 b_1 b_1^T P_1 + Q_1 = 0, \; (Q_1 \succ 0, r_1 > 0).$$

Similarly, minimizing J_2 subject to (15.16) gives

$$v_2 = -r_2^{-1} b_2^T P_2 \eta$$

where P_2 is the positive-definite solution to the Riccati equation with design parameters Q_2 and r_2,

$$A_2^T P_2 + P_2 A_2 - r_2^{-1} P_2 b_2 b_2^T P_2 + Q_2 = 0, \; (Q_2 \succ 0, r_2 > 0).$$

The nonlinear control law u is obtained by inserting $v = \begin{bmatrix} v_1 \; v_2 \end{bmatrix}^T$ into (15.11),

$$u = -(G^*(x))^{-1} f^*(x) + (G^* x)^{-1} \begin{bmatrix} -r_1^{-1} b_1^T P_1 \xi \\ -r_2^{-1} b_2^T P_2 \eta \end{bmatrix}.$$

In the next step we consider the system robustness subject to the variations of the uncertain aerodynamic parameters. Appropriate Q_1, r_1, Q_2, and r_2 in the intermediate objective functions are found by minimizing the stochastic robustness cost function (15.12). For simplicity, we choose the design parameters $Q_1 = \mathrm{diag}\{q_1, q_2, q_3, q_4\}$ and $Q_2 = \mathrm{diag}\{q_5, q_6, q_7, q_8, q_9\}$, and the design parameter vector is

$$d = \begin{bmatrix} q_1, \, q_2, \, \cdots, \, q_9, \, r_1, \, r_2 \end{bmatrix}. \tag{15.19}$$

Satisfactory values of the eleven design parameters in (15.19) are computed by applying a genetic algorithm to Monte Carlo estimates of the stochastic robustness cost function (15.12).

15.4.2 Stochastic Robustness Analysis of the Design Result

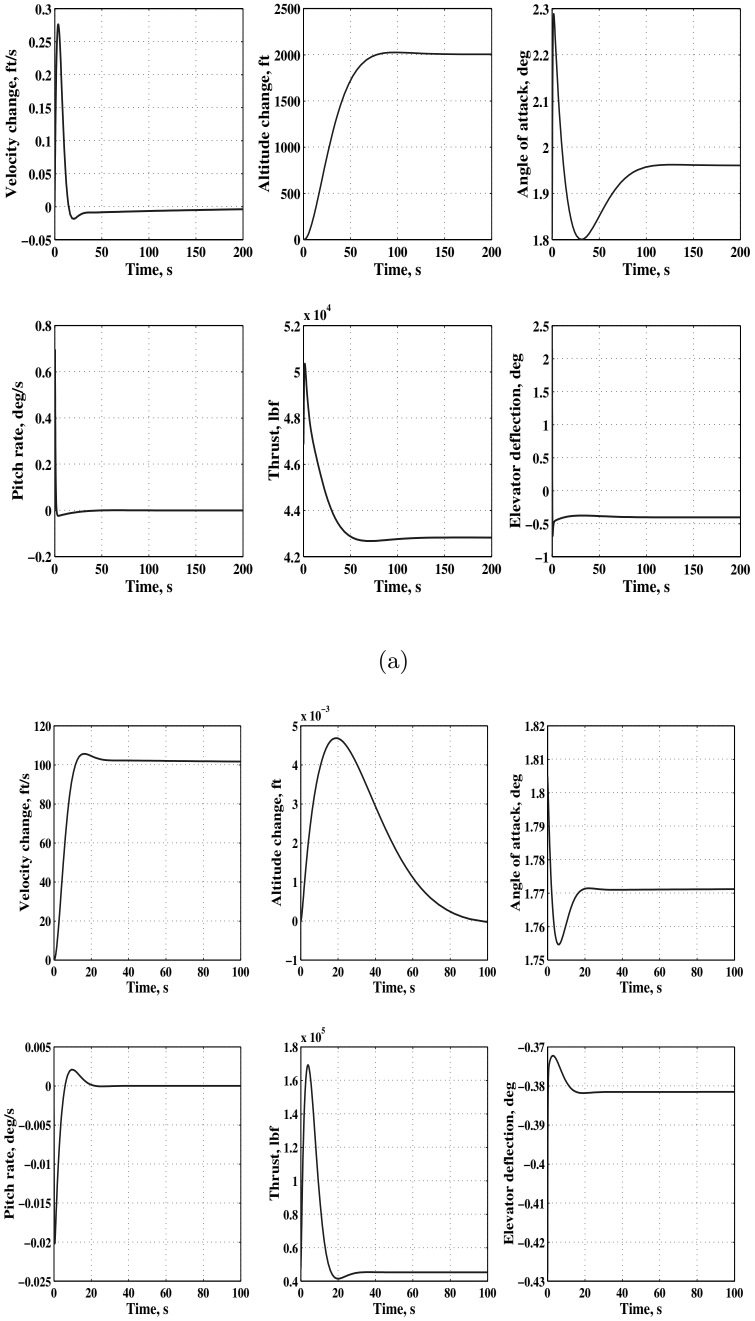
The design parameter vector (15.19), found by a genetic algorithm after 20 generations, is given as

$$\begin{aligned} d = [&8.54 \times 10^{-6}, 0.34, 0.86, 47.93, 1.1 \times 10^{-11}, 2.35 \times 10^{-3}, \\ &0.52, 220.6, 57.12, 0.89, 1.05]. \end{aligned} \tag{15.20}$$

The performance for the nominal closed-loop system is shown in Figure 15.2. Figure 15.2(a) shows the response due to a 100 ft/s step-velocity command from the trimmed condition ($V = 15,060$ ft/s, $h = 110,000$ ft). The velocity converges to the command value in 30 s with little change in altitude and with a change of angle of attack of less than 0.06°. We note that the use of thrust is unrealistically high, as there are limits to the thrust available. Nevertheless, the example illustrates the effectiveness of the design approach for the specified criteria. Figure 15.2(b) shows the velocity, altitude change, and control input time histories to a 2000 ft step-altitude command. The altitude converges to the command value in 75 s, with a change of angle of attack of less than 0.5°. Figure 15.2 demonstrates that the nominal system has good performance.

Figure 15.3 shows the robustness comparison of the current feedback linearization control (nonlinear dynamic inversion NDI) in (15.20) to a linear quadratic (LQ) design [219] based on 2000 Monte Carlo evaluations. The nonlinear design (NDI) has a cost of $J = 1.23$, while the linear design LQ has a cost of $J = 1.72$. The closed-loop probability of instability of the nonlinear design equals zero with a 95% confidence interval of $(0, 0.0018)$; it has 5% to 56% lower probability of exceeding settling time than the LQ design (Metrics 2-3 and 21-22) and 15% to 80% lower probability of exceeding rise time (Metrics 7-8 and 26-27). The nonlinear design has also reduced the probability of exceeding load factor by more than 80% compared to the LQ design (Metrics 13-14 and 32-33).

The NDI has larger probability of exceeding control effort corresponding to Metric 18, $I_{V,\delta T100}$, and Metric 36, $I_{h,\delta T50}$, due to the possibility that nonlinear dynamic inversion may cancel some useful nonlinearities. Furthermore, in (15.17) and (15.18), the weights r_1 and r_2 penalize large inputs v_1 and v_2 instead of penalizing thrust directly as in LQ. We can see that NDI performs better than LQ in Metric 19 (20), $I_{V,\delta E5}$ ($I_{V,\delta E10}$), and Metric 38 (39), $I_{h,\delta E5}$ ($I_{h,\delta E10}$). The robustness profiles can be adjusted by changing the weights in the robustness cost function. For example, trade-offs between using less thrust and accepting longer rise time, or putting heavy weight on P_{h,T_s} to decrease the probability of exceeding settling-time can be easily examined.



(a)

(b)

Figure 15.2. (a) Response to a 100 ft/s step-velocity command; (b) Response to a 2000 ft step-altitude command

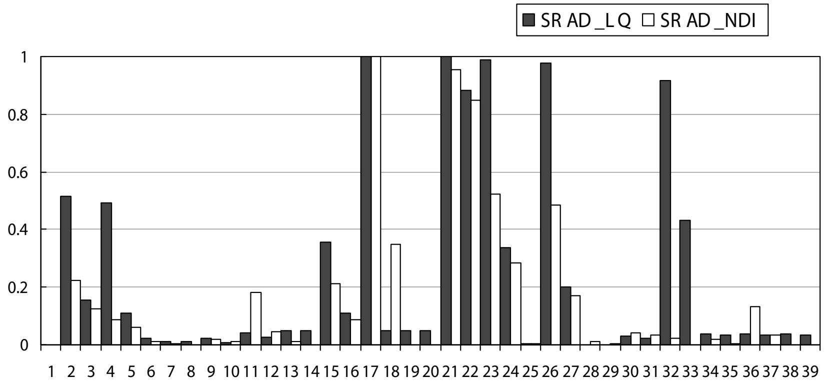


Figure 15.3. Comparison of the robustness profiles of the stochastic robust control based on linear quadratic regulator structure (LQ), and nonlinear dynamic inversion structure (NDI)

15.5 Stochastic Robust Control Design for the High Incidence Research Model

The HIRM aircraft configuration has canard and tailplane control surfaces plus an elongated nose. The mathematical model uses aerodynamic data obtained from wind tunnel and flight tests of an unpowered, scaled drop model. Engine, sensor, and actuator models have been added to the mathematical model to create a representative nonlinear simulation of a twin-engine modern fighter. The aircraft is basically stable both longitudinally and laterally, although there are some combinations of angle of attack and control surface deflections that cause the aircraft to be unstable. Reference [235] described in detail the six-degree-of-freedom nonlinear High Incidence Research Model including nonlinear actuator and sensor models. We first present the dynamic equations of motion for a general aircraft, and then address the aerodynamics for the HIRM problem.

15.5.1 System Model and Design Metrics

Equations of motion

The dynamic equations of motion for an aircraft in a combined wind and body axes are written as follows [123]:

$$\dot{V} = \frac{F_{wx}}{m} - g \sin \gamma \tag{15.21}$$

$$\dot{\alpha} = q - \frac{q_w}{\cos \beta} - p \cos \alpha \tan \beta - r \sin \alpha \tan \beta \tag{15.22}$$

$$
\begin{aligned}
\dot{\beta} &= r_w + p\sin\alpha - r\cos\alpha \qquad (15.23)\\
\dot{\gamma} &= q_w\cos\varphi - r_w\sin\varphi\\
\dot{\varphi} &= p_w + (q_w\sin\varphi + r_w\cos\varphi)\tan\gamma\\
\dot{\psi} &= \frac{q_w\sin\varphi + r_w\cos\varphi}{\cos\gamma}
\end{aligned}
$$

$$\dot{q} = \frac{1}{I_{yy}}[\mathcal{M} + I_{xz}(r^2 - p^2) + (I_{zz} - I_{xx})rp] \qquad (15.24)$$

$$\begin{bmatrix} \dot{p} \\ \dot{r} \end{bmatrix} = \begin{bmatrix} I_{xx} & -I_{xz} \\ -I_{xz} & I_{zz} \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{L} + I_{xz}pq + (I_{yy} - I_{zz})qr \\ \mathcal{N} - I_{xz}qr + (I_{xx} - I_{yy})pq \end{bmatrix} \qquad (15.25)$$

with

$$q_w = -\frac{F_{wz}}{mV} - \frac{g}{V}\cos\gamma\cos\varphi \qquad (15.26)$$

$$r_w = \frac{F_{wy}}{mV} + \frac{g}{V}\cos\gamma\sin\varphi \qquad (15.27)$$

$$p_w = p\cos\alpha\cos\beta + (q - \dot{\alpha})\sin\beta + r\sin\alpha\cos\beta \qquad (15.28)$$

$$\begin{bmatrix} \mathcal{L} \\ \mathcal{M} \\ \mathcal{N} \end{bmatrix} = \begin{bmatrix} \mathcal{L}_A \\ \mathcal{M}_A \\ \mathcal{N}_A \end{bmatrix} + \begin{bmatrix} \mathcal{L}_T \\ \mathcal{M}_T \\ \mathcal{N}_T \end{bmatrix} \qquad (15.29)$$

$$\begin{bmatrix} F_{wx} \\ F_{wy} \\ F_{wz} \end{bmatrix} = -\begin{bmatrix} D \\ S \\ L \end{bmatrix} + \begin{bmatrix} T_{wx} \\ T_{wy} \\ T_{wz} \end{bmatrix}. \qquad (15.30)$$

Here, V = flight path velocity, α = angle of attack, β = sideslip angle, (γ, φ, ψ) = wind-axis Euler angles, (p, q, r) = body-axis angular rates, (p_w, q_w, r_w) = wind-axis angular rates; $(\mathcal{L}, \mathcal{M}, \mathcal{N})$ = body-axis total rolling, pitching, and yawing moments, $(\mathcal{L}_A, \mathcal{M}_A, \mathcal{N}_A)$ = body-axis aerodynamic moments, $(\mathcal{L}_T, \mathcal{M}_T, \mathcal{N}_T)$ = body-axis moments due to engine thrust, (F_{wx}, F_{wy}, F_{wz}) = wind-axis total forces, (D, S, L) = drag, side, and lift force in wind axis, and (T_{wx}, T_{wy}, T_{wz}) = wind-axis thrust.

The transformation matrix from body axis to wind axis is defined as

$$L_{WB} = \begin{bmatrix} \cos\alpha\cos\beta & \sin\beta & \sin\alpha\cos\beta \\ -\cos\alpha\sin\beta & \cos\beta & -\sin\alpha\sin\beta \\ -\sin\alpha & 0 & \cos\alpha \end{bmatrix}.$$

The Mach number M is defined as the ratio of airspeed V and sound speed a, *i.e.* $M = V/a$.

Aerodynamics

Body-axis aerodynamic forces and moments, (F_{xA}, F_{yA}, F_{zA}) and $(\mathcal{L}_A, \mathcal{M}_A, \mathcal{N}_A)$, are represented in terms of the non-dimensional aerodynamic coefficients (C_X, C_Y, C_Z) and (C_l, C_m, C_n) as follows:

$$\begin{cases} F_x = \frac{1}{2}\rho V^2 S C_X \\ F_y = -\frac{1}{2}\rho V^2 S C_Y \\ F_z = \frac{1}{2}\rho V^2 S C_Z \end{cases}, \begin{cases} \mathcal{L}_A = \frac{1}{2}\rho V^2 S b C_l \\ \mathcal{M}_A = \frac{1}{2}\rho V^2 S \bar{c} C_m \\ \mathcal{N}_A = \frac{1}{2}\rho V^2 S b C n \end{cases}$$

where ρ denotes the air density, S denotes the aircraft's wing planform area, b denotes the span, and $\bar{c}$ denotes the mean aerodynamic chord. The aerodynamic force and moment coefficients are highly nonlinear functions of angle of attack α, sideslip angle β, airspeed V, angular rates p, q, r, and control deflections (symmetrical and differential taileron deflections δ_{TS} and δ_{TD}, symmetrical and differential canard deflections δ_{CS} and δ_{CD}, rudder deflection δ_R, and engine throttle δ_{TH}). Each component of the aerodynamic force and moment coefficients is represented by a look-up table. Details on the high-fidelity model can be found in [235].

Pilot commands

The pilot commands should control the responses as follows: lateral stick deflection commands velocity-vector roll rate p_{wc}, which is a roll performed at constant angle of attack and zero sideslip; longitudinal stick deflection commands pitch rate q_c; rudder pedal deflection commands sideslip angle β_c; throttle lever deflection commands velocity-vector air speed V_c, which represents a step command from its trim value V_{trim}.

Design envelope

The flight envelope that is specified by the GARTEUR/HIRM competition and used in comparison has Mach number within $(0.15, 0.5)$, angle of attack $(-10°, 30°)$, sideslip angle $(-10°\ , 10°)$, and altitude (100 ft, 20000 ft).

Modelling errors

The control system should be robust to the errors in the aerodynamic moment derivatives and to the biases in the total moment coefficients. The variation of C_{m_w} is within $(-0.001, 0.001)$, variation of C_{l_v} is within $(-0.01, 0.01)$, and the variation of C_{n_v} is within $(0.002, 0.002)$. The variations of C_{m_q}, C_{l_p}, C_{n_r}, C_{l_r}, C_{n_p}, $C_{m_{TS}}$, $C_{m_{CS}}$, $C_{l_{TD}}$, $C_{l_{CD}}$, $C_{l_{RUDDER}}$, $C_{n_{TD}}$, $C_{n_{CD}}$, and $C_{n_{RUDDER}}$ are within $(-10\%, 10\%)$ of the derivative's trim values. Though these uncertainties are proposed for linear analysis in [235], we include these aerodynamic-moment-derivative uncertainties in the assessment of nonlinear time responses. We assume that the uncertainties take uniform distributions in the designated ranges.

Formulation of the robustness metrics

In Table 15.2, we formulate robustness metrics in keeping with performance requirements in the assessments of a set of required maneuvers. All the robustness metrics are evaluated by Monte Carlo simulations with random number generators providing possible values of the uncertain aerodynamic parameters. It is assumed that the uncertain parameters take uniform distributions in the designated ranges.

Table 15.2. Formulation of robustness metrics

Metric	Weight in J	Indicator function	Design requirement
1	10	I_i	Stability at all flight conditions
Pitch rate command response			
2,3	1.0	$I_{3q\_qT_s}$ $I_{5q\_qT_s}$	10% settling time less than 2s for pitch rate command response at M=0.3 and 0.5
4,5,6	0.1	I_{2q_amax} I_{3q_amax} I_{5q_amax}	$-10^\circ < \alpha < 30^\circ$ for pitch rate demand response at M=0.2, 0.3, and 0.5
7,8,9	0.1	I_{2q_zmax} I_{3q_zmax} I_{5q_zmax}	$-3g < a_{nz} < 7g$ for pitch rate demand response at M=0.2, 0.3, and 0.5
Velocity command response			
10	1.0	$I_{3V\_VT_s}$	10% settling time less than 15s for velocity response at M=0.3
11	0.1	I_{3V_qt}	Pitch rate transient less than $10^\circ/s$ for velocity response at M = 0.3
Sideslip command response			
12,13,14	1.0	I_{2b} I_{3b} I_{5b}	Sideslip command response lies within specified boundaries at M=0.2, 0.3, and 0.5
Roll rate command response			
15,16	1.0	$I_{3p\_pT_s}$ $I_{5p\_pT_s}$	10% settling time less than 2s for roll rate command response at M=0.3 and 0.5
17,18	0.1	I_{3p_qt} I_{5p_qt}	Pitch rate transient less than $5^\circ/s$ for roll rate command response at M=0.3 and 0.5

In Table 15.2, the first indicator function, I_i, measures system stability. The system stability is evaluated in terms of the simulation of nonlinear time response. If all of the step command responses listed in Table 15.2 do not have finite escape time, we specify $I_i = 0$; otherwise, $I_i = 1$. Indicator functions 2-9 characterize the nonlinear time responses to step pitch-rate commands at different flight conditions. The angles of attack during pitch-rate commands should be within the specified limits with maximum overshoot less than 5°. The normal acceleration should be within the specified limits with maximum overshoot less than 0.5g. The settling time requirement is not specified for the pitch-rate response at $M = 0.2$ because the necessity of an angle-of-attack limiter could cause transients of the pitch rate. Indicator functions 10-11 characterize the step velocity command response at $M = 0.3$. Indicator functions 12-14 are for sideslip-angle command responses. The step response to sideslip command should lie within some specified boundaries [235]. Indicator functions 15-18 illustrate the requirements for roll-rate command responses.

The stochastic robustness cost function chosen to guide the design is a weighted quadratic sum of the eighteen probabilities of design metric violations:

$$J = \sum_{j=1}^{18} w_j P_j^2.$$

The weight for each probability is given in Table 15.2.

15.5.2 Controller Structure

The design of the controller structure is based on nonlinear dynamic inversion. It is possible to separate system dynamics into two time scales if one subset of the state components (referred to as 'fast dynamics') is known to evolve in a much faster time scale than the other subset (referred to as 'slow dynamics'). The inversion performed here is based on the assumption that the dynamics of angular rates are faster than those of angles of attack and sideslip. The design of controller structure is separated into two steps relating to the slow and fast dynamics.

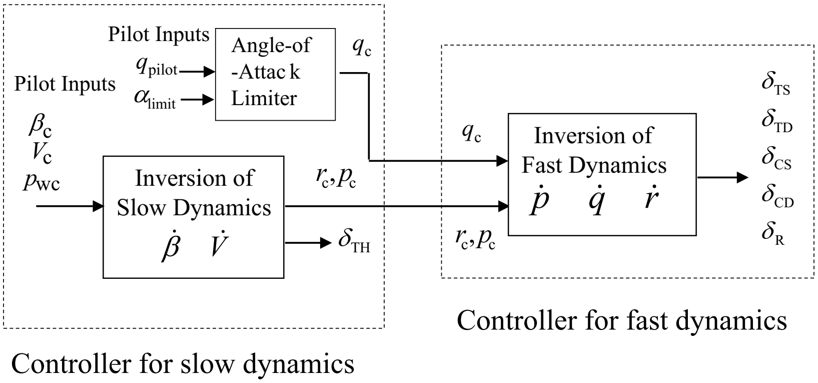


Figure 15.4. Controller structure designed using two-time-scale nonlinear dynamic inversion

For the slow dynamics, commanded angular rates are derived through either direct pilot inputs or the inversion of the force equations. The engine throttle position is derived through the inversion of the velocity dynamics. The values of yaw rate and engine throttle are obtained in terms of design parameters that characterize desired dynamics of sideslip angle and velocity. For the fast dynamics, control surface deflections are derived explicitly through the inversion of a first-order differentiation of angular velocities. They are defined in terms of design parameters that characterize desired dynamics of angular rates. The procedure of this two-time-scale nonlinear dynamic inversion is illustrated in Figure 15.4.

Slow dynamics

Design of the controller for slow dynamics shown in Figure 15.4 deals with force equations and the kinematics' equation for velocity-vector roll rate. The purpose of this inversion is to derive command angular rates (p_c, r_c) for the fast dynamics from the pilot commands (p_{wc}, β_c), and to derive engine throttle position δ_{TH} from the pilot command velocity V_c.

First, we rewrite the equations for $\dot{\alpha}$, $\dot{\beta}$, $\dot{V}$, and p_w in appropriate forms. The wind-axis thrust induced by the two engines is derived from the body-axis thrust:

$$\begin{bmatrix} T_{xw} \\ T_{yw} \\ T_{zw} \end{bmatrix} = L_{WB} \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} = L_{WB} \begin{bmatrix} 2F_E \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2F_E \cos\alpha \cos\beta \\ -2F_E \cos\alpha \sin\beta \\ -2F_E \sin\alpha \end{bmatrix}. \qquad (15.31)$$

By (15.31), equation (15.30) becomes

$$\begin{bmatrix} F_{wx} \\ F_{wy} \\ F_{wz} \end{bmatrix} = \begin{bmatrix} -D + 2F_E \cos\alpha \cos\beta \\ -S - 2F_E \cos\alpha \sin\beta \\ -L - 2F_E \sin\alpha \end{bmatrix}.$$

We define wind axis load factors as

$$n_{wx} = \frac{F_{wx}}{mg} = \frac{-D + 2F_E \cos\alpha \cos\beta}{mg}$$

$$n_{wy} = \frac{F_{wy}}{mg} = \frac{-S - 2F_E \cos\alpha \sin\beta}{mg}$$

$$n_{wz} = \frac{F_{wz}}{mg} = \frac{-L - 2F_E \sin\alpha}{mg}.$$

Equations (15.26) and (15.27) are rewritten in terms of wind-axis load factor as,

$$q_w = -\frac{g}{V}(\cos\gamma \cos\varphi + n_{wz}) \qquad (15.32)$$

$$r_w = \frac{g}{V}(\cos\gamma \cos\varphi + n_{wy}). \qquad (15.33)$$

By setting $\dot{\alpha}$ and β to zero in (15.28), we have

$$p_{wc} = (p \cos\alpha + r \sin\alpha). \tag{15.34}$$

With (15.33) and (15.34), equation (15.23) becomes

$$\dot{\beta} = -\frac{r}{\cos\alpha} + p_{wc} \tan\alpha + \frac{g}{V}(n_{wy} + \cos\gamma\cos\varphi)\cos\alpha. \tag{15.35}$$

By (15.30), equation (15.21) becomes

$$\dot{V} = \frac{2F_E \cos\alpha\cos\beta - D}{m} - g\sin\gamma. \tag{15.36}$$

Next, we formulate the state and control inputs for the slow dynamics. Integral compensation is used to minimize steady-state error of the command response. Therefore, we define new state variables

$$V_I = \int_0^t [V(\tau) - (V_{trim} + V_c)]d\tau$$

$$\beta_I = \int_0^t (\beta(\tau) - \beta_c)d\tau.$$

The corresponding augmented state vector for slow dynamics is defined as:

$$x_s = \begin{bmatrix} V_I \; V \; \beta_I \; \beta \end{bmatrix}^T.$$

The dynamic model for x_s is

$$\begin{bmatrix} \dot{V}_I \\ \dot{V} \\ \dot{\beta}_I \\ \dot{\beta} \end{bmatrix} = \begin{bmatrix} V - (V_{trim} + V_c) \\ -2\xi_V\omega_V[V - (V_{trim} + V_c)] - \omega_V^2 V_I \\ \beta - \beta_c \\ -2\xi_\beta\omega_\beta(\beta - \beta_c) - \omega_\beta^2\beta_I \end{bmatrix} \tag{15.37}$$

where ξ_V, ω_V, ξ_β, and ω_β are design parameters. ξ_V and ω_V denote the desired damping ratio and frequency for velocity dynamics, while ξ_β and ω_β denote the desired damping ratio and frequency for the dynamics of sideslip angle.

The control vector for slow dynamics consists of the thrust of each engine F_E and the commanded yaw rate r_c for the fast dynamics. Utilizing (15.35), equations (15.36) and (15.37), we derive the control vector $u_s = \begin{bmatrix} F_E \; r_c \end{bmatrix}^T$,

$$F_E = \frac{m}{2\cos\alpha\cos\beta}\left\{\frac{D}{m} + g\sin\gamma - 2\xi_V\omega_V[V - (V_{trim} + V_c)] - \omega_V^2 V_I\right\} \tag{15.38}$$

$$\begin{aligned} r_c = p_{wc}\cos\beta\sin\alpha + \frac{g}{V}(n_{wy} + \cos\gamma\sin\varphi)\cos\alpha \\ + [2\xi_\beta\omega_\beta(\beta - \beta_c) + \omega_\beta^2\beta_I]\cos\alpha. \end{aligned} \tag{15.39}$$

By (15.34) and (15.39), we derive the commanded roll rate p_c for the fast dynamics as

$$p_c = p_{wc}\cos\beta\cos\alpha - \frac{g}{V}(n_{wy} + \cos\gamma\sin\varphi)\sin\alpha \\ + [-2\xi_\beta\omega_\beta(\beta - \beta_c) - \omega_\beta^2\beta_I]\sin\alpha. \qquad (15.40)$$

In terms of the engine model in [235], the throttle position is

$$\delta_{TH} = \begin{cases} \frac{F_E\frac{\rho_0}{\rho} - F_{IDLE}}{F_{MD} - F_{IDLE}}, & F_E\frac{\rho_0}{\rho} < F_{MD} \\ 1 + \frac{F_E\frac{\rho_0}{\rho} - F_{MD}}{F_{MR} - F_{MD}}, & F_{MD} \le F_E\frac{\rho_0}{\rho} \le F_{MR} \end{cases}$$

with F_E given by (15.38). F_{IDLE}, F_{MD}, and F_{MR} denote the idle thrust, maximum dry thrust, and maximum reheat thrust for the engine.

The computation of r_c, p_c, and F_E is conducted as follows. Through the transformation from body axes to wind axes L_{WB}, the wind-axis load factor n_{wy} in (15.39) and (15.40) is calculated from the body-axis accelerations a_{nx}, a_{ny}, and a_{nz}, which are measured variables. Also through L_{WB}, drag D in (15.38) is calculated from body-axis aerodynamic forces F_{xA}, F_{yA}, and F_{zA}, which are computed in terms of the aerodynamic force coefficients C_X, C_Y and C_Z by (15.31). The calculation of C_X, C_Y and C_Z depends on the values of control surface deflections, which are unknown and are computed in the phase of fast dynamics. In this chapter, the values of control surface deflections of the previous time iteration are used in computing aerodynamic force coefficients C_X, C_Y, and C_Z.

An angle-of-attack limiter is important because the commanded pitch rate q_c, which is an input for the fast dynamics, should be chosen as the minimum of the pilot-commanded pitch rate q_{pilot} and the pitch rate q_{limit} that would induce the maximum allowable angle of attack α_{limit},

$$q_c = \min(q_{pilot}, q_{limit}).$$

In terms of equations (15.22) and (15.32), q_{limit} is derived as

$$q_{limit} = (p\cos\alpha + r\sin\alpha)\tan\beta \\ - \frac{g}{V}\frac{n_{wz} + \cos\gamma\cos\varphi}{\cos\beta} + \dot{\alpha}_{LIM} \qquad (15.41)$$

where the maximum allowable angle-of-attack rate, $\dot{\alpha}_{LIM}$, is calculated from

$$\dot{\alpha}_{LIM} = -\omega_\alpha(\alpha - \alpha_{limit})$$

where ω_α is a design parameter that denotes the bandwidth of the angle-of-attack control loop; α is the current angle of attack. The limit of angle of attack α_{limit} equals 30°.

Fast dynamics

Design of the controller corresponding to the fast dynamics in Figure 15.4 consists of the inversion of the moment equations. The purpose of this inversion is to derive a vector of control surface deflections for a given set of commanded angular rates p_c, q_c and r_c.

Integral compensation minimizes the steady-state error of the pitch rate command response; thus we define a new state variable,

$$q_I = \int_0^t (q(\tau) - q_c) d\tau.$$

The state vector for the fast dynamics is

$$x_f = \begin{bmatrix} p \; r \; q_I \; q \end{bmatrix}^T.$$

The dynamic model for angular rates is

$$\dot{p} = -\omega_p (p - p_c) \tag{15.42}$$

$$\dot{r} = -\omega_r (r - r_c) \tag{15.43}$$

$$\dot{q} = -2\xi_q \omega_q (q - q_c) - \omega_q^2 q_I \tag{15.44}$$

where ξ_q, ω_q, ω_p, and ω_r are design parameters. ξ_q and ω_q denote the desired damping ratio and frequency for the dynamics of pitch rate while ω_p and ω_r denote the desired bandwidths for p and r.

The vector of control inputs for the fast dynamics consists of control surface deflections of the taileron, canard, and rudder:

$$u_f = \begin{bmatrix} \delta_{TS} \; \delta_{TD} \; \delta_{CS} \; \delta_{CD} \; \delta_R \end{bmatrix}^T.$$

From (15.24), (15.25), (15.29), and (15.42), (15.43), (15.44), we have

$$\begin{bmatrix} \mathcal{L}_A \\ \mathcal{N}_A \\ \mathcal{M}_A \end{bmatrix} = - \begin{bmatrix} \mathcal{L}_T \\ \mathcal{N}_T \\ \mathcal{M}_T \end{bmatrix} + \begin{bmatrix} -I_{xz} pq + (I_{zz} - I_{yy}) qr \\ I_{xz} qr + (I_{yy} - I_{xx}) pq \\ (I_{xx} - I_{zz}) rp + I_{xz} (p^2 - r^2) \end{bmatrix} + \begin{bmatrix} I_{xx} & -I_{xz} & 0 \\ -I_{xz} & I_{zz} & 0 \\ 0 & 0 & I_{yy} \end{bmatrix} \begin{bmatrix} -\omega_p (p - p_c) \\ -\omega_r (r - r_c) \\ -2\xi_q \omega_q (q - q_c) - \omega_q^2 q_I \end{bmatrix}. \tag{15.45}$$

Note that the aerodynamic moments $\mathcal{L}_A$, $\mathcal{N}_A$, and $\mathcal{M}_A$ are nonlinear functions of the control surface deflections u_f; the inverse mappings of these nonlinear functions have to be calculated in order to derive the control surface deflections u_f. For simplicity of calculation, we approximate the aerodynamic moments by their first-order expansions with respect to control surface deflections around the values of control surface deflections at the previous time iteration:

$$\begin{bmatrix} \mathcal{L}_A \\ \mathcal{N}_A \\ \mathcal{M}_A \end{bmatrix} \cong \frac{1}{2} \rho V^2 \bar{S} \Lambda (u_f - u_f^*) + \frac{1}{2} \rho V^2 \bar{S} \Upsilon. \tag{15.46}$$

Matrices Λ and Υ, which are functions of the control surface deflections at the previous time iteration u_f^*, are given in [389].

Note that in (15.46), we have more unknown variables (u_f consists of five control surface deflections) than equations (three equations); hence, the solution of u_f is not unique. We derive the control u_f in terms of $\Lambda^\sharp$, which is the pseudo-inverse of matrix Λ,

$$u_f = u_f^* + \Lambda^\sharp \left\{ \frac{1}{\frac{1}{2}\rho V^2 \bar{S}} \begin{bmatrix} \mathcal{L}_A \\ \mathcal{N}_A \\ \mathcal{M}_A \end{bmatrix} - \Upsilon \right\}$$

where $\begin{bmatrix} \mathcal{L}_A \; \mathcal{N}_A \; \mathcal{M}_A \end{bmatrix}^T$ is given by (15.45). The (right) pseudo-inverse operation used here corresponds to a minimization of the normalized control surface deflections.

We concatenate the control design parameters in (15.37) and (15.42), (15.43), (15.44) into a single design vector as

$$d = \begin{bmatrix} \xi_V \; \omega_V \; \xi_\beta \; \omega_\beta \; \omega_\alpha \; \xi_q \; \omega_q \; \omega_p \; \omega_r \end{bmatrix}^T . \tag{15.47}$$

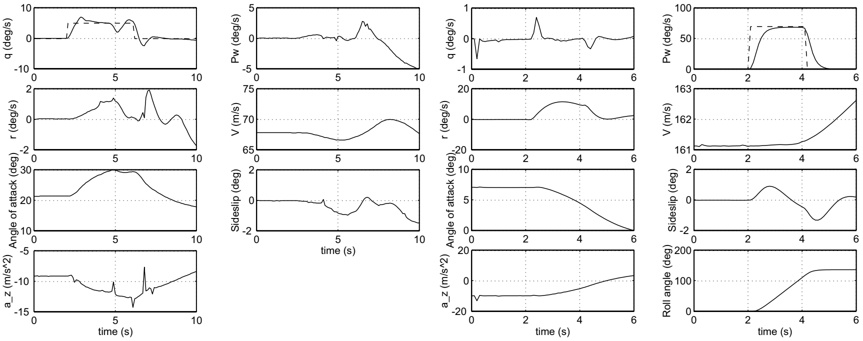
15.5.3 Control Design Results

The design parameter vector in (15.47) for our robust HIRM controller is found by using a genetic algorithm as follows:

$$d = \begin{bmatrix} 0.419 \; 1.046 \; 2.872 \; 0.489 \; 4.983 \; 1.448 \; 3.063 \; 4.023 \; 2.663 \end{bmatrix}^T . \tag{15.48}$$

The performance of the nominal closed-loop system is illustrated by a set of maneuvers in Figures 15.5 and Figure 15.6; the time responses for other maneuvers can be found in [210]. The figures show histories of the command variables and state variables of interest. The command values of pitch-rate, velocity-vector-roll-rate, airspeed, and sideslip angle are plotted using dashed lines. The response to command is good in all cases. For the $5°/s$ pitch rate commanded response at $M = 0.2$, Figure 15.5(a) shows angle of attack being limited to the maximum value, 30°. The pitch-rate transient that occurs at $t = 5$ s is due to this limiting. With the increase of the pitch attitude, the gravitational force component from the mass of the aircraft induces an additional force in the wind x-axis that results in the variation of the airspeed. The thrust is increased to compensate for the change in attitude. For the $70°/s$ roll rate commanded response at $M = 0.5$, Figure 15.5(b) shows good performance. The roll rate follows the command input quite well, with 10% settling time less than 2 s. The coupling to sideslip angle is less than 1.5°, and the coupling to pitch rate is less than $1°/s$.

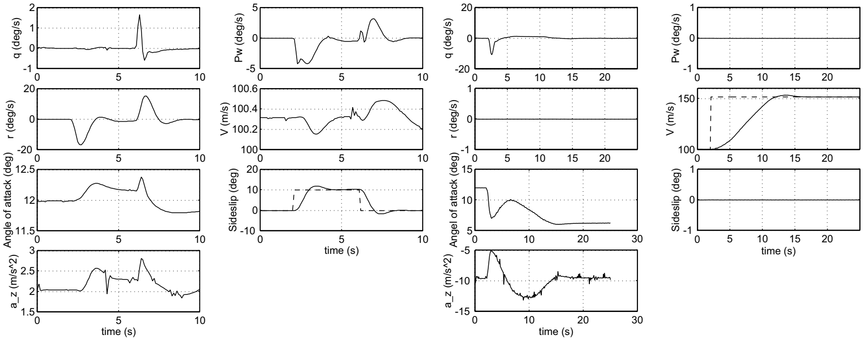
Figure 15.6(a) illustrates the responses due to a $10°/s$ step command on sideslip angle at $M = 0.3$. The time history of the sideslip angle is well within the specified boundaries. It follows the command input with 10% settling time of less than 2 s. The couplings into roll and pitch rate are low. Figure 15.6(b) shows a 51.48 m/s (100 kn) step on velocity commanded response at $M = 0.3$.



(a)

(b)

Figure 15.5. (a) Pitch rate command response at M = 0.2; (b) Roll rate command response at M = 0.5



(a)

(b)

Figure 15.6. (a) Sideslip angle command response at M = 0.3; (b) Velocity command response at M = 0.3

The 10% settling time is less than 15 s, and the overshoot is within 3%. The pitch rate transient is low and returns to zero quickly. The engine is fully used for the rapid speed command change. The maximum throttle position is attained. The noise in the time history of normal acceleration a_z is due to the relatively high bandwidth of the velocity. The control system shows good performance for the entire flight envelope including extreme flight conditions such as 30° angle of attack. It is demonstrated that the controller has strong ability to account for significant nonlinearities.

15.5.4 Comparison of Present Design with Controllers Developed for GARTEUR Competition

A set of control designs has been presented for the HIRM control challenges in the GARTEUR competition [210]. They include controllers based on linear-quadratic (LQ) methods [4], $\mathcal{H}_\infty$ loop-shaping approaches [258], μ-synthesis [149, 213], nonlinear dynamic inversion combined with linear-quadratic regulator (NDI/LQ) [122], and robust inverse dynamics estimation approaches (RIDE) [236]. The first three design approaches are linear techniques. Gain scheduling of linear feedback gains was utilized to cover the whole operating envelope of the aircraft. Reference [122] used a two-level controller structure consisting of a nonlinear-dynamic-inversion feedforward controller and a linear-quadratic feedback controller. In [122], the simulations for the nonlinear time responses were performed with the nonlinear-dynamic-inversion feedforward controller alone, without the linear-quadratic correction. Reference [236] combined dynamic inversion with proportional and integral feedback loops. Robustness issues were not directly taken into account in [236].

It is difficult to compare the present controller against the designs presented in the GARTEUR competition because they were not intended to minimize the probabilities of metric violations subjected to expected parameter variations, as is the present design. In the evaluation software provided by GARTEUR, a single set of values of uncertain parameters is used to test a control system's robustness (deterministic characterization of uncertainties). Furthermore, very limited simulation results were presented for each design. Nevertheless, we provide a comparison of the present controller with the earlier designs based on the available information.

Performance in nominal control responses

For each design in the GARTEUR competition, maneuver simulations are offered only at some of the flight conditions. There are no results shown for the commanded time responses in the presence of parameter uncertainties. A comparison of the performance of nominal time responses for a set of maneuvers between the present controller and previous designs is given in Table 15.3.

In Table 15.3, T_s denotes a 10% settling time for a command response. T_w represents the overshoot wash-out time for the angle of attack above its limiting value. A 2s wash-out time is required. We use '-' to denote unavailable results. Inadequate performances of each control design are highlighted.

The linear-quadratic design has quite good performance except that there is a slight excess of overshoot in the velocity command response, compared to the desired specification of less than 3%. The $\mathcal{H}_\infty$ loop-shaping controller has excess wash-out time for angle-of-attack overshoot above 30° in the pitch-rate command response, excess steady-state offsets of the roll-rate command response, and excess overshoot in the velocity command response. The first μ-synthesis design has large steady-state offsets for the pitch-rate command response and excess settling time for the roll-rate command response. The second μ-synthesis design has very good performance, except the settling time is longer than the required two seconds for the pitch rate command response. The NDI/LQ design has large overshoot in the velocity command response; otherwise, it demonstrates excellent nominal performance. The RIDE design has no overshoot in velocity, but there are slight steady-state offsets, and it has relatively long settling time for the sideslip-angle command response. Compared to previous designs in the GARTEUR competition, the controller presented in this chapter shows less overshoot in the velocity command response, faster response in all the maneuvers, and more accurate tracking of the commands without steady state offsets.

Table 15.3. Comparison of nominal performance for a set of controllers ('o.s.' denotes overshoot)

	LQ	$\mathcal{H}_\infty$	μ-1	μ-2	NDI/LQ	RIDE	Present
Pitch rate command responses							
M=0.2	-	$\alpha > 30°$ w/ 1.5° o.s., $T_w > 2s$	-	-	$\alpha \leq 30°$ w/o o.s.	-	$\alpha \leq 30°$ w/o o.s.
M=0.3	$T_s < 2s$	-	$T_s < 2s$	$T_s > 2s$	-	-	$T_s < 2s$
M=0.4	-	-	-	-	-	$\alpha \leq 30°$ w/o o.s.	-
M=0.5	-	-	q offset = 14%	-	-	-	$T_s < 2s$
Roll rate command responses							
M=0.3	$T_s < 2s$ $\lvert q\rvert < 7°/s$ $\lvert\beta\rvert < 4°$	$\lvert q\rvert < 5°/s$ $\lvert\beta\rvert < 1.2°$ p offset = 20%	$T_s > 2s$ $\lvert q\rvert < 8°/s$ $\lvert\beta\rvert < 0.7°$	-	-	- $\lvert\beta\rvert < 2°$	$T_s < 2s$ $\lvert q\rvert < 4°/s$
M=0.4	-	-	-	-	$T_s < 2s$ $\lvert q\rvert < 2°/s$	$T_s < 2s$, $\lvert q\rvert < 3°/s$	-
M=0.5	-	-	-	$T_s < 2s$ $\lvert q\rvert < 1°/s$	-	-	$T_s < 2s$ $\lvert q\rvert < 1°/s$
Sideslip angle command response							
M=0.2	-	-	-	-	-	-	$T_s < 2s$
M=0.3	$T_s < 2s$	-	-	-	-	-	$T_s < 2s$
M=0.4	-	-	-	-	$T_s < 2s$	$T_s > 2s$	-
M=0.5	-	$T_s < 2s$	-	-	-	-	$T_s < 2s$
Velocity command response							
M=0.3	6.7% o.s.	8% o.s.	-	-	20% o.s. 6% offset	w/o o.s. 4.5% offset	< 3% o.s.

Performance robustness in linear frequency responses with parametric uncertainties

In the GARTEUR competition, the evaluation software analyzes linear frequency responses of controllers in the presence of parametric uncertainties in moment derivatives. Linear frequency specifications have less value for our nonlinear control law; therefore, we do not include them in the formulation of our cost function. Nevertheless, our controller is evaluated against linear frequency requirements specified in the GARTEUR competition for comparison with earlier designs. The open-loop Nichols plot of the frequency response between each actuator demand u and the corresponding error signal e should avoid a gain-phase exclusion region. The evaluation is made in the presence of parametric uncertainties as: $C_{m_v} = -0.001$, $C_{l_v} = -0.01$, $C_{n_v} = -0.002$, C_{l_r}, $C_{n_p} = 10\%$, and C_{m_q}, C_{l_p}, C_{n_r}, $C_{m_{TS}}$, $C_{m_{CS}}$, $C_{l_{TD}}$, $C_{l_{CD}}$, $C_{l_{RUDDER}}$, $C_{n_{TD}}$, $C_{n_{CD}}$, $C_{n_{RUDDER}} = -10\%$.

Open-loop Nichols plots for the present controller with parametric uncertainties are plotted in Figure 15.7 for a flight condition at Mach 0.24, 20000 ft altitude, 28.9° angle of attack, and zero sideslip angle. This flight condition represents an edge of the flight envelope, which is likely to cause stability and actuator-limiting problems. Figure 15.7 shows that the frequency responses for all of the six control loops (differential and symmetrical taileron loops; differential and symmetrical canard loops; rudder loop, and thrust loop) avoid the specified gain-phase exclusion zone. In comparison to existing designs, for each of the controllers except the NDI/LQ and $\mathcal{H}_\infty$ (lack of robustness information) in the GARTEUR competition, one loop's linear frequency response cannot satisfy the robustness criteria. We conclude that the nonlinear controller of this chapter shows better performance robustness than the earlier designs as portrayed by linear frequency analysis.

Linear frequency analysis is inadequate for evaluating nonlinear dynamic systems and nonlinear control laws. Furthermore, a single set uncertainty that is not proved to be the worst case for the parametric uncertainties is not enough to quantify system robustness. Two thousand Monte Carlo evaluation of the present design with controller parameters in (15.48) give the probabilistic robustness profile in Figure 15.8. The confidence interval for each probability is not shown due to space limitations and can be found in [389]. In the Monte Carlo simulations, random number generators with uniform distributions provide the possible values of the system uncertain parameters. The design cost equals 1.14. The control system has a zero probability of instability (Metric 1) with a 95% confidence interval of $(0, 0.0018)$. For the pitch-rate command response at $M = 0.2$, adding the angle-of-attack limiter causes transients in pitch rate; therefore, the settling-time specification is not evaluated. The pitch-rate command response at $M = 0.3$ is quite good, with low probability of excess settling time (Metric 2, $I_{3q\text{-}qT_s}$). The probability of violating settling-time condition at $M = 0.5$ (Metric 3, $I_{5q\text{-}qT_s}$) is more than double the probability at $M = 0.3$. It is within expectation because $M = 0.2$

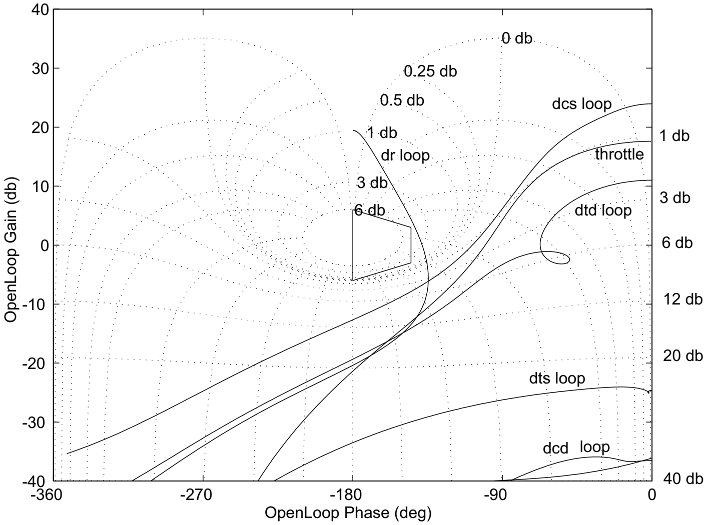


Figure 15.7. Open loop Nichols plots of the present controller in the presence of parametric uncertainties with a flight condition at M = 0.24. The trapezoid denotes the gain-phase exclusion region.

and 0.5 represent edge-of-the-envelope flight conditions, and $M = 0.3$ represents a nominal flight condition within the envelope. The probabilities of exceeding angle-of-attack and normal-acceleration limits in pitch-rate command responses equal zero (Metric 4-9) for all flight conditions with 95% confidence intervals of $(0, 0.0018)$. Figure 15.8 shows that the probability of exceeding settling time for the velocity-command response is relatively high (Metric 10, $I_{3V\_VT_s}$), which is caused by the uncertainties in yawing moments and derivatives. The probability of pitch-rate coupling for velocity command is low (Metric 11, I_{3V_qt}). The performance robustness for sideslip-angle command responses is fine for each flight condition. The probabilities of violating settling time condition are about 20% (Metrics 12-14). For roll-rate command responses, there are about 30% probability of excess settling time (Metrics 15-16) and less than 20% probability of pitch-rate coupling for all flight conditions (Metrics 17-18).

15.5.5 Effects on Robustness Profiles by Changing Weights in the Robustness Cost Function

Trade-offs between satisfying different aspects of robustness can be balanced through changing the weights in the robustness cost function. In this section, the controller structure is unchanged, and the weights for the pitch-rate settling-time metric $I_{5q\_qT_s}$, roll-rate settling-time metrics $I_{3p\_pT_s}$ and $I_{5p\_pT_s}$ are increased to 10. The new design based on the cost function with modified

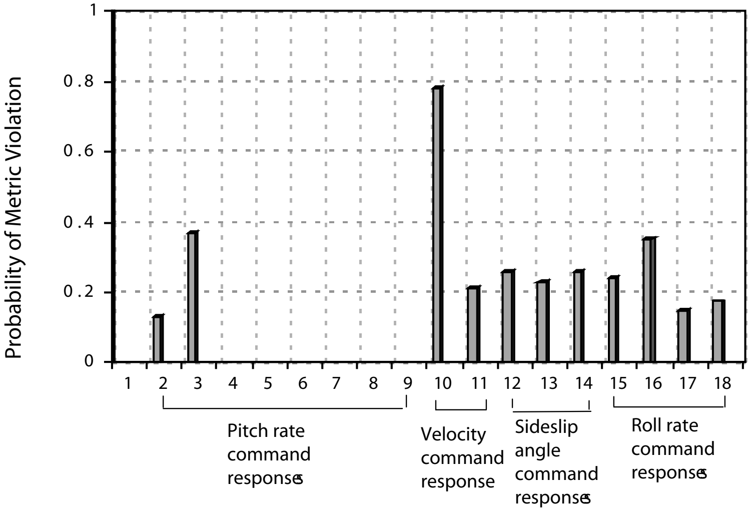


Figure 15.8. Robustness profile of the present controller for the HIRM challenge

weights is obtained as

$$d = \left[0.7529\ 0.6514\ 0.8099\ 0.5753\ 4.95\ 1.233\ 2.951\ 4.165\ 3.51\right]^T \qquad (15.49)$$

Figure 15.9 shows the variations in the robustness profile of designs due to different weights in the robustness cost function. In Figure 15.9, white bars (weight_1) represent the probabilities of violating design metrics for the design in (15.48), and dark bars (weight_2) denote the probabilities for the design in (15.49). Figure 15.9 shows that the probabilities of violating $I_{5q\_qT_s}$, $I_{3p\_pT_s}$, and $I_{5p\_pT_s}$ (Metrics 3, 15 and 16) have decreased by almost two thirds. The probability of violating $I_{3q\_qT_s}$ (Metric 2), and the probabilities of violating I_{3p_qt} and I_{5p_qt} (Metrics 17-18) have fallen to zero. However, the improvement in robustness for these metrics is achieved at the expense of increasing the probability of violating some other metrics. It is shown that the probabilities of violating requirements in sideslip-angle command responses (Metrics 12-14) are doubled, and the probability of violating the settling-time requirement in the velocity command response (Metric 10) has increased, too.

This comparison illustrates the limitation of redesign within a fixed controller structure. Changing cost function weights can improve specific responses, but it may do so at the expense of degrading the robustness of other responses. Comparing the original design vector (15.48) with the revised design vector (15.49), we see that the improved pitch and roll-rate responses led to higher airspeed and sideslip-angle damping, lower airspeed bandwidth, and stiffer yaw rate response. Further improvements would require revisions to the specified structures for slow and fast controllers.

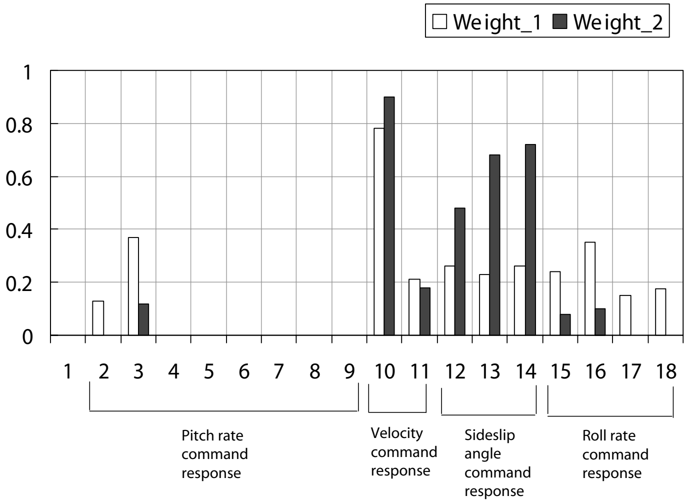


Figure 15.9. Comparison of the robustness profiles of two designs with different weights in the robustness cost function

15.6 Conclusion

Stochastic robustness analysis and synthesis break the computation complexity barrier suffered by deterministic worst-case approaches; Monte Carlo simulation and randomized search have polynomial complexity in computation. Instead of trying to guarantee that stability and performance specifications are satisfied for the worst-case scenario, the stochastic approach minimizes the likelihood of violating design requirements in the presence of expected variations in plant parameters. By focusing on the uncertainties most likely to occur in real engineering problems, the stochastic approach avoids undue conservativeness that could sacrifice nominal performance, cause extra controller complexity, and increase the possibility for control saturation. With Monte Carlo evaluation of probabilities of violating design metrics as an inherent feature of the control design process, a wide range of design specifications can be taken into account. Randomized algorithms such as genetic algorithms allow efficient tuning of design parameters for control problems formulated in a general and realistic fashion. The robustness profile of the final design and the choice of weights in the cost function provide sufficient information and flexibility for engineers to make tradeoffs between satisfying different aspects of robustness.

A stochastic robust nonlinear control design methodology is proposed by combining probabilistic robustness with feedback linearization (nonlinear dynamic inversion). The proposed approach is demonstrated through two design examples for robust flight control systems, where the high-fidelity models contain large dimensional uncertain parameters and complicated design specifications. The combination of stochastic robustness with nonlinear control design methodologies provides the ability to account for all significant nonlinearities and to produce better stability and performance robustness than linear robust control design with gain scheduling. The approach also reduces the complexity of control systems and the possibility of control saturation compared to the deterministic worst-case approaches to nonlinear robust control. It demonstrates engineering utility in addition to pure mathematical beauty, enhances the applicability of modern control theories, and reduces the gap between theory and practice.

16

Fast Randomized Algorithms for Probabilistic Robustness Analysis

Xinjia Chen, Kemin Zhou, and Jorge Aravena

Department of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, LA 70803
{chan,kemin,aravena}@ece.lsu.edu

Summary. In this chapter, we develop efficient randomized algorithms for estimating probabilistic robustness margin and constructing robustness degradation curve for uncertain dynamic systems. One remarkable feature of these algorithms is their universal applicability to robustness analysis problems with arbitrary robustness requirements and uncertainty bounding sets. We have developed efficient methods such as probabilistic comparison, probabilistic bisection, backward iteration and sample reuse to facilitate the computation. In particular, confidence interval for binomial random variables has been frequently used in the estimation of probabilistic robustness margin and in the accuracy evaluation of estimating robustness degradation function. Motivated by the importance of fast computation of binomial confidence interval in the context of probabilistic robustness analysis, we have recently derived an explicit formula for constructing the confidence interval of binomial parameter with guaranteed coverage probability. The formula overcomes the limitation of normal approximation which is asymptotic in nature and thus inevitably introduces unknown errors in applications. Moreover, the formula is extremely simple and very tight in comparison with classic Clopper-Pearson's approach.

16.1 Introduction

In recent years, there have been growing interests on the development of probabilistic methods for robustness analysis and design problems aimed at overcoming the computational complexity and the issue of conservatism of deterministic worst-case framework (see, *e.g.*, [22, 24, 26–28, 73, 74, 85–87, 129, 163, 176, 182, 190, 198, 200, 218, 251, 272, 290, 304, 347, 358, 359, 381, 384, 392], and the references therein). In the deterministic worst-case framework, one is interested in knowing if the robustness requirement is guaranteed for every value of the uncertainty. However, it should be borne in mind that the uncertainty set may include worst cases which never happen in reality. Instead of seeking the worst-case guarantee, it is sometimes 'acceptable' that the robustness requirement is satisfied for most of the cases. It has been demonstrated that the proportion of systems guaranteeing the robustness requirement can be close to 1 even if the radii of uncertainty set are much larger than the worst case deterministic robustness margin (see, *e.g.*, [26, 27, 74, 272] and the references therein). Therefore, it is of practical importance to construct a function which describes quantitatively the relationship between the proportion of systems guaranteeing the robustness requirement and the radius of uncertainty set. This function can serve as a guide for control engineers in evaluating the robustness of a control system once a controller design is completed. Such a function, referred as *robustness degradation function*, has been proposed by a number of researchers. For example, Barmish and Lagoa [24] have constructed a curve of robustness margin amplification *vs* risk in a probabilistic setting. In a similar spirit, Calafiore, Dabbene and Tempo [74] have constructed a probability degradation function in the context of real and complex parametric uncertainty. It is important to note that the robustness degradation function can be done in a distribution-free manner. This can be justified by the Truncation Theory established by Barmish, Lagoa and Tempo [26] and can also be illustrated by relaxing the deterministic worst-case paradigm.

In this work, we consider robustness analysis problems with arbitrary robustness requirement and uncertainty bounding set. To construct a robustness degradation curve of practical interest, the selection of uncertainty radius interval is itself a question. Clearly, the range of uncertainty radius for which robustness degradation curve is significantly below 1 is not of practical interest since only a small risk can be tolerated in reality. From application point of view, it is only needed to construct robustness degradation curve for the range of uncertainty radius such that the curve is above an *a-priori* specified level $1-\epsilon$ where risk parameter $\epsilon \in (0,1)$ is acceptably small. We develop efficient randomized algorithms for estimating probabilistic robustness margin ρ_ϵ which is defined as the maximal uncertainty radius such that the probability of guaranteeing the robust requirements is at least $1 - \epsilon$. We have also developed fast algorithms for constructing robustness degradation curve which is above an *a-priori* specified level $1 - \epsilon$. In particular, we have developed efficient mechanisms such as probabilistic comparison, probabilistic bisection and backward iteration to reduce the computational complexity.

In our algorithms, confidence interval for binomial random variables has been frequently used to improve the efficiency of estimating probabilistic robustness margin and in the accuracy evaluation of robustness degradation function. Obviously, fast construction of binomial confidence interval is important to the efficiency of the randomized algorithm. Therefore, we have derived an explicit formula for constructing the confidence interval of binomial parameter with guaranteed coverage probability. The formula overcomes the limitation of normal approximation which is asymptotic in nature and thus inevitably introduces unknown errors in applications. Moreover, the formula is extremely simple and very tight in comparison with classic Clopper-Pearson's approach.

This chapter is organized as follows. Section 16.2 is the problem formulation. Section 16.3 discusses binomial confidence interval. Section 16.4 is devoted to probabilistic robustness margin. Section 16.5 presents algorithms for constructing robustness degradation curve. Illustrative examples are given in Section 16.6. Section 16.7 is the conclusion.

16.2 Problem Formulations

We adopt the assumption, from the classical robust control framework, that the uncertainty is deterministic and bounded. We formulate a general robustness analysis problem in a similar way as [74, 87] as follows.

Let $\mathcal{R}$ denote a robustness requirement. The definition of $\mathcal{R}$ can be a fairly complicated combination of requirements such as stability or $\mathcal{D}$-stability, $\mathcal{H}_\infty$ (or $\mathcal{H}_2$) norm, overshoot, rise time, settling time, steady state error, *etc.* Let $\mathcal{B}(r)$ denote the set of uncertainties with size smaller than r. In applications, we are usually dealing with uncertainty sets such as l_p ball, spectral norm ball, homogeneous star-shaped bounding set, *etc.* In general, the norm of uncertainty X can be represented as $\ell(X)$ where function $\ell(.)$ guarantees $\ell(X) = \min\{r : X \in \mathcal{B}(r)\}$.

To allow the robustness analysis be performed in a distribution-free manner, we introduce the notion of *proportion* as follows. For any $\Delta \in \mathcal{B}(r)$ there is an associated system $G(\Delta)$. Define *proportion*

$$P(r) := \frac{\mathrm{vol}(\{\Delta \in \mathcal{B}(r) : \ G(\Delta) \text{ guarantees } \mathcal{R}\})}{\mathrm{vol}(\mathcal{B}(r))}$$

where the volume function vol(.) is a Lebesgue measure defined on uncertainty parameter space. It follows that $P(r)$ is a reasonable measure of the robustness of the system [74,359]. Since the uncertainty set in our model may include worst cases which never happen in reality, it would be 'acceptable' in many applications if the robustness requirement $\mathcal{R}$ is satisfied for most of the uncertainty instances. Hence, we should obtain the value of $P(r)$ for uncertainty radius r which exceeds the deterministic robustness margin.

Clearly, $P(r)$ is deterministic in nature. However, we can resort to a probabilistic approach to evaluate $P(r)$. To see this, one needs to observe that $P(r) = \mathbb{P}\{G(\mathbf{\Delta}^u) \text{ guarantees } \mathcal{R}\}$ where $\mathbf{\Delta}^u$ is a random variable with *uniform distribution* over $\mathcal{B}(r)$. Define a Bernoulli random variable X such that X takes value 1 if the associated system $G(\mathbf{\Delta}^u)$ guarantees $\mathcal{R}$ and takes value 0 otherwise. Then estimating $P(r)$ is equivalent to estimating binomial parameter $P_X \doteq \Pr\{X = 1\} = P(r)$. It follows that a Monte Carlo method can be employed to estimate $P(r)$ based on i.i.d. observations of X.

Obviously, the robustness analysis problem would be completely solved if we can efficiently estimate $P(r)$ for all $r \in (0, \infty)$. However, this is infeasible from computational perspective. In practice, only a small risk ϵ can be tolerated by a system. Therefore, what is really important to know is the value of $P(r)$ *over the range of uncertain radius r for which $P(r)$ is at least* $1-\epsilon$ where $\epsilon \in (0,1)$ is referred as the *risk* parameter in this chapter. Our strategy is to firstly estimate the *probabilistic robustness margin* $\rho(\epsilon) := \sup\{r : \ P(r) \geq 1-\epsilon\}$ and consequently construct the robust degradation curve in a backward direction (in which r is decreased) by choosing the estimate of $\rho(\epsilon)$ as the starting uncertainty radius.

To reduce computational burden, the estimation of probabilistic robustness margin relies on the frequent use of binomial confidence interval. The confidence interval is also served as a validation method for the accuracy of estimating robustness degradation function. Hence, it is desirable to quickly construct binomial confidence interval with guaranteed coverage probability.

16.3 Binomial Confidence Intervals

Clopper and Pearson [90] has provided a rigorous approach for constructing binomial confidence interval. However, the computational complexity involved with this approach is very high. The standard technique is to use normal approximation which is not accurate for rare events. The coverage probability of the confidence interval derived from normal approximation can be significantly below the specified confidence level even for very large sample size. In the context of robustness analysis, we are dealing with rare events because the probability that the robustness requirement is violated is usually very small. We shall illustrate these standard methods as follows.

16.3.1 Clopper-Pearson Confidence Limits

Let the sample size N and confidence parameter $\delta \in (0,1)$ be fixed. We refer to an observation of X with value 1 as a *successful trial.* Let K denote the number of successful trials during the N i.i.d. sampling experiments. Let k be a realization of K. The classic Clopper-Pearson lower confidence limit $L_{N,k,\delta}$ and upper confidence limit $U_{N,k,\delta}$ are given respectively by

$$L_{N,k,\delta} := \begin{cases} 0 & \text{if } k=0 \\ \underline{p} & \text{if } k>0 \end{cases} \quad \text{and} \quad U_{N,k,\delta} := \begin{cases} 1 & \text{if } k=N \\ \overline{p} & \text{if } k<N \end{cases}$$

where $\underline{p} \in (0,1)$ is the solution of the equation

$$\sum_{j=0}^{k-1} \binom{N}{j} \underline{p}^j (1-\underline{p})^{N-j} = 1 - \frac{\delta}{2} \tag{16.1}$$

and $\overline{p} \in (0,1)$ is the solution of the equation

$$\sum_{j=0}^{k} \binom{N}{j} \overline{p}^j (1-\overline{p})^{N-j} = \frac{\delta}{2}. \tag{16.2}$$

16.3.2 Normal Approximation

It is easy to see that equations (16.1) and (16.2) are hard to solve and thus the confidence limits are difficult to determine using Clopper-Pearson's approach. For large sample size, it is computationally intensive. To get around the difficulty, normal approximation has been widely used to develop simple approximate formulas (see, for example, [151] and the references therein). Let $\Phi(.)$ denote the normal distribution function and $Z_{\frac{\delta}{2}}$ denote the critical value such that $\Phi(Z_{\frac{\delta}{2}}) = 1 - \frac{\delta}{2}$. It follows from the Central Limit Theorem that, for sufficiently large sample size N, the lower and upper confidence limits can be estimated respectively as $\widetilde{L} \approx \frac{k}{N} - Z_{\frac{\delta}{2}} \sqrt{\frac{\frac{k}{N}(1-\frac{k}{N})}{N}}$ and $\widetilde{U} \approx \frac{k}{N} + Z_{\frac{\delta}{2}} \sqrt{\frac{\frac{k}{N}(1-\frac{k}{N})}{N}}$.

The critical problem with the normal approximation is that it is of asymptotic nature. It is not clear how large the sample size is sufficient for the approximation error to be negligible. Such an asymptotic approach is not good enough for studying the robustness of control systems.

16.3.3 Explicit Formula

It is desirable to have a simple formula which is rigorous and very tight for the confidence interval construction. Recently, we have derived the following simple formula for constructing the confidence limits.

Theorem 1. *Let* $\mathcal{L}(k) = \frac{k}{N} + \frac{3}{4}\,\frac{1-\frac{2k}{N}-\sqrt{1+4\theta\, k(1-\frac{k}{N})}}{1+\theta N}$ *and* $\mathcal{U}(k) = \frac{k}{N} + \frac{3}{4}\,\frac{1-\frac{2k}{N}+\sqrt{1+4\theta\, k(1-\frac{k}{N})}}{1+\theta N}$ *with* $\theta = \frac{9}{8\ln\frac{2}{\delta}}$. *Then* $\mathbb{P}\{\mathcal{L}(K) < P_X < \mathcal{U}(K)\} > 1-\delta$.

The proof of this result is reported in the Appendix.

As can be seen from Figure 16.1, our formula is very tight in comparison with the Clopper-Pearson's approach. Obviously, there is no comparison on the computational complexity. Our formula is simple enough for hand calculation. Simplicity is especially important when the confidence limits are frequently used in the context of robustness analysis.

16.4 Estimating Probabilistic Robustness Margin

In this section, we shall develop efficient randomized algorithms for constructing an estimate for $\rho(\epsilon)$.

16.4.1 Separable Assumption

We assume that *the robustness degradation curve of the system can be separated into two parts by a horizontal line with height* $1-\epsilon$, *i.e.*,

$$P(r) < 1-\epsilon \text{ for all } r > \rho(\epsilon).$$

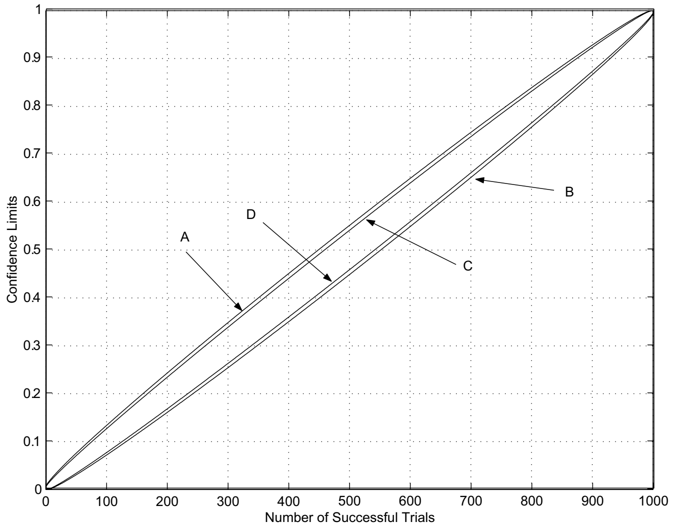


Figure 16.1. Confidence Interval ($N = 1000,\ \delta = 10^{-2}$. A and B are the upper and lower confidence limits by our formula; C and D are the upper and lower confidence limits by Clopper-Pearson's method)

We refer to such an assumption as the *Separable Assumption*. Our extensive simulation experience indicated that, for small risk parameter ϵ, most control systems guarantee the separable assumption. It should be noted that it is even much weaker than assuming that $P(r)$ is non-increasing (See illustrative Figure 16.2). Moreover, the non-increasing assumption is rather mild. This can be explained by a heuristic argument as follows. Let

$$\mathcal{B}^{\mathcal{R}}(r) \doteq \{\Delta \in \mathcal{B}(r) : \text{The associated system } G(\Delta) \text{ guarantees } \mathcal{R}\}.$$

Then

$$P(r) = \frac{\text{vol}(\mathcal{B}^{\mathcal{R}}(r))}{\text{vol}(\mathcal{B}(r))}$$

and

$$\frac{dP(r)}{dr} = \frac{1}{\text{vol}(\mathcal{B}(r))} \left[\frac{d\, \text{vol}(\mathcal{B}^{\mathcal{R}}(r))}{dr} - P(r)\, \frac{d\, \text{vol}(\mathcal{B}(r))}{dr} \right]. \qquad (16.3)$$

In the range of uncertainty radius such that $P(r)$ is close to 1, $\text{vol}(\mathcal{B}(r))$ increases (as r increases) much faster than $\text{vol}(\mathcal{B}^{\mathcal{R}}(r))$ due to the constraint of robust requirement $\mathcal{R}$. Hence inequality

$$\frac{d\, \text{vol}(\mathcal{B}^{\mathcal{R}}(r))}{dr} \leq \frac{d\, \text{vol}(\mathcal{B}(r))}{dr} \approx P(r)\, \frac{d\, \text{vol}(\mathcal{B}(r))}{dr}$$

can be easily satisfied. It follows from equation (16.3) that $\frac{dP(r)}{dr} \leq 0$ can be readily guaranteed.

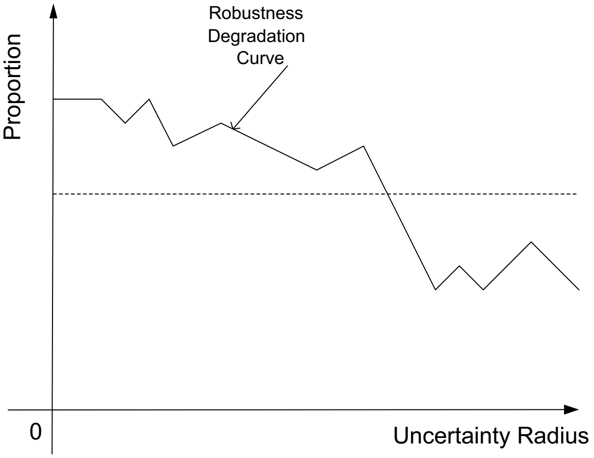


Figure 16.2. Illustration of separable assumption. The robustness degradation curve can be separated as the upper segment and lower segment by the dash horizontal line with height $1 - \epsilon$. In this example, the separable assumption is satisfied, while the non-increasing assumption is violated.

When the separable assumption is guaranteed, an interval which includes $\rho(\epsilon)$ can be readily found by starting from uncertainty radius $r = 1$ and then successively doubling r or cutting r in half based on the comparison of $P(r)$ with $1-\epsilon$. Moreover, bisection method can be employed to refine the estimate for $\rho(\epsilon)$. Of course, the success of such methods depends on the reliable and efficient comparison of $P(r)$ with $1-\epsilon$ based on Monte Carlo method. In the following subsection, we illustrate a fast method of comparison.

16.4.2 Probabilistic Comparison

In general, P_X can only be estimated by a Monte Carlo method. The conventional method is to compare directly $\frac{K}{N}$ with $1 - \epsilon$ where K is the number of successful trials during N i.i.d. sampling experiments. There are several problems with the conventional method. First, the comparison of $\frac{K}{N}$ with $1 - \epsilon$ can be very misleading. Second, the sample size N is required to be very large to obtain a reliable comparison. Third, we don't know how reliable the comparison is. In this subsection, we present a new approach which allows for a reliable comparison with much fewer samples. The key idea is to compare binomial confidence limits with the fixed probability $1 - \epsilon$ and hence reliable judgement can be made in advance.

Algorithm 16.1 (Probabilistic comparison) *Given risk parameter ϵ and confidence parameter δ, returns the index d.*

Step 1. Let $d \leftarrow 0$.
Step 2. While $d = 0$ do the following:

- *Sample X.*
- *Update N and K.*
- *Compute lower confidence limit L and upper confidence limit U by Theorem 1.*
- *If $U < 1 - \epsilon$ then let $d \leftarrow -1$. If $L > 1 - \epsilon$ then let $d \leftarrow 1$.*

The confidence parameter δ is used to control the reliability of the comparison. A typical value is $\delta = 0.01$. The implication of output d is interpreted as follows: $d = 1$ indicates that $P_X > 1 - \epsilon$ is true with high confidence; $d = -1$ indicates that $P_X < 1 - \epsilon$ is true with high confidence.

Obviously, the sample size is random in nature. For $\epsilon = \delta = 0.01$, we simulated the Probabilistic Comparison Algorithm identically and independently 100 times for different values of P_X. We observe that, for each value of P_X, the Probabilistic Comparison Algorithm makes correct judgement among all 100 simulations. Figure 16.3 shows the average sample size and the 95%-quantile of the sample size estimated from our simulation. It can be seen from the figure that, as long as P_X is not very close to $1-\epsilon$, the Probabilistic Comparison Algorithm can make a reliable comparison with a small sample size.

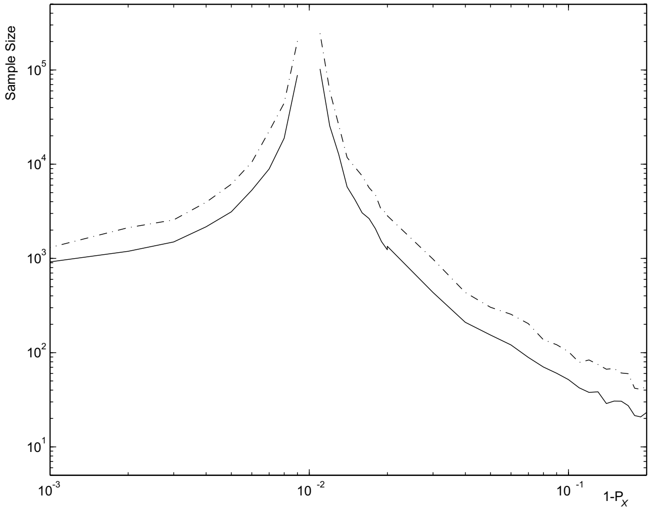


Figure 16.3. Complexity of Probabilistic Comparison. The horizontal axis represents $1-P_X$. The vertical axis represents sample size. The solid line and the dash-dot line respectively show the average sample size and the 95%-quantile of the sample size.

16.4.3 Computing Initial Interval

Under the separable assumption, an interval $[a, b]$ which includes $\rho(\epsilon)$ can be quickly determined by the following algorithm.

Algorithm 16.2 (Initial interval) *Given risk parameter ϵ and confidence parameter δ, returns an initial interval $[a, b]$.*

```
Step 1. Let $r \leftarrow 1$. Apply Probabilistic comparison algorithm to compare $P(1)$
    with $1 - \epsilon$. Let the outcome be $d_1$.
Step 2. If $d_1 = 1$ then let $d \leftarrow d_1$ and do the following:
    • While $d = 1$ do the following:
        – Let $r \leftarrow 2r$. Apply Probabilistic comparison algorithm to compare
          $P(r)$ with $1 - \epsilon$. Let the outcome be $d$.
    • Let $a \leftarrow \frac{r}{2}$ and $b \leftarrow r$.
Step 3. If $d_1 = -1$ then let $d \leftarrow d_1$ and do the following:
    • While $d = -1$ do the following:
        – Let $r \leftarrow \frac{r}{2}$. Apply Probabilistic comparison algorithm to compare
          $P(r)$ with $1 - \epsilon$. Let the outcome be $d$.
    • Let $a \leftarrow r$ and $b \leftarrow 2r$.
```

16.4.4 Probabilistic Bisection

Once an initial interval $[a, b]$ is obtained, an estimate $\widehat{R}$ for the probabilistic robustness margin $\rho(\epsilon)$ can be efficiently computed as follows.

Algorithm 16.3 (Bisection) *Given risk parameter ϵ, confidence parameter δ, initial interval $[a, b]$, and relative tolerance γ, returns $\widehat{R}$.*

```
Step 1. While $b - a > \gamma a$ do the following:
    • Let $r \leftarrow \frac{a+b}{2}$. Apply Probabilistic comparison algorithm to compare
      $P(r)$ with $1 - \epsilon$. Let the outcome be $d$.
    • If $d = -1$ then let $b \leftarrow r$, else let $a \leftarrow r$.
Step 2. Return $\widehat{R} = b$.
```

16.5 Constructing Robustness Degradation Curve

We shall develop efficient randomized algorithms for constructing robustness degradation curve, which provide more insight for the robustness of the uncertain system than probabilistic robustness margin. First we introduce the sample reuse algorithm developed in [87] for constructing robustness degradation curve for a given range of uncertainty radius.

Algorithm 16.4 (Sample reuse) *Given sample size N, confidence parameter $\delta \in (0,1)$, uncertainty radius interval $[a,b]$, and number of uncertainty radii l, returns the proportion estimate $\widehat{P}_i$ and the related confidence interval for $r_i = b - \frac{(b-a)(i-1)}{l-1}$, $i = 1,2,\cdots,l$. In the following, m_{i1} denotes the number of sampling experiments conducted at r_i and m_{i2} denotes the number of observations guaranteeing $\mathcal{R}$ during the m_{i1} sampling experiments.*

Step 1. Let $M = [m_{ij}]_{l \times 2}$ be a zero matrix.
Step 2. (Backward iteration) For $i = 1$ to $i = l$ do the following:

- *Let $r \leftarrow r_i$.*
- *While $m_{i1} < N$ do the following:*
  - *Generate uniform sample q from $\mathcal{B}(r)$. Evaluate the robustness requirement R for q.*
  - *Let $m_{s1} \leftarrow m_{s1} + 1$ for any s such that $r \geq r_s \geq \ell(q)$.*
  - *If robustness requirement R is satisfied for q then let $m_{s2} \leftarrow m_{s2}+1$ for any s such that $r \geq r_s \geq \ell(q)$.*
- *Let $\widehat{P}_i \leftarrow \frac{m_{i2}}{N}$ and construct the confidence interval of confidence level $100(1-\delta)\%$ by Theorem 1.*

It should be noted that the idea of the sample reuse algorithm is not simply a save of sample generation. It is actually a backward iterative mechanism. In the algorithm, the most important save of computation is usually the evaluation of the complex robustness requirements $\mathcal{R}$ (see, *e.g.*, [87] for details).

Now we introduce the global strategy for constructing robustness degradation curve. The idea is to apply successively the sample reuse algorithm for a sequence of intervals of uncertainty radius. Each time the size of interval is reduced by half. The lower bound of the current interval is defined to be the upper bound of the next consecutive interval. The algorithm is terminated once the robustness requirement $\mathcal{R}$ is guaranteed for all N samples of an uncertainty set of which the radius is taken as the lower bound of an interval of uncertainty radius. More precisely, the procedure is presented as follows.

Algorithm 16.5 (Global strategy) *Given sample size N, risk parameter ϵ and confidence parameter $\delta \in (0,1)$, returns the proportion estimate $\widehat{P}_i$, and the related confidence interval.*

Step 1. Compute an estimate $\widehat{R}$ for probabilistic robustness margin $\rho(\epsilon)$.
Step 2. Let $STOP \leftarrow 0$. Let $a \leftarrow \frac{\widehat{R}}{2}$ and $b \leftarrow \widehat{R}$.
Step 3. (Backward iteration) While $STOP = 0$ do the following:

- *Apply sample reuse algorithm to construct robustness degradation curve for uncertainty radius interval $[a,b]$.*
- *If the robustness property $\mathcal{R}$ is guaranteed for all N samples of uncertainty set $\mathcal{B}(r)$ with radius $r = a$ then let $STOP \leftarrow 1$, otherwise let $b \leftarrow a$ and $a \leftarrow \frac{b}{2}$.*

For given risk parameter ϵ and confidence parameter δ, the sample size is chosen as

$$N > \frac{2(1-\epsilon+\frac{\alpha\epsilon}{3})(1-\frac{\alpha}{3})\ln\frac{2}{\delta}}{\alpha^2\epsilon} \tag{16.4}$$

with $\alpha \in (0,1)$. It follows from Massart's inequality [221] that such a sample size ensures $\mathbb{P}\left\{\left|P_X - \frac{K}{N}\right| < \alpha\epsilon\right\} > 1-\delta$ with $P_X = 1-\epsilon$ (See also Lemma 1 in Appendix). It should be noted that Massart's inequality is less conservative than the Chernoff bounds in both multiplicative and additive forms.

We would like to remark that the algorithms for estimating the probabilistic robustness margin and constructing robustness degradation curve are susceptible of further improvement. For example, the idea of sample reuse is not employed in computing the initial interval and in the probabilistic bisection algorithm. Moreover, in constructing the robustness degradation curve, the sample reuse algorithm is independently applied for each interval of uncertainty radius. Actually, the simulation results can be saved for the successive intervals.

16.6 Illustrative Examples

In this section we demonstrate through examples the power of randomized algorithms in solving complicated robustness analysis problems which are not tractable in the classical deterministic framework.

We consider a system which has been studied in [133] by a deterministic approach. The system is as shown in Figure 16.4.

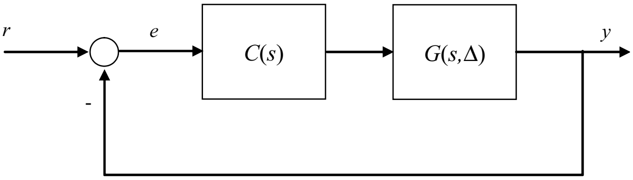


Figure 16.4. Uncertain system

The compensator is $C(s) = \frac{s+2}{s+10}$ and the plant is

$$G(s,\Delta) = \frac{800(1+0.1\delta_1)}{s(s+4+0.2\delta_2)(s+6+0.3\delta_3)}$$

with parametric uncertainty $\Delta = [\delta_1, \delta_2, \delta_3]^{\mathrm{T}}$. The nominal system is stable. The closed-loop poles of the nominal system are: $z_1 = -15.9178$, $z_2 = -1.8309$, $z_3 = -1.1256+7.3234i$, $z_4 = -1.1256-7.3234i$. The peak value, rise time, settling time of step response of the nominal system, are respectively, $P^0_{peak} = 1.47$, $t^0_r = 0.185$, $t^0_s = 3.175$.

We first consider the robust $\mathcal{D}$-stability of the system. The robustness requirement $\mathcal{R}$ is defined as $\mathcal{D}$-stability with the domain of poles specified as: real part < -1.5, or fall within one of the two disks centered at z_3 and z_4 with radius 0.3. The uncertainty set is defined as the polytope

$$\mathcal{B}_H(r) := \left\{ r\Delta + (1-r)\,\frac{\sum_{i=1}^4 \Delta^i}{4} : \Delta \in \operatorname{co}\{\Delta^1, \Delta^2, \Delta^3, \Delta^4\} \right\}$$

where co denotes the convex hull of

$$\Delta^i = \left[\frac{1}{2}\sin\left(\frac{2i-1}{3}\pi\right),\ \frac{1}{2}\cos\left(\frac{2i-1}{3}\pi\right),\ -\frac{\sqrt{3}}{2} \right]^{\mathrm{T}}$$

for $i = 1, 2, 3$ and $\Delta^4 = [0,\ 0,\ 1]^{\mathrm{T}}$.

Obviously, there exists no effective method for computing the deterministic robustness margin in the literature. However, our randomized algorithms can efficiently construct the robustness degradation curve. See Figure 16.5.

In this example, the risk parameter is *a-priori* specified as $\epsilon = 0.001$. The procedure for estimating the probabilistic robustness margin is explained as follows. Let N denote the dynamic sample size which is random in nature. Let K denote the number of successful trials among N i.i.d. sampling experiments as defined in Section 16.2 and Subsection 16.3.1 (*i.e.*, a successful trial is equivalent to an observation that the robustness requirement is guaranteed). Let confidence parameter $\delta = 0.01$ and choose tolerance $\gamma = 0.05$. Starting from $r = 1$, after $N = 7060$ simulations we obtain $K = 7060$, the probabilistic comparison algorithm determined that $P(1) > 1-\epsilon$ since the lower confidence limit $L > 1 - \epsilon$. The simulation is thus switched to uncertainty radius $r = 2$. After $N = 65$ times of simulation, it is found that $K = 61$. The probabilistic comparison algorithm detected that $P(2) < 1 - \epsilon$ because the upper confidence limit $U < 1 - \epsilon$. So, initial interval $[1, 2]$ is readily obtained. Now the probabilistic bisection algorithm is invoked. Staring with the middle point of the initial interval (*i.e.*, $r = \frac{1+2}{2} = \frac{3}{2}$), after $N = 613$ times of simulations, it is found that $K = 607$, the probabilistic comparison algorithm concluded that $P(\frac{3}{2}) < 1 - \epsilon$ since the upper confidence limit $U < 1 - \epsilon$. Thus simulation is moved to $r = \frac{1+\frac{3}{2}}{2} = \frac{5}{4}$. It is found that $K = 9330$ among $N = 9331$ times of simulations. Hence, the probabilistic comparison algorithm determined that $P(\frac{5}{4}) > 1-\epsilon$ since the lower confidence limit $L > 1-\epsilon$. Now the simulation is performed at $r = \frac{\frac{5}{4}+\frac{3}{2}}{2} = \frac{11}{8}$. After $N = 6653$ simulations, it is discovered that $K = 6636$. The probabilistic comparison algorithm judged that $P(\frac{11}{8}) < 1-\epsilon$ based on calculation that the upper confidence limit $U < 1 - \epsilon$. At this point the interval is $[\frac{5}{4}, \frac{11}{8}]$ and the bisection is terminated since tolerance condition $b-a \leq \gamma a$ is satisfied. The evolution of intervals produced by the probabilistic bisection algorithm is as follows:

$$[1, 2] \ \longrightarrow\ \left[1, \frac{3}{2}\right] \ \longrightarrow\ \left[\frac{5}{4}, \frac{3}{2}\right] \ \longrightarrow\ \left[\frac{5}{4}, \frac{11}{8}\right].$$

Now we have obtained an interval $[\frac{5}{4}, \frac{11}{8}]$ which includes $\rho(0.001)$, so the sample reuse algorithm can be employed to construct robustness degradation curve. In this example, the number of uncertainty radii is $l = 100$ and the confidence parameter is chosen as $\delta = 0.001$. A constant sample size is computed by formula (16.4) with $\alpha = 0.5$ as $N = 50,631$. The interval from which we start constructing robustness degradation curve is $[\frac{11}{16}, \frac{11}{8}]$. It is determined that $K = N = 50,632$ at uncertainty radius $r = \frac{11}{16}$. Therefore, the Sample reuse algorithm is invoked only once and the overall algorithm is terminated (If $K \neq N$ for $r = \frac{11}{16}$, then the next interval would be $[\frac{11}{32}, \frac{11}{16}]$). Although $P(r)$ is evaluated for $l = 100$ uncertainty radii with the same sample size N, the total number of simulation is only $153,358 << Nl = 100N$. To give an accuracy for all the estimates of $P(r)$, confidence limits are computed by Theorem 1.

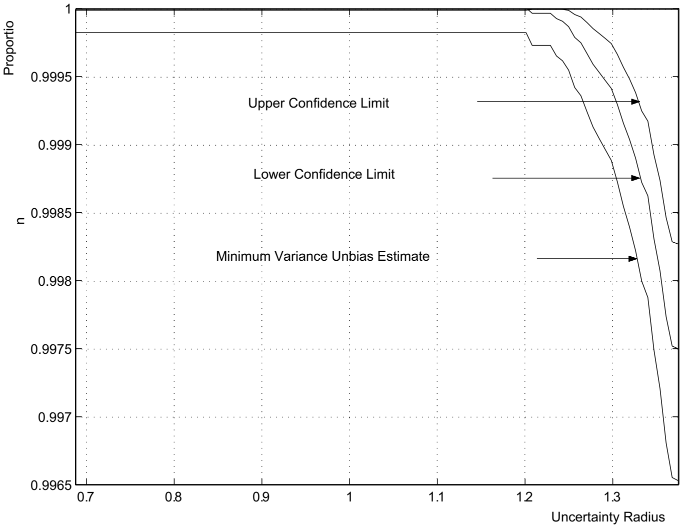


Figure 16.5. Robustness degradation curve

We now apply our algorithms to a robustness problem with time specifications. Specifically, the robustness requirement $\mathcal{R}$ is: stability, and rise time $t_r < 135\%\ t_r^0 = 0.25$, settling time $t_s < 110\%\ t_s^0 = 3.5$, overshoot $P_{peak} < 116\%\ P_{peak}^0 = 1.7$. The uncertainty set is $\mathcal{B}_\infty(r) := \{\Delta : ||\Delta||_\infty \leq r\}$.

In this case, the risk parameter is *a-priori* specified as $\epsilon = 0.01$. It is well known that, for this type of problem, there exists no effective method for computing the deterministic robustness margin in the literature. However, our randomized algorithms can efficiently construct the robustness degradation curve. See Figure 16.6.

We choose $\gamma = 0.25$ and $\delta = 0.01$ for estimating $\rho(0.01)$. Starting from uncertainty radius $r = 1$, the initial interval is easily found as $[\frac{1}{8}, \frac{3}{16}]$ through the following evolution:

$$\left[\frac{1}{2}, 1\right] \longrightarrow \left[\frac{1}{4}, \frac{1}{2}\right] \longrightarrow \left[\frac{1}{8}, \frac{1}{4}\right].$$

The sequence of intervals produced by the probabilistic bisection algorithm is as follows:

$$\left[\frac{1}{8}, \frac{1}{4}\right] \longrightarrow \left[\frac{1}{8}, \frac{3}{16}\right] \longrightarrow \left[\frac{1}{8}, \frac{5}{32}\right].$$

So, we obtained an estimate for the probabilistic robustness margin $\rho(0.01)$ as $\frac{5}{32}$. To construct robustness degradation curve, the number of uncertainty radii is $l = 100$ and the confidence parameter is chosen as $\delta = 0.01$. A constant sample size is computed by formula (16.4) with $\alpha = 0.2$ as $N = 24,495$. The interval from which we start constructing robustness degradation curve is $[\frac{5}{64}, \frac{5}{32}]$. We found that this is also the last interval of uncertainty radius because it is determined that $K = N$ at uncertainty radius $r = \frac{5}{64}$.

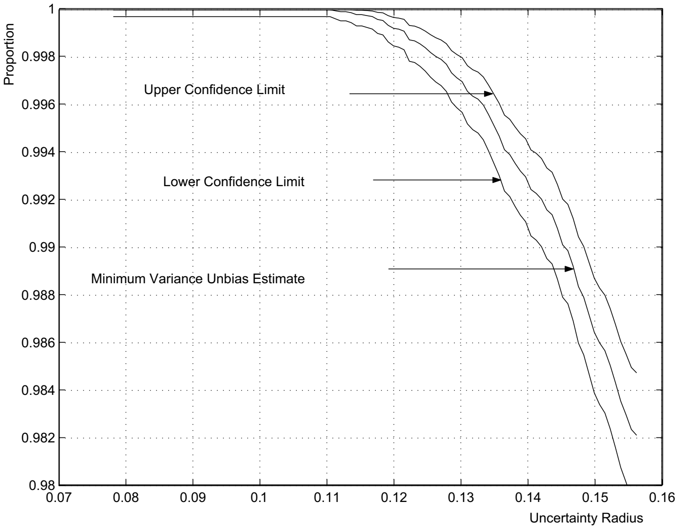


Figure 16.6. Robustness degradation curve

16.7 Conclusions

In this contribution, we have established efficient techniques which apply to robustness analysis problems with arbitrary robustness requirements and uncertainty bounding set. The key mechanisms are probabilistic comparison, probabilistic bisection and backward iteration. Motivated by the crucial role of binomial confidence interval in reducing the computational complexity, we have derived an explicit formula for computing binomial confidence limits. This formula overcomes the computational issue and inaccuracy of standard methods.

16.8 Appendix: Proof of Theorem 1

To show Theorem 1, we need some preliminary results. The following Lemma 1 is due to Massart [221].

Lemma 1. $\mathbb{P}\left\{\frac{K}{N} \geq P_X + \epsilon\right\} \leq \exp\left(-\frac{N\epsilon^2}{2(P_X+\frac{\epsilon}{3})\,(1-P_X-\frac{\epsilon}{3})}\right)$ *for all* $\epsilon \in (0, 1 - P_X)$.

Of course, the above upper bound holds trivially for $\epsilon \geq 1 - P_X$. Thus, Lemma 1 is actually true for any $\epsilon > 0$.

Lemma 2. $\mathbb{P}\left\{\frac{K}{N} \leq P_X - \epsilon\right\} \leq \exp\left(-\frac{N\epsilon^2}{2(P_X-\frac{\epsilon}{3})\,(1-P_X+\frac{\epsilon}{3})}\right)$ *for all* $\epsilon > 0$.

Proof. Define $Y = 1 - X$. Then $P_Y = 1 - P_X$. At the same time when we are conducting N i.i.d. experiments for X, we are also conducting N i.i.d. experiments for Y. Let the number of successful trials of the experiments for Y be denoted as K_Y. Obviously, $K_Y = N - K$. Applying Lemma 1 to Y, we have $\mathbb{P}\left\{\frac{K_Y}{N} \geq P_Y + \epsilon\right\} \leq \exp\left(-\frac{N\epsilon^2}{2(P_Y+\frac{\epsilon}{3})\,(1-P_Y-\frac{\epsilon}{3})}\right)$. It follows that $\mathbb{P}\left\{\frac{N-K}{N} \geq 1 - P_X + \epsilon\right\} \leq \exp\left(-\frac{N\epsilon^2}{2(1-P_X+\frac{\epsilon}{3})\,[1-(1-P_X)-\frac{\epsilon}{3}]}\right)$. The proof is thus completed by observing that $\mathbb{P}\left\{\frac{N-K}{N} \geq 1 - P_X + \epsilon\right\} = \mathbb{P}\left\{\frac{K}{N} \leq P_X - \epsilon\right\}$. □

The following lemma can be found in [91].

Lemma 3. $\sum_{j=0}^{k} \binom{N}{j} x^j (1-x)^{N-j}$ *decreases monotonically with respect to* $x \in (0, 1)$ *for* $k = 0, 1, \cdots, N$.

Lemma 4. $\sum_{j=0}^{k} \binom{N}{j} x^j (1-x)^{N-j} \leq \exp\left(-\frac{N(x-\frac{k}{N})^2}{2\,(\frac{2}{3}x+\frac{k}{3N})\,(1-\frac{2}{3}x-\frac{k}{3N})}\right) \quad \forall x \in (\frac{k}{N}, 1)$ *for* $k = 0, 1, \cdots, N$.

Proof. Consider binomial random variable X with parameter $P_X > \frac{k}{N}$. Let K be the number of successful trials during N i.i.d. sampling experiments. Then $\sum_{j=0}^{k} \binom{N}{j} P_X^j (1-P_X)^{N-j} = \Pr\{K \leq k\}$. Note that $\Pr\{K \leq k\} = \Pr\left\{\frac{K}{N} \leq P_X - \left(P_X - \frac{k}{N}\right)\right\}$. Applying Lemma 2 with $\epsilon = P_X - \frac{k}{N} > 0$, we have

$$\sum_{j=0}^{k} \binom{N}{j} P_X^j (1-P_X)^{N-j} \leq \exp\left(-\frac{N(P_X-\frac{k}{N})^2}{2(P_X - \frac{P_X-\frac{k}{N}}{3})\,(1-P_X+\frac{P_X-\frac{k}{N}}{3})}\right)$$

$$= \exp\left(-\frac{N(P_X-\frac{k}{N})^2}{2\,(\frac{2}{3}P_X+\frac{k}{3N})\,(1-\frac{2}{3}P_X-\frac{k}{3N})}\right).$$

Since the argument holds for arbitrary binomial random variable X with $P_X > \frac{k}{N}$, thus the proof of the lemma is completed. □

Lemma 5. $\sum_{j=0}^{k-1} \binom{N}{j} x^j (1-x)^{N-j} \geq 1 - \exp\left(- \frac{N(x-\frac{k}{N})^2}{2\,(\frac{2}{3}x + \frac{k}{3N})\,(1-\frac{2}{3}x-\frac{k}{3N})} \right)$ $\forall x \in (0, \frac{k}{N})$ *for* $k = 1, \cdots, N$.

Proof. Consider binomial random variable X with parameter $P_X < \frac{k}{N}$. Let K be the number of successful trials during N i.i.d. sampling experiments. Then

$$\sum_{j=0}^{k-1} \binom{N}{j} P_X^j (1-P_X)^{N-j} = \Pr\{K < k\} = \Pr\left\{ \frac{K}{N} < P_X + \left(\frac{k}{N} - P_X \right) \right\}.$$

Applying Lemma 1 with $\epsilon = \frac{k}{N} - P_X > 0$, we have that

$$\begin{aligned}\sum_{j=0}^{k-1} \binom{N}{j} P_X^j (1-P_X)^{N-j} &\geq 1 - \exp\left(- \frac{N(\frac{k}{N} - P_X)^2}{2(P_X + \frac{\frac{k}{N}-P_X}{3})\,(1 - P_X - \frac{\frac{k}{N}-P_X}{3})} \right) \\ &= 1 - \exp\left(- \frac{N(P_X - \frac{k}{N})^2}{2\,(\frac{2}{3}P_X + \frac{k}{3N})\,(1 - \frac{2}{3}P_X - \frac{k}{3N})} \right).\end{aligned}$$

Since the argument holds for arbitrary binomial random variable X with $P_X < \frac{k}{N}$, thus the proof of the lemma is completed. □

Lemma 6. *Let* $0 \leq k \leq N$. *Then* $L_{N,k,\delta} < U_{N,k,\delta}$.

Proof. Obviously, the lemma is true for $k = 0, N$. We consider the case that $1 \leq k \leq N-1$. Let $\mathcal{S}(N,k,x) = \sum_{j=0}^{k} \binom{N}{j} x^j (1-x)^{N-j}$ for $x \in (0,1)$. Notice that $\mathcal{S}(N,k,\overline{p}) = \mathcal{S}(N,k-1,\overline{p}) + \binom{N}{k}\overline{p}^k(1-\overline{p})^{N-k} = \frac{\delta}{2}$. Thus

$$\mathcal{S}(N,k-1,\underline{p}) - \mathcal{S}(N,k-1,\overline{p}) = 1 - \frac{\delta}{2} - \left[\frac{\delta}{2} - \binom{N}{k} \overline{p}^k (1-\overline{p})^{N-k} \right].$$

Notice that $\delta \in (0,1)$ and that $\overline{p} \in (0,1)$, we have

$$\mathcal{S}(N,k-1,\underline{p}) - \mathcal{S}(N,k-1,\overline{p}) = 1 - \delta + \binom{N}{k} \overline{p}^k (1-\overline{p})^{N-k} > 0.$$

By Lemma 3, $\mathcal{S}(N,k-1,x)$ decreases monotonically with respect to x, we have that $\underline{p} < \overline{p}$. □

We are now in position to prove Theorem 1. It can be checked that $U_{N,k,\delta} \leq \mathcal{U}(k)$ for $k = 0,\ N$. We need to show that $U_{N,k,\delta} \leq \mathcal{U}(k)$ for $0 < k < N$. Straightforward computation shows that $\mathcal{U}(k)$ is the only root of equation

$$\exp\left(- \frac{N(x - \frac{k}{N})^2}{2\,(\frac{2}{3}x + \frac{k}{3N})\,(1 - \frac{2}{3}x - \frac{k}{3N})} \right) = \frac{\delta}{2}$$

with respect to $x \in (\frac{k}{N}, \infty)$. There are two cases: $\mathcal{U}(k) \geq 1$ and $\mathcal{U}(k) < 1$. If $\mathcal{U}(k) \geq 1$ then $U_{N,k,\delta} \leq \mathcal{U}(k)$ is trivially true. We only need to consider the case that $\frac{k}{N} < \mathcal{U}(k) < 1$. In this case, it follows from Lemma 4 that $\sum_{j=0}^{k} \binom{N}{j} [\mathcal{U}(k)]^j [1 - \mathcal{U}(k)]^{N-j} \leq \exp\left(-\frac{N(\mathcal{U}(k) - \frac{k}{N})^2}{2\,(\frac{2}{3}\mathcal{U}(k) + \frac{k}{3N})\,(1 - \frac{2}{3}\mathcal{U}(k) - \frac{k}{3N})}\right) = \frac{\delta}{2}$. Recall that $\sum_{j=0}^{k} \binom{N}{j} U_{N,k,\delta}^{j} (1 - U_{N,k,\delta})^{N-j} = \frac{\delta}{2}$, we have

$$\sum_{j=0}^{k} \binom{N}{j} U_{N,k,\delta}^{j} (1 - U_{N,k,\delta})^{N-j} \geq \sum_{j=0}^{k} \binom{N}{j} [\mathcal{U}(k)]^j [1 - \mathcal{U}(k)]^{N-j}.$$

Therefore, by Lemma 3, we have that $U_{N,k,\delta} \leq \mathcal{U}(k)$ for $0 < k < N$. Thus, we have shown that $U_{N,k,\delta} \leq q$ for all k.

Similarly, by Lemma 5 and Lemma 3 we can show that $L_{N,k,\delta} \geq \mathcal{L}(k)$. Hence, by Lemma 6, we have $\mathcal{L}(k) \leq L_{N,k,\delta} < U_{N,k,\delta} \leq \mathcal{U}(k)$. Finally, the proof of Theorem 1 is completed by invoking the probabilistic implication of the Clopper-Pearson confidence interval. □

References

1. C. Acerbi and D. Tasche. On the coherence of expected shortfall. *Journal of Banking and Finance*, 26:1487–1503, 2002.
2. J.E. Ackermann, D. Kaesbauer, and W. Sienel. Design by search. In *1st IFAC Symposium on Design Methodologies and Control Systems*, 1991.
3. A. Alessandri, M. Baglietto, and G. Battistelli. Receding-horizon estimation for discrete-time linear systems. *IEEE Transactions on Automatic Control*, 48(3):473–478, 2003.
4. F. Amato, M. Mattei, S. Scala, and L. Verde. Design via LQ methods. In J-F. Magni, S. Bennani, and J. Terlouw, editors, *Robust Flight Control, A Design Challenge (GARTEUR)*, volume 224 of *Lecture Notes in Control and Information Sciences*, pages 444–463. Springer-Verlag, Berlin, 1997.
5. B.D.O. Anderson and J.B. Moore. *Optimal Control: Linear Quadratic Methods.* Prentice-Hall, Englewood Cliffs, 1990.
6. L.R. Anderson. Fine tuning of aircraft control laws using pro-Matlab software. In *AIAA-91-2600-CP*, 1991.
7. T.W. Anderson and H. Burnstein. Approximating the upper binomial confidence limit. *Journal of American Statistic Association*, 62:857–861, 1967.
8. T.W. Anderson and H. Burnstein. Approximating the lower binomial confidence limit. *Journal of American Statistic Association*, 63:1413–1415, 1968.
9. P.J. Antsaklis and A.N. Michel. *Linear Systems.* McGraw-Hill, Singapore, 1997.
10. P. Apkarian and R.J. Adams. Advanced gain-scheduling techniques for uncertain systems. *IEEE Transactions on Control Systems Technology*, 6:21–32, 1998.
11. P. Apkarian, P.C. Pellanda, and H.D. Tuan. Mixed H_2/H_∞ multi-channel linear parameter-varying control in discrete time. *Systems & Control Letters*, 41(5):333–346, 2000.
12. P. Apkarian and H.D. Tuan. Parameterized LMIs in control theory. *SIAM Journal on Control and Optimization*, 38(4):1241–1264, 2000.
13. P. Apkarian, H.D. Tuan, and J. Bernussou. Continuous-time analysis, eigenstructure assignment, and H_2 synthesis with enhanced linear matrix inequalities (LMI) characterizations. *IEEE Transactions on Automatic Control*, 46(12):1941–1946, 2001.

14. E. Arjas and D. Gasbarra. Bayesian inference of survival probabilities under stochastic ordering constraints. *J. Amer. Statist. Assoc.*, 91:1101–1109, 1996.
15. D.V. Arnold. *Noisy Optimization with Evolution Strategies.* Kluwer Academic Publishers, Boston, 2002.
16. P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.
17. P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, and H. Ku. Coherent multiperiod risk measurement. *Manuscript, ETH Zürich*, 2003.
18. T. Asai and S. Hara. A unified approach to LMI-based reduced order self-scheduling control synthesis. *Systems & Control Letters*, 36:75–86, 1999.
19. K.B. Athreya and S.G. Pantula. Mixing properties of Harris chains and autoregressive processes. *Journal of Applied Probability*, 23:880–892, 1986.
20. D.M. Auslander, R.C. Spear, and G.E. Young. A simulation-based approach to the design of control systems with uncertain parameters. *Journal of Dynamic Systems, Measurement, and Control*, 104(1):20–26, 1982.
21. T. Bäck, F. Hoffmeister, and H.-P. Schwefel. A survey of evolution strategies. In *Proceedings of the Fourth International Conference on Genetic Algorithms*, 1991.
22. E.-W. Bai, R. Tempo, and M. Fu. Worst-case properties of the uniform distribution and randomized algorithms for robustness analysis. *Mathematics of Control, Signals, and Systems*, 11:183–196, 1998.
23. J.A. Ball, I. Gohberg, and L. Rodman. Interpolation of rational matrix functions. In *Operator Theory: Advances and Applications*, volume 45. Birhauser, Basel, 1990.
24. 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*, 10:203–222, 1997.
25. B.R. Barmish and C.M. Lagoa. On convexity of the probabilistic design problem for quadratic stabilizability. In *Proceedings of the American Control Conference*, 1999.
26. B.R. Barmish, C.M. Lagoa, and R. Tempo. Radially truncated uniform distributions for probabilistic robustness of control systems. In *Proceedings of the American Control Conference*, 1997.
27. B.R. Barmish and B.T. Polyak. A new approach to open robustness problems based on probabilistic prediction formulae. In *Proceedings of the IFAC World Congress*, 1996.
28. B.R. Barmish and P.S. Scherbakov. A dilation method for robustness problems with nonlinear parameter dependence. *Proceedings of the American Control Conference*, pages 3834–3839, 2003.
29. O. Barndorff-Nielsen. Unimodality and exponential families. *Comm. Statistics*, 1:189–216, 1973.
30. C. Barnhart, E.L. Johnson, G.L. Nemhauser, M.V.P. Savelsbergh, and P.H. Vance. Branch-and-price: Column generation for solving huge integer programs. *Operations Research*, 46:316–329, 1998.
31. G. Becker and A. Packard. Robust performance of linear parametrically varying systems using parametrically-dependent linear feedback. *Systems & Control Letters*, 23:205–215, 1994.
32. T. Becker and V. Weispfenning. *Gröbner Bases.* Springer-Verlag, New York, 1993.

33. R. Bellman. *Adaptive Control Processes.* Princeton University Press, Princeton, 1961.
34. A. Ben-Tal and A. Nemirovski. Robust truss topology design via semidefinite programming. *SIAM Journal on Optimization*, 7(4):991–1016, 1997.
35. A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23:769–805, 1998.
36. A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.
37. A. Ben-Tal and A. Nemirovski. On tractable approximations of uncertain linear matrix inequalities affected by interval uncertainty. *SIAM Journal on Optimization*, 12(3):811–833, 2002.
38. A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximation.* Springer-Verlag, 1990.
39. P. Beraldi and A. Ruszczyński. The probabilistic set covering problem. *Operations Research*, 50:956–967, 1999.
40. P. Beraldi and A. Ruszczyński. A branch and bound method for stochastic integer problems under probabilistic constraints. *Optimization Methods and Software*, 17:359–382, 2002.
41. D.P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, Belmont, MA, 1995.
42. D.P. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.
43. D.P. Bertsekas and S. Ioffe. Temporal differences-based policy iteration and applications in neuro-dynamic programming. Technical Report LIDS-P-2349, MIT Laboratory for Information and Decision Systems, 1996.
44. D.P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.
45. H.-G. Beyer. Toward a theory of evolution strategies: On the benefits of sex – the $(\mu/\mu,\lambda)$ theory. *Evolutionary Computation*, 3:81–111, 1995.
46. H.-G. Beyer and D.V. Arnold. Theory of evolution strategies - a tutorial. In *Theoretical Aspects of Evolutionary Computing (L. Kallel and B. Naudis, eds.)*, pages 109–133. Springer-Verlag, New York, 2001.
47. K. Binder. *Monte Carlo Methods in Statistical Physics.* Springer-Verlag, 1986.
48. F. Blanchini. The gain scheduling and the robust state feedback stabilization problems. *IEEE Transactions on Automatic Control*, 45(11):2061–2070, November 2000.
49. F. Blanchini and S. Miani. Stabilization of LPV systems: state feedback, state estimation and duality. *SIAM Journal on Control and Optimization*, 32(1):76–97, 2003.
50. R.G. Bland, D. Goldfarb, and M.J. Todd. The ellipsoid method: A survey. *Operations Research*, 29(6):1039–1091, 1981.
51. C. Borell. Convex measures on locally convex spaces. *Ark. Mat.*, 12:239–252, 1974.
52. C. Borell. The Brunn-Minkowski inequality in Gauss space. *Inventiones Mathematicae*, 30(2):207–216, 1975.
53. C. Borell. Convex set functions in d-space. *Periodica Mathematica Hungarica*, 6:111–136, 1975.
54. V.S. Borkar and S.P. Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38:447–469, 2000.

55. E. Boros and A. Prékopa. Closed form two-sided bounds for probabilities that exactly r and at least r out of n events occur. *Math. Oper. Res.*, 14:317–342, 1989.
56. N.K. Bose. *Applied Multidimentional System Theory.* Electrical Computer Science and Engineering. Van Nostrand Reinhold, New York, 1982.
57. S. Boyd and C.H. Barrat. *Linear Controller Design - Limits of Performance.* Prentice-Hall, Englewood Cliffs, 1991.
58. S. Boyd and L. El Ghaoui. Method of centers for minimizing generalized eigenvalues. *Linear Algebra and its Applications, Special Issue on Systems and Control*, 188:63–111, 1993.
59. S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory.* SIAM, Philadelphia, 1994.
60. S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge, 2003.
61. S. Boyd and Q. Yang. Structured and simultaneous Lyapunov functions for system stability problems. *International Journal of Control*, 49:2215–2240, 1989.
62. R.P. Braatz, P.M. Young, J.C. Doyle, and M. Morari. Computational complexity of μ calculation. *IEEE Transactions on Automatic Control*, 39:1000–1002, 1994.
63. H.J. Brascamp and E.H. Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems including inequalities for log concave functions, and with an application to the diffusion equations. *Journal of Functional Analysis*, 22:366–389, 1976.
64. L. Breiman. *Probability.* Addison-Wesley, Reading, 1968.
65. I.N. Bronshtein and K.A. Semendyayev. *Handbook of Mathematics.* Springer-Verlag, Berlin, 1998.
66. G.W. Brown. Monte Carlo methods. In E.F. Beckenbach, editor, *Modern Mathematics for the Engineer.* McGraw-Hill, New York, 1956.
67. H. Burgiel. How to lose at Tetris. *Mathematical Gazette*, page 194, July, 1997.
68. P.E. Caines. Stationary linear and nonlinear system identification and predictor set completeness. *IEEE Transactions on Automatic Control*, 23:583–594, 1978.
69. G. Calafiore and M.C. Campi. Robust convex programs: randomized solutions and confidence levels. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
70. G. Calafiore and M.C. Campi. A new bound on the generalization rate of sampled convex programs. In *Proceedings of the IEEE Conference on Decision and Control*, 2004.
71. G. Calafiore and M.C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.
72. G. Calafiore and F. Dabbene. Control design with hard/soft performance specifications: a Q-parameter randomisation approach. *International Journal of Control*, 77:461–471, 2004.
73. G. Calafiore, F. Dabbene, and R. Tempo. Radial and uniform distributions in vector and matrix spaces for probabilistic robustness. In D.E. Miller and L. Qiu, editors, *Topics in Control and its Applications*, pages 17–31. Springer-Verlag, New York, 1999.
74. G. Calafiore, F. Dabbene, and R. Tempo. Randomized algorithms for probabilistic robustness with real and complex structured uncertainty. *IEEE Transactions on Automatic Control*, 45:2218–2235, 2000.

75. G. Calafiore and L. El Ghaoui. Ellipsoidal bounds for uncertain linear equations and dynamical systems. *Automatica*, 50(5):773–787, 2004.
76. G. Calafiore and B.T. Polyak. Stochastic algorithms for exact and approximate feasibility of robust LMIs. *IEEE Transactions on Automatic Control*, 46:1755–1759, 2001.
77. M. Campi and P.R. Kumar. Learning dynamical systems in a stationary environment. In *Proceedings of the IEEE Conference on Decision and Control*, 1996.
78. M.C. Campi and E. Weyer. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47:1329–1334, 2002.
79. S. Chandrasekaran, G.H. Golub, M. Gu, and A.H. Sayed. Parameter estimation in the presence of bounded data uncertainties. *SIAM Journal on Matrix Analysis and Applications*, 19:235–252, 1998.
80. A. Charnes and W.W. Cooper. Chance constrained programming. *Management Sci.*, 6:73–79, 1959.
81. A. Charnes, W.W. Cooper, and G.H. Symonds. Cost horizons and certainty equivalents; an approach to stochastic programming of heating oil. *Management Science*, 4:235–263, 1958.
82. J. Chen and G. Gu. *Control-oriented System Identification: an $\mathcal{H}_\infty$ Approach.* Wiley, New York, 2000.
83. J. Chen and S. Wang. Validation of linear fractional uncertain models: solutions via matrix inequalities. *IEEE Transactions on Automatic Control*, 41(6):844–849, 1996.
84. X. Chen and K. Zhou. A probabilistic approach to robust control. In *Proceedings of the IEEE Conference on Decision and Control*, 1997.
85. X. Chen and K. Zhou. Order statistics and probabilistic robust control. *Systems & Control Letters*, 35, 1998.
86. X. Chen and K. Zhou. Constrained robustness analysis and synthesis by randomized algorithms. *IEEE Transactions on Automatic Control*, 45:1180–1186, 2000.
87. X. Chen, K. Zhou, and J.L. Aravena. Fast construction of robustness degradation function. *SIAM Journal on Control and Optimization*, 42:1960–1971, 2004.
88. D.C. Chin. Comparative study of stochastic algorithms for system optimization based on gradient approximations. *IEEE Transactions on Systems, Man and Cybernetics, Part B*, 27:244–249, 1997.
89. V. Chvátal. *Linear Programming.* W.H. Freeman, New York, 1983.
90. C.J. Clopper and E.S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.
91. C.W. Clunies-Ross. Interval estimation for the parameter of a binomial distribution. *Biometrika*, 45:275–279, 1958.
92. W.J. Conover. *Practical Nonparametric Statistics.* Wiley, New York, 1980.
93. M. Cottrell, J.-C. Fort, and G. Malgouyres. Large deviations and rare events in the study of stochastic algorithms. *IEEE Transactions on Automatic Control*, 28(9):907–920, 1983.
94. L. Creamean, W.B. Dunbar, D. van Gogh, J. Hickey, E. Klavins, J. Meltzer, and R.M. Murray. The Caltech multi-vehicle wireless testbed. In *Proceedings of the IEEE Conference on Decision and Control*, 2002.
95. F. Dabbene. *Randomized Algorithms for Probabilistic Robustness Analysis and Design.* PhD thesis, Politecnico Di Torino, Italy, 1999.

96. G.B. Dantzig. Linear programming under uncertainty. *Management Sci.*, 1:197–206, 1955.
97. Y.A. Davydov. Mixing conditions for Markov chains. *Thy. Probab. Appl.*, 18:312–328, 1973.
98. D.P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.
99. D.P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29(3):462–478, 2004.
100. M.C. de Oliveira, J. Bernussou, and J.C. Geromel. A new discrete-time robust stability condition. *Systems & Control Letters*, 37:261–265, 1999.
101. M.C. de Oliveira, J.C. Geromel, and J. Bernussou. Extended H_2 and H_∞ norm characterizations and controller parametrizations for discrete-time systems. *International Journal of Control*, 75(9):666–679, 2002.
102. I. Deák. Three digit accurate multiple normal probabilities. *Numerische Mathematik*, 35:369–380, 1980.
103. D. Dentcheva. Regular castaing representations with application to stochastic programming. *SIAM Journal on Optimization*, 10:732–749, 2000.
104. D. Dentcheva, B. Lai, and A Ruszczyński. Dual methods for probabilistic optimization problems. *Mathematical Methods of OR*, 60(2):331–346, 2004.
105. D. Dentcheva, A. Prékopa, and A. Ruszczyński. Concavity and efficient points of discrete distributions in probabilistic programming. *Mathematical Programming*, 89:55–77, 2000.
106. D. Dentcheva, A. Prékopa, and A. Ruszczyński. Bounds for probabilistic integer programming problems. *Discrete Applied Mathematics*, 124:55–65, 2002.
107. D. Dentcheva and A. Ruszczyński. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization*, 14:548–566, 2003.
108. D. Dentcheva and A Ruszczyński. Convexification of stochastic ordering constraints. *Comptes Rendus de l'Academie Bulgare des Sciences*, 57:11–16, 2004.
109. D. Dentcheva and A. Ruszczyński. Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints. *Mathematical Programming*, 99:329–350, 2004.
110. D. Dentcheva and A. Ruszczyński. Portfolio optimization under stochastic dominance constraints. *Journal of Banking and Finance*, to appear.
111. D. Dentcheva and A Ruszczyński. Semi-infinite probabilistic constraints: optimality and convexification. *Optimization*, to appear.
112. J. Dippon and J. Renz. Weighted means in stochastic approximation of minima. *SIAM Journal on Control and Optimization*, 35:1811–1827, 1997.
113. K. Dowd. *Beyond Value at Risk. The Science of Risk Management.* Wiley, New York, 1997.
114. J. Doyle. Analysis of feedback systems with structured uncertainties. *IEE Proceedings*, 129(D):242–250, 1982.
115. M.E. Dyer, A.M. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM*, 38:1–17, 1991.
116. Y. Ebihara and T. Hagiwara. New dilated LMI characterizations for continuous-time control design and robust multiobjective control. In *Proceedings of the American Control Conference*, 2002.

117. S. Hohenberger E.D. Demaine and D. Liben-Nowell. Tetris is Hard, Even to Approximate. In *Proceedings of the 9th International Computing and Combinatorics Conference*, 2003.
118. A.E. Eiben, E.H.L. Aarts, and K.M. van Hee. Global convergence of genetic algorithms: A Markov chain analysis. In *Parallel Problem Solving from Nature (H.-P. Schwefel and R. Männer, eds.)*, pages 4–12. Springer-Verlag, Berlin and Heidelberg, 1991.
119. A. Eichhorn and W. Römisch. Polyhedral risk measures in stochastic programming. *Stochastic Programming E-Print Series*, 2004.
120. E.J. and A. Goldenberg. Robust control of a general servomechanism. *Automatica*, 11(5):461–471, 1975.
121. L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18:1035–1064, 1997.
122. B. Escande. Nonlinear dynamic inversion and LQ techniques. In J-F. Magni, S. Bennani, and J. Terlouw, editors, *Robust Flight Control, A Design Challenge (GARTEUR)*, volume 224 of *Lecture Notes in Control and Information Sciences*, pages 523–540. Springer-Verlag, Berlin, 1997.
123. B. Etkin. *Dynamics of Atmospheric Flight*. Wiley, New York, 1972.
124. E. Feron, P. Apkarian, and P. Gahinet. Analysis and synthesis of robust control systems via parameter-dependent Lyapunov functions. *IEEE Transactions on Automatic Control*, 41(7):1041–1046, 1996.
125. P.C. Fishburn. *Utility Theory for Decision Making*. Wiley, New York, 1970.
126. D.S.K. Fok and D. Crevier. Volume estimation by Monte Carlo methods. *Journal of Statistical Computation and Simulation*, 31(4):223–235, 1989.
127. H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6:429–447, 2002.
128. L.R Ford and D.R. Fulkerson. A suggested computation for maximal multicommodity network flows. *Management Science*, 5:97–101, 1958.
129. Y. Fujisaki, F. Dabbene, and R. Tempo. Probabilistic robust design of LPV control systems. *Automatica*, 39:1323–1337, 2003.
130. Y. Fujisaki and Y. Kozawa. Probabilistic robust controller design: Probable near minmax value and randomized algorithms. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
131. P. Gahinet, P. Apkarian, and M. Chilali. Affine parameter-dependent Lyapunov functions and real parametric uncertainty. *IEEE Transactions on Automatic Control*, 41(3):436–442, 1996.
132. R.E. Gangnon and W.N. King. Minimum distance estimation of the distribution functions of stochastically ordered random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51:485–498, 2002.
133. R. De Gaston and M.G. Safonov. Exact calculation of the multiloop stability margin. *IEEE Transactions on Automatic Control*, 33:156–171, 1988.
134. S. Gelfand and S.K. Mitter. Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 31:111–131, 1993.
135. J.E. Gentle. *Random Number Generation and Monte Carlo Methods*. Springer-Verlag, New York, 1998.
136. T.T. Georgiou and M.C. Smith. Optimal robustness in the gap metric. *IEEE Transactions on Automatic Control*, 35:673–686, 1992.

137. L. Gerencsér. Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting. *IEEE Transactions on Automatic Control*, 44:894–905, 1999.
138. L. Gerencsér and Z. Z. Vago. The mathematics of noise-free SPSA. In *Proceedings of the IEEE Conference on Decision and Control*, 2001.
139. J.C. Geromel, J. Bernussou, G. Garcia, and M.C. de Oliveira. H_2 and H_1 robust filtering for discrete-time linear systems. *SIAM Journal on Control and Optimization*, 38(5):1353–1368, 2000.
140. L. El Ghaoui and G. Calafiore. Robust filtering for discrete-time systems with bounded noise and parametric uncertainty. *IEEE Transactions on Automatic Control*, 46(7):1084–1089, July 2001.
141. L. El Ghaoui and H. Lebret. Robust solutions to uncertain semidefinite programs. *SIAM J. Optim.*, 9(1):33–52, 1998.
142. L. El Ghaoui, M. Oks, and F. Oustry. Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, 51(4):543–556, July 2003.
143. E.G. Gilbert. Conditions for minimizing the norm sensitivity of characteristic roots. *IEEE Transactions on Automatic Control*, 29(7):658–661, 1984.
144. P. Glasserman. *Monte Carlo Methods in Financial Engineering.* Applications of Mathematics. Springer-Verlag, 2003.
145. M.A. Goberna and M.A. Lopez. *Linear Semi-Infinite Optimization.* Wiley, 1998.
146. A. Göpfert. *Mathematische Optimierung in allgemeinen Vektorraumen.* Teubner, Leipzig, 1973.
147. G.R. Grimmett and D.R. Stirzaker. *Probability and Random Processes.* The Clarendon Press Oxford University Press, New York, second edition, 1992.
148. C. Guestrin, D. Koller, and R. Parr. Efficient Solution Algorithms for Factored MDPs. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.
149. K.S. Gunnarsson. Design of stability augmentation system using μ-synthesis. In J-F. Magni, S. Bennani, and J. Terlouw, editors, *Robust Flight Control, A Design Challenge (GARTEUR)*, volume 224 of *Lecture Notes in Control and Information Sciences*, pages 484–502. Springer-Verlag, Berlin, 1997.
150. J. Hadar and W. Russell. Rules for ordering uncertain prospects. *The American Economic Review*, 59:25–34, 1969.
151. A. Hald. *Statistical Theory with Engineering Applications.* John Wiley and Sons, 1952.
152. W.K. Klein Haneveld. *Duality in Stochastic Linear and Dynamic Programming*, volume 274 of *Lecture Notes in Economics and Mathematical Systems.* Springer-Verlag, New York, 1986.
153. W.E. Hart. A stationary point convergence theory for evolutionary algorithms. In R.K. Belew and M.D. Vose, editors, *Foundations of Genetic Algorithms, 4*, pages 325–342. Morgan Kauffmann, 1997.
154. D. Henrion, R. Tarbouriech, and D. Arzelier. LMI approximations for the radius of the interconnection of ellipsoids: Survey. *Journal of Optimization Theory and Applications*, 108(1):10–21, 2001.
155. R. Henrion. On the connectedness of probabilistic constraint sets. *Journal of Optimization Theory and Applications*, 112:657–663, 2002.
156. R. Henrion. Perturbation analysis of chance-constrained programs under variation of all constraint data. In K. Marti *et al.*, editor, *Dynamic Stochastic*

Optimization, Lecture Notes in Economics and Mathematical Systems, pages 257–274. Springer-Verlag, Heidelberg, 2003.

157. R. Henrion and W. Römisch. Hölder and lipschitz stability of solution sets in programs with probabilistic constraints. *Mathematical Programming*, 100:589–611, 2004.
158. R. Hettich and K. Kortanek. Semi-infinite programming: Theory, methods and applications. *SIAM Review*, 35(3):380–429, 1993.
159. J.L. Higle and S. Sen. On the convergence of algorithms with implications for stochastic and nondifferentiable optimization. *Mathematics of Operations Research*, 17:112–131, 1992.
160. H.A. Hindi and S.P. Boyd. Robust solutions to l_1, l_2, and l_∞ uncertain linear approximation problems using convex optimization. In *Proceedings of the American Control Conference*, volume 6, 1998.
161. J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I and II.* Springer-Verlag, New York, 1993.
162. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
163. P.F. Hokayem, C.T. Abdallah, and P. Dorato. Quasi-Monte Carlo methods in robust control. In *Proceedings of the Mediterranean Conference on Control and Automation*, 2003.
164. P.F. Hokayem, C.T. Abdallah, and P. Dorato. Quasi-Monte Carlo methods in robust control design. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
165. N. Hovakimyan, F. Nardi, A. Calise, and N. Kim. Adaptive output feedback control of uncertain systems using single hidden layer neural networks. *IEEE Transactions on Neural Networks*, 13(6):1420–1431, 2002.
166. J.W. Howze and R.K. Cavin III. Regulator design with model insensitivity. *IEEE Transactions on Automatic Control*, 24(3):466–469, 1979.
167. L.K. Hua and Y. Wang. *Applications of Number Theory to Numerical Analysis.* Springer-Verlag, Berlin, 1981.
168. M. Iosifescu. *Finite Markov Processes and Their Applications.* Wiley, New York, 1980.
169. K. Ito. On stochastic differential equations. *Mem. Am. Math. Soc.*, 4, 1951.
170. Z. Jarvis-Wloszek, R. Feeley, W. Tan, K. Sun, and A.Packard. Some controls applications of sum of square programming. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
171. N.L. Johnson and S. Kotz. *Continuous Univariate Distributions - 1.* Houghton Mifflin, Boston, 1970.
172. W.B. Johnson and G. Schechtman. Remarks on Talagrand's deviation inequality for Rademacher functions. *Springer Lecture Notes on Mathematics*, 1470:72–77, 1991.
173. T. Kailath, A. Sayed, and B. Hassibi. *Linear Estimation.* Information and System Science. Prentice Hall, Upper Saddle River, NJ, 2000.
174. S. Kakade. A Natural Policy Gradient. In *Advances in Neural Information Processing Systems 14*, 2002.
175. M.H Kalos and P.A. Whitlock. *Monte Carlo Methods.* Wiley, New York, 1986.
176. S. Kanev, B. De Schutter, and M. Verhaegen. An ellipsoid algorithm for probabilistic robust controller design. *Systems & Control Letters*, 49:365–375, 2003.

177. S. Kanev and M. Verhaegen. Robust output-feedback integral MPC: A probabilistic approach. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
178. V. Kankova. On the convergence rate of empirical estimates in chance constrained stochastic programming. *Kybernetika (Prague)*, 26:449–461, 1990.
179. R. Kannan, L. Lovász, and M. Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms*, 11:1–50, 1997.
180. R.L. Karandikar and M. Vidyasagar. Probably approximately correct learning with beta-mixing input sequences. *Submitted for publication*, 2005.
181. M. Karpinski and A.J. Macintyre. Polynomial bounds for VC dimension of sigmoidal neural networks. In *Proc. 27th ACM Symp. Thy. of Computing*, 1995.
182. P. Khargonekar and A. Tikku. Randomized algorithms for robust control analysis and synthesis have polynomial complexity. In *Proceedings of the IEEE Conference on Decision and Control*, 1996.
183. V.L. Kharitonov. Asymptotic stability of an equilibrium position of a family of systems of linear differential equations. *Differentsial'nye Uravneniya*, 14:2086–2088, 1978 (in Russian).
184. A.I. Kibzun and G.L. Tretyakov. Differentiability of the probability function (russian). *Doklady Akademii Nauk*, 354:159–161, 1997.
185. A.I. Kibzun and S. Uryasev. Differentiability of probability function. *Stochastic Analysis and Applications*, 16:1101–1128, 1998.
186. A.J. King and R.T. Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic optimization. *Mathematics of Operations Research*, 18:148–162, 1993.
187. S. Kirkpatrick, C.D. Gelatt, and M.P Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
188. K.C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization*, volume 1133 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1985.
189. V. Koltchinskii, C.T. Abdallah, M. Ariola, P. Dorato, and D. Panchenko. Statistical learning control of uncertain systems: It is better than it seems. Technical Report 99-001, EECE Dept., The University of New Mexico, Albuquerque, NM, 1999.
190. V. Koltchinskii, C.T. Abdallah, M. Ariola, P. Dorato, and D. Panchenko. Improved sample complexity estimates for statistical learning control of uncertain systems. *IEEE Transactions on Automatic Control*, 46:2383–2388, 2000.
191. E. Komáromi. A dual method of probabilistic constrained problem. *Mathematical Programming Study*, 28:94–112, 1986.
192. F. Kozin. A survey of stability of stochastic systems. *Automatica*, 5:95–112, 1969.
193. K. Krishnakumar and D.E. Goldberg. Control system optimization using genetic algorithms. *Journal of Guidance, Control, and Dynamics*, 15(3):735–740, 1992.
194. L. Kuipers and H. Niederreiter. *Uniform Distribution of Sequences.* Wiley, 1974.
195. H.J. Kushner. *Stochastic Stability and Control.* Academic Press, New York, 1967.
196. H.J. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Springer-Verlag, New York, 2003.

197. P.I. Kuznetsov, R.L. Stratonovich, and V.I. Tikhonov. *Non-Linear Transformations of Stochastic Processes.* Pergamon Press, Oxford, 1965.
198. C.M. Lagoa. Probabilistic enhancement of classic robustness margins: A class of non symmetric distributions. *Proceedings of the American Control Conference*, pages 3802–3806, 2000.
199. C.M. Lagoa. A convex parametrization of risk-adjusted stabilizing controllers. *Automatica*, 39(10):1829–1835, October 2003.
200. C.M. Lagoa, X. Li, M.C. Mazzaro, and M. Sznaier. Sampling random transfer functions. *Proceedings of the IEEE Conference on Decision and Control*, pages 2429–2434, 2003.
201. C.M. Lagoa, X. Li, and M. Sznaier. On the design of robust controllers for arbitrary uncertainty structures. In *Proceedings of the American Control Conference*, 2003.
202. C.M. Lagoa, X. Li, and M. Sznaier. Probabilistically constrained linear programs and risk-adjusted controller design. *SIAM Journal on Optimization*, page to appear, 2005.
203. E. Lehmann. Ordered families of distributions. *Annals of Mathematical Statistics*, 26:399–419, 1955.
204. E.L. Lehmann. *Theory of Point Estimation.* Wiley, New York, 1983.
205. D. Liberzon and R. Tempo. Gradient algorithms for finding common Lyapunov functions. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
206. D. Liberzon and R. Tempo. Common Lyapunov functions and gradient algorithms. *IEEE Transactions on Automatic Control*, 49:990–994, 2004.
207. K.B. Lim and J.L. Junkins. Probability of stability: New measures of stability robustness for linear dynamical systems. *J. of Astro. Sci.*, 35(4):383–397, 1987.
208. L. Ljung. Convergence analysis of parametric identification methods. *IEEE Transactions on Automatic Control*, 23:770–783, 1978.
209. L. Ljung, G. Pflug, and H. Walk. *Stochastic Approximation and Optimization of Random Systems.* Birkhuser, Boston, 1992.
210. J.-F. Magni, S. Bennani, and J. Terlouw (editors). *Robust Flight Control, A Design Challenge (GARTEUR)*, volume 224 of *Lecture Notes in Control and Information Sciences.* Springer-Verlag, Berlin, 1997.
211. W.-K. Mak, D.P. Morton, and R.K. Wood. Monte-Carlo bounding techniques for determining solution quality in stochastic programs. *Mathematics of Operations Research*, 24:47–56, 1999.
212. H.B. Mann and D.R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60, 1947.
213. J.A. Markerink. Design of a robust scheduled controller using μ-synthesis. In J-F. Magni, S. Bennani, and J. Terlouw, editors, *Robust Flight Control, A Design Challenge (GARTEUR)*, volume 224 of *Lecture Notes in Control and Information Sciences*, pages 503–522. Springer-Verlag, Berlin, 1997.
214. H.M. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952.
215. C.I. Marrison. *The Design of Control Laws for Uncertain Dynamic Systems Using Stochastic Robustness Metrics.* PhD thesis, Princeton University, 1995.
216. C.I. Marrison and R.F. Stengel. The use of random search and genetic algorithms to optimize stochastic robustness functions. In *Proceedings of the American Control Conference*, 1994.

217. C.I. Marrison and R.F. Stengel. Stochastic robustness synthesis applied to a benchmark control problem. *International Journal of Robust and Nonlinear Control*, 5(1):13–31, 1995.
218. C.I. Marrison and R.F. Stengel. Robust control system design using random search and genetic algorithms. *IEEE Transactions on Automatic Control*, 42:835–839, 1997.
219. C.I. Marrison and R.F. Stengel. Design of robust control systems for a hypersonic aircraft. *Journal of Guidance, Control, and Dynamics*, 21(1):58–63, 1998.
220. J.L. Maryak and D.C. Chin. Global random optimization by simultaneous perturbation stochastic approximation. In *Proceedings of the American Control Conference*, 2001.
221. P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18:1269–1283, 1990.
222. S. Mastellone, P.F. Hokayem, C.T. Abdallah, and P. Dorato. Nonlinear stability analysis for non-polynomial systems. In *Proceedings of the American Control Conference*, 2004.
223. J. Mayer. Computational techniques for probabilistic constrained optimization problems. In K. Marti, editor, *Stochastic Optimization, Numerical Methods and Technical Applications*, volume 379 of *Lecture Notes in Economics and Math. Systems*, page 141164. Springer-Verlag, Berlin, 1992.
224. J. Mayer. On the numerical solution of jointly chance constrained problems. In S. Uryasev, editor, *Probabilistic Constrained Optimization: Methodology and Applications*, pages 220–233. Kluwer Academic Publishers, 2000.
225. M. Mayer and S. Reisner. Characterizations of affinely-rotation-invariant log-concave measures by section-centroid location. In *Geometric Aspects of Functional Analysis*, Lecture Notes in Mathematics, pages 145–152. Springer-Verlag, Berlin, 1991.
226. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A. Teller, and H. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.
227. S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, 1996.
228. G.A. Mikhailov. *New Monte Carlo Methods with Estimating Derivatives*. Brill Academic Publishers, The Netherlands, 1995.
229. I. Miller and J.E. Freund. *Probability and Statistics for Engineers*. Prentice-Hall, Englewood Cliffs, 1977.
230. L.B. Miller and H. Wagner. Chance-constrained programming with joint constraints. *Operations Research*, 13:930–945, 1965.
231. B.S. Morgan Jr. Sensitivity analysis and synthesis of multivariable systems. *IEEE Transactions on Automatic Control*, 11(4):506–512, 1966.
232. K. Mosler and M. Scarsini. *Stochastic Orders and Decision Under Risk*. Institute of Mathematical Statistics, Hayward, California, 1991.
233. T. Motoda, R.F. Stengel, and Y. Miyazawa. Robust control system design using simulated annealing. *Journal of Guidance, Control, and Dynamics*, 25(2):267–274, 2002.
234. R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, 1995.

235. E. Muir. Robust flight control design challenge problem formulation and manual: The High Incidence Research Model (HIRM). In J-F. Magni, S. Bennani, and J. Terlouw, editors, *Robust Flight Control, A Design Challenge (GARTEUR)*, volume 224 of *Lecture Notes in Control and Information Sciences*, pages 419–443. Springer-Verlag, Berlin, 1997.
236. E. Muir. The robust inverse dynamics estimation approach. In J-F. Magni, S. Bennani, and J. Terlouw, editors, *Robust Flight Control, A Design Challenge (GARTEUR)*, volume 224 of *Lecture Notes in Control and Information Sciences*, pages 541–563. Springer-Verlag, Berlin, 1997.
237. A. Müller and D. Stoyan. *Comparison Methods for Stochastic Models and Risks.* Wiley, Chichester, 2002.
238. K. Mulmuley. *Computational Geometry: An Introduction through Randomization Algorithms.* Prentice-Hall, Englewood Cliffs, 1994.
239. G.L. Nemhauser and L.A. Wolsey. *Integer and Combinatorial Optimization.* Wiley, New York, 1988.
240. A. Nemirovski. On tractable approximations of randomly perturbed convex constraints. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
241. A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, New York, 1983.
242. Y. Nesterov and J.-Ph. Vial. *Confidence level solutions for stochastic programming.* Technical Report, Université Catholique de Louvain, 2000.
243. J. Neveu. *Discrete Parameter Martingales.* North Holland, 1975.
244. H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods.* SIAM, Philadelphia, 1992.
245. V.I. Norkin and N.V. Roenko. α-concave functions and measures and their applications. *Kibernet. Sistem. Anal.*, pages 77–88, 1991.
246. W. Ogryczak and A. Ruszczyński. From stochastic dominance to mean-risk models: semideviations as risk measures. *European Journal of Operational Research*, 116:33–50, 1999.
247. W. Ogryczak and A. Ruszczyński. On consistency of stochastic dominance and mean-semideviation models. *Mathematical Programming*, 89:217–232, 2001.
248. W. Ogryczak and A. Ruszczyński. Dual stochastic dominance and related mean risk models. *SIAM Journal on Optimization*, 13:60–78, 2002.
249. Y. Oishi. Probabilistic design of a robust state-feedback controller based on parameter-dependent Lyapunov functions. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
250. Y. Oishi. Polynomial-time algorithms for probabilistic solutions of parameter-dependent linear matrix inequalities. Technical Report METR 2004-23, Department of Mathematical Informatics, The University of Tokyo, Japan, 2004. http://www.keisu.t.u-tokyo.ac.jp/Research/techrep.0.html. Submitted to Automatica.
251. Y. Oishi and H. Kimura. Randomized algorithms to solve parameter-dependent linear matrix inequalities. Technical Report METR 2001-04, Department of Mathematical Informatics, The University of Tokyo, Japan, 2001. http://www.keisu.t.u-tokyo.ac.jp/Research/techrep.0.html.
252. Y. Oishi and H. Kimura. Computational complexity of randomized algorithms for solving parameter-dependent linear matrix inequalities. *Automatica*, 39:2149–2156, 2003.

253. A. Packard and J. Doyle. The complex structured singular value. *Automatica*, 29:71–109, 1993.
254. F. Paganini. *Sets and Constraints in the Analysis of Uncertain Systems.* PhD thesis, Caltech, Pasadena, California, 1996.
255. B. Palka. *An Introduction to Complex Function Theory.* Springer-Verlag, New York, 1992.
256. A. Papachristodoulou. Analysis of nonlinear time-delay systems using the sum of squares decomposition. In *Proceedings of the American Control Conference*, 2004.
257. A. Papachristodoulou and S. Prajna. On the construction of Lyapunov functions using the sum of squares decomposition. In *Proceedings of the IEEE Conference on Decision and Control*, 2002.
258. G. Papageorgiou, K. Glover, and R.A. Hyde. The H_∞ loop shaping approach. In J-F. Magni, S. Bennani, and J. Terlouw, editors, *Robust Flight Control, A Design Challenge (GARTEUR)*, volume 224 of *Lecture Notes in Control and Information Sciences*, pages 464–483. Springer-Verlag, Berlin, 1997.
259. A. Papoulis and S.U. Pillai. *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, New York, 2002.
260. P.A. Parrilo. *Structural semidefinite programs and semialgebraic geometry methods in robustness and optimization.* PhD thesis, California Institute of Technology, Pasadena, California, 2000.
261. P.A. Parrilo. An explicit construction of distinguished representations of polynomials nonnegative over finite sets. IfA Technical Report AUT02-02, Automatic Control Laboratory Swiss Federal Institute of Technology, Zürich, Switzerland, March 2002.
262. P.A. Parrilo. Exploiting structure in sum of square programs. In *Proceedings of the IEEE Conference on Decision and Control*, 2003.
263. K.R. Parthasarathy. *Probability Measures on Metric Spaces.* Academic Press, New York, 1967.
264. D. Peaucelle and D. Arzelier. Robust performance analysis with LMI-based methods for real parametric uncertainty via parameter-dependent Lyapunov functions. *IEEE Transactions on Automatic Control*, 46(4):624–630, 2001.
265. G. Pflug. *Optimization of Stochastic Models.* Kluwer Academic Publishers, Boston, 1996.
266. G. Pflug and A. Ruszczyński. A risk measure for income processes. In *Risk Measures for the 21st Century, G. Szegö (Ed.).* John Wiley & Sons, 2004.
267. R.R. Phelps. *Convex Functions, Monotone Operators, and Differentiability*, volume 1364 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, 1989.
268. C. Pierre. Root sensitivity to parameter uncertainties: A statistical approach. *International Journal of Control*, 49(2):521–532, 1989.
269. E. Polak. On the use of consistent approximations in the solution of semi-infinite optimization and optimal control problems. *Mathematical Programming*, 62:385–414, 1993.
270. E. Polak. *Optimization: Algorithms and Consistent Approximations.* Springer-Verlag, New York, 1997.
271. B.T. Polyak. Random algorithms for solving convex inequalities. In D. Butnariu, Y. Censor, and S. Reich, editors, *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications.* Elsevier, Amsterdam, 2001.
272. B.T. Polyak and P.S. Shcherbakov. Random spherical uncertainty in estimation and robustness. *IEEE Transactions on Automatic Control*, 45:2145–2150, 2000.

273. B.T. Polyak and R. Tempo. Probabilistic robust design with linear quadratic regulators. *Systems & Control Letters*, 43:343–353, 2001.

274. K. Poolla, P. Khargonekar, A. Tikku, J. Krause, and K. Nagpal. A time-domain approach to model validation. *IEEE Transactions on Automatic Control*, 39(5):951–959, 1994.

275. A. Prékopa. On probabilistic constrained programming. In *Proceedings of the Princeton Symposium on Mathematical Programming*, pages 113–138. Princeton University Press, Princeton, New Jersey, 1970.

276. A. Prékopa. Logarithmic concave measures with applications to stochastic programming. *Acta Scientiarium Mathematicarum (Szeged)*, 32:301–316, 1971.

277. A. Prékopa. On logarithmic concave measures and functions. *Acta Scientiarium Mathematicarum (Szeged)*, 34:335–343, 1973.

278. A. Prékopa. Dual method for the solution of a one-stage stochastic programming problem with random RHS obeying a discrete probality distribution. *ZOR-Methods and Models of Operations Research*, 34:441–461, 1990.

279. A. Prékopa. Sharp bound on probabilities using linear programming. *Operations Research*, 38:227–239, 1990.

280. A. Prékopa. *Stochastic Programming.* Kluwer Academic Publishers, Dordrecht, 1995.

281. A. Prékopa. Probabilistic programming. In A. Rusczyński and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science.* Elsevier, Amsterdam, 2003.

282. A. Prékopa, B. Vízvári, and T. Badics. Programming under probabilistic constraint with discrete random variable. In L. Grandinetti *et al.*, editor, *New Trends in Mathematical Programming*, pages 235–255. Kluwer, Boston, 2003.

283. X. Qi and F. Palmeiri. Theoretical analysis of evolutionary algorithms with infinite population size in continuous space, Part I: Basic properties. *IEEE Transactions on Neural Networks*, 5:102–119, 1994.

284. J.P Quirk and R. Saposnik. Admissibility and measurable utility functions. *Review of Economic Studies*, 29:140–146, 1962.

285. G. Rappl. On linear convergence of a class of random search algorithms. *Zeitschrift für angewandt Mathematik und Mechanik (ZAMM)*, 69:37–45, 1989.

286. L.R. Ray. *Stochastic Robustness of Linear Multivariable Control System: Towards Comprehensive Robust Analysis.* PhD thesis, Princeton University, 1991.

287. L.R. Ray and R.F. Stengel. Application of stochastic robustness to aircraft control systems. *Journal of Guidance, Control, and Dynamics*, 14(6):1251–1259, 1991.

288. L.R. Ray and R.F. Stengel. Stochastic measures of performance robustness in aircraft control systems. *Journal of Guidance, Control, and Dynamics*, 15(6):1381–1387, 1992.

289. L.R. Ray and R.F. Stengel. Computer-aided analysis of linear control system robustness. *Mechatronics*, 3(1):119–124, 1993.

290. L.R. Ray and R.F. Stengel. A Monte Carlo approach to the analysis of control system robustness. *Automatica*, 29:229–236, 1993.

291. R. Reemtsen. Numerical methods for semi-infinite programming: A survey. In R. Reemtsen and J.-J. Rückmann, editors, *Semi-Infinite Programming.* Kluwer Academic Publishers, Dordrecht, 1998.

292. R. Reemtsen and J.-J. Rückmann (editors). *Semi-Infinite Programming.* Kluwer Academic Publishers, Dordrecht, 1998.

293. C.R. Reeves and J.E. Rowe. *Genetic Algorithms - Principles and Perspectives: A Guide to GA Theory.* Kluwer Academic Publishers, Boston, 2003.
294. F. Riedel. Dynamic coherent risk measures. *Stochastic Processes and Their Applications*, 112:185–200, 2004.
295. Y. Rinott. On convexity of measures. *Annals of Probability*, 4:1020–1026, 1976.
296. C.P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer-Verlag, 1999.
297. G.O. Roberts and J.S. Rosenthal. General state space Markov chains and MCMC algorithms. *Submitted for publication*, 2004.
298. R.T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, NJ, 1970.
299. R.T. Rockafellar. *Conjugate Duality and Optimization.* SIAM, Regional Conference Series in Applied Mathematics, Philadelphia, 1974.
300. R.T. Rockafellar and S. Uryasev. Optimization of conditional value at risk. *The Journal of Risk*, 2:21–41, 2000.
301. R.T. Rockafellar, S. Uryasev, and M. Zabarankin. Deviation measures in risk analysis and optimization. Technical Report 2002-7, Department of Industrial and Systems Engineering, University of Florida, 2002.
302. R.T. Rockafellar and R.J-B. Wets. *Variational Analysis.* Springer-Verlag, Berlin, 1998.
303. W. Römisch. Stability of stochastic programming problems. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 483–554. Elsevier, Amsterdam, 2003.
304. S.R. Ross and B.R. Barmish. Distributionally robust gain analysis for systems containing complexity. *Proceedings of the IEEE Conference on Decision and Control*, pages 5020–5025, 2001.
305. H. Rotstein. A Nevanlinna-Pick approach to time-domain constrained $\mathcal{H}_\infty$ control. *SIAM Journal on Control and Optimization*, 34(4):1329–1341, 1996.
306. G. Rudolph. Convergence analysis of canonical genetic algorithms. *IEEE Transactions on Neural Networks*, 5:96–101, 1994.
307. G. Rudolph. *Convergence Properties of Evolutionary Algorithms.* Kovac, Hamburg, 1997.
308. G. Rudolph. Convergence rates of evolutionary algorithms for a class of convex objective functions. *Control and Cybernetics*, 26:375–390, 1997.
309. G. Rudolph. Finite Markov chain results in evolutionary computation: A tour d'horizon. *Fundamenta Informaticae*, 34:1–22, 1998.
310. W.J. Rugh and J.S. Shamma. Research on gain-scheduling. *Automatica*, 36:1401–1425, 2000.
311. A. Ruszczyński. Probabilistic programming with discrete distributions and precedence constrained knapsack polyhedra. *Mathematical Programming*, 93:195–215, 2002.
312. A. Ruszczyński and A. Shapiro (editors). *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science.* Elsevier, Amsterdam, 2003.
313. A. Ruszczyński and A. Shapiro. Conditional risk mappings. *E-print available at* `http://www.optimization-online.org`, 2004.
314. A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *E-print available at* `http://www.optimization-online.org`, 2004.

315. K.K. Sabelfeld and N.A. Simonov. *Random Walks on Boundary for Solving PDEs.* VSP Intl. Science Publishers, 1994.
316. M.G. Safonov. Stability margins of diagonally perturbed multivariable feedback systems. *IEE Proceedings*, 129(D):251–256, 1982.
317. G. Salinetti. Approximations for chance constrained programming problems. *Stochastics*, 10:157–169, 1983.
318. R.S. Sánchez-Peña and M. Sznaier. *Robust Systems: Theory and Applications.* John Wiley, New York, 1998.
319. A.H. Sayed, V.H. Nascimento, and S. Chandrasekaran. Estimation and control with bounded data uncertainties. *Linear Algebra and Applications*, 248:259–306, 1999.
320. C.W. Scherer. LPV control and full block multipliers. *Automatica*, 37:361–375, 2001.
321. W.M. Schubert and R.F. Stengel. Parallel stochastic robustness synthesis for control system design. *IEEE Transactions on Control Systems Technology*, 6(6):701–706, 1998.
322. D. Schuurmans and R. Patrascu. Direct value-approximation for factored MDPs. In *Advances in Neural Information Processing Systems*, volume 14, 2001.
323. H.-P. Schwefel. *Evolution and Optimum Seeking.* Wiley, New York, 1995.
324. P. Schweitzer and A. Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110:568–582, 1985.
325. A.A. Schy and D.P. Geisy. Multiobjective insensitive design of airplane control systems with uncertain parameters. In *Proceedings of the AIAA Guidance, Navigation and Control Conference*, 1981.
326. S. Sen. Relaxations for the probabilistically constrained programs with discrete random variables. *Operations Research Letters*, 11:81–86, 1992.
327. R.J. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley, New York, 1980.
328. U. Shaked. Improved LMI representation for the analysis and the design of continuous-time systems with polytopic type uncertainty. *IEEE Transactions on Automatic Control*, 46(4):652–656, 2001.
329. D.F. Shanno and R.J. Vanderbei. Interior point methods for noncovex nonlinear programming: Orderings and higher order methods. *Mathematical Programming, Ser. B*, 87:303–316, 2000.
330. A. Shapiro. Asimptotic properties of statistical estimators in stochastic programming. *Annals of Statistics*, 17:841–858, 1989.
331. A. Shapiro. Duality, optimality conditions, and perturbation analysis. In H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors, *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, pages 68–92. Kluwer, Boston, USA, 2000.
332. A. Shapiro. Monte Carlo sampling methods. In A. Rusczyński and A. Shapiro, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science.* Elsevier, Amsterdam, 2003.
333. J.D. Shaughnessy, S.Z. Pinckney, J.D. McMinn, C.I. Cruz, and M.-L. Kelley. Hypersonic vehicle simulation model: Winged-cone configuration. Technical report, NASA TM 102610, 1990.

334. T. Shimomura, M. Takahashi, and T. Fujii. Extended-space control design with parameter-dependent Lyapunov functions. In *Proceedings of the 40th IEEE Conference on Decision and Control*, 2001.

335. A.N. Shiryaev. *Probability.* Springer, New York, USA, second edition, 1996.

336. D.D. Siljak. Parameter space methods for robust control design: A guided tour. Technical Report EECS-031588, University of California, Santa Clara, 1988.

337. R.S. Smith and J.C. Doyle. Model validation: a connection between robust control and identification. *IEEE Transactions on Automatic Control*, 37(7):942–952, 1992.

338. F.J. Solis and J.B. Wets. Minimization by random search techniques. *Mathematics of Operations Research*, 6:19–30, 1981.

339. J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37:332–341, 1992.

340. J.C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45:1839–1853, 2000.

341. J.C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control.* Wiley, New York, 2003.

342. R.C. Spear. The application of Kolmogorov-Rényi statistics to problems of parameter uncertainty in systems design. *International Journal of Control*, 11(5):771–778, 1970.

343. W.M. Spears. A compression algorithm for probability transition matrices. *SIAM Journal on Matrix Analysis and Applications*, 20:60–77, 1998.

344. D.R. Stark and J.C. Spall. Rate of convergence in evolutionary computation. In *Proceedings of the American Control Conference*, 2003.

345. R.F. Stengel. Some effects of parameter variations on the lateral-directional stability of aircraft. *AIAA Journal of Guidance and Control*, 3:124–131, 1980.

346. R.F. Stengel. *Optimal Control and Estimation.* Dover, New York, 1994.

347. R.F. Stengel and L.R. Ray. Stochastic robustness of linear time-invariant control systems. *IEEE Transactions on Automatic Control*, 36:82–87, 1991.

348. J. Suzuki. A Markov chain analysis on simple genetic algorithms. *IEEE Transactions on Systems, Man and Cybernetics*, 25:655–659, 1995.

349. T. Szántai. Evaluation of a special multivariate gamma distribution. *Mathematical Programming Study*, 27:1–16, 1986.

350. T. Szántai. A computer code for solution of probabilistic-constrained stochastic programming problems. In Y. Ermoliev and R.J.-B. Wets, editors, *Numerical Techniques for Stochastic Optimization*, pages 267–352. Springer-Verlag, Berlin, 1988.

351. T. Szántai. Improved bounds and simulation procedures on the value of the multivariate normal probability distribution function. *Annals of Oper. Res.*, 100:85–101, 2000.

352. R. Szekli. *Stochastic Ordering and Dependence in Applied Probability.* Springer-Verlag, New York, 1995.

353. L. Szirmay-Kalos, T. Fóris, L. Neumann, and B. Csébfalvi. An analysis of quasi-Monte Carlo integration applied to the transillumination radiosity method. *Eurographics*, 16(3), 1997.

354. L. Szirmay-Kalos and W. Purgathofer. Analysis of the quasi-Monte Carlo integration of the rendering equation. Technical report, Dept. of Control Eng. and Information Technology, Technical University of Budapest, Aug. 1998.

355. V.B. Tadić. Stochastic approximation with random truncations, state dependent noise and discontinuous dynamics. *Stochastics and Stochastics Reports*, 64:283–325, 1998.
356. K. Takagi and H. Nishimura. Gain-scheduled control of a tower crane considering varying load-rope length. *JSME International Journal, Ser. C*, 42(4):914–921, 1999.
357. E. Tamm. On g-concave functions and probability measures. *Eesti NSV Teaduste Akademia Toimetised (News of the Estonian Academy of Sciences) Füüs. Mat.*, 26:376–379, 1977.
358. R. Tempo, E.-W. Bai, and F. Dabbene. Probabilistic robustness analysis: Explicit bounds for the minimum number of samples. *Systems & Control Letters*, 30:237–242, 1997.
359. R. Tempo, G. Calafiore, and F. Dabbene. *Randomized Algorithms for Analysis and Control of Uncertain Systems.* Communications and Control Engineering Series. Springer-Verlag, London, 2004.
360. R. Tempo and F. Dabbene. Randomized algorithms for analysis and control of uncertain systems: An overview. In S.O. Moheimani, editor, *Perspectives in Robust Control*, Lecture Notes in Control and Information Science, pages 347–362. Springer-Verlag, London, 2001.
361. A. Tikhonov and V. Arsenin. *Solution to Ill-posed Problems.* Wiley, New York, 1977.
362. O. Toker and J. Chen. Time domain validation of structured uncertainty model sets. In *Proceedings of the IEEE Conference on Decision and Control*, volume 1, 1996.
363. Y.L. Tong. *Probability Inequalities in Multivariate Distributions.* Academic Press, New York, 1980.
364. J.F. Traub and A.G. Werschulz. *Complexity and Information.* Cambridge University Press, Cambridge, 1998.
365. M. Trick and S. Zin. A linear programming approach to solving dynamic programs. Unpublished manuscript, 1993.
366. M. Trick and S. Zin. Spline approximations to value functions: A linear programming approach. *Macroeconomic Dynamics*, 1, 1997.
367. A. Trofino and C.E. de Souza. Biquadratic stability of uncertain linear systems. *IEEE Transactions on Automatic Control*, 46(8):1303–1307, 2001.
368. K.C. Tsai and D.M. Auslander. A statistical methodology of designing controllers for minimum sensitivity of parameter variations. *Journal of Dynamic Systems, Measurement, and Control*, 110(6):126–133, 1984.
369. J.N. Tsitsiklis and B. Van Roy. Feature-Based Methods for Large Scale Dynamic Programming. *Machine Learning*, 22:59–94, 1996.
370. J.S. Tyler and F.B. Tuteur. The use of a quadratic performance index to design multivariable invariant plants. *IEEE Transactions on Automatic Control*, 11:84–92, 1966.
371. V.A. Ugrinovskii, R. Tempo, and Y. Fujisaki. A primal-dual setting for quadratic stability of uncertain systems. *Systems & Control Letters*, 52(1):39–48, 2004.
372. S. Uryasev. A differentiation formula for integrals over sets given by inclusion. *Numerical Functional Analysis and Optimization*, 10:827–841, 1989.
373. S. Uryasev. Derivatives of probability and integral functions. In P.M. Pardalos and C.M. Floudas, editors, *Encyclopedia of Optimization*, pages 267–352. Kluwer Academic Publishers, 2001.

374. S. Vajda. *Probabilistic Programming.* Academic Press, New York, 1972.
375. V.N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.
376. V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies to their probabilities. *Theory of Probability and Its Applications*, 16:264–280, 1971.
377. V.N. Vapnik and A.Ya. Chervonenkis. Necessary and and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and Its Applications*, 26:532–553, 1981.
378. M. Verhaegen and V. Verdult. *Filtering and System Identification: An Introduction.* TU-Delft/ITS, lecture notes for the course SC4040(ET4094) edition, 2003.
379. M. Vidyasagar. The graph metric for unstable plants and robustness estimates for feedback stability. *IEEE Transactions on Automatic Control*, 29:403–418, 1984.
380. M. Vidyasagar. *A Theory of Learning and Generalization.* Springer-Verlag, London, 1997.
381. M. Vidyasagar. Randomized algorithms for robust controller synthesis using statistical learning theory. *Automatica*, 37:1515–1528, 2001.
382. M. Vidyasagar. *Learning and Generalization: With Applications to Neural Networks (second edition).* Springer-Verlag, New York, 2002.
383. M. Vidyasagar. *Nonlinear Systems Analysis.* SIAM Journal on Applied Mathematics, Philadelphia, 2003.
384. M. Vidyasagar and V. Blondel. Probabilistic solutions to some NP-hard matrix problems. *Automatica*, 37:1397–1405, 2001.
385. J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, 1947.
386. M. Vose. Logarithmic convergence of random heuristic search. *Evolutionary Computation*, 4:395–404, 1997.
387. M. Vose. *The Simple Genetic Algorithm.* MIT Press, Cambridge, 1999.
388. J. Žáčková. On minimax solutions of stochastic linear programming problems. *Čas. Pěst. Mat.*, 91:423–430, 1966.
389. Q. Wang. *Stochastic Robust Control of Nonlinear Dynamic Systems.* PhD thesis, Princeton University, 2001.
390. Q. Wang and R.F. Stengel. Robust control of nonlinear systems with parametric uncertainty. In *Proceedings of the IEEE Conference on Decision and Control*, 1998.
391. Q. Wang and R.F. Stengel. Searching for robust minimal-order compensators. *ASME Journal of Dynamic Systems, Measurement, and Control*, 123(2):233–236, 2001.
392. Q. Wang and R.F. Stengel. Robust control of nonlinear systems with parametric uncertainty. *Automatica*, 38:1591–1599, 2002.
393. R.J.-B. Wets. Stochastic programming. In G.L. Nemhauser, A.H.G. Rinnoy Kan, and M.J. Todd, editors, *Optimization.* North-Holland, Amsterdam, 1989.
394. E. Weyer. Finite sample properties of system identification of ARX models under mixing conditions. *Automatica*, 36:1291–1299, 2000.
395. E. Weyer and M.C. Campi. Non-asymptotic confidence ellipsoids for the least squares error estimate. *Automatica*, 38:1529–1547, 2002.
396. E. Weyer, R.C. Williamson, and I. Mareels. Sample complexity of least squares identification of FIR models. In *Proceedings of the IFAC World Congress*, 1996.

397. E. Weyer, R.C. Williamson, and I. Mareels. Finite sample properties of linear model identification. *IEEE Transactions on Automatic Control*, AC44:1370–1383, 1999.
398. B. Wie and D.S. Bernstein. A benchmark problem for robust control design. In *Proceedings of the American Control Conference*, 1990.
399. D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1:67–82, 1997.
400. F. Wu and K.M. Grigoriadis. LPV systems with parameter-varying time delays: Analysis and control. *Automatica*, 37:221–229, 2001.
401. L. Xie, Y.C. Soh, and C.E. de Souza. Robust Kalman filtering for uncertain discrete-time systems. *IEEE Transactions on Automatic Control*, 39(6):1310–1314, 1994.
402. E. Yaz. Deterministic and stochastic robustness measures for discrete systems. *IEEE Transactions on Automatic Control*, 33(10):952–955, 1988.
403. G.G. Yin. Rates of convergence for a class of global stochastic optimization algorithms. *SIAM Journal on Optimization*, 10:99–120, 1999.
404. A. Yoon and P. Khargonekar. Computational experiments in robust stability analysis. In *Proceedings of the IEEE Conference on Decision and Control*, 1997.
405. A. Yoon, P. Khargonekar, and K. Hebbale. Design of computer experiments for open-loop control and robustness analysis of clutch-to-clutch shifts in automatic transmissions. In *Proceedings of the American Control Conference*, 1997.
406. G.E. Young and D.M. Auslander. A design methodology for nonlinear systems containing parameter uncertainty. *Journal of Dynamic Systems, Measurement, and Control*, 110(6):126–133, 1984.
407. G. Zames. Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms and approximate inverses. *IEEE Transactions on Automatic Control*, 26:301–320, 1981.
408. G. Zames and A.K. El-Sakkary. Unstable systems and feedback: The gap metric. In *Proceedings of the Allerton Conference*, 1980.
409. K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control.* Prentice-Hall, Upper Saddle River, 1996.
410. T. Zhou. Unfalsified probability estimation for a model set based on frequency domain data. *International Journal of Control*, 73(5):391–406, 2000.
411. X. Zhu, Y. Huang, and J.C. Doyle. Soft vs. hard bounds in probabilistic robustness analysis. In *Proceedings of the IEEE Conference on Decision and Control*, 1996.

Index