

Springer Series in Operations Research
and Financial Engineering

Georg Ch. Pflug
Alois Pichler

Multistage Stochastic Optimization

Springer Series in Operations Research and Financial Engineering

Series Editors:

Thomas V. Mikosch
Sidney I. Resnick
Stephen M. Robinson

More information about this series at
<http://www.springer.com/series/3182>

Georg Ch. Pflug • Alois Pichler

Multistage Stochastic Optimization



Springer

Georg Ch. Pflug
Department of Statistics and Operations
Research
University of Vienna
Vienna
Austria

Alois Pichler
Department of Industrial Economics
and Technology Management
Norwegian University of Science
and Technology
Trondheim
Norway

ISSN 1431-8598 ISSN 2197-1773 (electronic)
ISBN 978-3-319-08842-6 ISBN 978-3-319-08843-3 (eBook)
DOI 10.1007/978-3-319-08843-3
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014951437

© Springer International Publishing Switzerland 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Esther and Fumie

Preface

The topic of this book is multistage stochastic optimization. *Multistage* reflects the fact that an optimal decision is an entire strategy or policy, which is executed during subsequent instants of time, or in successive stages. The term *stochastic* emphasizes the fact that unforeseen events may happen during this process. This can be in favor of the initial goal or even jeopardizing the success of the entire task. The decision maker then will react to this new situation and take respective measures to ensure the initial goal.

These types of optimization problems are natural and common in managing and planning processes. Especially in an economic environment it is an everyday situation that incidents occur or new information is revealed, which influence the original goal. Those events require a respective response action by the decision maker: in logistics and production, incidents like machine or transport systems breakdowns, as well as new information about demands will require adaptive decisions. Fund managers will adjust the strategy on a regular (say weekly) basis in order to achieve the long-term investment goal. More generally, many managerial or planning supervisions are designed to keep track and to meet a predefined goal. Meetings on a regular basis are typically held to evaluate the current situation and to resolve actions or reactions.

The stochastic character of the problem might imply that there is some risk that the initial goal is not met. This respective risk is addressed as well and strategies to comply with this intrinsic risk are part of the optimization problem.

The present book is organized as follows. Chapter 1 of this book (Introduction) addresses typical situations, which give evidence that multistage decision making is superior to myopic single-stage decisions. It also sets the mathematical outline and addresses important principles and problem solution techniques.

Uncertainty in decision relevant data is expressed in terms of discrete stochastic processes, which can be represented as scenario trees. As trees are the basic data structures, concepts of distances are introduced and studied in detail in Chap. 2.

These distances allow measuring the deviation of the simplified decision model from the underlying, more complex model, and they allow comparing two decision models. Measuring this deviation gives a kind of quality control.

Risk and utility functionals are introduced in Chap. 3. It is their role to quantify the risk, which is associated with the current problem; they provide either goals (minimize risk) or constraints (risk should not exceed a given risk limit) for the decision situation at hand.

Chapter 4 demonstrates how the process should be organized which leads from observed data to viable optimization models, which are mostly just approximations of the reality. The quality of approximations is crucial for the quality of the decisions.

An important aspect in multistage optimization is time consistency. That is, for short, that optimal strategies, found at the beginning of the process, remain optimal in later stages and no change of strategies is necessary. Chapter 5 addresses time consistent decisions as well as some situations of time inconsistency.

A basic assumption in stochastic optimization is that the random outcomes of the scenario process are not known at the time of decision making, but the model for the random distributions is known. This assumption is relaxed in Chap. 7. Inexact knowledge of the probability model is called model ambiguity. In ambiguous situations the decisions are not only subject to outcome risk (aleatoric risk) but also to modeling errors (epistemic risk, i.e., the risk of having chosen the wrong distribution model). The decision maker is confronted not only with the random errors (aleatoric errors) but also with the fact that observations are obtained empirically and thus do not reflect the real world situation. The final chapter (Chap. 8) contains some examples of larger multistage optimization problems.

It is the intention of the book to summarize modern modeling aspects, provide theoretic foundations as well as some solution techniques. The book moreover summarizes the current status of research in multistage stochastic optimization. For this reason and in order to provide a comprehensive and complete presentation this book adapts content and results from previous publications and from currently unpublished manuscripts. This is made evident by proper citations wherever possible.

We wish to express our thanks to the University of Vienna, where it was possible to conduct respective research over years. The second author moreover is indebted to *Asgeir Tomasdard* and the Norwegian University of Science and Technology. They provided a generous and comfortable environment for completing this book.

Many colleagues helped improving the autograph with proofreading and important suggestions. We thank *Eric Laas-Nesbitt*, *Bita Analui*, *Peter Gross*, *Anna Timonina*, and *Raimund Kovacevic*.

Vienna, Austria
Trondheim, Norway
October 2014

Georg Ch. Pflug
Alois Pichler

A list of typos and errata will be maintained at Georg Pflug's homepage
<http://homepage.univie.ac.at/georg.pflug>.

Contents

1	Introduction	1
1.1	Multistage Decision Models	1
1.2	Fundamentals of Decision Making	2
1.2.1	Stochastic Problem Formulation	3
1.2.2	From Single-Stage to Multistage Decision Problems	7
1.3	Multistage Stochastic Optimization Versus Dynamic Optimization	21
1.4	Scenario Trees and Nested Distributions	23
1.4.1	Nested Distributions	29
1.4.2	Equivalence and Minimality	33
1.4.3	Convex Structures for Scenario Models	37
2	The Nested Distance	41
2.1	Distances of Probability Measures	42
2.1.1	Semi-Distances Generated by a Class of Test Functions	42
2.2	The Wasserstein Distance	46
2.3	Elementary Properties of the Wasserstein Distance	51
2.3.1	The Wasserstein Distance on the Real Line	53
2.4	Alternative Distances as Basis for the Wasserstein Distance	55
2.4.1	The Role of the Distance on the Underlying Space	55
2.4.2	Transformation of the Axis, and Fortet–Mourier Distances	55
2.5	Estimates Involving the Wasserstein Distance	58
2.6	Approximations in the Wasserstein Metric	62
2.7	The Wasserstein Distance in a Discrete Framework	63
2.8	Duality for the Wasserstein Metric	65
2.9	Continuity of the Dual Variables, and the Kantorovich–Rubinstein Theorem	69

2.10	Multistage Generalization: The Nested Distance	71
2.10.1	The Inherited Distance	71
2.10.2	The Nested Distance	74
2.10.3	The Nested Distance for Trees	79
2.11	Dual Representation of the Nested Distance	88
2.11.1	Martingale Representation of the Nested Distance	91
3	Risk and Utility Functionals	95
3.1	Single-Period Risk and Utility Functionals	95
3.2	Examples of Risk and Utility Functionals	97
3.3	Dual Representation of Risk Functionals	103
3.3.1	Kusuoka's Representation	103
3.3.2	The Dual Representation	105
3.4	An Alternative Description of Distortion Risk Functionals	110
3.5	The Impact of the Probability Measure on Risk Functionals	114
3.5.1	Compound Concavity and Convex-Concavity	114
3.5.2	Continuity with Respect to the Probability Measure	117
3.6	Conditional Risk Functionals	119
3.6.1	Properties of Conditional Risk Functionals	122
4	From Data to Models	125
4.1	Approximations of Single-Period Probability Distributions	126
4.1.1	Approximation Quality of the Monte Carlo Generation Method	127
4.1.2	Quasi-Monte Carlo Approximations	130
4.1.3	Optimal and Nearly Optimal Single-Period Discretizations	132
4.1.4	The Stochastic Approximation Algorithms for Multidimensional Quantization	142
4.1.5	Asymptotic Distribution of Optimal Quantizers	145
4.2	Approximations of Multiperiod Distributions	149
4.3	Construction of Scenario Trees	154
4.3.1	Distance Calculation	155
4.3.2	The Construction of Large Trees	157
4.4	Scenario Tree Reduction	163
4.5	Improvement of Approximating Trees	166
4.5.1	Improvement of the Probability Measure	167
4.5.2	Improvement of the Paths	168
4.6	An Alternative View on Approximations	172
5	Time Consistency	175
5.1	Time Consistency in Stochastic Decision Problems	176
5.2	Time Consistent Risk Functionals	179

5.3	Time Consistency and Decomposition	187
5.3.1	Composition of Risk Functionals	187
5.3.2	Multistage Decomposition of Risk Functionals: The Decomposition Theorem	188
5.4	Martingale Formulations of Time Inconsistent Stochastic Programs	196
5.4.1	Verification Theorems	199
5.4.2	An Algorithm for Sequential Improvement.....	202
5.4.3	Numerical Experiments	203
5.5	Dualization of Nonanticipativity Constraints	205
6	Approximations and Bounds	209
6.1	Two-Stage Problems, and Approximation in the Wasserstein Distance	209
6.2	Approximation in the Nested Distance Sense	211
6.3	Bounds	218
6.3.1	Lower Bounds by Changing the Probability Measure	219
6.3.2	Lower Bounds for Replacing the Scenario Process by Its Expectation	225
6.3.3	Bounds for Changing the Filtration	227
6.3.4	Upper Bounds by Inserting (Sub)Solutions	227
6.4	Martingale Properties	228
7	The Problem of Ambiguity in Stochastic Optimization	229
7.1	Single- or Two-Stage Models: Wasserstein Balls	234
7.2	Solution Methods for the Single- or Two-Stage Case.....	237
7.3	The Multistage Case	238
7.3.1	A Minimax Theorem.....	241
7.3.2	Ambiguity Sets Defined by Nested Transportation Kernels.....	245
7.3.3	Algorithmic Solution	247
7.4	Example: A Multiperiod Production / Inventory Control Problem	249
7.4.1	Mathematical Modeling Summary	251
7.4.2	Computational Results	252
8	Examples	257
8.1	Thermal Electricity Production	257
8.2	Hydro Electricity Production.....	261
8.3	Budget Management for Risk-Prone Countries	270
A	Risk Functionals: Definitions and Notations	275
A.1	Multiperiod Risk Functionals	279
A.2	Information Monotonicity	280

B Minimax Theorems	283
C Comparison of Weighted Norms	287
D The Canonical Construction for Nested Distributions	289
Bibliography	291
Index	299

Nomenclature

(Ω, \mathcal{F}, P)	probability space
\mathbb{E}	expectation operator
Ξ	state space of ξ , general metric space; often, $\Xi = \mathbb{R}^m$
\mathcal{F}	sigma-algebra
$(\Omega, \mathfrak{F}, P)$	filtered probability space (stochastic basis)
$t \in \{0, \dots, T\}$	stage, often referred to as <i>time</i> ; T is the final stage
x_t, ξ_t	the subscript t typically indicates the stage
$x_{s:t}$	the substring (x_s, \dots, x_t) of the vector $x = (x_0, x_1, \dots, x_T)$
\mathfrak{F}	filtration, $\mathfrak{F} = (\mathcal{F}_0, \dots, \mathcal{F}_T)$
$\mathfrak{F}_{s:t}$	the subfiltration $(\mathcal{F}_s, \dots, \mathcal{F}_t)$
$\xi_t \triangleleft \mathcal{F}_t$	ξ_t is measurable with respect to the sigma algebra \mathcal{F}_t
$\xi \triangleleft \mathfrak{F}$	$\xi_t \triangleleft \mathcal{F}_t$ for all stages t
\mathbb{P}	nested distribution
$\mathbb{P}^{v_t=n}$	nested distribution, conditioned on $v_t = n$
d	distance function on a metric space
d_r	Wasserstein distance of order r
\mathbf{d}_r	nested distance of order r ; also multistage, or process distance
\mathbb{T}	a tree structure, defined by the node sets and the precedence relations
$\mathcal{N}, \mathcal{N}_t$	collection of all nodes (all nodes at stage t , resp.) of a tree
\mathcal{N}_T	the leaf set of a tree, often identified with the sample space Ω
$n-$	direct predecessor of node n (ancestor node, parent node)
$n+$	collection of all direct successors of node n (children)
$m \prec n$	m is a predecessor of n
$n_t = \text{pred}_t(n)$	predecessor of the node n at stage t . The notation $\text{pred}_t(\cdot)$ emphasizes the mapping character of the predecessor
$\xi(n)$	state of the random variable ξ_t at node $n \in \mathcal{N}_t$
id, id_Ξ	identity function on Ξ ; $\text{id}(\xi) = \xi$ for all $\xi \in \Xi$

$\mathbb{1}_\Xi$	indicator function: $\mathbb{1}_\Xi(\xi) = 1$ if $\xi \in \Xi$, and $\mathbb{1}_\Xi(\xi) = 0$ otherwise
\mathbb{I}_Ξ	indicator function of optimization: $\mathbb{I}_\Xi(\xi) = 0$ if $\xi \in \Xi$, and $\mathbb{I}_\Xi(\xi) = \infty$ otherwise
L^p	space of p -integrable functions
$Q(x, \xi)$	random cost function
\mathcal{R}	risk functional, also written as $\mathcal{R}_P(\cdot)$
$\mathcal{U} = -\mathcal{R}$	utility functional, $\mathcal{U}(\cdot) = -\mathcal{R}(\cdot)$
$\mathcal{R}(\cdot, \dots)$	multiperiod risk functional also written as $\mathcal{R}_{\mathbb{P}}(Y_0, \dots, Y_T)$
$v(\mathbb{P})$	optimal value of the optimization problem $v(\mathbb{P}) = \min_{x \ll \mathbb{P}} \mathcal{R}_{\mathbb{P}}[Q(\xi, x)]$
$x^*(\mathbb{P})$	solution of the optimization problem for the nested distribution \mathbb{P}
a_+	$\max\{a, 0\}$
$a^\top b$	inner product of vectors a and b
$\int f(u) P(du)$	integration of f with respect to the (probability) measure P
$\int f(s) dG(s)$	Riemann–Stieltjes integral of f with respect to the integrator function G

Chapter 1

Introduction

1.1 Multistage Decision Models

To find optimal decisions is the fundamental problem of management. This goal sounds easier than it actually is, because some basic questions have to be answered when formulating the decision problem. These basic questions are:

- What is the decision goal?
- Which actions can be taken?
- Which information is available for each decision?
- How and when may some decisions be corrected or adapted at later stages?

This book deals with decision problems for which there is uncertainty about the relevant parameters (stochastic problems) and for which a sequence of decisions can be planned (multistage problems). Some examples may illustrate the topic.

Example (An Inventory Control Problem). Inventories must satisfy an uncertain demand, while the costs should be as low as possible. Three types of costs have to be considered: fixed costs for ordering, holding costs for pieces kept in the inventory, and shortage costs, if a demand cannot be satisfied (e.g., extra costs for quick reorder or costs for disappointed customers). Typically there is a time lag between order and delivery, called the lead time. Quick orders have shorter or zero lead times but come at extra costs.

- The decision goal could be to minimize the long-term expected costs or to minimize the risk of shortfall under cost constraints.
- Possible actions are, e.g., to issue regular orders or to place quick reorders at a higher price.
- The past demands can be recorded and this information may be used to forecast the demand.
- Order times can be fixed at regular intervals or may be chosen in an adaptive way.

The well-known newsboy problem is a special inventory problem (see Example 1.2 below). This problem is inherently single stage, since newspapers loose their value the next day and it makes no sense to store them for later selling. A multistage extension is the flowergirl problem (see Example 1.4 below), since flowers can be stored at least for some time. It is demonstrated below that the multistage view is superior to the myopic view of repeated single stages. This is somehow evident: if the distribution of future demands and/or future prices can be estimated based on the past observations, then the inventory management should incorporate longer term forecasts and optimize the costs over several periods.

Example (A Hydropower Generation Problem). Hydrostorages are buffers between rainfall or inflow at some times and energy production at later times. For this reason, the optimal management of hydrostorages is a multistage problem. It is stochastic since both, the future inflow of rain as well as the future electricity spot prices, are random. A hydro power generation system consists of a set of interconnected hydro plants, reservoirs and pump-storage plants for a short term planning horizon. It is very crucial to identify the optimal scheduling of the production due to the fact that satisfying the production portfolio might be difficult in some months of the year. A stochastic multistage programming framework can handle this decision problem. Chapter 8 illustrates and discusses such a model in detail.

Example (An Investment/ Insurance Problem). The budget planning process of every government is characterized by the problem of short term consumption and long-term investment. Consider in particular a country which is subject to natural hazards and has to protect its infrastructure by special mitigation measures or insurance plans. The optimal long-term development is a multistage stochastic problem, where the uncertainties are losses due to natural hazard events like earthquakes or floods. Also this example is presented in detail in Chap. 8.

1.2 Fundamentals of Decision Making

A general (deterministic) decision problem consists of

- a real valued *objective function* $x \mapsto F(x)$, which quantifies the losses associated with the decision x , and
- a *feasible set* \mathbb{X} , which contains all candidates x for the decision. We typically assume that $\mathbb{X} \subseteq \mathbb{R}^d$ for some dimension d .

The decision problem is written as

$$\text{minimize } \{F(x) : x \in \mathbb{X}\}. \quad (1.1)$$

The problem is called *feasible*, if \mathbb{X} is nonempty. For infeasible problems, we set $\min \{F(x) : x \in \emptyset\} = +\infty$. The problem (1.1) may be

- *bounded*, if there exists a $K \in \mathbb{R}$ such that $F(x) \geq K$ for all $x \in \mathbb{X}$, or
- *unbounded* otherwise.

If a problem is bounded, we may distinguish between

- problems for which the minimum is attained. In this case there is at least one $x^+ \in \mathbb{X}$ such that $F(x^+) \leq F(x)$ for all $x \in \mathbb{X}$ and the *argmin set*

$$\operatorname{argmin}\{F(x) : x \in \mathbb{X}\} = \{x \in \mathbb{X} : F(x) \leq F(x^+) \text{ for all } x \in \mathbb{X}\}$$

is nonempty. The problem has a unique solution x^* , if the argmin set is a singleton, i.e.,

$$\operatorname{argmin}\{F(x) : x \in \mathbb{X}\} = \{x^*\};$$

- problems for which the minimum is not attained. In this case the argmin set is empty, since

$$F(x) > \inf\{F(w) : w \in \mathbb{X}\}$$

for all $x \in \mathbb{X}$.

As decision makers we are interested in finding at least one element of the argmin set. Problems with empty argmin set are ill-posed from the application point of view and are not further considered in this book.

Typically the feasible set \mathbb{X} is described by inequalities like $\mathbb{X} = \{x : G_i(x) \leq 0, i = 1, \dots, k\}$. Such deterministic optimization problems

$$\min\{F(x) : G_i(x) \leq 0 \text{ for all } i = 1, \dots, k\} \quad (1.2)$$

require, however, complete knowledge of the objective function F as well as all constraint functions G_i . Already in the early days of mathematical optimization it has been observed that in the majority of applications this precise information is lacking and some uncertainty is present when the problem is posed. For instance, an investment has to be made now, but the possible returns from this investment will become known only later. Future prices, demands, or other economic or environmental conditions are typical examples of uncertain parameters.

1.2.1 Stochastic Problem Formulation

We extend the decision problem (1.2) by adding some parameters ξ to it and consider

$$\min\{Q(x, \xi) : G_i(x, \xi) \leq 0, i = 1, \dots, k\}.$$

If the decision maker only knows some range Ξ of these parameters, but not their true value exactly, she or he may use *robust optimization* by solving the following minimax problem (see Ben-Tal and Nemirovski [10]):

$$\min \left\{ F(x) := \max \{Q(x, \xi) : \xi \in \Xi\} \mid \begin{array}{l} \max \{G_i(x, \xi) : \xi \in \Xi\} \leq 0 \\ \text{for every } i = 1, \dots, k \end{array} \right\}. \quad (1.3)$$

This model is of worst case type. If x^* is the solution of (1.3) and $\bar{\xi}$ is the true parameter value, then one can be sure that $G_i(x^*, \bar{\xi}) \leq 0$ and that $Q(x^*, \bar{\xi}) \leq v^*$, where v^* is the optimal value of (1.3). In robust optimization all unknown parameters are treated equally and are not weighted according to their importance. Robust optimization is not only based on a pessimistic world view, it also ignores the fact that the relevant scenarios, in particular the most costly ones, may depend on the decision x and cannot be chosen beforehand. In contrast, *stochastic optimization* makes the assumption that the parameters ξ follow a probability law P^ξ such that the functions $Q(x, \xi)$ and $G_i(x, \xi)$ become random variables. While robust optimization looks at the worst case, stochastic optimization considers a summarizing functional, which maps the distribution of $Q(x, \xi)$ and $G_i(x, \xi)$, respectively, to the real line. The most common type of a summarizing functional is the expectation \mathbb{E} , leading to a *risk-neutral decision problem*

$$\min \{\mathbb{E}[Q(x, \cdot)] : \mathbb{E}[G_i(x, \cdot)] \leq 0, i = 1, \dots, k\}.$$

If, however, risk functionals \mathcal{R} and \mathcal{R}_i are employed, which penalize high risk, then the decision problem

$$\min \{\mathcal{R}[Q(x, \cdot)] : \mathcal{R}_i[G_i(x, \cdot)] \leq 0, i = 1, \dots, k\}$$

is called *risk-averse*. See Chap. 3 for a thorough treatment of risk functionals.

If the constraining functionals are of the form

$$P\{G_i(x, \xi) \leq 0\} \geq \alpha,$$

the problem is called *chance constrained*. Notice the difference between a robust program and a chance constrained program: a problem in which the constraints should be fulfilled in $100\alpha\%$ of the cases (say) can be formulated in both ways:

- For a robust formulation, one would *first* choose a set Ξ such that $P(\xi \in \Xi) \geq \alpha$ and use this set in the formulation of (1.3). The probability P does not appear in the optimization problem any longer.
- In stochastic chance constrained formulation, the probability P appears explicitly in the formulation, e.g.,

$$\max \{\mathcal{R}[Q(x, \cdot)] : P\{G(x, \cdot) \leq 0\} \geq \alpha\}.$$

Notice that in the latter case the exception set, which is the set where $G(x, \xi) > 0$, is not determined beforehand and may depend on x . For each x , a different set of ξ 's may form the exception, while the exceptional set is always the same for the robust formulation. This is a notable advantage of the chance constrained optimization over the robust formulation in the present context, since only if the decision is determined, the exceptional set can be determined as well.

Example 1.1 (A Simple Portfolio Optimization Problem). Suppose that an investment can be made in 3 different assets and that 5 equally probable scenarios for the returns are given. The 5×3 data matrix

$$\begin{pmatrix} 1.10 & 0.96 & 0.96 \\ 1.08 & 1.06 & 1.05 \\ 1.02 & 1.06 & 1.05 \\ 0.98 & 1.01 & 1.00 \\ 1.00 & 1.00 & 0.90 \end{pmatrix} \begin{matrix} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \\ \omega_5 \end{matrix}$$

$$\begin{matrix} \xi_1 & \xi_2 & \xi_3 \end{matrix}$$

collects the 5 possible outcomes of the return variable $\xi = (\xi_1, \xi_2, \xi_3)$.

The classical mean–variance (Markovitz-type) optimization problem reads

$$\begin{aligned} \min \quad & x^\top \text{cov}(\xi) x \\ \text{subject to} \quad & \mathbb{E}(\xi) x \geq r_{\min}, \\ & x \geq 0. \end{aligned} \tag{1.4}$$

Here, $\mathbb{E}(\xi)$ is the mean return vector (columnwise sums divided by 5) and $\text{cov}(\xi)$ is the covariance matrix of ξ . The threshold r_{\min} is a minimally required expected return, which we set to 1 here. Suppose in addition that we want to safeguard ourselves against large drops in portfolio value by requiring that the probability that the return is larger than 0.9750 (say) is at least 80 %. We would then add the constraint

$$P(\xi^\top x \geq 0.975) \geq 0.8. \tag{1.5}$$

This additional constraint reduces the feasible set.

In robust optimization, one would first choose a subset $\Omega' \subset \Omega$ of scenarios with probability $P(\Omega') = 0.8$ and require then the validity of

$$\xi(\omega)^\top x \geq 0.975 \quad \text{for all } \omega \in \Omega'. \tag{1.6}$$

In our example, obviously 4 out of 5 scenarios should fulfill this inequality. Which scenario to be left out? Of course a bad one. It is reasonable to exclude the last row, i.e., to choose $\Omega' = \{\omega_1, \dots, \omega_4\}$, since ξ_3 may drop to 0.9 in this scenario

ω_5 , the overall worst value. Solving the problem (1.4) with the additional constraint (1.6) leads to the solution $x^+ = (0.4031, 0.5731, 0)$. If, however, the full chance constrained stochastic problem (1.4) + (1.5) is solved, one gets a different and better solution $x^* = (0.4102, 0.5647, 0)$. Why can this happen? Simply because when choosing the bad scenario beforehand in the robust setup, one could not know that the optimal solution will not pick asset 3 at all ($x_3 = 0$) so that bad cases for asset 3 are irrelevant. If, however, one allows to choose the bad scenarios dependent on the decision, one finds that for the optimal decision x^* the scenario ω_4 is the worst and this one should form the exception set in the chance constrained stochastic problem.

We summarize that a stochastic optimization model is based upon

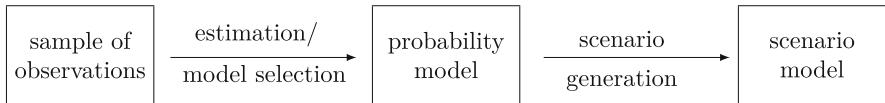
- a probability model P for the uncertain values,
- a feasible set \mathbb{X} for the decisions, typically $\mathbb{X} \subseteq \mathbb{R}^d$,
- a cost function Q depending on the decision variables and the uncertainties, and
- a probability functional \mathcal{R} , like expectation, median, etc. to summarize the random costs in a real valued objective.

Simplifying the constraints as $x \in \mathbb{X}$ while making the probability P explicit we denote the stochastic decision model by¹

$$\min \{F_P(x) = \mathcal{R}_P[Q(x, \xi)] : x \in \mathbb{X}\}. \quad (1.7)$$

The most difficult part in establishing a formalized decision model of type (1.7) for a real decision situation is to find the appropriate probability model P . Typically, there is a sample of past data available, but not more. For the solution of the decision problem two steps are needed to come from the sample of observations to the scenario model.

- In the first step, a *probability model* is identified, i.e., the description of the uncertainties as random variables or random processes by identifying the probability distribution. This step is based on statistical methods of model selection and parameter estimation.
- In the following scenario generation step, a *scenario model* is found, which is an approximation of (1.7) by a *finite model* of lower complexity than the probability model.



¹An utility maximization problem

$$\max \{\mathcal{U}_P[H(x, \xi)] : x \in \mathbb{X}\}$$

can be brought to form (1.7) by setting $Q(x, \xi) = -H(x, \xi)$ and $\mathcal{R}_P[Y] = -\mathcal{U}_P[-Y]$.

The second step, scenario generation, is treated in detail in Chap. 4. We emphasize that the basic assumption in *stochastic optimization* is that the distribution of ξ is known, while the actual values are unknown. In Chap. 7 (on ambiguity) we will relax this assumption and allow the distribution P^ξ to be also not totally known, but a family \mathcal{P} of possible models is known, to which P^ξ belongs.

Remark About Notation. If (Ω, \mathcal{F}, P) is a probability space and ξ is a random variable defined on it with values in \mathbb{R}^m , then the image measure (or pushforward measure) of ξ on \mathbb{R}^m is denoted by $P^\xi = P \circ \xi^{-1}$ (i.e., $P^\xi(A) = P(\{\omega : \xi(\omega) \in A\})$ for $A \subset \mathbb{R}^m$ measurable). Stochastic optimization problems are always defined by the distributions P^ξ of the random scenario variable(s), the concrete sample space is irrelevant. However, we will often consider both, the probability space (Ω, \mathcal{F}, P) and the random variable ξ and not just the image measure P^ξ . The reason is that for finite sample spaces Ω it is convenient to keep the random variable ξ if one wants to change the probability P to a new one, say P' . This change of measure changes also the image measure P^ξ to P'^ξ , but the consideration of the basic probabilities P and P' is simply more convenient.

1.2.2 From Single-Stage to Multistage Decision Problems

Single-Stage Stochastic Optimization Problems. In the simplest situation a decision is made at time 0 and the result is observed later (at time 1, say), but no further decisions can be made. Let $Q(x, \xi)$ be a cost function, which assigns a real valued loss to any pair consisting of a decision x and a random variable ξ . Since $Q(x, \xi)$ is a random variable, it cannot be minimized as such. We need a functional \mathcal{R} , which maps random variables to the real line. Such *risk functionals* must exhibit certain properties to qualify for reasonable stochastic optimization problems (see Chap. 3 for an overview on appropriate risk functionals).

The single-stage stochastic optimization problem reads

$$\text{minimize } \{F(x) := \mathcal{R}[Q(x, \xi)] : x \in \mathbb{X}\}.$$

Typical risk functionals are

- the expectation $R[\cdot] = \mathbb{E}[\cdot]$,
- the α -quantiles $G_Y^{-1}(\alpha)$, where G_Y is the distribution function of Y , i.e., $G_Y(q) = P\{Y \leq q\}$ and G_Y^{-1} is the inverse distribution function (quantile function), i.e.,

$$G_Y^{-1}(\alpha) := \inf\{q : G_Y(q) \geq \alpha\}$$

($V@R_\alpha(Y) := G_Y^{-1}(\alpha)$ is also called Value-at-Risk), or

- the upper Average Value-at-Risk,

$$\text{AV@R}_\alpha(Y) = \frac{1}{1-\alpha} \int_{\alpha}^1 G_Y^{-1}(p) dp. \quad (1.8)$$

An overview over a variety of risk functionals is given in Chap. 3.

Example 1.2 (The Flowergirl Problem). As an example we consider the single-stage flowergirl problem, which is also called the newsboy problem: a flowergirl has to decide how many flowers she orders from the wholesaler. She buys for the price of b per flower and sells them for a price of $s > b$. Unsold flowers may be returned for a price of $r < b$. The demand is ξ . If the demand is higher than the available stock, she may procure additional flowers for an extra price $e > b$. What is the optimal order quantity x , if the expected profit should be maximized?

We formulate the profit as negative costs. We set

$$\begin{aligned} \text{total costs} := & \text{ initial purchase} \\ & - \text{revenue from sales} \\ & + \text{extra procurement costs} \\ & - \text{revenue from returns}. \end{aligned}$$

The cost function therefore is

$$Q(x, \xi) = bx - s\xi + e[\xi - x]_+ - r[\xi - x]_-.$$

Here, $[a]_+ = \max\{a, 0\}$ is the positive part of a and $[a]_- = \max\{-a, 0\}$ is the negative part of a . Since $a = [a]_+ - [a]_-$, the cost function may be rewritten as

$$\begin{aligned} Q(x, \xi) &= (b - r)x - (s - r)\xi + (e - r)[\xi - x]_+ \\ &= (b - r) \left\{ x + \frac{1}{1 - \frac{e-b}{e-r}} [\xi - x]_+ \right\} - (s - r)\xi. \end{aligned}$$

Since $\text{AV@R}_\alpha(Y) = \min \{q + \frac{1}{1-\alpha} \mathbb{E}[Y - q]_+ : q \in \mathbb{R}\}$ (see (3.3) later) it follows that

$$\min \{\mathbb{E}[Q(x, \xi)] : x \in \mathbb{R}\} = (b - r)\text{AV@R}_\alpha(\xi) - (s - r)\mathbb{E}(\xi),$$

where $\alpha = \frac{e-b}{e-r}$. The optimal procurement quantity of the flowergirl is $x^* = \text{V@R}_\alpha(\xi)$ (see, e.g., Pflug and Römisch [97, page 56]).

Probability Functionals in Constraints. In some decision problems the feasible set contains probability functionals as well. The constraint set is then of the form

$$\mathbb{X} = \{x : \mathcal{R}_1[Q_1(x, \xi)] \leq \rho_1, \dots, \mathcal{R}_k[Q_k(x, \xi)] \leq \rho_k\}.$$

A special case are *chance constraints* of the form

$$P\{Q(x, \xi) \leq t\} \geq \alpha.$$

Chance constraints introduce some algorithmic difficulties for the numerical solution, since they may lead to nonconvex feasible sets. However, conceptually they are just constraints of a special type. Convexity of the constraint set can, for instance, be ensured, if P , the distribution of ξ , is a log-concave probability measure on \mathbb{R}^m , and $\{(x, \omega) : Q(x, \omega) \leq t\}$ is a convex set in $\mathbb{R}^d \times \mathbb{R}^m$. A probability measure P on \mathbb{R}^m is called log-concave if for all convex sets A, B and all $0 < \lambda < 1$,

$$P(\lambda A + (1 - \lambda)B) \geq P(A)^\lambda \cdot P(B)^{1-\lambda}.$$

Conditions to ensure log-concavity of probability measures in terms of their densities have been investigated by Prekopa [103]. Further conditions for convexity of chance constraints have been found by Henrion and Strugarek [56, 57]. In this book, the special features of chance constrained problems are not treated.

Two-Stage Stochastic Optimization Problems (Recourse Problems). If the decision at time 0, now written as x_0 (the *here-and-now* decision), can be corrected at a later time (for simplicity at time 1) by some corrective decision x_1 (the *wait-and-see* decision), we speak of a *recourse problem*. The corrective decision is made by solving a second stage decision problem

$$\text{minimize } \{Q_1(x_0, x_1, \xi) : x_1 \in \mathbb{X}_1(x_0, \xi)\}.$$

The total cost function is the sum of the first stage costs $Q_0(x_0)$ and the second stage costs, i.e., $Q_0(x_0) + Q(x_0)$, where

$$Q(x_0) = \min \{\mathcal{R}[Q_1(x_0, x_1, \xi)] : x_1 \in \mathbb{X}_1(x_0, \xi)\}.$$

$Q(\cdot)$ is called the *recourse function*. The complete two-stage problem reads

$$\begin{aligned} \min \{Q_0(x_0) + \min \{\mathcal{R}[Q_1(x_0, x_1, \xi)] : x_1 \in \mathbb{X}_1(x_0, \xi)\} : x_0 \in \mathbb{X}_0\} \\ = \min \{Q_0(x_0) + Q(x_0) : x_0 \in \mathbb{X}_0\}. \end{aligned} \quad (1.9)$$

If \mathcal{R} is translation-equivariant, i.e., $\mathcal{R}[c + Y] = c + \mathcal{R}[Y]$ for each constant c , then the problem (1.9) can be reformulated as

$$\min \{\mathcal{R}[Q(x_0, x_1, \xi)] : x_0 \in \mathbb{X}_0, x_1 \in \mathbb{X}_1(x_0, \xi)\}$$

with $Q(x_0, x_1, \xi) = Q_0(x_0) + Q_1(x_0, x_1, \xi)$.

The problem is said to possess

- *complete recourse*, if $Q(x_0) < \infty$ for all $x_0 \in \mathbb{R}^d$,
- *relatively complete recourse*, if $Q(x_0) < \infty$ for all feasible $x_0 \in \mathbb{X}_0$.

In complete or relatively complete recourse problems, second stage infeasibility does not lead to complications. If, however, the problem does not have relatively complete recourse, then the feasibility of the second stage problem induces additional constraints to the first stage problem, called *implicit constraints*. The total feasible set for the first stage is

$$\mathbb{X}_0 \cap \{x_0 : \mathcal{Q}(x_0) < \infty\},$$

the intersection of the explicit constraints \mathbb{X}_0 and the implicit constraints.

Linear Recourse Problems. An important class of problems is given by linear expectation problems, in which the functions \mathcal{Q}_0 and \mathcal{Q}_1 are linear in the decisions x_0 and x_1 and the constraint sets \mathbb{X}_0 and \mathbb{X}_1 are polyhedral. These risk-neutral problems are written as

$$\text{minimize} \left\{ c^\top x_0 + \mathbb{E} \left[\min \left\{ q^\top x_1 : A_1(\omega) x_0 + W_1 x_1 = h_1(\omega), x_1 \in \mathbb{X}_1 \right\} \right] : A_0 x_0 \leq b \right\}. \quad (1.10)$$

Here, $\xi(\omega) = (A_1(\omega), h(\omega))$ is the random parameter. In principle, the *recourse costs* q and the *recourse matrix* W could also be random, but (1.10) is the standard model. If the *technology matrix* A_1 is nonrandom, the model is said to have only a *random right-hand side*. If W is the identity matrix, the recourse is called *simple*.

Example (A Network Design and Operation Problem). Suppose that a communications network, which connects n units, has to be designed before the actual demands become known (see Fig. 1.1 for illustration). In the first stage, the capacities x_0 of the links are planned and in stage two (after the demands became available), the network

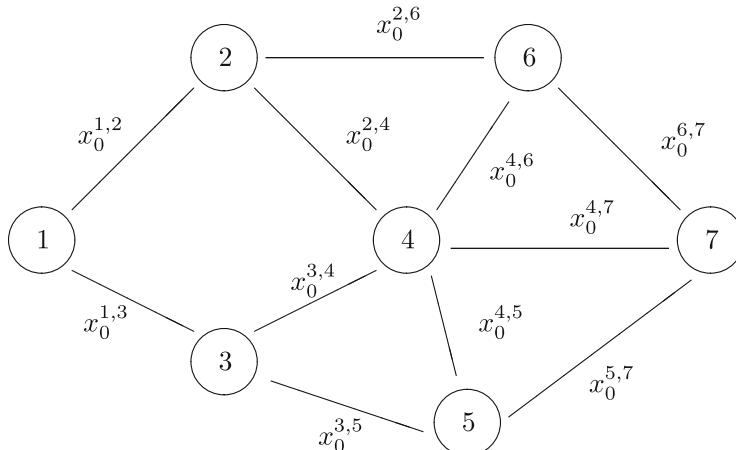


Fig. 1.1 A network problem: the first stage variables are the link capacities $x_0^{i,j}$. The second stage variables are the network flows

flows x_1 are determined. The costs to build a link of capacity $x_0^{i,j}$ connecting units i and j are $c^{i,j} \cdot x_0^{i,j}$. The demands for the origin-destination relation k to ℓ are given by the random values $\xi^{k,\ell}$. The second stage decision variables $x_1^{i,l,k,\ell}$ denote the flow from k to ℓ which runs over the link i, j . Given proportional operating costs $d^{i,j}$ per link, the objective is

$$\sum_{i,j} c^{i,j} \cdot x_0^{i,j} + \mathbb{E} \left[\sum_{i,j,k,\ell} d^{i,j} \cdot x_1^{i,j,k,\ell} \right],$$

which has to be minimized under the capacity constraints

$$\sum_{k,\ell} x_1^{i,j,k,\ell} \leq x_0^{i,j}$$

and the flow constraints

$$\begin{aligned} \sum_j x_1^{i,j,k,\ell} &= \sum_{j'} x_1^{j',i,k,\ell} \quad i \neq k, \ell \\ \sum_j x_1^{k,j,k,\ell} &= \xi^{k,\ell} \text{ and} \\ \sum_{j'} x_1^{j',k,k,\ell} &= \xi^{k,\ell}. \end{aligned}$$

This problem is a linear, two-stage risk-neutral stochastic program with random right-hand side.

Multistage Stochastic Optimization Problems. In recourse problems corrective decisions can be made at just one moment in time. In a straightforward generalization one may consider multistage problems for which corrective decisions may be taken at several times $1, 2, \dots, T$. We use the numbering of decisions and observations such that the initial decision is x_0 , it is followed by a random observation ξ_1 , a subsequent decision x_1 and so on, up to the terminal time T . Some multistage decision problems allow a terminal decision x_T ,

$$x_0 \rightarrow \xi_1 \rightarrow x_1 \rightarrow \dots \rightarrow x_{T-1} \rightarrow \xi_T \rightarrow x_T, \quad (1.11)$$

others finish with the last observation ξ_T ,

$$x_0 \rightarrow \xi_1 \rightarrow x_1 \rightarrow \dots \rightarrow x_{T-1} \rightarrow \xi_T.$$

Notice that single-stage problems, i.e.,

$$x_0 \rightarrow \xi_1$$

and the standard two-stage problems, i.e.,

$$x_0 \rightarrow \xi_1 \rightarrow x_1$$

are special cases. Notice as well that the problem

$$x_0 \rightarrow \xi_1 \rightarrow x_1 \rightarrow \xi_2$$

could also be called two-stage, but this form of two-stage decisions and two-period observations is not the standard form of a two-stage problem.

The decision times may not be equidistant and could be just $0, \tau_1, \dots, \tau_T$. But in order to simplify the language we call x_t the *decision at time t*, even if the real times are different.

Information. A crucial notion in multistage decision making under uncertainty is the notion of available information. As stochastic optimization is based on probability theory, information is represented by σ -algebras (σ -fields). It is assumed that, together with the values of the decision relevant parameters ξ_t , a σ -algebra \mathcal{F}_t is given, which models the available information at time t . It is always assumed that the observed process ξ_t is measurable with respect to \mathcal{F}_t , for which the symbol

$$\xi_t \triangleleft \mathcal{F}_t$$

is used. Note that we allow that \mathcal{F}_t is larger than the σ -algebra generated by ξ_t , i.e., $\mathcal{F}_t \supset \sigma(\xi_t)$.

Remark 1.3. The distinction between the information contained in the process (ξ_t) and the (possibly larger) information \mathcal{F}_t , on which the decisions (x_t) may depend, is important for tree models, where the same values of the ξ -process can sit on different nodes. However, the distinction is only a matter of easy notation. In fact, if \mathcal{F}_t is larger than $\sigma(\xi_t)$, then one may augment the process ξ_t by some additional component η_t such that $\mathcal{F}_t = \sigma(\xi_t, \eta_t)$. The new process (ξ_t, η_t) generates the informations \mathcal{F}_t and the cost function $Q(x, \xi)$ is a function of (ξ_t, η_t) (but in fact only of its first component ξ_t).

The decisions x_t can incorporate only information available before or at time t , the time of decision, i.e., they must satisfy

$$x_t \triangleleft \mathcal{F}_t. \tag{1.12}$$

Condition (1.12) is called the *nonanticipativity constraint*. Obviously, the σ -algebras must be increasing, i.e., $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$, since information cannot be lost. An increasing sequence of σ -algebras $(\mathcal{F}_1, \dots, \mathcal{F}_T)$ is called a *filtration*.

By adding the trivial σ -algebra $\mathcal{F}_0 = \{\emptyset, \Omega\}$ at the beginning of this sequence one gets the (extended) filtration

$$\tilde{\mathfrak{F}} = (\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_T),$$

and by prepending the deterministic observation ξ_0 available at time 0, the extended process $\xi = (\xi_0, \dots, \xi_T)$ is obtained. This extends the initial sequence (1.11) to

$$\xi_0 \rightarrow x_0 \rightarrow \xi_1 \rightarrow x_1 \rightarrow \dots \rightarrow x_{T-1} \rightarrow \xi_T \rightarrow x_T,$$

although the deterministic observation ξ_0 and the decision x_T are usually not considered (or omitted) in this book.

Since $\xi_t \triangleleft \mathcal{F}_t$ for all $t = 1, \dots, T$, the process $\xi = (\xi_1, \dots, \xi_T)$ is *adapted* to the filtration $\tilde{\mathfrak{F}}$, which is written as

$$\xi \triangleleft \tilde{\mathfrak{F}}.$$

Similarly, the nonanticipativity constraints (1.12) can be written in a compact way as

$$x = (x_0, \dots, x_T) \triangleleft \tilde{\mathfrak{F}},$$

or in extensive form as

$$x_t \triangleleft \mathcal{F}_t \quad \text{for all } t.$$

Notice that for the first stage decision (at time 0), the condition $x_0 \triangleleft \mathcal{F}_0 = \{\emptyset, \Omega\}$ means that x_0 must be a constant and cannot depend on any future information. Using these notations, a multistage stochastic program can be stated in compact form as

$$\text{minimize } \{F(x) := \mathcal{R}[Q(x, \xi)] : x \in \mathbb{X} \text{ and } x \triangleleft \tilde{\mathfrak{F}}\}, \quad (1.13)$$

or in its extended form

$$\begin{aligned} & \text{minimize } \left\{ F(x_0, x_1, \dots, x_{T-1}) := \mathcal{R}[Q(x_0, x_1, \dots, x_{T-1}, \xi_1, \dots, \xi_T)] \right. \\ & \quad \left. \begin{array}{l} x \in \mathbb{X}, \text{ and} \\ x_t \triangleleft \mathcal{F}_t, t = 0, \dots, T \end{array} \right\}. \end{aligned}$$

For simplicity of notation, but with no restriction on generality, we assume that all stochastic constraints are incorporated in the overall cost function and that the constraint set \mathbb{X} does not contain random variables. This is possible, since constraints can be lifted to the objective using the indicator function of optimization

$$\mathbb{I}_{\mathbb{X}}(x) = \begin{cases} 0 & \text{if } x \in \mathbb{X}, \\ \infty & \text{otherwise.} \end{cases}$$

If $x \in \mathbb{X}$ is a constraint, which has to be fulfilled with probability 1, we consider the extended cost function

$$\bar{Q}(x, \xi) = Q(x, \xi) + \mathbb{I}_{\mathbb{X}}(x).$$

This trick does only simplify the notation, it is of no help for finding solution algorithms. The functional \mathcal{R} appearing in the objective quantifies the risk of the final costs.

Some problems may extend the structure (1.13) by considering the multiperiod risk of a sequence of cost functions. They are formulated as

$$\begin{aligned} \text{minimize } & \{F(x) := \mathcal{R}[Q_1(x_0, \xi_1), \dots, Q_T(x_0, x_1 \dots, x_{T-1}, \xi_1, \dots, \xi_T)] : \\ & x_t \triangleleft \mathcal{F}_t; x \in \mathbb{X}\}. \end{aligned}$$

Here, $\mathcal{R} = \mathcal{R}(Y_1, \dots, Y_T)$ is a multiperiod risk functional, which is applied to the intermediate, period-wise costs Q_1, \dots, Q_T .

Risk Neutral Linear Multistage Problems. A multistage stochastic optimization problem with expectation objective is called linear if both, the objective and the constraints, are linear in the decision variables, i.e., if it is of the following form

$$\begin{aligned} \min \Bigg\{ & c_0 x_0 + \mathbb{E} \Big[\min c_1(\xi_1) x_1 + \mathbb{E} \Big[\min c_2(\xi_2) x_2 \\ & + \mathbb{E} [\dots + \mathbb{E} [\min c_T(\xi_T) x_T] \dots] \Big] \Big] \Bigg] : x \in \mathbb{X}(\xi) \Bigg\}, \end{aligned}$$

where the feasible set $\mathbb{X}(\xi)$ is given by

$$\begin{aligned} A_0 x_0 &= h_0, \quad x_0 \in \bar{\mathbb{X}}_0, \quad (x_0 \triangleleft \mathcal{F}_0) \\ A_1(\xi_1) x_0 + W_1 x_1 &= h_1(\xi_1), \quad x_1 \in \bar{\mathbb{X}}_1, \quad x_1 \triangleleft \mathcal{F}_1, \\ A_2(\xi_2) x_1 + W_2 x_2 &= h_2(\xi_2), \quad x_2 \in \bar{\mathbb{X}}_2, \quad x_2 \triangleleft \mathcal{F}_2, \\ &\vdots \quad \vdots \quad \vdots \\ A_T(\xi_T) x_{T-1} + W_T x_T &= h_T(\xi_T), \quad x_T \in \bar{\mathbb{X}}_T, \quad x_T \triangleleft \mathcal{F}_T, \end{aligned}$$

where $\bar{\mathbb{X}}_1, \dots, \bar{\mathbb{X}}_T$ are nonrandom polyhedral sets, typically just box or sign constraints. Notice that the solutions x_t depend on the random parameters and are therefore random variables. Using the indicator function one may rewrite the total cost function $Q(x, \xi)$ as

$$\begin{aligned}
Q(x, \xi) = & c_0 x_0 + c_1 x_1 + \cdots + c_T x_T \\
& + \mathbb{I}_{\{A x_0 = h_0\}} + \mathbb{I}_{\{A_1(\xi_1) x_0 + W_1 x_1(\xi_1) = h_1(\xi_1)\}} \\
& + \cdots + \mathbb{I}_{\{A_T(\xi_T) x_{T-1} + W_T x_T(\xi_T) = h_T(\xi_T)\}},
\end{aligned}$$

so that all randomness appears in the cost function and the remaining constraint sets $\bar{\mathbb{X}}_t$ are nonrandom. Notice that with this formulation the random cost function $Q(x, \xi)$ is no longer linear, but convex in the decisions x for each fixed ξ . In Chap. 5 we will discuss the special form of linear multistage problems in view of the time-consistency of the decisions.

Solving a Stochastic Problem by Discrete Approximation. A multistage stochastic optimization problem consists in finding appropriate measurable solution functions $x_t \triangleleft \mathcal{F}_t$. While in very rare cases analytical methods may be used to find the solution functions, in the vast majority of cases this approach fails and one has to use approximation techniques. For a numerical solution the original problem is replaced by a simpler one, where the general probability distribution P is replaced by a simpler distribution \tilde{P} , which has a finite support (i.e., sits only on finitely many points). In finite probability models, integrals become sums and measurable functions become vectors. Most importantly, finite filtrations can be represented as tree structures, and processes, adapted to filtrations, are just vectors sitting on the nodes of the tree. Figure 1.2 illustrates how numerical solutions of multistage stochastic programs are obtained: the solution \tilde{x}^* of the approximate problem does not directly qualify as a solution of the original problem, since it is defined on a different probability space. An extension function $e_{\mathbb{X}} : \tilde{x} \mapsto x \in \mathbb{X}$ transforms a solution of the approximate problem into a solution of the original problem.

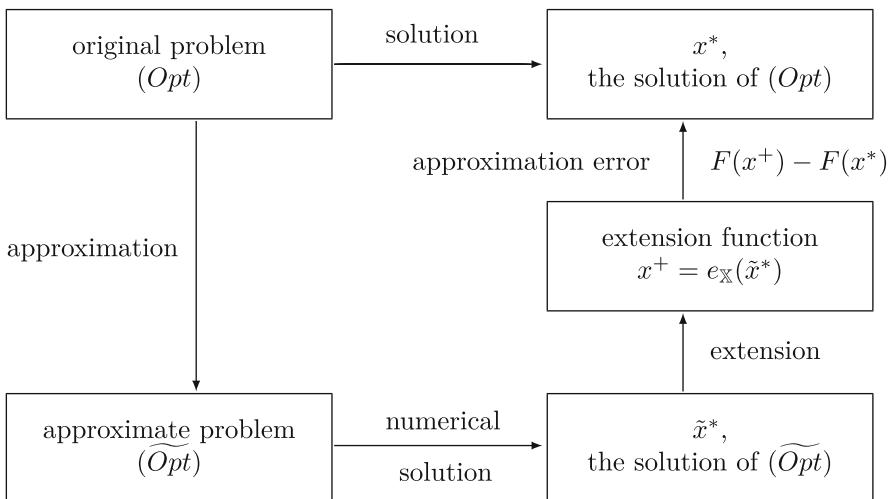


Fig. 1.2 Approximation of stochastic optimization problems

Here is how extension functions are typically constructed. Suppose that the filtration $\tilde{\mathfrak{F}}$ is generated by the process $\tilde{\xi}$ (this is no restriction, see Remark 1.3). Suppose that $\tilde{\xi}_t$ takes finitely many values $\tilde{z}_1, \dots, \tilde{z}_k$. Then the solution \tilde{x}^+ of the approximate problem is a function of $\tilde{\xi}$, i.e., $\tilde{x}_t^+ = \tilde{x}_t^+(\tilde{z})$ and the extension can be defined for $\xi_t = z$ as

$$x^+(z) = \tilde{x}^+(\tilde{z}_i) \quad \text{if } \tilde{z}_i \text{ is the point which is nearest to } z.$$

This extension is a step function, but one may use as well some smoothed versions of it.

Since this solution x^+ is obtained in an indirect way, there is an approximation error

$$F(x^+) - F(x^*) \geq 0.$$

In Chap. 6, bounds for the approximation error will be presented. The approximation error decreases, if the finite model gets larger so that for quite large scenario tree models one may ignore the fact that the model is just an approximation. In most large applications one simply pretends that the tree model *is* the underlying problem itself.

If the underlying problem is a linear risk-neutral multistage model, the approximate problem is just a classical linear program (LP), for which many very sophisticated solvers exist. In other cases, nonlinear solvers would do the job.

The Value of the Multistage Stochastic Solution. One might ask whether it is really necessary to plan ahead in a multistage way. Why not follow a simpler strategy and do repeated planning over shorter periods? Although repeated myopic decision making is often done in practice, it is typically outperformed by multistage planning. The reason is simple: myopic decisions optimize for a shorter horizon and somehow pretend that the world ends at the end of the planning horizon. With scarce resources for instance, one would exhaust all resources at the end and not think about the further consequences. But even if a final valuation of the remaining resources enters the objective, myopic planning is still suboptimal, as the following example shows.

Example 1.4 (The Flowergirl Problem, Continued). We extend the flowergirl problem (Example 1.2) to become a linear multistage problem. In this extension, unsold flowers are not returned, but may be kept for the next day. However, a certain percentage of the flowers fades and is lost. If the demand exceeds the actual stock, the necessary flowers can be procured from another retailer, but for a higher price than from the wholesaler. Introduce the following quantities:

- b_t the buy price at time t ,
- s_t the selling price at time t , here set to $s_t = 1$,
- ℓ_t the proportion of flowers which survive from time t to $t + 1$,
- u_t the procurement costs for extra flowers from another retailer,
- ξ_t the demand at time t .

Let ζ_t be the (hypothetical) stock of flowers after all sales of the stock are effectuated at time t . If ζ_t is positive (i.e., $\zeta_t = [\zeta_t]_+$), then $\ell_t \cdot [\zeta_t]_+$ is the amount of flowers which will survive to time $t + 1$. If ζ_t is negative (i.e., $\zeta_t = -[\zeta_t]_-$), then extra procurement costs of $u_t \cdot [\zeta_t]_-$ occur at time t . The flowers, which survived up to the final time T , are valued with the value $\ell_T \cdot [\zeta_T]_+$. The expected profit should be maximized.

The objective function can be written as

$$\text{maximize } \mathbb{E} \left[\sum_{t=1}^T \xi_t - \sum_{t=0}^{T-1} b_t x_t - \sum_{t=1}^T u_t \cdot [\zeta_t]_- + \ell_T \cdot [\zeta_T]_+ \right],$$

and the constraints are

$$x_{t-1} + \ell_{t-1} \cdot [\zeta_{t-1}]_+ - \xi_t = \zeta_t = [\zeta_t]_+ - [\zeta_t]_-, \quad (1.14)$$

$$[\zeta_t]_+ \geq 0 \text{ and} \quad (1.15)$$

$$[\zeta_t]_- \geq 0 \quad (1.16)$$

for $t = 1, \dots, T$.

As an illustration, this problem was implemented on a simple finite binary scenario tree of height 3 for the random demands ξ (how to get form some given data to a simple scenario tree will be discussed in detail in Chap. 4). The example tree in Fig. 1.3 displays the random demand. The other parameters were chosen as

$$\ell = (0.5, 0.5, 0.5), b = (0.85, 0.85, 0.85) \text{ and } u = (1.25, 1.25, 1.25).$$

The problem is a linear program with optimal value

$$v^* = 8.58.$$

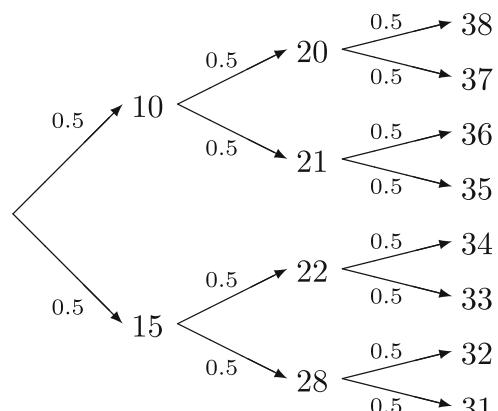


Fig. 1.3 The scenario process modeling the demand ξ

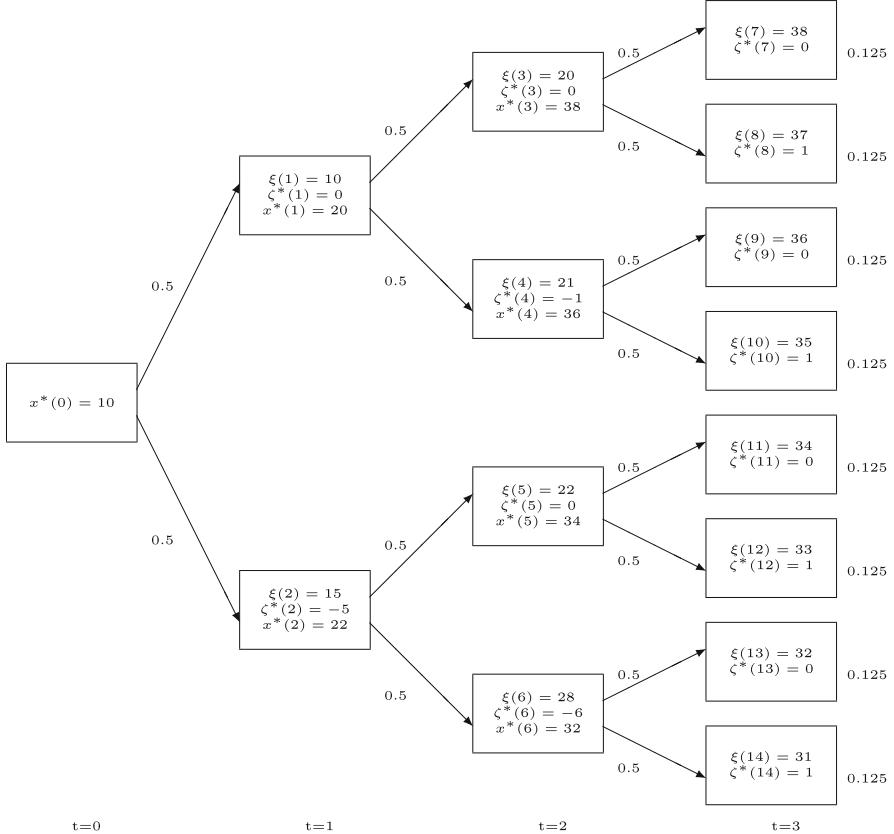


Fig. 1.4 The flowergirl problem—multistage solution

The optimal decisions x^* and the stocks ζ^* at each node are shown in Fig. 1.4.

In a further step, the same problem was solved in a repeatedly myopic way: for the first decision x_0 , a single-stage problem was solved, which maximizes the expected profit only at stage 1. Then the optimal decisions at stage 2 were calculated for the expected profit up to stage 2 and under the assumption that the first stage decision x_0 is kept fixed. Next, the decisions at the first two stages were fixed to the previously found values and the optimal decisions at stage 3 were calculated, but only considering the profit at the final time. The myopic solutions x^+ , as well as the corresponding stocks ζ^+ , are shown in Fig. 1.5.

Notice that the stepwise myopic optimizations lead to a decision vector x^+ , which is different from the optimal decisions x^* . Inserting the myopic decisions into the original problem leads to an expected profit of

$$v^+ = 7.97,$$

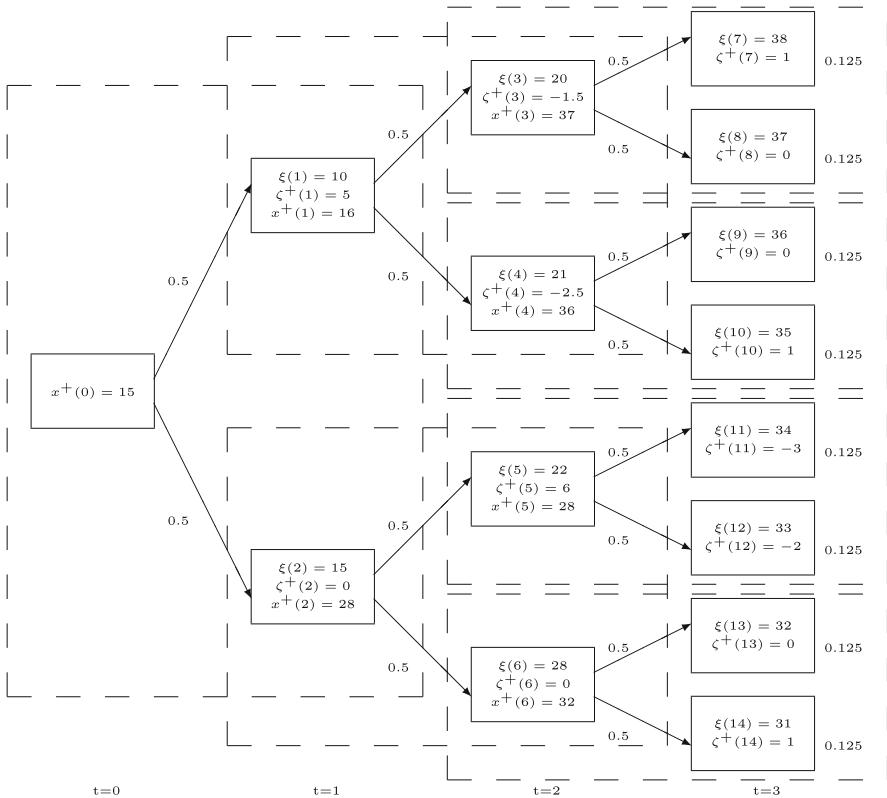


Fig. 1.5 The flowergirl problem—repeated myopic solution: the *dashed boxes* indicate the subproblems which were solved one after the other

which is smaller than $v^* = 8.58$. Thus problems, which are formulated for several steps, cannot be solved by repeated solution of single-stage problems. Although this method of solving repeatedly short sighted problems in a rolling horizon manner is used by many practical decision makers, it cannot replace the consideration of multistage problems.

The superiority of the multistage formulation is easy to understand: if the decisions at stage t have some aftereffects at later stages, then the myopic view cannot anticipate these effects and this leads to suboptimal decisions.

Decision Times and Decision Stages. Quite often the decision stages do not correspond to equally spaced time intervals. In this case a distinction between decision stages and decision times is necessary. We denote the decision stages by $t = 0, 1, \dots, T$, but the decision times by $\tau_0 = 0, \tau_1, \dots, \tau_T$. The time intervals between decision times are

$$\Delta\tau_t = \tau_{t+1} - \tau_t. \quad (1.17)$$

Notice that a single-stage problem has only stage 0, a two-stage problem has stages 0 and 1. Longer term multistage decision models often consider increasing decision intervals, as, for instance, decisions in 1, 3, 6 months, 1 and 2 years from now in an investment or production model. However, in many illustrative examples in this book, the decision times and the decision stages coincide to make things easier.

Decision Times and Decision Variables Depending on Randomness. In some decision problems the times between decision stages, as well as the selection of decision variables, are random. An example of such a situation is contained in the book *Modeling with Stochastic Programming* by A. King and S. Wallace [67], where they have adapted an example from Anderson et al. We describe this example briefly here.

A production line consisting of two machines has to be serviced. Each machine should be overhauled and adjusted and then the whole system should be tested. See Fig. 1.6 for a network plan of the necessary activities: (A) overhaul machine I, (B) adjust machine I, (C) overhaul machine II, (D) adjust machine II, and (E) test system. The duration of each activity (except E) in days is a random variable with mean given in Table 1.1.² The activities may be speeded up, but this causes extra costs. The decisions x_A, x_B, \dots, x_E indicate the amount of reduction in days. The costs for reduction are also indicated in Table 1.1.

The total project should be finished in 10 days, each extra day results in 275 units as penalty costs. The objective is to minimize the total expected costs, which consist of the costs for reduction and the penalty costs. For details see the given reference.

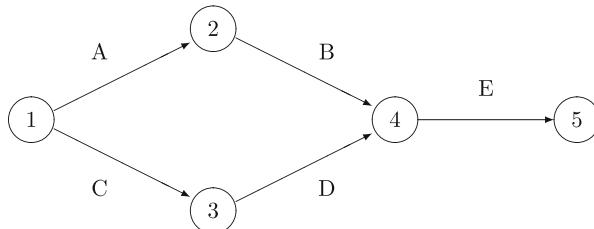


Fig. 1.6 The machine overhaul problem

Table 1.1 The data of the machine overhaul problem

Activity	Duration	Reduction variable	Maximal reduction	Reduction cost
A: Overhaul machine I	$7 + \xi_A$	x_A	$0 \leq x_A \leq 3$	$100 x_A$
B: Adjust machine II	$3 + \xi_B$	x_B	$0 \leq x_B \leq 1$	$150 x_B$
C: Overhaul machine I	$6 + \xi_C$	x_C	$0 \leq x_C \leq 2$	$200 x_C$
D: Adjust machine II	$3 + \xi_D$	x_D	$0 \leq x_D \leq 2$	$175 x_D$
E: Test system	2	x_E	$0 \leq x_E \leq 1$	$250 x_E$

²The ξ 's are uniform $[-1,1]$ random variables.

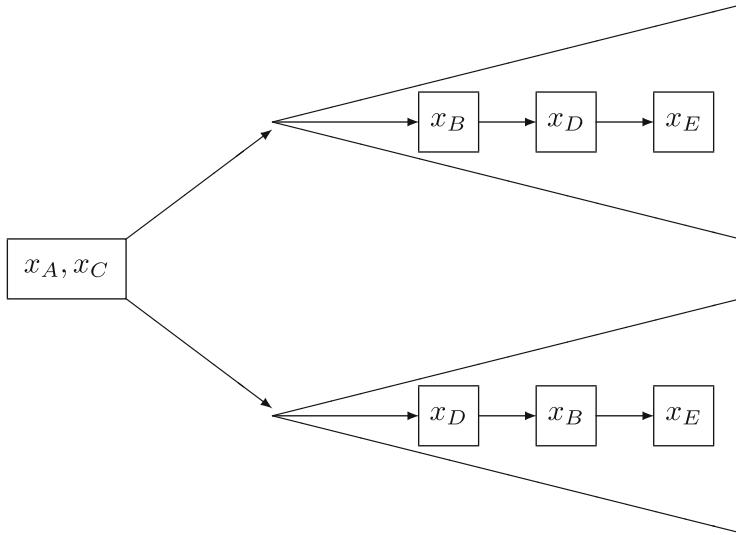


Fig. 1.7 Subtrees may have different decision variables

This problem may be seen as a multistage problem, where at the first stage the decisions x_A and x_C are made. The next decision stage is at the time, when the first overhaul is finished, followed by the time when the second overhaul is finished and finally when both adjustments are finished.

The times of decisions, as well as the name of the variable, which has to be decided upon, are random. In tree structured problems, this, however, does not cause a principal problem: see Fig. 1.7 for an illustration: depending on the random event of determining which activity finished first, the order of decisions is determined and this results in differently structured subtrees. But the concept of stages extends to this case as well.

1.3 Multistage Stochastic Optimization Versus Dynamic Optimization

There is a close relation between multistage stochastic optimization and dynamic stochastic optimization, but there are also essential differences. Dynamic programming is based on the idea of the state of a system, which describes its relevant characteristics. The state of the system evolves as a controlled Markov process, the decision being the controls. The decision at time t must be a function of the state.

Denote by ξ_t the state of the system. In the state-space representation, the state variable follows a dynamic model (cf. Fig. 1.8)

$$\xi_{t+1} = k_{t+1}(\xi_t, x_t, \xi_{t+1}) \quad (1.18)$$

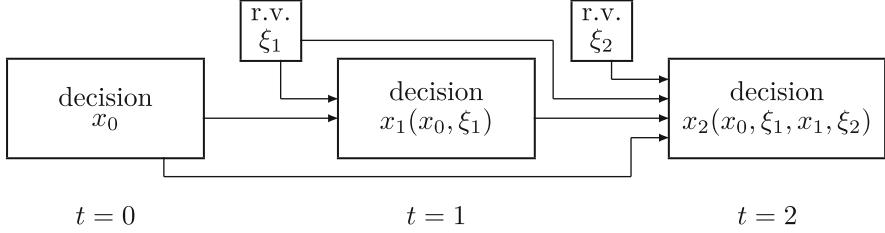


Fig. 1.8 The dynamics of multistage stochastic optimization

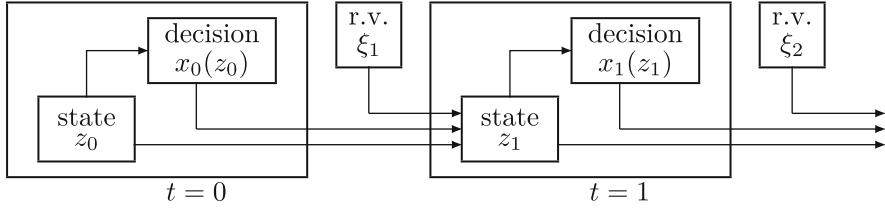


Fig. 1.9 The dynamics of the state-space decision model

and the decision functions $x_t = x_t(\xi_t)$ are functions of the actual state ξ_t . In this formulation, the notion of filtration disappears, since information is not accumulated, but concentrated in the state vector. Of course the special form (1.18) is not really a restriction, since one may enlarge the stochastic process ξ_t ($t = 1, \dots, T$) such that it generates the filtration, i.e., $\mathcal{F}_t = \sigma(\xi_t)$, by defining the state as the tuple (cf. Fig. 1.9)

$$\xi_t = (x_0, \xi_1, x_1, \dots, \xi_t).$$

However, typical stochastic dynamic models have a restricted state space, which does not grow in dimension with time. For instance, a hydrostorage management model could consider the filling heights of the reservoirs as the only relevant states, no history of the level patterns is needed for decisions.

The final costs may depend on the last stage via a terminal profit function $Q(\xi_T)$ or on intermediate stages via a set of cost functions

$$Q_1(\xi_1), Q_2(\xi_2), \dots, Q_T(\xi_T).$$

The problem (1.13) in state-space formulation reads

$$\text{minimize} \left\{ F(x) := \mathcal{R}[Q(\xi_T)] \mid \begin{array}{l} \xi_0 = z_0, x \in \mathbb{X}, \\ \xi_{t+1} = k_{t+1}(\xi_t, x_t, \xi_{t+1}), t = 0, \dots, T-1 \end{array} \right\},$$

or with intermediate cost functions

$$\begin{aligned} \text{minimize } & \left\{ F(x) := \mathcal{R}[Q_1(\xi_1), \dots, Q_T(\xi_T)] \right. \\ & \left. \begin{array}{l} \xi_0 = z_0, x \in \mathbb{X}, \\ \xi_{t+1} = k_{t+1}(\xi_t, x_t, \xi_{t+1}), t = 0, \dots, T-1 \end{array} \right\}. \end{aligned}$$

As before, we may assume that all additional constraints are absorbed in definitions of the functions k_t and Q_t .

1.4 Scenario Trees and Nested Distributions

Multistage stochastic optimization programs on finite probability spaces are defined using *scenario trees*. Scenario trees are circle-free directed graphs with a single root, for which the distance of all leaves (i.e., nodes with no outgoing edge) are at the same level; to each node, a k -vector of values is assigned and to the leaf nodes probabilities are assigned such that the leaf nodes can be seen as a discrete probability space.

While the topology of the tree is most relevant, the ordering or numbering of the nodes is irrelevant, and any order or numbering of the nodes generates the same tree. Mathematically spoken, trees are equivalence classes with respect to bijective mappings, which preserve the precedence topology. However, one may always take one representative of a class and assign labels or numbers to the nodes. To do so, let us assume that the tree consists of N nodes $\{1, \dots, N\}$, where 1 is the root. To each node n , except the root, a predecessor $\text{pred}(n)$ is defined. To each node, a stage is defined, which is its distance from the root. The nodes of the tree are dissected into the node sets at each stage \mathcal{N}_t , such that (cf. Fig. 1.10)

$\mathcal{N}_0 = \{1\}$ is the root,

\mathcal{N}_T are the leaves, and

$\mathcal{N}_{0:T-1}$ are the inner nodes (the complement set of the leaves).

Evidently, for all $n \in \mathcal{N}_t$, $\text{pred}(n) \in \mathcal{N}_{t-1}$ for $t = 1, \dots, T$. While $\text{pred}(n)$ represents the immediate predecessor, one may also define the predecessor of a node at each earlier stage. For $n \in \mathcal{N}_t$ and $s < t$ define

$$\text{pred}_s(n) = m,$$

if $m \in \mathcal{N}_s$ and there is a sequence of pred operations, such that

$$m = \text{pred}(\text{pred}(\dots \text{pred}(n) \dots)).$$

Of course, these predecessor mapping satisfies the relation

$$\text{pred}_s(\text{pred}_t(\cdot)) = \text{pred}_s(\cdot) \quad \text{for } s < t \quad (1.19)$$

and

$$\text{pred}_0(\cdot) = 1$$

is identically the root.

We also denote the direct predecessor of a node n by $n-$ and the set of all direct successors $n+$. If a node n is any successor (direct or not) of node m , we write $m \prec n$ (or $n \succ m$). Thus

$m \prec n$ is equivalent to: there exists a t such that $m = \text{pred}_t(n)$.

Scenario Trees. Scenario trees are special circle-free directed finite graphs, which do have a unique root and for which the distance of all terminal nodes from the root are equal to T , the height of the tree. In addition, scenario trees carry probability valuations on nodes and arcs: the unconditional probabilities are sitting on the nodes and the conditional branching probabilities are sitting on the arcs. Notice that it suffices to assign the unconditional probabilities $P(n)$ to the leaf nodes $n \in \mathcal{N}_T$, since the unconditional probabilities for other nodes are given by

$$P(m) = \sum_{\substack{n \succ m, \\ n \in \mathcal{N}_T}} P(n).$$

The conditional arc probabilities are defined as $Q(n) = P(n|n-) = P(n)/P(n-)$. Notice that $P(1) = 1$. Conversely, P can be gotten from Q by

$$P(n) = Q(n) \cdot \prod_{m \prec n} Q(m).$$

If every node also carries a value or vector of scenario values $\xi(n) \in \mathbb{R}^m$, then the tree is *fully valued*.

Finite Filtered Probability Spaces Are Equivalent to Scenario Trees. Alternatively one may begin with a probability space (Ω, \mathcal{F}, P) on which mappings are defined in the following way:

$$\text{pred} : \Omega \rightarrow \Omega_{T-1},$$

$$\text{pred} : \Omega_{T-1} \rightarrow \Omega_{T-2},$$

⋮

$$\text{pred} : \Omega_1 \rightarrow \Omega_0.$$

It is required that all spaces $\Omega = \Omega_T, \Omega_{T-1}, \dots, \Omega_0$ are distinct and that the mappings pred are all different, since they are defined on different spaces. We require that Ω_0 is a singleton. Let the compositions pred_t be defined as

$$\text{pred}_t := \underbrace{\text{pred} \circ \cdots \circ \text{pred}}_{t \text{ times}} : \Omega \rightarrow \Omega_{T-t}.$$

These, and the intermediary compositions fulfill the *projection property* (1.19). Let \mathcal{F}_t be the σ -algebra generated by pred_t . By the projection property, the sequence of σ -algebras \mathcal{F}_t is a filtration $\mathfrak{F} = (\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_t)$ and $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is the trivial σ -algebra, since Ω_0 is a singleton.

Notice that this construction does not necessarily require that Ω is finite. However, if Ω is finite, we may identify it with the leaf set \mathcal{N}_T of a tree and the sets Ω_t may be identified with the node sets \mathcal{N}_t at the respective stage t .

Tree Processes. Yet another completely equivalent description is based on the notion of tree processes:

Definition 1.5. A stochastic process $(v_t), t = 0, \dots, T$ with values in some state space $\mathcal{N}_t, t = 0, \dots, T$, where the \mathcal{N}_t are pairwise disjoint and \mathcal{N}_0 is a singleton, is called *tree process*, if the generated σ -algebras $\sigma(v_t)$ satisfy

$$\sigma(v_t) = \sigma(v_0, \dots, v_t) \quad (1.20)$$

for all t .³ A tree process can be equivalently characterized by the fact that the conditional distribution of (v_0, \dots, v_{t-1}) given v_t is degenerate (i.e., sits on just one value).

The tree process induces a probability distribution P on \mathcal{N}_T and we may introduce the image space \mathcal{N}_T as the basic probability space, i.e., we may set without loss of generality $\Omega = \mathcal{N}_T$. The generated sigma algebras $\mathcal{F}_t := \sigma(v_t)$ form a filtration $\mathfrak{F} = (\mathcal{F}_0, \dots, \mathcal{F}_T)$, which is evident from (1.20).

The Leaf–Path Correspondence. Obviously, the leaves of a tree uniquely determine the paths, which lead from the root to the leaf nodes. An exemplary tree is displayed in Fig. 1.10. It has 10 nodes and 6 leaves, the nodes are numbered from 1 to 10. This tree represents a probability space Ω with a filtration structure \mathfrak{F} . The elements of Ω are the leaves, which can be renamed $\omega_1, \dots, \omega_6$. At the same time, the leaves determine the whole path from the root and therefore it makes sense to use the following correspondence:

³If \mathcal{N}_0 is not a singleton, then the process is called a forest process.

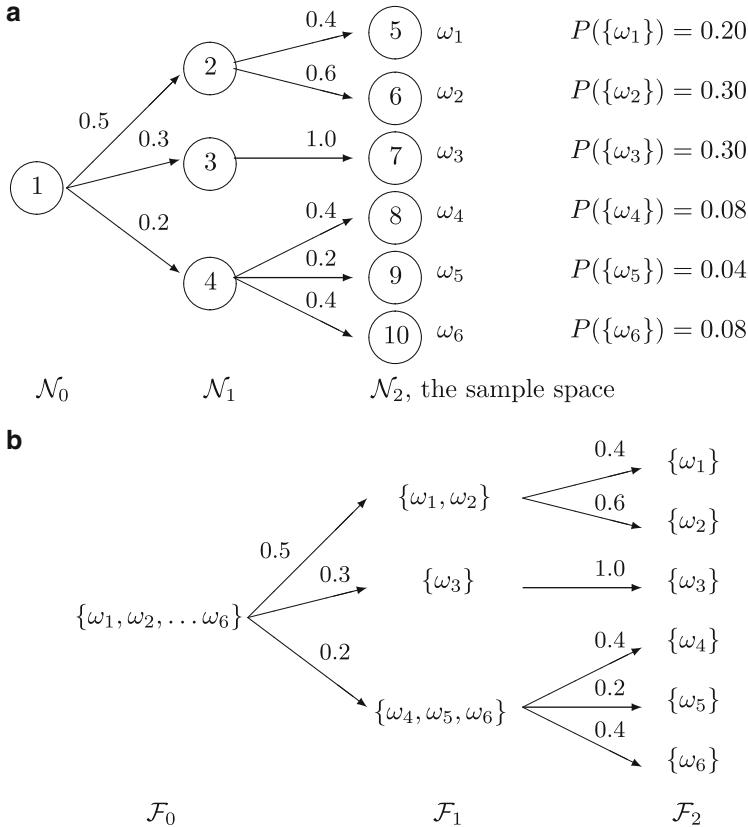


Fig. 1.10 An example tree illustrating the leaf-path correspondence (a), and the filtration induced (b)

- ω_1 corresponds to the path $(1, 2, 5)$,
 - ω_2 corresponds to the path $(1, 2, 6)$,
 - ω_3 corresponds to the path $(1, 3, 7)$,
 - ω_4 corresponds to the path $(1, 4, 8)$,
 - ω_5 corresponds to the path $(1, 4, 9)$ and
 - ω_6 corresponds to the path $(1, 4, 10)$.
- (1.21)

The pertaining filtration $\mathfrak{F} = (\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2)$ consists of the σ -algebras generated by the sets

$$\begin{aligned}\mathcal{F}_0 &= \sigma(\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}) = \{\emptyset, \Omega\}, \\ \mathcal{F}_1 &= \sigma(\{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4, \omega_5, \omega_6\}) \text{ and} \\ \mathcal{F}_2 &= \sigma(\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_5\}, \{\omega_6\}),\end{aligned}$$

as Fig. 1.10b indicates.

The Value Process Sitting on the Tree Process. While tree processes represent the available information structure in a decision process, the observable values of the scenarios are random variables or random vectors ξ_t , which are adapted to the filtration \mathfrak{F} , i.e., which are functions of the tree process v_t .

While the values of the tree process (i.e., the names assigned to the nodes of a finite scenario tree) are not relevant and are only determined up to bijective transformation, the values of the scenario process ξ_t are the basis of the decisions. We call the structure $(\Omega, \mathfrak{F}, P, \xi)$ the *value-and-information structure*.

Let us summarize our approaches.

- One may consider the filtered probability space $(\Omega, \mathfrak{F}, P)$ as the primary structure. On this space the scenario process is defined as a stochastic process $\xi = (\xi_1, \dots, \xi_T)$, which is adapted to the filtration \mathfrak{F} and what is denoted by $\xi \triangleleft \mathfrak{F}$. When considering this approach one emphasizes the role of the fixed probability space, but the concrete probability space plays a minor role in stochastic optimization. All that matters is the *distribution* of the process together with the information structure, i.e., the filtration. It may seem that the filtration cannot be defined without reference to a concrete probability space. But this is not true: one may define the filtration as the one generated by a history process $v = (v_1, \dots, v_T)$, where $v_t = (\xi_1, \dots, \xi_t)$, and study the joint distribution of v and ξ . This is when tree processes enter the picture and only their distributions matter.
- One may start with defining a tree process $v = (v_1, \dots, v_T)$. Recall that the peculiarity of a tree process is that the sequence of generated σ -algebras is increasing, i.e., that $\sigma(v_t) \subseteq \sigma(v_{t+1})$. Notice that any stochastic process η_t can be considered as the basis of a tree process by considering the history process

$$v_t = (\eta_1, \dots, \eta_t).$$

The reason why we emphasize the notion of tree processes is that this notion directly leads to scenario trees: any tree carrying probabilities can be interpreted as a tree process. It is important to notice that the “names” of the nodes do not matter: any renaming of the nodes would not change the generated filtration and the “renamed” tree process would also do the job. This observation is called *tree equivalence* and will be treated in the next subsection. If the scenario process ξ is adapted to the filtration generated by v , then there must exist functions g_t such that

$$\xi_t = g_t(v_t). \quad (1.22)$$

In the finite case, the relation (1.22) can be spelled out as “the scenario process ξ sits on the tree.”

- One may go one step further: since only the distribution of the tree process matters (and not on which probability space it is defined) and since the names of the values of the node process do not matter, one may choose a standard

representation of the values of the tree process as the conditional distributions of the scenario process ξ . This standard representation is completely Ω -free and is called the *nested distribution*. It will be presented in detail in the next subsection. For the canonical construction see Appendix D.

Related to the different paradigms of how the relevant processes can be defined is the question how the optimization model should be formulated. In a setup with a filtered probability space $(\Omega, \mathfrak{F}, P)$ the *time-oriented* modeling paradigm can be used. If, however, a concrete finite tree process v is already defined, one may call its node set \mathcal{N} and use the *node-oriented* modeling paradigm: while for the time-oriented setup we use a notation like ξ_t ($t = 1, \dots, T$) for the random scenario process, we use the notation $\xi(n)$ ($n \in \mathcal{N}$) for the values of this process on the nodes $n \in \mathcal{N}$ in the node-oriented setup. Here is an illustrative example.

Example. Consider the problem of optimal management of hydrostorages for electricity production, a typical multistage stochastic decision problem. Hydrosystems contain usually one or several cascades of hydrostorages, possibly with pumping facilities. We consider here the simplest model, where only one hydrostorage is present. Periodic decisions have to be made about the amount of electricity to be produced in the next period by turbining. The uncertain parameters concern the inflow to the storage (by rivers, rain or snow melt) as well as the market prices for energy. Here is the detailed model:

Let V_0 be the volume of water in the storage at beginning. For each time step t , the volume x_t to be turbined is determined. However, the market price η for energy, as well as the inflows ξ to the reservoir, is observed only later. The reservoir balance is

$$V_{t+1} = \min \{V_t - x_t + \xi_{t+1}, V_{\max}\} \quad (1.23)$$

where V_{\max} is the maximal storage volume. Equation (1.23) can be reformulated as

$$\begin{aligned} V_{t+1} &\leq V_t - x_t + \xi_{t+1}, \\ V_{t+1} &\leq V_{\max}. \end{aligned}$$

The profit is $Y = \sum_{t=0}^{T-1} x_t \eta_{t+1}$. In order to introduce risk aversion into the decision process we consider the upper Average Value-at-Risk (see (1.8)) of the negative profit as objective.

$$\text{minimize} \left\{ \text{AV@R} \left(- \sum_{t=0}^{T-1} x_t \eta_{t+1} \right) : x \geq 0, \text{ Eq. (1.23)} \right\}.$$

This is the time-oriented formulation. It does not refer to a discrete structure and may be considered as a general infinite variational problem, which typically cannot be solved without approximation. The finite approximation replaces the general stochastic processes ξ and η by processes sitting on a tree with node set \mathcal{N} . Let $P(n)$

be the unconditional node probabilities. In the node-oriented formulation indices do not indicate decision stages, but nodes of the tree. For instance, Eq. (1.23) looks in node-oriented formulation as

$$V(n) = \min \{V(n-) - x(n-) + \xi(n), V_{\max}\}, \quad n \geq 1.$$

Some people call the node-oriented program the “deterministic equivalent.” This is, however, not appropriate, since (i) every stochastic program has a real valued objective and real valued constraints and can be seen as a deterministic program (with a special structure, where random variables appear in the formulation of the objective and the constraints); (ii) the relation between the general time-oriented formulation and the concretization on some tree is not an equivalence, but just an approximation.

1.4.1 Nested Distributions

In the basic setup a filtered probability space $(\Omega, \mathfrak{F}, P)$ is given, on which the scenario process ξ is defined in an adapted way. Recall that the quadruple $(\Omega, \mathfrak{F}, P, \xi)$ was called a value-and-information structure. It was already argued that the basic decision problem is defined in a Ω -free way: typically the distribution of a vector-valued process is given, which represents the available information and a subvector of it represents the scenario process. While the problem setting is Ω -free, the modeling with value-and-information structures is not. That is why we introduce a complete *in-distribution*-setting, which represents all equivalent models: the nested distribution.⁴

Let d be a metric on \mathbb{R}^m , which makes it a complete separable metric space. A typical metric is

$$d(u, v) = \left(\sum_{i=1}^m w_i |u_i - v_i|^p \right)^{1/p},$$

the p -order distance with weights w_i . Let $\mathcal{P}_r(\mathbb{R}^m)$ be the collection of all Borel probability measures P on (\mathbb{R}^m, d) , such that

$$\int_{\mathbb{R}^m} d(u, u_0)^r P(du) < \infty$$

for some $u_0 \in \mathbb{R}^m$ and $r \geq 1$.

⁴Cf. also Pflug [92] for a detailed construction of the nested distribution.

For defining the nested distribution of a scenario process (ξ_1, \dots, ξ_T) with values in \mathbb{R}^m adapted to a filtration \mathfrak{F} , we define the following spaces in a recursive way:

$$\mathcal{X}_1 := \mathbb{R}^m,$$

with the interpretation that this is just one fixed scenario value of a deterministic problem. The next space is

$$\mathcal{X}_2 := \mathbb{R}^m \times \mathcal{P}_r(\mathcal{X}_1),$$

representing the scenario value at the root node, say $\xi_0 \in \mathbb{R}^m$, and the scenario distribution $\xi_1 \in \mathcal{P}_r(\mathbb{R}^m)$ at time 1 in a joint object. Ξ_2 is the basic space for single or two-stage models. A three-stage model can be interpreted as a two-stage model for which the second stage is again two-stage. Therefore the correct scenario space for a three-stage model is

$$\mathcal{X}_3 := \mathbb{R}^m \times \mathcal{P}_r(\mathcal{X}_2) = \mathbb{R}^m \times \mathcal{P}_r(\mathbb{R}^m \times \mathcal{P}_r(\mathbb{R}^m))$$

with the interpretation of the initial deterministic value ξ_0 , the distribution of ξ_1 and the conditional distribution of ξ_2 . For T -stage models this construction can be iterated T times and finally one gets the scenario space

$$\mathcal{X}_T := \mathbb{R}^m \times \mathcal{P}_r(\mathcal{X}_{T-1}).$$

Definition 1.6 (Nested Distribution). A probability distribution \mathbb{P} on \mathcal{X}_T is called a *nested distribution of depth T* .

In discrete cases, a nested distribution may be written in a recursive way. We represent discrete probabilities by a list of probabilities (in the first row) and the values (in the subsequent rows).

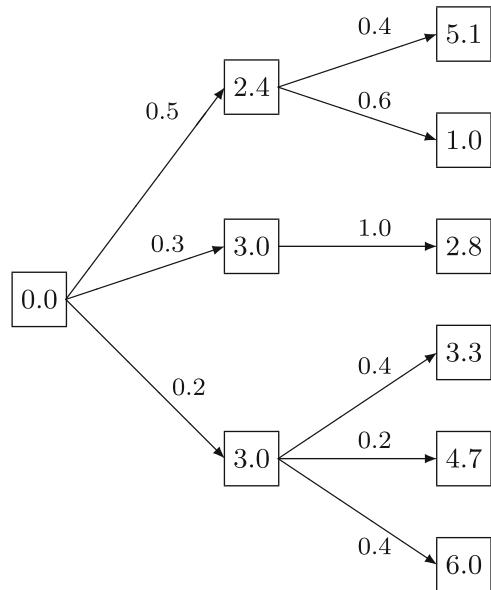
Notice that *one* value or vector of values may only carry *one* probability; the following structure on the left does not represent a valid distribution, while the structure on the right is correct:

$$\left[\begin{array}{cccc} 0.1 & 0.2 & 0.4 & 0.3 \\ \hline 3.0 & 3.0 & 1.0 & 5.0 \end{array} \right] \quad \left[\begin{array}{ccc} 0.3 & 0.4 & 0.3 \\ \hline 3.0 & 1.0 & 5.0 \end{array} \right].$$

If the discrete distribution is on \mathbb{R}^m , say on \mathbb{R}^3 , we use the notation

$$\left[\begin{array}{ccc} 0.1 & 0.6 & 0.3 \\ \hline \begin{pmatrix} 3.0 \\ 2.0 \\ 1.5 \end{pmatrix} & \begin{pmatrix} 1.0 \\ 1.2 \\ 1.3 \end{pmatrix} & \begin{pmatrix} 5.0 \\ 4.0 \\ 3.0 \end{pmatrix} \end{array} \right].$$

Fig. 1.11 A nested distribution displayed by a tree (cf. Fig. 1.10)



In the same manner, but in a recursive way, we may now represent a nested distribution as a structure, where some values are distributions themselves.

$$\begin{bmatrix} & 0.2 & 0.3 & 0.5 \\ \hline & 3.0 & 3.0 & 2.4 \\ \left[\begin{array}{c} 0.4 \\ 0.2 \\ 0.4 \end{array} \right] & \left[\begin{array}{c} 1.0 \\ 2.8 \end{array} \right] & \left[\begin{array}{c} 0.6 \\ 0.4 \end{array} \right] \\ \hline & 6.0 & 4.7 & 3.3 \end{bmatrix}$$

This recursive structure encodes the tree shown in Fig. 1.11. The random variable ξ_1 takes the values 3.0 and 2.4, while the random variable ξ_2 takes the values 6.0, 4.7, 3.3, 2.8, 1.0, and 5.1, however, also the conditional distributions of ξ_2 given \mathcal{F}_1 .

Notice that this structure encodes also the filtration and this filtration may be larger than the one generated by the scenario process itself: if one would reduce the σ -algebra \mathcal{F}_1 to the one generated by ξ_1 , one would get the following different nested distribution:

$$\begin{bmatrix} & 0.5 & 0.5 \\ \hline & 3.0 & 2.4 \\ \left[\begin{array}{c} 0.16 \\ 0.08 \\ 0.16 \\ 0.60 \end{array} \right] & \left[\begin{array}{c} 0.6 \\ 0.4 \end{array} \right] \\ \hline & 6.0 & 4.7 & 3.3 & 2.8 \end{bmatrix} \cdot \left[\begin{array}{c} 1.0 \\ 5.1 \end{array} \right]$$

There is no need to reduce the filtration to the one generated by the scenario values and this is the reason why a distinction between the tree process (ν_t) and the scenario process (ξ_t) is useful.

For any nested distribution \mathbb{P} , there is an embedded multivariate distribution P , i.e., the distribution of (ξ_1, \dots, ξ_T) without reference to the filtration. For instance, the scenario tree of Fig. 1.11 can be mapped to the pertaining multivariate distribution

$$\left[\begin{array}{cccccc} 0.08 & 0.04 & 0.08 & 0.30 & 0.30 & 0.20 \\ \hline (3.0) & (3.0) & (3.0) & (3.0) & (2.4) & (2.4) \\ (6.0) & (4.7) & (3.3) & (2.8) & (1.0) & (5.1) \end{array} \right].$$

Evidently, this multivariate distribution has lost the information about the nested structure.

The Nested Distance. Nested distributions can be metricized in a natural way by the nested distance. This distance will be defined and studied in detail in Chap. 2. Here, we give a brief introduction. Recall that the spaces \mathcal{X} were constructed recursively to hold the values of the scenario process and the distributions of the subsequent subtrees. All these spaces are metric and therefore the nested structure of spaces comes along with a nested structure of distances. We have equipped \mathbb{R}^m with a metric d which makes it a complete separable metric space. On $\mathcal{P}_r(\mathbb{R}^m, d)$ (the family of all Borel probability measures P on (\mathbb{R}^m, d) such that $\int d(u, u_0)^r dP(u) < \infty$), the Wasserstein distance d_r is well defined, which makes $\mathcal{P}_r(\mathbb{R}^m, d)$ a complete separable metric space (for the definition, properties and a discussion of the Wasserstein distance see Sect. 2.2 of Chap. 2). An important property of this distance is the following: for the point masses δ_u (δ_v , resp.) sitting on the points u (v , resp.), we have that $d_r(\delta_u, \delta_v) = d(u, v)$ and this means that the space (\mathbb{R}^m, d) is embedded in $\mathcal{P}_r(\mathbb{R}^m, d_r)$ in an isometric way.

For introducing the nested distance in a systematic way we use the new notation

$$\begin{aligned} \mathcal{X}_1 &:= \mathbb{R}^m, \\ d^{(1)} &:= d. \end{aligned}$$

On the next space

$$\mathcal{X}_2 := \mathbb{R}^m \times \mathcal{P}_r(\mathcal{X}_1),$$

which is a product of an Euclidean space and a space of measures, the distance

$$d^{(2)}((u_1, P_1), (u_2, P_2)) := d(u_1, u_2) + d_r(P_1, P_2)$$

is defined. The elements of \mathcal{X}_2 are nested distributions of depth 1. Continuing the same way we get the next step

$$\mathcal{X}_3 := \mathbb{R}^m \times \mathcal{P}_r(\mathcal{X}_2) = \mathbb{R}^m \times \mathcal{P}_r(\mathbb{R}^m \times \mathcal{P}_r(\mathbb{R}^m))$$

with distance $\mathbf{d}^{(3)}((u_1, \mathbb{P}_1), (u_2, \mathbb{P}_2)) := \mathbf{d}(u_1, u_2) + \mathbf{d}_r(\mathbb{P}_1, \mathbb{P}_2)$, where \mathbb{P}_1 and \mathbb{P}_2 are nested distributions of depth 1 and \mathbf{d}_r is their Wasserstein distance of measures on the basis of $\mathbf{d}^{(2)}$.

This construction is iterated until

$$\mathcal{X}_T := \mathbb{R}^m \times \mathcal{P}_r(\mathcal{X}_{T-1})$$

with distance $\mathbf{d}^{(T)}((u_1, \mathbb{P}_1), (u_2, \mathbb{P}_2)) := \mathbf{d}(u_1, u_2) + \mathbf{d}_r(\mathbb{P}_1, \mathbb{P}_2)$ is reached.

Definition 1.7 (Cf. [92]). The distance $\mathbf{d}\mathbf{l}_T$ is called the *nested distance* on \mathcal{X}_T .

There are alternate ways of defining the nested distance, details will be discussed in Chap. 2.

1.4.2 Equivalence and Minimality

While the distribution and hence the values of the scenario process ξ are always given, the tree process v serves as the generator of the filtration describing the information flow and hence its values are irrelevant up to one-to-one renaming transformations. We may therefore introduce the notion of equivalence. Recall that a pair $(v, \xi) = (v_1, \dots, v_T, \xi_1, \dots, \xi_T)$ of random processes on (Ω, P) generates a value-and-information structure $(\Omega, \mathfrak{F}, P, \xi)$, if

- v is a tree process in the sense of Definition 1.5 defined on Ω such that the filtration \mathfrak{F} is the one generated by (v) , and
- ξ_t are \mathbb{R}^m -valued functions of v_t , i.e., $\xi_t = f_t(v_t)$.

Recall that if $v = (v_1, \dots, v_T)$ is a tree process, then its state spaces, say $\mathcal{N}_1, \dots, \mathcal{N}_T$, are pairwise disjoint ($\mathcal{N}_t \cap \mathcal{N}_s = \emptyset$ for $s \neq t$).

Definition 1.8. Let two processes (ξ, v) , defined on (Ω, P) and $(\bar{\xi}, \bar{v})$, defined on $(\bar{\Omega}, \bar{P})$ be given, such that they generate value-and-information structures. These structures are *equivalent* if there are bijective functions k_t mapping the state spaces \mathcal{N}_t of v_t to the state spaces $\bar{\mathcal{N}}_t$ of \bar{v}_t such that the two processes

$$(\bar{v}, \bar{\xi}) = (\bar{v}_1, \dots, \bar{v}_T, \bar{\xi}_1, \dots, \bar{\xi}_T) \quad \text{and} \quad (k_1(v_1), \dots, k_T(v_T), \xi_1, \dots, \xi_T)$$

have the same distribution. It is easy to see that equivalent value-and-information structures generate the same nested distribution. As an example, consider Fig. 1.12.

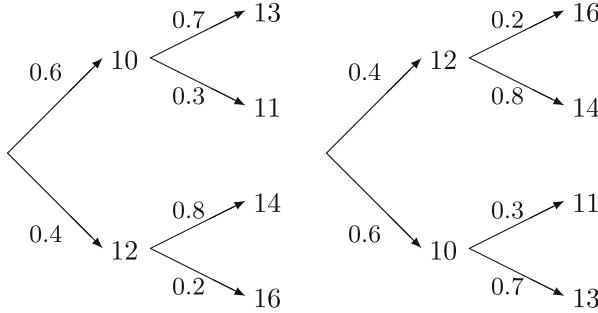


Fig. 1.12 These two trees are equivalent and the pertaining nested distributions are identical. The nested distance of the trees is 0

The two trees are equivalent, since they differ only by the (irrelevant) numbering and ordering of subtrees. Therefore they generate the same nested distribution.

We summarize what we have achieved so far:

- To each value-and-information structure $(\Omega, \mathfrak{F}, P, \xi)$ generated by (v, ξ) , there corresponds a nested distribution \mathbb{P} on the standard space \mathcal{X}_T . Equivalent value-and-information structures induce the same nested distributions.
- For every nested distribution \mathbb{P} on \mathcal{X}_T one may construct a filtered probability space $(\Omega, \mathfrak{F}, P)$, being the image measure of a tree process and an adapted process $\xi \triangleleft \mathfrak{F}$ on it such that the nested distribution of the structure $(\Omega, \mathfrak{F}, P, \xi)$ is \mathbb{P} . The canonical construction is contained in Appendix D.

The nested distributions are defined in a pure distributional concept. The relation between the nested distribution \mathbb{P} and the value-and-information structure $(\Omega, \mathfrak{F}, P, \xi)$ is comparable to the relation between a probability measure P on \mathbb{R}^m and a \mathbb{R}^m -valued random variable ξ with distribution P^ξ .

Bearing this in mind, we may alternatively consider either the nested distribution \mathbb{P} or its realization on some probability space $(\Omega, \mathfrak{F}, P, \xi)$ with \mathfrak{F} being the filtration generated by a tree process v and ξ being the value process. We symbolize this fact by writing

$$(\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}.$$

Not only equivalent trees generate the same nested distribution, also different trees may generate the same nested distribution. This may happen if a tree has two completely equivalent subtrees, which are indistinguishable from the standpoint of distribution. We call a tree *minimal*, if no smaller tree may represent the same nested distribution. Consider the two trees in Fig. 1.13: these trees \mathbb{T}_1 and \mathbb{T}_2 are not equivalent, but they generate the same nested distribution, i.e.,

$$\mathbb{T}_1 \sim \mathbb{P}, \quad \mathbb{T}_2 \sim \mathbb{P} \text{ (but } \mathbb{T}_1 \neq \mathbb{T}_2\text{)}.$$

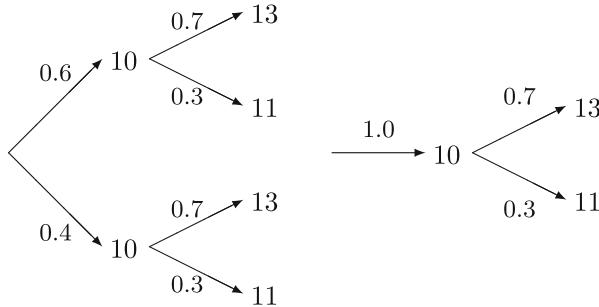


Fig. 1.13 The left tree and the right tree are not identical, but the induced nested distribution is that of the right tree, which is minimal. The left tree is not minimal. The nested distance of the trees is 0 (cf. Example 2.46 below)

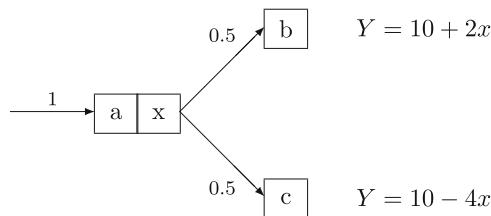


Fig. 1.14 At node (a) the decision x ($0 \leq x \leq 1$) has to be made. The return is Y

One may ask whether something is lost when considering only nested distributions, i.e., minimal trees. If a decision process is defined on a non-minimal tree, then different decisions may be taken at nodes, which belong to identical subtrees. To put it differently, on non-minimal trees randomized decisions are allowed, i.e., decisions which depend on additional random draws, which are independent of the scenario process: for instance, at some node n , decision $\dot{x}(n)$ is taken with probability α (say) and decision $\ddot{x}(n)$ is chosen with probability $1 - \alpha$. It is not difficult to construct examples, for which random decisions outperform nonrandom ones.

Example 1.9. A value-at-risk minimization. Consider the decision problem shown in Fig. 1.14. The decision x results with probability 0.5 in costs of $Y_x = 10 + 2x$ and with probability 0.5 in $Y_x = 10 - 4x$. The objective is to minimize $\mathbb{E}(Y_x) + V@R_{0.9}(Y_x) = 10 - x + 10 + 2x = 20 + x$ for $0 \leq x \leq 1$. The optimal decision is $x = 0$ with an objective value of 20.

If, however, the first node is split in two with respective probabilities of 0.8 and 0.2, then the decision situation is as in Fig. 1.15. Now the best decision is $x_1 = 0$ and $x_2 = 1$ resulting in a better objective value of 19.8.

However, if the probability functional $\mathcal{R}_P(\cdot)$ is concave in the probability measure (i.e., the mapping $P \mapsto \mathcal{R}_P(Y)$ in the objective is concave for all Y ; see also Definition 3.25 later), then the optimal solution can always be chosen as a nonrandomized one.

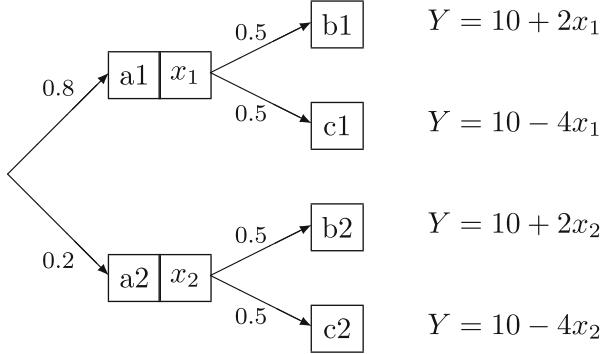


Fig. 1.15 An extension of Fig. 1.14: the node (a) is split into two nodes with identical subtrees. At node (a1) the decision x_1 has to be made, at the parallel node (a2) the decision is x_2 . Notice that the filtration in this tree is not minimal

Lemma 1.10. Consider the basic problem

$$\text{minimize } \{\mathcal{R}[Q(x, \xi)] : x \triangleleft \mathfrak{F}; x \in \mathbb{X}\}.$$

If the functional \mathcal{R}_P is concave in P , then among the set of optimal solutions there are nonrandomized ones. Randomization does not lead to better solutions.

Proof. Start with solving the basic problem on a minimal tree \mathbb{T}_1 (with nested distribution \mathbb{P}_1) and call the optimal solution $x^{(1)}$. Let n be a specific node and let \mathbb{T}_+ be the subtree, which has n as root. Let $Q(n)$ be the conditional probability of n . Suppose now that the node n at stage t is split into two nodes n' and n'' having conditional probabilities $\lambda Q(n)$ ($(1 - \lambda)Q(n)$, resp.) and identical subtrees \mathbb{T}'_+ and \mathbb{T}''_+ . Call this extended tree \mathbb{T}_2 (with nested distribution \mathbb{P}_2) and let $x^{(2)}$ be the optimal solution of the basic problem on this extended tree. Let the solutions x' (x'' , resp.) of the problem on \mathbb{T}_1 be the identical to $x^{(1)}$ on all nodes except on the subtree \mathbb{T}_+ , where x' is identical to $x^{(2)}$ on \mathbb{T}' and x'' is identical to $x^{(2)}$ on \mathbb{T}'' . Then, by concavity for compounds $\mathcal{C}(\cdot, \cdot, \lambda)$ (see (1.24) below),

$$\begin{aligned} \mathcal{R}_{\mathbb{P}_2}[Q(x^{(2)}, \xi)] &= \mathcal{R}[\mathcal{C}(Q(x', \xi), Q(x'', \xi), \lambda)] \\ &\geq \lambda \cdot \mathcal{R}_{\mathbb{P}_1}[Q(x', \xi)] + (1 - \lambda) \mathcal{R}_{\mathbb{P}_1}[Q(x'', \xi)] \\ &\geq \lambda \cdot \min_x \mathcal{R}_{\mathbb{P}_1}[Q(x, \xi)] + (1 - \lambda) \min_x \mathcal{R}_{\mathbb{P}_1}[Q(x, \xi)] \\ &= \mathcal{R}_{\mathbb{P}_1}[Q(x^{(1)}, \xi)]. \end{aligned}$$

Thus the solution on the extended tree, i.e., the randomized solution on the original tree cannot lead to a better objective value than the nonrandomized one.

1.4.3 Convex Structures for Scenario Models

Suppose that structures $(\Omega, \mathfrak{F}, P, \xi)$ are given. There are at least three ways of defining convex combinations.

- (i) Convex combination of the value process: for two given processes $\xi^{(1)}$ and $\xi^{(2)}$ one may form the convex combination

$$(\Omega, \mathfrak{F}, P, \lambda \xi^{(1)} + (1 - \lambda) \xi^{(2)}).$$

- (ii) Convex combination of the probability measure: for two given probability measures $P^{(1)}$ and $P^{(2)}$ on (Ω, \mathfrak{F}) one may form the convex combination

$$(\Omega, \mathfrak{F}, \lambda P^{(1)} + (1 - \lambda) P^{(2)}, \xi).$$

- (iii) Compounding two nested distributions. Let $(\Omega^{(1)}, \mathfrak{F}^{(1)}, P^{(1)}, \xi^{(1)}) \sim \mathbb{P}^{(1)}$ and $(\Omega^{(2)}, \mathfrak{F}^{(2)}, P^{(2)}, \xi^{(2)}) \sim \mathbb{P}^{(2)}$ be two nested distributions. The compound of the two is the nested distribution

$$\mathbb{P} := \mathcal{C}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \lambda) := \begin{cases} \mathbb{P}^{(1)} & \text{with probability } \lambda \\ \mathbb{P}^{(2)} & \text{with probability } 1 - \lambda. \end{cases} \quad (1.24)$$

Notice that in models, the tree corresponding to the compound distribution has an additional node, a new root, which has the trees $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ as subtrees (see Fig. 1.16 for illustration).

Denote by $\mathbf{d}\ell$ the nested distance between two nested distributions. By writing $(\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}$ we mean that Ω is a concrete probability space endowed with a filtration $\mathfrak{F} = (\mathcal{F}_1, \dots, \mathcal{F}_T)$, carrying a probability measure P and a stochastic process $\xi = (\xi_1, \dots, \xi_T)$. \mathbb{P} is only defined in distributional terms, i.e., many different concrete probability models may lead to the same nested distribution.

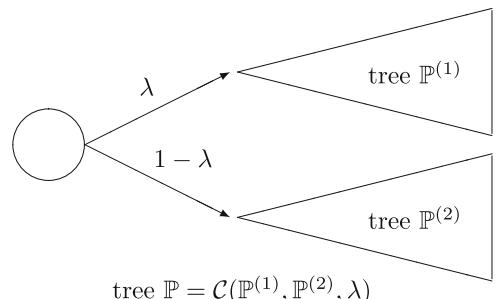


Fig. 1.16 Compounding two subtrees

$$\text{tree } \mathbb{P} = \mathcal{C}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \lambda)$$

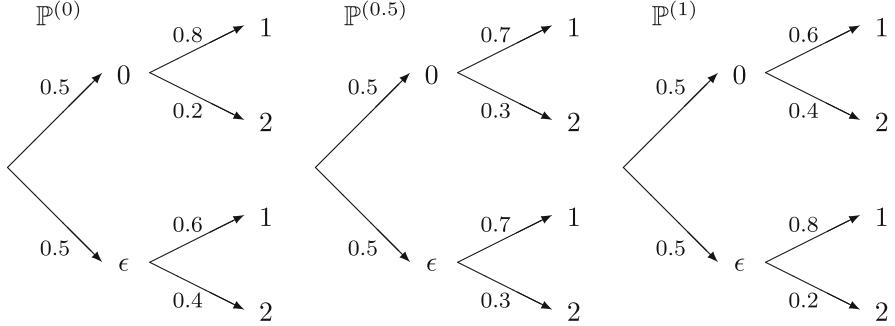


Fig. 1.17 Left: $\mathbb{P}^{(0)}$, right: $\mathbb{P}^{(1)}$, middle: $\mathbb{P}^{(0.5)}$. For $0 \leq \epsilon \leq 0.2$ it holds that $\text{dl}(\mathbb{P}^{(0)}, \mathbb{P}^{(0.5)}) = 0.1$ (and $\text{dl}(\mathbb{P}^{(0.5)}, \mathbb{P}^{(1)}) = 0.1$), but $\text{dl}(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}) = \epsilon$

If we assign the nested distance to value-and-information structures,

$$\text{dl}_r \left((\Omega^{(1)}, \mathfrak{F}^{(1)}, P^{(1)}, \xi^{(1)}), (\Omega^{(2)}, \mathfrak{F}^{(2)}, P^{(2)}, \xi^{(2)}) \right),$$

we mean $\text{dl}_r(\mathbb{P}^{(1)}, \mathbb{P}^{(2)})$ with $(\Omega^{(1)}, \mathfrak{F}^{(1)}, P^{(1)}, \xi^{(1)}) \sim \mathbb{P}^{(1)}$ and $(\Omega^{(2)}, \mathfrak{F}^{(2)}, P^{(2)}, \xi^{(2)}) \sim \mathbb{P}^{(2)}$.

Convexity of Nested Balls. We investigate the relation of convex structures of valuated trees to their nested distance. It turns out that not all ways to form convex combinations is in a natural way compatible with the nested distance. Notice that the nested distance depends only on the nested distribution induced by the scenario tree and is invariant with respect to equivalence of the tree. However, the mapping $P \mapsto (\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}$ is not invariant for fixed ξ , since only by changing both P and ξ simultaneously preserves the nested distribution. As illustration, consider Fig. 1.17. The left tree model is $(\Omega, \mathfrak{F}, P^{(0)}, \xi) \sim \mathbb{P}^{(0)}$, the right model is $(\Omega, \mathfrak{F}, P^{(1)}, \xi) \sim \mathbb{P}^{(1)}$ and the middle model is $(\Omega, \mathfrak{F}, \frac{1}{2}P^{(0)} + \frac{1}{2}P^{(1)}, \xi) \sim \mathbb{P}^{(0.5)}$. Set first $\epsilon = 0$. Then evidently $\mathbb{P}^{(0)} = \mathbb{P}^{(1)}$, but $\mathbb{P}^{(0.5)} \neq \mathbb{P}^{(0)}$ and therefore

$$\begin{aligned} \text{dl}_r(\mathbb{P}^{(0)}, \mathbb{P}^{(0.5)}) &= \text{dl}_r \left((\Omega, \mathfrak{F}, P^{(0)}, \xi), \left(\Omega, \mathfrak{F}, \frac{1}{2}P^{(0)} + \frac{1}{2}P^{(1)}, \xi \right) \right) \\ &> \frac{1}{2} \text{dl}_r \left((\Omega, \mathfrak{F}, P^{(0)}, \xi), (\Omega, \mathfrak{F}, P^{(0)}, \xi) \right) \\ &\quad + \frac{1}{2} \text{dl}_r \left((\Omega, \mathfrak{F}, P^{(0)}, \xi), (\Omega, \mathfrak{F}, P^{(1)}, \xi) \right) \\ &= 0 + 0 = \frac{1}{2} \text{dl}_r(\mathbb{P}^{(0)}, \mathbb{P}^{(0)}) + \frac{1}{2} \text{dl}_r(\mathbb{P}^{(0)}, \mathbb{P}^{(1)}). \end{aligned} \tag{1.25}$$

However, for $\epsilon = 0$, the model $\mathbb{P}^{(0)}$ is not minimal. But by choosing ϵ small enough, e.g., $\epsilon = 0.001$, the inequality (1.25) is still valid, demonstrating that convexity of nested balls w.r.t. the probability measure does not hold.

A similar example may also exhibit the possible relation

$$\begin{aligned} \mathbf{d}\mathbf{l}_r \left((\Omega, \mathfrak{F}, P, \xi), \left(\Omega, \mathfrak{F}, \frac{1}{2}\xi^{(1)} + \frac{1}{2}\xi^{(2)}, P \right) \right) \\ > \frac{1}{2} \mathbf{d}\mathbf{l}_r ((\Omega, \mathfrak{F}, P, \xi), (\Omega, \mathfrak{F}, P, \xi^{(1)})) + \frac{1}{2} \mathbf{d}\mathbf{l}_r ((\Omega, \mathfrak{F}, P, \xi), (\Omega, \mathfrak{F}, P, \xi^{(2)})). \end{aligned}$$

Thus also no convexity of nested balls with respect to the scenario values may hold. On the other hand, the correct way to combine nested distributions as convex combinations is the compounding operation (1.24). In fact it is easy to see that

$$\mathbf{d}\mathbf{l}_r(\mathbb{P}_+^{(0)}, \mathcal{C}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \lambda))^r \leq \lambda \mathbf{d}\mathbf{l}_r(\mathbb{P}^{(0)}, \mathbb{P}^{(1)})^r + (1 - \lambda) \mathbf{d}\mathbf{l}_r(\mathbb{P}^{(0)}, \mathbb{P}^{(2)})^r \quad (1.26)$$

where $\mathbb{P}_+^{(0)}$ is the nested distribution of $\mathbb{P}^{(0)}$ augmented by an additional root, which is connected to the old root by an arc with probability one. The proof of (1.26) uses the notion of transportation plans, which will be introduced in (2.35): if π_1 is a transportation plan between $\mathbb{P}^{(0)}$ and $\mathbb{P}^{(1)}$ and π_2 is a transportation plan between $\mathbb{P}^{(0)}$ and $\mathbb{P}^{(2)}$, then it can be composed to a transportation plan between $\mathbb{P}_+^{(0)}$ and $\mathcal{C}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \lambda)$. The plan $\lambda\pi_1$ is responsible for the transport between $\mathbb{P}^{(0)}$ and the $\mathbb{P}^{(1)}$ -subtree of $\mathcal{C}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \lambda)$, while $(1 - \lambda)\pi_2$ is responsible for the transport between $\mathbb{P}^{(0)}$ and the $\mathbb{P}^{(2)}$ -subtree of $\mathcal{C}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \lambda)$. Since this is a valid transportation plan, the optimal can only be better and thus the inequality (1.26) is proved.

Chapter 2

The Nested Distance

In the present context of stochastic optimization we are interested in approximations of stochastic processes. To quantify the quality of an approximation, a concept of distance between stochastic processes is necessary. This is accomplished by the nested distance, which was introduced in Chap. 1 and is systematically treated in what follows. To this end we review different concepts of distances for probability measures first. The Wasserstein distance will be generalized to the nested distance between discrete time stochastic processes.

The distance of stochastic processes is based on the distance of the induced probability measures. There exists a broad variety of different concepts of distances on probability spaces in the literature. Some of them metricize convergence in probability or other variants of different topologies on random variables or probability measures. Rachev [105, 108] lists 76 metrics for measures, and many of them are adapted to concrete and particular problems.

A useful distance, which is adapted to stochastic optimization, should comprise various properties: it

- should measure distances of distributions and be independent of different, underlying probability spaces,
- should allow reasonable computational implementations,
- should represent a version of the weak* topology¹ for random variables to enable approximations by discrete measures and, above all,
- should extend to general stochastic processes.

The Wasserstein distance, which is a solution of an optimization problem itself, covers the desired properties in a natural way. As an extra, there is a close, almost intimate relation between the Wasserstein distance and risk functionals. In addition,

¹Recall that $P_n \rightarrow P$ in the weak* topology, if $\int h dP_n \rightarrow \int h dP$ for all bounded and continuous functions h .

this distance is the basis for its multistage generalization, the nested distance and is therefore discussed in more detail below.

2.1 Distances of Probability Measures

In this section we work with ordinary probability distributions P on \mathbb{R}^m , say. When replacing a probability model P by another (typically simpler) model \tilde{P} , the basic question arises: how close is \tilde{P} to P ? Obviously, distances quantify the notion of closeness. We review here some ways of dealing with the concept of closeness for probability measures.

Let \mathcal{P} be a set of probability measures on \mathbb{R}^m .

Definition 2.1. A *semi-distance* d on $\mathcal{P} \times \mathcal{P}$ satisfies the following three conditions:

(i) *Nonnegativity*: for all $P_1, P_2 \in \mathcal{P}$,

$$d(P_1, P_2) \geq 0;$$

(ii) *Symmetry*: for all $P_1, P_2 \in \mathcal{P}$,

$$d(P_1, P_2) = d(P_2, P_1);$$

(iii) *Triangle Inequality*: for all $P_1, P_2, P_3 \in \mathcal{P}$,

$$d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2).$$

A semi-distance $d(\cdot, \cdot)$ is called a *distance* if it satisfies the strictness property:

(iv) *Strictness*: if $d(P_1, P_2) = 0$, then $P_1 = P_2$.

2.1.1 Semi-Distances Generated by a Class of Test Functions

A general principle for defining semi-distances and distances consists in choosing a family of integrable functions \mathcal{H} (i.e., a family of functions such that the integral $\int h(w) P(dw)$ exists for all $P \in \mathcal{P}$) and defining

$$d_{\mathcal{H}}(P_1, P_2) := \sup_{h \in \mathcal{H}} \left| \int h \, dP_1 - \int h \, dP_2 \right|.$$

$d_{\mathcal{H}}$ is called the (semi-)distance *generated by* \mathcal{H} .

In general, $d_{\mathcal{H}}$ is only a semi-distance. If \mathcal{H} is *separating*, i.e., if for every pair $P_1, P_2 \in \mathcal{P}$ there is a function $h \in \mathcal{H}$ such that $\int h dP_1 \neq \int h dP_2$, then $d_{\mathcal{H}}$ is strict and thus is a distance.

The Moment Matching Semi-Distance. Let \mathcal{P}_q be the set of all probability measures on \mathbb{R}^1 which possess the q -th moment, i.e., for which $\int \max\{1, |w|^q\} P(dw) < \infty$. The moment matching semi-distance on \mathcal{P}_q is

$$d_{M_q}(P_1, P_2) = \sup \left\{ \left| \int w^s P_1(dw) - \int w^s P_2(dw) \right| : s \in \{1, 2, \dots, q\} \right\}. \quad (2.1)$$

The Moment Matching Caveat. The moment matching semi-distance is not a distance, even if q is chosen to be large or even infinity. In fact, there are examples of different probability measures on \mathbb{R}^1 , which have the same moments of all orders. For instance, there is a manifold of probability measures, which have all moments equal to those of the lognormal distribution, but are not lognormal (the lognormal distribution is often present in mathematical finance). Figure 2.1b displays two (of infinitely many) distributions with all moments coinciding, cf. Heyde [58]. Indeed, there are also distributions taking values on the negative axis having the same moments as the lognormal distribution, which itself has nonnegative support.

Ignoring these facts it is a widespread method in applications to match the first four moments, i.e., to work with d_{M_4} . The following example displays two further densities, coinciding in their first four moments, but exhibiting very different properties in a drastic way.

Example 2.2 (See [93]). Let P_1 and P_2 be the two probability measures on \mathbb{R} with densities g_1 and g_2 , where

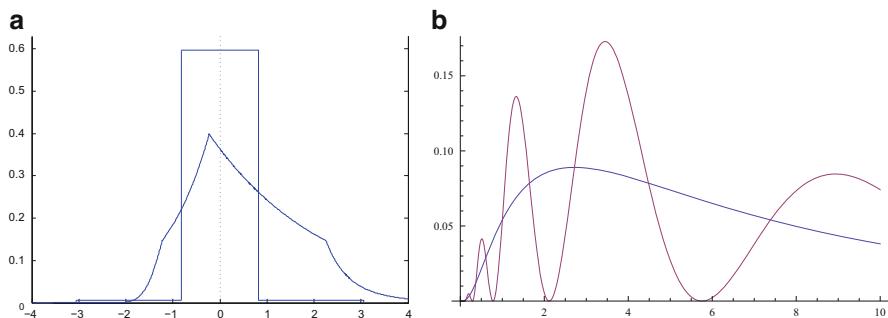


Fig. 2.1 The moment matching caveat. (a) Two densities with identical first four moments. (b) Two densities with all moments coinciding

$$\begin{aligned}
g_1(w) = & 0.3988 [\exp(-|w + 0.2297|^3) \cdot \mathbb{1}_{\{w \leq -1.2297\}} \\
& + \exp(-|w + 0.2297|) \cdot \mathbb{1}_{\{-1.2297 < w \leq -0.2297\}} \\
& + \exp(-0.4024 \cdot (w + 0.2297)) \cdot \mathbb{1}_{\{-0.2297 < w \leq 2.2552\}} \\
& + 1.0985 \cdot (0.4024w + 0.2925)^{-6} \cdot \mathbb{1}_{\{2.2552 < w\}}] \text{ and} \\
g_2(w) = & 0.5962 \cdot \mathbb{1}_{\{|w| \leq 0.8163\}} + 0.00595 \cdot \mathbb{1}_{\{0.8163 < |w| \leq 3.0588\}}
\end{aligned}$$

(see Fig. 2.1a). Both densities are unimodal and coincide in the first four moments, which are $m_1 = 0$, $m_2 = 0.3275$, $m_3 = 0$, and $m_4 = 0.7230$ ($m_q(P) = \int w^q dP(w)$). Their fifth moment, however, could not differ more: while the fifth moment of P_2 is zero, P_1 has infinite fifth moment. The density g_1 is asymmetric, has a sharp cusp at -0.2297 and unbounded support; in contrast, g_2 is symmetric around 0, has a flat density there, has finite support, and possesses all moments. The distribution functions and quantiles differ drastically as well: we have that $G_{P_1}(0.81) = 0.6257$ and $G_{P_1}(-0.81) = 0.1098$, while $G_{P_2}(0.81) = 0.9807$ and $G_{P_2}(-0.81) = 0.0133$. Thus the probability of the interval $[-0.81, 0.81]$ is only 51 % under P_1 , while it is 95 % under P_2 .

Summarizing, matching moments do not match the distributions. The moment matching semi-distance is not well suited for approximating probability distributions, since it is not fine enough to capture the relevant quality of an approximation (cf. also the additional Example 2.22 below.)

Variational Distance. The other extreme would be to choose as the generating class \mathcal{H} all measurable functions h such that $|h| \leq 1$. This class generates a distance, which is called the variational distance (more precisely, twice the variational distance). It is easy to see that if P_1 (resp. P_2) has density g_1 (resp. g_2), then

$$\begin{aligned}
& \sup \left\{ \left| \int h dP_1 - \int h dP_2 \right| : |h| \leq 1, h \text{ measurable} \right\} \\
&= \int |g_1(w) - g_2(w)| dw \\
&= 2 \cdot \sup \{|P_1(A) - P_2(A)| : A \text{ a measurable set}\}.
\end{aligned}$$

The distance

$$\mathbf{d}_V(P_1, P_2) := \sup \{|P_1(A) - P_2(A)| : A \text{ a measurable set}\} \quad (2.2)$$

is called the *variational distance* between P_1 and P_2 .

The variational distance is a very fine distance, too fine for our applications: if P_1 has a density and P_2 sits on at most countably many points, then $\mathbf{d}_V(P_1, P_2) = 1$, independently of the number of mass points of P_2 . Thus there is no hope to approximate any continuous distribution by a discrete one with respect to the variational distance.

Uniform Distance. One may restrict the class of sets in (2.2) to a certain subclass. If one employs the class of half-unbounded rectangles in \mathbb{R}^m of the form $(-\infty, w_1] \times (-\infty, w_2] \times \cdots \times (-\infty, w_m]$ one obtains the *uniform distance*, also called *Kolmogorov distance*

$$\mathbf{d}_U(P_1, P_2) := \sup \{|G_{P_1}(w) - G_{P_2}(w)| : w \in \mathbb{R}^m\},$$

where $G_P(\cdot)$ is the distribution function of P ,

$$G_P(w) = P\{(-\infty, w_1] \times \cdots \times (-\infty, w_m]\}.$$

Notice that a unit mass at point x and at point y are at a distance 1 both in the \mathbf{d}_V distance and in the \mathbf{d}_U distance, irrespective of how close x is to y . Especially when dealing with continuous baseline models and approximating discrete models, these distances are too fine.

Bounded Lipschitz Distance. Reducing the class \mathcal{H} to the class of all bounded, Lipschitz functions leads to the bounded Lipschitz metric, which metricizes the weak convergence of probability measures.

The *bounded Lipschitz distance* is defined as

$$\mathbf{d}_{BL}(P_1, P_2) := \sup \left\{ \int h \, dP_1 - \int h \, dP_2 : |h(w)| \leq 1, |h(w) - h(v)| \leq \|w - v\| \right\},$$

it involves the class \mathcal{H} of functions h which are uniformly bounded by 1, and which are Lipschitz continuous with Lipschitz constant 1.

Kantorovich Distance. The *Kantorovich distance* (also *Wasserstein distance of order 1*, cf. Definition 2.4 below) is the bounded Lipschitz distance, where the requirement of boundedness of h is dropped:

$$\mathbf{d}_1(P_1, P_2) := \sup \left\{ \int h \, dP_1 - \int h \, dP_2 : h(w) - h(v) \leq \|w - v\| \right\}.$$

This distance metricizes weak convergence on sets of probability measures which possess uniformly a first moment, as is elaborated in Theorem 2.23 below. On the real line, the Kantorovich metric may also be written as

$$\mathbf{d}_1(P_1, P_2) = \int_{-\infty}^{\infty} |G_{P_1}(w) - G_{P_2}(w)| \, dw = \int_0^1 |G_{P_1}^{-1}(p) - G_{P_2}^{-1}(p)| \, dp, \quad (2.3)$$

where $G_P^{-1}(p) = \inf\{w : G_P(w) \geq p\}$ (see Vallander [135]).

Fortet–Mourier Distance. If \mathcal{H} is the class of Lipschitz functions of order q (q -Lipschitz), the *Fortet–Mourier distance* is obtained:

$$\mathbf{d}_{\text{FM}_q}(P_1, P_2) := \sup \left\{ \int h \, dP_1 - \int h \, dP_2 : L_q(h) \leq 1 \right\}, \quad (2.4)$$

where the Lipschitz constant of order q is defined as

$$L_q(h) = \inf \left\{ L : |h(w) - h(v)| \leq L \cdot \|w - v\| \cdot \max(1, \|w\|^{q-1}, \|v\|^{q-1}) \right\}. \quad (2.5)$$

Notice that $L_{q'}(h) \leq L_q(h)$ for $q \leq q'$; in particular, $L_q(h) \leq L_1(h)$ for all $q \geq 1$ and therefore

$$\mathbf{d}_1(P_1, P_2) \leq \mathbf{d}_{\text{FM}_q}(P_1, P_2) \leq \mathbf{d}_{\text{FM}_{q'}}(P_1, P_2) \quad \text{for } 1 \leq q \leq q'.$$

The Fortet–Mourier distance metricizes weak convergence on sets of probability measures possessing uniformly a q -th moment. Notice that the function $w \mapsto \|w\|^q$ is q -Lipschitz with Lipschitz constant $L_q = q$. On \mathbb{R}^1 , the Fortet–Mourier distance may be equivalently written as

$$\mathbf{d}_{\text{FM}_q}(P_1, P_2) = \int \max \{1, |u|^{q-1}\} \cdot |G_{P_1}(u) - G_{P_2}(u)| \, du$$

(see Rachev [105, page 93]). For $q = 1$, the Fortet–Mourier distance coincides with the Kantorovich distance.

Further distances on probability measures and their relations can be found, e.g., in the review of Gibbs and Su [45].

2.2 The Wasserstein Distance

The Wasserstein distance generalizes the Kantorovich distance, although it is not generated by a set of test functions \mathcal{H} (except in special cases).

Importantly, the Wasserstein distance allows a generalization for stochastic processes. This generalization, the nested distance, is of particular interest in multistage stochastic optimization, and addressed in Sect. 2.10 below.

We adapt and augment the common concept of the Wasserstein distance here to prepare it for multistage stochastic optimization. For this we consider a general, real valued and measurable function

$$c: \Omega \times \tilde{\Omega} \rightarrow \mathbb{R} \quad (2.6)$$

linking two sample spaces Ω and $\tilde{\Omega}$.

The function c is often associated with the interpretation that moving a particle $\omega \in \Omega$ to $\tilde{\omega} \in \tilde{\Omega}$ costs $c(\omega, \tilde{\omega})$, therefore c is often called a *cost function*.

The common definition of the Wasserstein distance considers the cost function

$$c(\cdot, \cdot) := d(\cdot, \cdot)^r : \Omega \times \Omega \rightarrow \mathbb{R},$$

where d is a distance on Ω and $r \geq 1$. The notable difference is that the function c in (2.6) deals with two *different* spaces Ω and $\tilde{\Omega}$, whereas the distance function d involves just a single space, i.e., $\tilde{\Omega} = \Omega$. In this situation the transportation costs $c(\omega, \tilde{\omega})$ are assumed to be proportional to the transported distance $d(\omega, \tilde{\omega})$, or to $d(\omega, \tilde{\omega})^r$.

Inheriting a Distance from Random Variables. Typically, the probability space (Ω, \mathcal{F}, P) does not carry a topology or distance. However, in all applications in this book, we assume that a distance or semi-distance is inherited on Ω from a random variable $\xi : \Omega \rightarrow \mathbb{R}^m$ by

$$d(\omega_1, \omega_2) := \|\xi(\omega_1) - \xi(\omega_2)\|,$$

where $\|\cdot\|$ is some norm in \mathbb{R}^m .² The notion can be extended to the case of two different probability spaces Ω and $\tilde{\Omega}$:

Definition 2.3. If ξ is an \mathbb{R}^m -valued random variable on Ω and $\tilde{\xi}$ is an \mathbb{R}^m -valued random variable on $\tilde{\Omega}$, then the *inherited distance* between elements of Ω and $\tilde{\Omega}$ can be defined by the transportation cost function $c(\omega, \tilde{\omega})$

$$d(\omega, \tilde{\omega}) := c(\omega, \tilde{\omega}) = d(\xi(\omega), \tilde{\xi}(\tilde{\omega})) \quad (2.7)$$

for some distance d in \mathbb{R}^m ; often $d(w, v) = \|w - v\|$ for some norm $\|\cdot\|$ in \mathbb{R}^m .

The transportation costs or distances between elements of Ω and $\tilde{\Omega}$ can be extended to transportation costs or distances between probabilities P on Ω and \tilde{P} on $\tilde{\Omega}$.

Definition 2.4 (Optimal Transportation Cost). Given two probability spaces (Ω, \mathcal{F}, P) and $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ and a transportation cost function c , the optimal transportation cost is

$$\inf_{\pi} \iint_{\Omega \times \tilde{\Omega}} c(\omega, \tilde{\omega}) \pi(d\omega, d\tilde{\omega}), \quad (2.8)$$

where the infimum is taken over all (bivariate) probability measures π on $\Omega \times \tilde{\Omega}$ having the marginals P and \tilde{P} , that is

²Notice that it might happen that two different elements ω_1 and ω_2 are at distance 0, namely if $\xi(\omega_1) = \xi(\omega_2)$. In this case the distance is only a semi-distance, but it can as well be taken as the basis of a Wasserstein distance construction, which will then also turn out to be a semi-distance.

$$\pi(A \times \tilde{\Omega}) = P(A) \text{ and } \pi(\Omega \times B) = \tilde{P}(B) \quad (2.9)$$

for all measurable sets $A \in \mathcal{F}$ and $B \in \tilde{\mathcal{F}}$. The optimal measure π is called the *optimal transport plan*. It exists under the conditions of Remark 2.6 below.

Specializing to the case where the costs are given by an inherited distance between elements of Ω and $\tilde{\Omega}$ one obtains the Wasserstein distance.

Definition 2.5 (Wasserstein Distance). The *Wasserstein distance* of order r ($r \geq 1$) is

$$d_r(P, \tilde{P}) := \left(\inf_{\pi} \iint_{\Omega \times \tilde{\Omega}} d(\omega, \tilde{\omega})^r \pi(d\omega, d\tilde{\omega}) \right)^{1/r}, \quad (2.10)$$

where the infimum is among all joint probability measures π on $\Omega \times \tilde{\Omega}$ (more precisely: on the product $\mathcal{F} \otimes \tilde{\mathcal{F}}$ of the σ -algebras) which satisfy (2.9).

Remark 2.6. The infimum in (2.8) is attained, if both measures P and \tilde{P} are tight, i.e., for every $\epsilon > 0$ there are compact sets K and \tilde{K} such that $P(K^c) \leq \epsilon$ and $\tilde{P}(\tilde{K}^c) \leq \epsilon$.³ Under this condition, the family of all measures π with marginals P and \tilde{P} is uniformly tight, since for all these measures

$$\pi((K \times \tilde{K})^c) \leq \pi(K^c \times \tilde{\Omega}) + \pi(\Omega \times \tilde{K}^c) \leq 2\epsilon,$$

i.e., is arbitrarily small if K and K' are chosen appropriately. Closed families of uniformly tight probability measures are compact (Prohorov's Theorem, see, e.g., Parthasarathy and Kalyanapuram [85]). Since the integrand of (2.10) is continuous in π , the infimum is attained.

Definition 2.5 is used in two different situations:

- either there are given two abstract probability spaces (Ω, \mathcal{F}, P) and $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ and two random variables $\xi : \Omega \rightarrow \mathbb{R}^m$ and $\tilde{\xi} : \tilde{\Omega} \rightarrow \mathbb{R}^m$ such that the distance is the induced distance according to Definition 2.3. In this case one may write

$$d_r(P, \tilde{P}) := \left(\inf_{\pi} \iint_{\Omega \times \tilde{\Omega}} d(\xi(\omega), \tilde{\xi}(\tilde{\omega}))^r \pi(d\omega, d\tilde{\omega}) \right)^{1/r},$$

where the infimum is over all joint probability measures with marginals P (\tilde{P} , resp.);

- or the two probabilities P and \tilde{P} are defined on \mathbb{R}^m endowed with a distance d (e.g., $d(u, v) = \|u - v\|$ for $u, v \in \mathbb{R}^m$). In the latter case one may write

³ K^c denotes the complement of the set K .

$$\mathbf{d}_r(P, \tilde{P}) := \left(\inf_{\pi} \iint_{\mathbb{R}^m \times \mathbb{R}^m} \mathbf{d}(u, v)^r \pi(du, dv) \right)^{1/r},$$

where the infimum is over all probability measures on $\mathbb{R}^m \times \mathbb{R}^m$ with marginals P (\tilde{P} , resp.).

The second case can be seen as a special case of the first one considering the identical random variables $\xi = \text{id}$ and $\tilde{\xi} = \text{id}$. Both cases are considered in the following. The context always makes clear whether we consider probabilities on abstract spaces endowed with the induced distance or their image measures on \mathbb{R}^m .

The collection of all probability measures P , which satisfy for some—and thus for any $\omega_0 \in \Omega$ —the moment-like condition

$$\int_{\Omega} \mathbf{d}(\omega, \omega_0)^r P(d\omega) < \infty$$

is denoted by $\mathcal{P}_r(\Omega; \mathbf{d})$. It is immediate from the inequality

$$\mathbf{d}(\omega, \tilde{\omega})^r \leq 2^{r-1} (\mathbf{d}(\omega, \omega_0)^r + \mathbf{d}(\tilde{\omega}, \omega_0)^r)$$

(this is the triangle inequality when $r = 1$) that the problem (2.10) is feasible and well defined whenever $P \in \mathcal{P}_r(\Omega; \mathbf{d})$ and $\tilde{P} \in \mathcal{P}_r(\tilde{\Omega}; \mathbf{d})$, because the product measure⁴

$$\pi := P \otimes \tilde{P}$$

has the required marginals and

$$\mathbf{d}_r(P, \tilde{P})^r \leq \int_{\Omega} \int_{\tilde{\Omega}} \mathbf{d}(\omega, \tilde{\omega})^r P(d\omega) \tilde{P}(d\tilde{\omega}) < \infty.$$

Notice that if \mathbf{d} is inherited from ξ and the distance on \mathbb{R}^m is given by a norm $\|\cdot\|$, then $P \in \mathcal{P}_r(\Omega; \mathbf{d})$ iff $\int \|\xi(\omega)\|^r P(d\omega) < \infty$, i.e., if ξ has finite r -th moment.

Remark 2.7. A comprehensive and intensive discussion of the Wasserstein distance is provided in the books by Rachev and Rüschorf [107] and the book by Villani [137]. We shall use the properties that the infimum in (2.10) is actually attained, and $\mathbf{d}_r(\cdot, \cdot)$ turns out to be a metric on the space $\mathcal{P}_r(\Omega; \mathbf{d})$.

⁴ $(P \otimes \tilde{P})(A \times B) := P(A) \cdot \tilde{P}(B)$ defines a σ -additive measure due to the Hahn–Kolmogorov theorem.

Remark 2.8 (Remark on Naming). The terms in Definition 2.4 are not used consistently in the literature: in honor of G. Monge⁵ (cf. [79]) and Leonid Kantorovich⁶ (cf. [64]) the distance d_r is sometimes called *Monge–Kantorovich distance* of order r . The term *Vasershtein distance*⁷ appears the first time in Dobrushin [29]. d_2 is sometimes called *quadratic Wasserstein distance*. Moreover, the distance d_1 is also called *Kantorovich–Rubinstein distance* and sometimes denoted by $d_{KA} := d_1$. In Russian literature the term Kantorovich distance (cf. Vershik [136]) is used instead of Wasserstein distance.

The terms in Definition 2.4 apparently became accepted in recent years, particularly due to Villani’s before-mentioned book [137] and other authors. We follow this general trend, in particular we reserve the term Kantorovich distance for $d_{KA} = d_1$ ($r = 1$).

Notational Convenience. We are using the symbol d for the distance in the original space Ω , and the same symbol $d_r(\cdot, \cdot)$ with subscript r to account for the distance on probabilities in $\mathcal{P}_r(\Omega; d)$ induced by d . This is justified in view of the following proposition, which identifies (Ω, d) as a closed subspace of (\mathcal{P}_r, d_r) .

Proposition 2.9 (Embedding). *It holds that*

$$d_r(P, \delta_{\omega_0})^r = \int_{\Omega} d(\omega, \omega_0)^r P(d\omega),$$

and the mapping

$$\begin{aligned} i: (\Omega, d) &\rightarrow (\mathcal{P}_r(\Omega; d), d_r), \\ \omega &\mapsto \delta_{\omega} \end{aligned}$$

assigning to each point $\omega \in \Omega$ its point measure δ_{ω} (Dirac measure⁸) is an isometric embedding for all $1 \leq r < \infty$ ($(\Omega, d) \hookrightarrow \mathcal{P}_r(\Omega; d)$).

Proof. There is just one single measure with marginals P and δ_{ω_0} , which is the transport plan $\pi = P \otimes \delta_{\omega_0}$. Hence

$$d_r(P, \delta_{\omega_0})^r = \int_{\Omega} \int_{\Omega} d(\omega, \tilde{\omega})^r \delta_{\omega_0}(d\tilde{\omega}) P(d\omega) = \int_{\Omega} d(\omega, \omega_0)^r P(d\omega),$$

the first assertion.

⁵Gaspard Monge (1746–1818) investigated how to efficiently construct dugouts.

⁶L. Kantorovich was awarded the price in Economic Sciences in Memory of Alfred Nobel in 1975.

⁷In honor of Leonid N. Vaserteň.

⁸ $\delta_{\omega}(A) := \mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$ is the usual Dirac measure.

For the particular choice $P = \delta_{\tilde{\omega}_0}$ the latter formula simplifies to

$$\mathbf{d}_r(\delta_{\tilde{\omega}_0}, \delta_{\omega_0})^r = \int_{\Omega} \mathbf{d}(\omega, \omega_0)^r \delta_{\tilde{\omega}_0}(d\omega) = \mathbf{d}(\tilde{\omega}_0, \omega_0)^r,$$

and hence $\omega \mapsto \delta_\omega$ is an isometry. \square

Notice that if \mathbf{d} is inherited by ξ , then $\mathbf{d}_r(P, \delta_{\omega_0})^r = \int_{\Omega} \|\xi(\omega) - \xi(\omega_0)\|^r P(d\omega)$.

2.3 Elementary Properties of the Wasserstein Distance

The following properties of the Wasserstein distance \mathbf{d}_r will be employed frequently.

Lemma 2.10 (Monotonicity and Convexity).

- (i) If $r_1 \leq r_2$, then $\mathbf{d}_{r_1}(P, \tilde{P}) \leq \mathbf{d}_{r_2}(P, \tilde{P})$.
- (ii) The Wasserstein distance is r -convex⁹ in any of its components, that is for $0 \leq \lambda \leq 1$ it holds that

$$\mathbf{d}_r(P, (1-\lambda)P_0 + \lambda P_1)^r \leq (1-\lambda) \mathbf{d}_r(P, P_0)^r + \lambda \mathbf{d}_r(P, P_1)^r,$$

and

$$\begin{aligned} \mathbf{d}_r(P, (1-\lambda)P_0 + \lambda P_1) &\leq (1-\lambda)^{\frac{1}{r}} \mathbf{d}_r(P, P_0) + \lambda^{\frac{1}{r}} \mathbf{d}_r(P, P_1) \\ &\leq \max\{\lambda, 1-\lambda\}^{\frac{1}{r}-1} \cdot ((1-\lambda) \mathbf{d}_r(P, P_0) + \lambda \mathbf{d}_r(P, P_1)). \end{aligned} \quad (2.11)$$

- (iii) \mathbf{d}_r is a distance, it satisfies the triangle inequality $\mathbf{d}_r(P, \tilde{P}) \leq \mathbf{d}_r(P, \tilde{P}) + \mathbf{d}_r(\tilde{P}, \tilde{\tilde{P}})$.

Remark 2.11. Convexity in the traditional sense is actually achieved for the Kantorovich distance ($r = 1$), it follows from (2.11) that

$$\mathbf{d}_1(P, (1-\lambda)P_0 + \lambda P_1) \leq (1-\lambda) \mathbf{d}_1(P, P_0) + \lambda \mathbf{d}_1(P, P_1).$$

For the general Wasserstein distance ($r > 1$), however, a correction factor

$$1 \leq \max\{\lambda, 1-\lambda\}^{\frac{1}{r}-1} \leq 2^{\frac{r-1}{r}} < 2$$

has to be accepted in (2.11).

⁹For the notion of r -concavity (r -convexity) see Dentcheva [129].

Proof. Observe that $\frac{1}{r_2} + \frac{1}{\frac{r_2}{r_2 - r_1}} = 1$. By use of Hölder's inequality

$$\int d^{r_1} d\pi = \int d^{r_1} \cdot 1 d\pi \leq \left(\int d^{r_1 \frac{r_2}{r_1}} d\pi \right)^{\frac{r_1}{r_2}} \cdot \left(\int 1^{\frac{r_2}{r_2 - r_1}} d\pi \right)^{\frac{r_2 - r_1}{r_2}} = \left(\int d^{r_2} d\pi \right)^{\frac{r_1}{r_2}}.$$

Thus, $(\int d^{r_1} d\pi)^{\frac{1}{r_1}} \leq (\int d^{r_2} d\pi)^{\frac{1}{r_2}}$ for every measure π , which proves the first assertion.

As for the second let π_0 and π_1 be measures chosen with adequate marginals in such way that the infimum is attained,

$$d_r(P, P_0)^r = \int d(\omega, \tilde{\omega})^r \pi_0(d\omega, d\tilde{\omega}) \text{ and } d_r(P, P_1)^r = \int d(\omega, \tilde{\omega})^r \pi_1(d\omega, d\tilde{\omega}).$$

The probability measure $\pi_\lambda := (1 - \lambda)\pi_0 + \lambda\pi_1$ then has the marginals P and $P_\lambda := (1 - \lambda)P_0 + \lambda P_1$, and

$$\begin{aligned} d_r(P, (1 - \lambda)P_0 + \lambda P_1)^r &\leq \int d(\omega, \tilde{\omega})^r \pi_\lambda(d\omega, d\tilde{\omega}) \\ &= (1 - \lambda) \int d(\omega, \tilde{\omega})^r \pi_0(d\omega, d\tilde{\omega}) + \lambda \int d(\omega, \tilde{\omega})^r \pi_1(d\omega, d\tilde{\omega}) \\ &= (1 - \lambda) d_r(P, P_0)^r + \lambda d_r(P, P_1)^r. \end{aligned}$$

The assertion follows from monotonicity and concavity of $x \mapsto x^{\frac{1}{r}}$ and as $(x + y)^{\frac{1}{r}} \leq x^{\frac{1}{r}} + y^{\frac{1}{r}}$.

The other statements follow by employing Hölder's $L^1 - L^\infty$ inequality. For (iii) we refer to the proof involving the gluing lemma in Villani [137]. \square

Remark 2.12. To note an important consequence: all functions are continuous with respect to d_r , provided they are continuous with respect to $d_1 = d_{KA}$, the Kantorovich distance. A simple and useful example is provided by the following well-known lemma.

Lemma 2.13. *If the distance d is inherited from ξ and $\tilde{\xi}$ and based on a norm $\|\cdot\|$ (see (2.7)), then*

$$\left\| \mathbb{E}_P(\xi) - \mathbb{E}_{\tilde{P}}(\tilde{\xi}) \right\| \leq d_r(P, \tilde{P}) \quad (2.12)$$

for $r \geq 1$.

In an alternative notation, let P_1 (\tilde{P}_1 , resp.) be probability measure on \mathbb{R}^m (for instance $P_1 = P^\xi$, the image or pushforward measure of P) and let the point

$\mu_{P_1} := \mathbb{E}_{P_1}(\text{id}) = \int \xi P_1(d\xi)$ be the expectation (barycenter) of measure P_1 ¹⁰ (provided it exists) and the same for \tilde{P}_1 , then

$$\|\mu_{P_1} - \mu_{\tilde{P}_1}\| \leq d_r(P_1, \tilde{P}_1).$$

Proof. The proof for the Kantorovich distance ($r = 1$) is an application of Jensen's inequality as the norm is a convex function:

$$\begin{aligned} \left\| \mathbb{E}_P(\xi) - \mathbb{E}_{\tilde{P}}(\tilde{\xi}) \right\| &= \left\| \int \xi(\omega) P(d\omega) - \int \tilde{\xi}(\tilde{\omega}) \tilde{P}(d\tilde{\omega}) \right\| \\ &= \left\| \int (\xi - \tilde{\xi}) \pi(d\omega, d\tilde{\omega}) \right\| \leq \int \|\xi - \tilde{\xi}\| \pi(d\omega, d\tilde{\omega}). \end{aligned}$$

Taking the infimum over all measures π with appropriate marginals P and \tilde{P} gives the assertion, as $\left\| \mathbb{E}_P(\xi) - \mathbb{E}_{\tilde{P}}(\tilde{\xi}) \right\| \leq d_1(P, \tilde{P}) \leq d_r(P, \tilde{P})$. \square

Remark 2.14. Formula (2.12) gives rise to the interpretation, that particles have to be transported—on average—at least the distance of the barycenters $\mathbb{E}_P(\xi) - \mathbb{E}_{\tilde{P}}(\tilde{\xi})$.

2.3.1 The Wasserstein Distance on the Real Line

The Wasserstein distance for probability measures on the real line allows a closed form representation, which turns out to be useful in many situations. We cite the statement from Ambrosi et al. [3, Theorem 6.0.2], see also Vallander [135] and (2.3).

Theorem 2.15. *The Wasserstein distance of order $r \geq 1$ for measures P and \tilde{P} on the real line \mathbb{R} is*

$$d_r(P, \tilde{P})^r = \int_0^1 \left| G_P^{-1}(\alpha) - G_{\tilde{P}}^{-1}(\alpha) \right|^r d\alpha,$$

where $G_P(y) = P((-\infty, y])$ is the associated cumulative distribution function and $G_P^{-1}(\alpha) = \inf\{y : G_P(y) \geq \alpha\}$ its generalized inverse.

Example 2.16 (Normal Distribution). If $P = N(\mu, \sigma^2)$ and $\tilde{P} = N(\tilde{\mu}, \tilde{\sigma}^2)$ are normally distributed, then the explicit value for the Wasserstein distance of order $r = 2$ is

$$d_2(P, \tilde{P})^2 = (\mu - \tilde{\mu})^2 + (\sigma - \tilde{\sigma})^2,$$

¹⁰ $\text{id}(\xi) := \xi$ is the identity.

while the Kantorovich distance is bounded by

$$d_1(P, \tilde{P}) \leq |\mu - \tilde{\mu}| + \sqrt{\frac{2}{\pi}} |\sigma - \tilde{\sigma}|.$$

The statement follows by considering $G_P^{-1}(u) = \mu + \sigma \cdot \Phi^{-1}(u)$, where $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^u e^{-\frac{1}{2}v^2} dv$ is the cdf of the standard normal distribution. It follows from Theorem 2.15 that

$$\begin{aligned} d_2(P, \tilde{P})^2 &= \int_0^1 (\mu - \tilde{\mu} + (\sigma - \tilde{\sigma}) \Phi^{-1}(u))^2 du \\ &= (\mu - \tilde{\mu})^2 + 2(\mu - \tilde{\mu})(\sigma - \tilde{\sigma}) \int_0^1 \Phi^{-1}(u) du \\ &\quad + (\sigma - \tilde{\sigma})^2 \int_0^1 (\Phi^{-1}(u))^2 du \\ &= (\mu - \tilde{\mu})^2 + (\sigma - \tilde{\sigma})^2. \end{aligned}$$

Moreover,

$$\begin{aligned} d_1(P, \tilde{P}) &= \int_0^1 |\mu + \sigma \Phi^{-1}(u) - \tilde{\mu} - \tilde{\sigma} \Phi^{-1}(u)| du \\ &\leq |\mu - \tilde{\mu}| + |\sigma - \tilde{\sigma}| \int_0^1 |\Phi^{-1}(u)| du \end{aligned}$$

provides the second assertion, as $\int_0^1 |\Phi^{-1}(u)| du = \int_{-\infty}^{\infty} |u| \Phi'(u) du = \sqrt{\frac{2}{\pi}}$.

Example 2.17. Evidently, explicit expressions are also available for even integer orders, an example is

$$d_4(P, \tilde{P})^4 = (\mu - \tilde{\mu})^4 + 6(\mu - \tilde{\mu})^2(\sigma - \tilde{\sigma})^2 + 3(\sigma - \tilde{\sigma})^4,$$

etc.

A further, general upper bound is provided by the following example.

Example 2.18. For two real valued random variables $\xi \sim P$ and $\tilde{\xi} \sim \tilde{P}$ with finite second moments, means μ ($\tilde{\mu}$, resp.) and variances σ^2 ($\tilde{\sigma}^2$, resp.), it follows from the elementary expansion

$$\begin{aligned} (x - y)^2 &= (\mu - \tilde{\mu})^2 + (x - \mu)^2 + (y - \tilde{\mu})^2 \\ &\quad - 2(xy - \mu\tilde{\mu}) + 2\mu(x - \mu) + 2\tilde{\mu}(y - \tilde{\mu}), \end{aligned}$$

together with (2.12), that

$$(\mu - \tilde{\mu})^2 \leq \mathbf{d}_2(P, \tilde{P})^2 \leq (\mu - \tilde{\mu})^2 + \sigma^2 + \tilde{\sigma}^2, \quad (2.13)$$

because $\pi := P \otimes \tilde{P}$ is a feasible bivariate measure. The upper bound (2.13) is rather conservative, although attained if one of the measures is a Dirac measure.

2.4 Alternative Distances as Basis for the Wasserstein Distance

2.4.1 The Role of the Distance on the Underlying Space

To every metric \mathbf{d} on \mathbb{R}^m there corresponds a Wasserstein distance according to Definition 2.5. A special situation occurs for the discrete metric

$$\mathbf{d}_0(u, v) := \begin{cases} 0 & \text{if } u = v \\ 1 & \text{if } u \neq v. \end{cases}$$

The set of all Lipschitz functions with respect to the discrete metric \mathbf{d}_0 coincides with the set of all measurable functions h such that $0 \leq h \leq 1$ or its translates. Consequently the pertaining Kantorovich distance coincides with the variational distance (see (2.2))

$$\mathbf{d}_1(P, \tilde{P} | d_0) = \mathbf{d}_V(P, \tilde{P}),$$

(we write $\mathbf{d}_r(P, \tilde{P} | d_0)$ to emphasize the dependency on the metric d_0 of the basic space).

2.4.2 Transformation of the Axis, and Fortet–Mourier Distances

Alternative metrics on \mathbb{R}^1 are obtained by a nonlinear transform of the axis. Let χ be any bijective, monotone transformation, which maps \mathbb{R} into \mathbb{R} . Then $\mathbf{d}_\chi(u, v) := |\chi(u) - \chi(v)|$ defines a new metric on \mathbb{R}^1 . Notice that the family of functions, which are Lipschitz with respect to the distance \mathbf{d}_χ and to the Euclidean distance $|u - v|$, may be quite different.

To establish a relation between the Fortet–Mourier distance and a transformation on the real line we consider the bijective transformation (for $q > 0$)

$$\chi_q(u) = \begin{cases} u & \text{if } |u| \leq 1 \\ |u|^q \cdot \text{sign}(u) & \text{otherwise,} \end{cases} \quad (2.14)$$

which introduces the metric $d_{\chi_q}(u, v) = |\chi_q(u) - \chi_q(v)|$. On bounded intervals the distances $\chi_1(u, v) = |u - v|$ and d_{χ_q} are equivalent, since

$$|u - v| \leq |\chi_q(u) - \chi_q(v)| \leq q \cdot K^{q-1} |u - v| \quad \text{whenever } |u| \leq K \text{ and } |v| \leq K.$$

Denote by $d_1(\cdot, \cdot | d_{\chi_q})$ the Kantorovich distance based on the distance d_{χ_q} ,

$$d_1(P, \tilde{P} | d_{\chi_q}) = \sup \left\{ \int h \, dP - \int h \, d\tilde{P} : |h(u) - h(v)| \leq d_{\chi_q}(u, v) \right\}.$$

Notice that $d_{\chi_{q'}}(u, v) \leq d_{\chi_q}(u, v)$ for $q' < q$ and therefore

$$d_1(P, \tilde{P} | d_{\chi_{q'}}) \leq d_1(P, \tilde{P} | d_{\chi_q}) \quad \text{for } q' < q. \quad (2.15)$$

Let P^{χ_q} be the image measure of P under χ_q , that is $P^{\chi_q}(A) = P(\chi_{1/q}(A))$, as $\chi_q^{-1}(u) = \chi_{1/q}(u)$, and note that P^{χ_q} has distribution function

$$G_{P^{\chi_q}}(x) = G_P(\chi_{1/q}(x)),$$

where G_P is the distribution function of P . This leads to the identity

$$d_1(P, \tilde{P} | d_{\chi_q}) = d_1(P^{\chi_q}, \tilde{P}^{\chi_q} | d_{\chi_1}),$$

as $d_{\chi_1}(u, v) = |u - v|$.

To relate the Fortet–Mourier distance d_{M_q} to the distance $d_1(\cdot, \cdot | d_{\chi_q})$ we show first the relations

$$L_q(h \circ \chi_q) \leq q \cdot L_1(h) \quad (2.16)$$

and

$$L_1(h \circ \chi_{1/q}) \leq 2 \cdot L_q(h) \quad (2.17)$$

for the Lipschitz constants of order q defined in (2.5).

Indeed, if $L_1(h) < \infty$, then

$$\begin{aligned} |h(\chi_q(u)) - h(\chi_q(v))| &\leq L_1(h) \cdot |\chi_q(u) - \chi_q(v)| \\ &\leq L_1(h) \cdot q \cdot \max\{1, |u|^{q-1}, |v|^{q-1}\} \cdot |u - v|, \end{aligned}$$

which implies (2.16). On the other hand, if $L_q(h) < \infty$, then (2.17) holds by

$$\begin{aligned} |h(\chi_{1/q}(u)) - h(\chi_{1/q}(v))| &\leq L_q(h) \cdot \max\{1, |\chi_{1/q}(u)|^{q-1}, |\chi_{1/q}(v)|^{q-1}\} \\ &\quad \cdot |\chi_{1/q}(u) - \chi_{1/q}(v)| \\ &\leq 2 \cdot L_q(h) \cdot |u - v|, \end{aligned}$$

where we have used that

$$\max\{1, |\chi_{1/q}(u)|^{q-1}, |\chi_{1/q}(v)|^{q-1}\} \cdot \frac{|\chi_{1/q}(u) - \chi_{1/q}(v)|}{|u - v|} \leq 2. \quad (2.18)$$

This latter inequality is clear if $|v| \leq |u| \leq 1$. If $|v| \leq |u|$ and $|u| > 1$, then the left-hand side of (2.18) is bounded by 2.

As a consequence of (2.16) and (2.17) the relations

$$\begin{aligned} \frac{1}{q} \mathbf{d}_1(P, \tilde{P} | \mathbf{d}_{\chi_q}) &= \frac{1}{q} \mathbf{d}_1(G_P \circ \chi_{1/q}, G_{\tilde{P}} \circ \chi_{1/q}) \\ &\leq \mathbf{d}_{\text{FM}_q}(P, \tilde{P}) \\ &\leq 2 \mathbf{d}_1(G_P \circ \chi_{1/q}, G_{\tilde{P}} \circ \chi_{1/q}) = 2 \mathbf{d}_1(P, \tilde{P} | \mathbf{d}_{\chi_q}) \quad (2.19) \end{aligned}$$

are obtained.

One thus sees that the Fortet–Mourier distance of order q and the Kantorovich distance (i.e., the Fortet–Mourier distance of order 1) with the alternative metric \mathbf{d}_{χ_q} are topologically equivalent.

A further relation can be based on the function $\psi_r(u) = |u|^r \cdot \text{sign}(u)$ and the distance $\mathbf{d}_{\psi_r}(u, v) = |\psi_r(u) - \psi_r(v)|$. Notice that by an easy geometric consideration, for $r \geq 1$,

$$|\psi_r(u) - \psi_r(v)| \geq 2 \left(\frac{|u - v|}{2} \right)^r$$

and therefore $|u - v|^r \leq 2^{r-1} |\psi(u) - \psi(v)|$, which implies that

$$\mathbf{d}_r(P, \tilde{P})^r \leq 2^{r-1} \cdot \mathbf{d}_1(P, \tilde{P} | \psi_r). \quad (2.20)$$

Lemma 2.19. *On the set of probability distributions, which have uniformly bounded r -th moments, the topologies generated by the distances \mathbf{d}_r and $\mathbf{d}_1(\cdot, \cdot | \mathbf{d}_{\psi_r})$ are equivalent.*

Proof. Inequality (2.20) shows that $\mathbf{d}_1(\cdot, \cdot | \psi_r)$ is finer than \mathbf{d}_r . For the inverse relation, let $\xi \sim P$ and $\tilde{\xi} \sim \tilde{P}$. Notice that the Lipschitz constant of order r of ψ_r is $L_r(\psi_r) = r$ and therefore

$$|\psi_r(\xi) - \psi_r(\tilde{\xi})| \leq r \cdot |\xi - \tilde{\xi}| \cdot \max \left\{ 1, |\xi|^{r-1}, |\tilde{\xi}|^{r-1} \right\}.$$

Using Hölder's inequality, we get¹¹

$$\begin{aligned} \mathbb{E}|\psi_r(\xi) - \psi_r(\tilde{\xi})| &\leq r \mathbb{E}^{1/r} |\xi - \tilde{\xi}|^r \cdot \mathbb{E}^{\frac{r-1}{r}} \left[(1 + |\xi|^{r-1} + |\tilde{\xi}|^{r-1})^{\frac{r}{r-1}} \right] \\ &\leq r \mathbb{E}^{1/r} |\xi - \tilde{\xi}|^r \cdot \left[1 + \mathbb{E}^{\frac{r-1}{r}} (|\xi|^r) + \mathbb{E}^{\frac{r-1}{r}} (|\tilde{\xi}|^r) \right] \end{aligned}$$

and consequently considering the minima with respect to the joint distribution of ξ and $\tilde{\xi}$ one gets

$$\mathbf{d}_1(P, \tilde{P} \mid \psi_r) \leq r \cdot \mathbf{d}_r(P, \tilde{P}) \left(1 + \mathbb{E}^{\frac{r-1}{r}} (|\xi|^r) + \mathbb{E}^{\frac{r-1}{r}} (|\tilde{\xi}|^r) \right),$$

which shows that \mathbf{d}_r is finer than $\mathbf{d}_1(\cdot, \cdot \mid \psi_r)$. \square

2.5 Estimates Involving the Wasserstein Distance

In this section we ask the question: How close are some important statistical parameters, if the Wasserstein distances are small? Suppose that a probability distribution P on \mathbb{R}^m and some (typically discrete) approximation \tilde{P} , which is close to P in Wasserstein distance, are given. One may ask the following questions:

- Do P and \tilde{P} have a similar mean?
- Do P and \tilde{P} have a similar variance?
- If P and \tilde{P} are multidimensional, do P and \tilde{P} have a similar covariance matrix?
- Are the higher moments of P and \tilde{P} similar?

Precise answers to these questions are given below in Proposition 2.20. Some authors argue that a close approximation \tilde{P} to P should have at least the same first and second moments. Since we aim at approximating the distribution as a whole, there is not much reason in trying to match some specific moments (as it is done by *moment matching*, cf. Example 2.2 and (2.1)). In some applications one might be interested in the median and mean matching would not help. Also matching some Pearson correlation (product-moment correlation) would not help in matching Spearman's or Kendall's correlation.

If $\xi \sim P$ and $\tilde{\xi} \sim \tilde{P}$, it is evident that for functions h with Lipschitz constant L the Wasserstein distance controls the distance of their integrals,

$$\left| \mathbb{E}[h(\xi)] - \mathbb{E}[h(\tilde{\xi})] \right| \leq L \cdot \mathbf{d}_1(P, \tilde{P}).$$

¹¹We use the shorthand notation $\mathbb{E}^p[\xi]$ for $(\mathbb{E}[\xi])^p$.

The following proposition collects results involving distances of probability measures and Lipschitz constants.

Proposition 2.20 (Bounds Involving Lipschitz Constants). *Assume that $\xi \sim P$ and $\tilde{\xi} \sim \tilde{P}$. Then*

- (i) $|\mathbb{E}\xi - \mathbb{E}\tilde{\xi}| \leq d_1(P, \tilde{P})$,
- (ii) $|\mathbb{E}|\xi| - \mathbb{E}|\tilde{\xi}|| \leq d_1(P, \tilde{P})$,
- (iii) $|\mathbb{E}(\xi - a)_+ - \mathbb{E}(\tilde{\xi} - a)_+| \leq d_1(P, \tilde{P})$,
- (iv) $|\mathbb{E}(\xi^q) - \mathbb{E}(\tilde{\xi}^q)| \leq q \cdot d_{\text{FM}_q}(P, \tilde{P})$ for integer q and
- (v) $|\mathbb{E}(|\xi|^q) - \mathbb{E}(|\tilde{\xi}|^q)| \leq q \cdot d_{\text{FM}_q}(P, \tilde{P})$.

Proof. The functions $u \mapsto u$, $u \mapsto |u|$ and $u \mapsto (u - a)_+$ are Lipschitz continuous (with Lipschitz constant 1). For the proof of (iv) and (v) recall the definition of the Fortet–Mourier distance $d_{\text{FM}_q}(P, \tilde{P})$ in (2.4) and use the fact that the Lipschitz constant of order q (see (2.5)) of $x \mapsto x^q$ is $L_q = q$. The same is true for the function $x \mapsto |x|^q$. \square

The following proposition collects examples to demonstrate how the Wasserstein distance controls also higher moments, provided that they exist.

Proposition 2.21 (Wasserstein Distance Controls All Moments). *Assume that $\xi \sim P$ and $\tilde{\xi} \sim \tilde{P}$. Then*

- (i) $|\mathbb{E}|\xi|^p - \mathbb{E}|\tilde{\xi}|^p| \leq p \cdot d_r(P, \tilde{P}) \cdot \max \left\{ \mathbb{E}^{\frac{r-1}{r}} \left[|\xi|^{r \cdot \frac{p-1}{r-1}} \right], \mathbb{E}^{\frac{r-1}{r}} \left[|\tilde{\xi}|^{r \cdot \frac{p-1}{r-1}} \right] \right\}$,
- (ii) $|\mathbb{E}(\xi^p) - \mathbb{E}(\tilde{\xi}^p)| \leq p \cdot d_r(P, \tilde{P}) \cdot \left\{ \mathbb{E}^{\frac{r-1}{r}} \left[|\xi|^{r \cdot \frac{p-1}{r-1}} \right] + \mathbb{E}^{\frac{r-1}{r}} \left[|\tilde{\xi}|^{r \cdot \frac{p-1}{r-1}} \right] \right\}$ for p an integer,
- (iii) $|\mathbb{E}\xi^2 - \mathbb{E}\tilde{\xi}^2| \leq 2 \cdot d_2(P, \tilde{P}) \cdot \max \left\{ \mathbb{E}^{\frac{1}{2}} \left[[\xi]^2 \right], \mathbb{E}^{\frac{1}{2}} \left[[\tilde{\xi}]^2 \right] \right\}$,
- (iv) $|\mathbb{E}|\xi|^r - \mathbb{E}|\tilde{\xi}|^r| \leq r \cdot d_r(P, \tilde{P}) \cdot \max \left\{ \mathbb{E}^{\frac{r-1}{r}} \left[|\xi|^r \right], \mathbb{E}^{\frac{r-1}{r}} \left[|\tilde{\xi}|^r \right] \right\}$ and
- (v) $|\mathbb{E}|\xi|^p - \mathbb{E}|\tilde{\xi}|^p| \leq p \cdot d_2(P, \tilde{P}) \cdot \max \left\{ \mathbb{E}^{\frac{1}{2}} \left[|\xi|^{2(p-1)} \right], \mathbb{E}^{\frac{1}{2}} \left[|\tilde{\xi}|^{2(p-1)} \right] \right\}$,

where $p \geq 1$ and $r > 1$.

Proof. By convexity of the function $x \mapsto |x|^p$ for $p \geq 1$ it holds that

$$|\tilde{x}|^p \geq |x|^p + p \cdot \text{sign}(x) |x|^{p-1} (\tilde{x} - x),$$

and consequently

$$|\xi|^p - |\tilde{\xi}|^p \leq p \cdot \text{sign}(\xi) |\xi|^{p-1} (\xi - \tilde{\xi}) \leq p |\xi - \tilde{\xi}| |\xi|^{p-1}.$$

Taking expectations with respect to π , where π has marginals P and \tilde{P} , and employing Hölder's inequality for the conjugate parameters $\frac{1}{r} + \frac{1}{r'} = 1$ (i.e., $r' = \frac{r}{r-1}$) reveals that

$$\mathbb{E}|\xi|^p - \mathbb{E}|\tilde{\xi}|^p \leq p \cdot \left\| \xi - \tilde{\xi} \right\|_r \cdot \left\| |\xi|^{p-1} \right\|_{r'} = p \cdot \left\| \xi - \tilde{\xi} \right\|_r \cdot \left\| \xi \right\|_{r, \frac{p-1}{r-1}}^{p-1}.$$

Taking the infimum with of all bivariate probability measures π with marginals P and \tilde{P} it follows that

$$\mathbb{E}|\xi|^p - \mathbb{E}|\tilde{\xi}|^p \leq p \cdot d_r(P, \tilde{P}) \cdot \mathbb{E}^{\frac{p-1}{r}} \left[|\xi|^{r, \frac{p-1}{r-1}} \right].$$

The assertion (i) follows now for general $r > 1$ and $p > 1$ by interchanging ξ and $\tilde{\xi}$.

The second inequality has only to be proved for odd p since for even p it is a consequence of (i). Using the monotonicity of the odd function $x \mapsto p \cdot x^{p-1}$ one gets

$$\xi^p - \tilde{\xi}^p \leq p \cdot \left(|\xi|^{p-1} + |\tilde{\xi}|^{p-1} \right) \left| \xi - \tilde{\xi} \right|,$$

and in analogy to the proof of (i) the inequality (ii) follows.

The other assertions can be derived from (i) as special cases ($r = p = 2$, $r = p$, etc.). \square

Further inequalities of the type addressed in Proposition 2.20 and in Proposition 2.21 can be derived, if one considers the Wasserstein norms with alternative distances on \mathbb{R} . Again, let $\xi \sim P$ and $\tilde{\xi} \sim \tilde{P}$. Using the functions (see (2.14))

$$\chi_q(u) = \begin{cases} u & \text{if } |u| \leq 1 \\ |u|^q \operatorname{sign}(u) & \text{otherwise} \end{cases}$$

and noticing (2.15) one gets that for $q' \leq q$

$$\mathbb{E} \left[\max \left\{ |\xi|, |\xi|^{q'} \right\} \right] \leq d_1(P, \tilde{P} | d_{\chi_q}).$$

Also, using the Fortet–Mourier metric (2.4) we get the basic inequality

$$\left| \mathbb{E}[h(\xi)] - \mathbb{E}[h(\tilde{\xi})] \right| \leq L_q(h) \cdot d_{\text{FM}_q}(P, \tilde{P}),$$

where L_q is the Lipschitz constant of order q (see (2.5)). A special case is

$$\left| \mathbb{E}|\xi|^{q'} - \mathbb{E}|\tilde{\xi}|^{q'} \right| \leq q \cdot d_{\text{FM}_q}(P, \tilde{P}) \leq 2q \cdot d_1(P, \tilde{P} | d_{\chi_q})$$

for $q' < q$, since the Lipschitz constant of order q of $x \mapsto |x|^q$ equals q (cf. also inequality (2.19)).

Example 2.22 (Approximation of a Bivariate Normal Distribution). As an illustration consider the best approximation of a normal distribution

$$N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \right) \quad (2.21)$$

by a discrete distribution located at s points in \mathbb{R}^2 . Notice that we may generate random variates from this distribution by

$$\xi_1 = Z_1 + Z_2,$$

$$\xi_2 = Z_2,$$

where Z_1 and Z_2 are independent standard normals. We approximate these distributions by discrete ones, sitting on $s = 5, 7, 9$, and 25 points. The discrete approximations are displayed in Fig. 2.2, where the little circles around each point

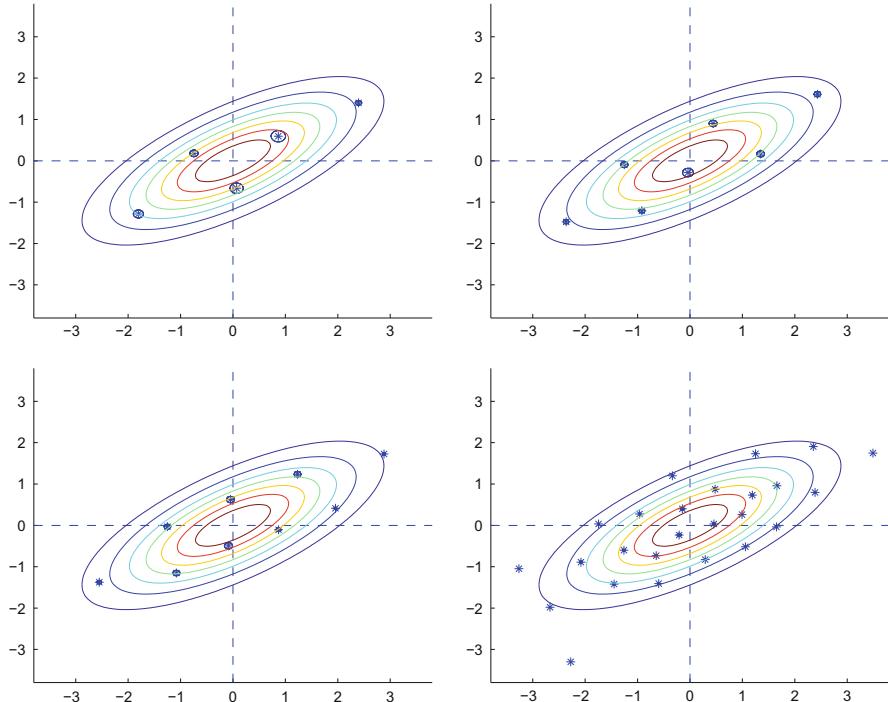


Fig. 2.2 Discrete approximations of a two-dimensional normal distribution with 5, 7, 9, and 25 points. Some statistical parameters of these approximations are given in Table 2.1

Table 2.1 Approximations of a bivariate normal distribution (2.21) by 5, 7, 9, and 25 points. Their Wasserstein distance (of order 1) is shown in the first column

	Distance	$\mathbb{E}(\xi_1)$	$\mathbb{E}(\xi_2)$	$\text{Var}(\xi_1)$	$\text{Var}(\xi_2)$	Cov	$\mathbb{E}(\xi_1^3)$	$\mathbb{E}(\xi_1^4)$
True value		0.	0.	2.	1.	1.	0.	12.
5 points	0.587	0.056	-0.064	1.59	0.76	0.96	0.54	6.06
7 points	0.454	0.043	0.004	1.95	0.87	1.105	0.47	8.63
9 points	0.359	-0.03	-0.009	1.90	0.83	1.02	0.22	9.52
25 points	0.147	0.01	0.01	2.04	0.99	1.07	0.41	11.6

(a) Approximation quality of selected moments. Cov is the covariance between ξ_1 and ξ_2

	Distance	Med(ξ_1)	Med(ξ_2)	$\mathbb{E} \xi_1 $	$\mathbb{E} \xi_2 $	Spm	$P(\xi_1 > \xi_2)$
True value		0.	0.	1.128	0.797	0.695	0.50
5 points	0.587	0.07	0.18	1.005	0.773	0.88	0.642
7 points	0.454	-0.03	0.094	1.11	0.73	0.83	0.60
9 points	0.359	-0.08	-0.03	1.07	0.75	0.76	0.502
25 points	0.147	-0.14	0.03	1.14	0.81	0.76	0.53

(b) Continuation of Table 2.1a: here the medians, the first absolute moments, the Spearman correlation coefficient (Spm) as well as the probability of a particular event are shown

symbolize the respective probability mass (these approximations were found by the Stochastic Approximation (SA) Algorithm 4.5, which is explained in Chap. 4). The results for comparison are collected in Table 2.1.

It is not claimed that these approximations are optimal, however, they are good ones. As one can see, they approximate the true distribution in many aspects, not just for the first two or four moments. While it is not possible to derive from closeness with respect to some moments the closeness with respect to other aspects, the closeness in the Wasserstein distance implies closeness for moments and other statistics in a natural way (cf. also Example 2.2).

2.6 Approximations in the Wasserstein Metric

This subsection provides the foundations for approximations of probability measures by probability measures with finite support. This is quite relevant, because only probability measures with finite support are eligible for numerical computations and algorithmic treatment.

Suppose that the supports of all considered probabilities are contained in some closed set $\Xi \subseteq \mathbb{R}^m$, which is endowed with some metric d . The elements of Ξ are denoted by ξ (these are here points and not random variables). We discuss the important and necessary theorems. The precise proofs of some results of this subsection are beyond the scope of this book, we give the references instead.

The crucial tool to identify the topology induced by the metric d_r with the topology of weak* convergence is the uniform tightness condition (2.22) below.

Theorem 2.23 (Wasserstein Metricizes the Weak* Topology). *Let $(P_n)_{n \geq 1}$ be a sequence of measures in $\mathcal{P}_r(\Xi)$, and let $P \in \mathcal{P}_r(\Xi)$. Then the following are equivalent:*

- (i) $d_r(P_n, P) \xrightarrow{n \rightarrow \infty} 0$,
- (ii) $P_n \xrightarrow{n \rightarrow \infty} P$ in weak* sense, and P_n satisfies the following uniform tightness condition: for some (and thus any) $\xi_0 \in \Omega$,

$$\limsup_{n \rightarrow \infty} \int_{\{d(\xi_0, \xi) \geq R\}} d(\xi_0, \xi)^r P_n(d\xi) \xrightarrow{R \rightarrow \infty} 0. \quad (2.22)$$

Proof. For a proof we refer to Theorem 7.12 in Villani [137]. \square

Remark 2.24. One may always replace the metric d by the uniformly bounded distance $d'(\xi, \tilde{\xi}) := \frac{d(\xi, \tilde{\xi})}{1+d(\xi, \tilde{\xi})}$ or $d'(\xi, \tilde{\xi}) := \min\{1, d(\xi, \tilde{\xi})\}$ without changing the topology of Ξ . In this situation, however, the uniform tightness condition (2.22) is trivial, and d' thus metricizes weak* convergence on the whole of $\mathcal{P}_r(\Xi)$.

The following theorem is essential for our intentions to approximate probability measures by measures with finite support.

Theorem 2.25. *If (Ξ, d) is separable, then $(\mathcal{P}_r(\Xi), d_r)$ is separable and all measures $\sum_{\xi \in \tilde{\Xi}} P_\xi \cdot \delta_\xi$ with finite support $\tilde{\Xi} \subset \Xi$ ($P_\xi \geq 0$ and $\sum_{\xi \in \tilde{\Xi}} P_\xi = 1$) are dense.*

Proof. A proof by elementary means is contained in Bolley [13]. Initial proofs of the statement, however, involve the weaker Prohorov distance and deep results of Kolmogorov; cf. Ambrosi et al. [3]. \square

To complete the essential characteristics we mention that the space $(\mathcal{P}_r(\Xi; d), d_r)$ is not only separable and metrizable, but also complete, hence a Polish space.

Theorem 2.26. *Let (Ξ, d) be a Polish space, then $(\mathcal{P}_r(\Xi; d), d_r)$ is a Polish space again.*

Proof. The space is metrizable and separability is established by Theorem 2.25. Completeness is proved in Bolley [13]. \square

2.7 The Wasserstein Distance in a Discrete Framework

In many applications and in implementations the measures considered are discrete measures (measures with finite support) of the form $P = \sum_{i=1}^n P_i \delta_{\xi_i}$ (where $P_i \geq 0$, $\sum_{i=1}^n P_i = 1$ and the support $\{\xi_i : i = 1, 2, \dots, n\} \subset \Xi$ is finite).

Given two discrete measures $P = \sum_{i=1}^n P_i \delta_{\xi_i}$ and $\tilde{P} = \sum_{j=1}^{\tilde{n}} \tilde{P}_j \delta_{\tilde{\xi}_j}$ the computation of the Wasserstein distance (2.10) corresponds to solving the linear program

$$\underset{(\text{in } \pi)}{\text{minimize}} \quad \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \quad (2.23)$$

$$\text{subject to } \sum_{j=1}^{\tilde{n}} \pi_{i,j} = P_i \quad (i = 1, 2, \dots, n),$$

$$\sum_{i=1}^n \pi_{i,j} = \tilde{P}_j \quad (j = 1, 2, \dots, \tilde{n}),$$

$$\pi_{i,j} \geq 0, \quad (2.24)$$

where $d_{i,j} = d(\xi_i, \tilde{\xi}_j)$ is an $n \times \tilde{n}$ -matrix carrying the distances. The $n \times \tilde{n}$ -matrix $\pi_{i,j}$ in (2.23) corresponds to the bivariate probability measure

$$\pi = \sum_{i,j} \pi_{i,j} \cdot \delta_{(\xi_i, \tilde{\xi}_j)}$$

on the product $\Xi \times \tilde{\Xi}$; π is a probability measure as $\pi_{i,j} \geq 0$ and $\sum_{i,j} \pi_{i,j} = \sum_i \sum_j \pi_{i,j} = \sum_i P_i = 1$.

Figure 2.3 exhibits the structure of this linear program (2.23), where the matrix can be written in the form

$$\begin{pmatrix} \mathbb{1}_{\tilde{n}} \otimes I_n \\ I_{\tilde{n}} \otimes \mathbb{1}_n \end{pmatrix}$$

$$n \begin{Bmatrix} \left(\begin{array}{cccccc} 1 & 0 & 1 & 0 & 1 & 0 \\ & \ddots & & \ddots & & \ddots \\ 0 & 1 & 0 & 1 & 0 & 1 \end{array} \right) \\ \vdots \\ \left(\begin{array}{ccc} 1 \cdots 1 & 0 \cdots 0 & 0 \cdots 0 \\ 0 \cdots 0 & 1 \cdots 1 & \ddots & 0 \cdots 0 \\ \underbrace{0 \cdots 0}_{n} & \underbrace{0 \cdots 0}_{n} & \underbrace{1 \cdots 1}_{n} \end{array} \right) \end{Bmatrix} \begin{pmatrix} \pi_{1,1} \\ \vdots \\ \pi_{n,1} \\ \pi_{1,2} \\ \vdots \\ \pi_{n,\tilde{n}} \end{pmatrix} = \begin{pmatrix} P_1 \\ \vdots \\ P_n \\ \tilde{P}_1 \\ \vdots \\ \tilde{P}_{\tilde{n}} \end{pmatrix}$$

Fig. 2.3 Structure of the linear constraints of the linear program (2.23). The $(n + \tilde{n}) \times (n \cdot \tilde{n})$ matrix is totally unimodular

(\otimes denotes the Kronecker product, $\mathbb{1}_n = (\underbrace{1, 1, \dots, 1}_{n \text{ times}})$ and I_n is the $n \times n$ -identity matrix). From this figure it becomes evident that the constraints are *linearly dependent*, because the sum of the first n lines equals the sum of the following \tilde{n} lines. As a consequence, one of all $n + \tilde{n}$ constraints in (2.23) can be removed. For efficiency reasons in numerical implementations a line *should* be removed for most numerical solvers.

It follows from complementary slackness conditions of linear programs that the optimal transport plan π in (2.23) is sparse, it has at most $n + \tilde{n} - 1$ nonzero entries, because (2.23) has not more than $n + \tilde{n} - 1$ linearly independent equality constraints.

Remark 2.27 (Transport Plans and Their Relation to Bipartite Graphs). One may define the bipartite graph $G = (U \cup V, E)$ with distinct nodes

$$U = \{\xi_i : i = 1, \dots, n\} \text{ and } V = \{\tilde{\xi}_j : j = 1, \dots, \tilde{n}\}$$

and vertices $E = \{(\xi_i, \tilde{\xi}_j) : \pi_{i,j} > 0, i = 1, \dots, n, j = 1, \dots, \tilde{n}\}$. The linear constraints in (2.23) correspond to the incidence matrix of this graph G , which is a *totally unimodular* matrix (i.e., every square non-singular submatrix is invertible over the integers, cf. Hoffman and Krukskal [60]). It follows from Cramer's rule that each entry of the matrix π has the specific form

$$\pi_{i,j} = \sum_{k=1}^n \epsilon_{i,j}^k P_k + \sum_{\ell=1}^{\tilde{n}} \tilde{\epsilon}_{i,j}^\ell \tilde{P}_\ell,$$

where $\epsilon_{i,j}^k, \tilde{\epsilon}_{i,j}^\ell \in \{-1, 0, 1\}$.

2.8 Duality for the Wasserstein Metric

The linear program (2.23) to compute $d_r(P, \tilde{P})$ naturally—as any linear program—has a dual linear program. It is given by

$$\begin{aligned} & \underset{(\lambda, \mu)}{\text{maximize}} \quad \sum_{i=1}^n P_i \lambda_i + \sum_{j=1}^{\tilde{n}} \tilde{P}_j \mu_j \\ & \text{(in } \lambda, \mu \text{)} \end{aligned} \tag{2.25}$$

$$\text{subject to } \lambda_i + \mu_j \leq d_{i,j}^r \quad \text{for all } i = 1, \dots, n \text{ and } j = 1, \dots, \tilde{n}. \tag{2.26}$$

By the vanishing duality gap of the primal (2.23) and its dual (2.25) it follows that

$$\sum_{i=1}^n \sum_{j=1}^{\tilde{n}} \pi_{i,j} d_{i,j}^r \leq \sum_{i=1}^n P_i \lambda_i + \sum_{j=1}^{\tilde{n}} \tilde{P}_j \mu_j = \sum_{i=1}^n \sum_{j=1}^{\tilde{n}} \pi_{i,j} (\lambda_i + \mu_j) \leq \sum_{i=1}^n \sum_{j=1}^{\tilde{n}} \pi_{i,j} d_{i,j}^r,$$

from which further follows, by (2.24) and (2.26), that

$$\pi_{i,j} (\lambda_i + \mu_j) = \pi_{i,j} \cdot d_{i,j}^r,$$

which is the complementary slackness condition.

Recalling the fact that $P = \sum_{i=1}^n P_i \delta_{\xi_i}$ ($\tilde{P} = \sum_{j=1}^{\tilde{n}} \tilde{P}_j \delta_{\tilde{\xi}_j}$, resp.) one may extend the dual variables

$$\lambda(\xi) := \begin{cases} \lambda_i & \text{if } \xi = \xi_i \\ -\infty & \text{else} \end{cases} \quad \text{and} \quad \mu(\tilde{\xi}) := \begin{cases} \mu_j & \text{if } \tilde{\xi} = \tilde{\xi}_j \\ -\infty & \text{else.} \end{cases}$$

Then the dual program (2.25) can be rewritten as

$$\begin{aligned} & \text{maximize } (\lambda, \mu) \mathbb{E}_P \lambda + \mathbb{E}_{\tilde{P}} \mu \\ & \text{subject to } \lambda(\xi) + \mu(\tilde{\xi}) \leq d(\xi, \tilde{\xi})^r \text{ for all } \xi \in \Xi \text{ and } \tilde{\xi} \in \tilde{\Xi}, \end{aligned} \tag{2.27}$$

and the complementary slackness reads

$$\pi \left(\left\{ (\xi, \tilde{\xi}) : \lambda(\xi) + \mu(\tilde{\xi}) = d(\xi, \tilde{\xi})^r \right\} \right) = 1.$$

This means that

$$\lambda(\xi) + \mu(\tilde{\xi}) = d(\xi, \tilde{\xi})^r \quad \pi \text{ almost everywhere,}$$

the inequality in (2.27) is thus replaced by equality on the support set of the optimal measure π .

A pair (λ, μ) of feasible dual variables can moreover be replaced by (λ, λ^*) or (μ^*, μ) , where

$$\lambda^*(\tilde{\xi}) := \inf_{\xi \in \Xi} d(\xi, \tilde{\xi})^r - \lambda(\xi)$$

and

$$\mu^*(\xi) := \inf_{\tilde{\xi} \in \tilde{\Xi}} d(\xi, \tilde{\xi})^r - \mu(\tilde{\xi}),$$

because

$$\lambda(\xi) + \mu\left(\tilde{\xi}\right) \leq \lambda(\xi) + \lambda^*\left(\tilde{\xi}\right) \leq d\left(\xi, \tilde{\xi}\right)^r$$

and

$$\lambda(\xi) + \mu\left(\tilde{\xi}\right) \leq \mu^*(\xi) + \mu\left(\tilde{\xi}\right) \leq d\left(\xi, \tilde{\xi}\right)^r.$$

For an arbitrary function λ the pair (λ, λ^*) is feasible. By the same reasoning, given μ , the pair (μ^*, μ) is feasible. This gives an improved objective, as

$$\mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\mu) \leq \mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\lambda^*) \leq d_r(P, \tilde{P})^r, \quad (2.28)$$

and analogously for the pair (μ, μ^*) .

Rapid Computation of the Wasserstein Distance. The cascading property (2.28) can be exploited in algorithms to quickly compute the Wasserstein distance of discrete probability measures. By duality the objective of both problems,

$$\text{maximize (in } \lambda) \mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\lambda^*) \quad \text{and} \quad \text{maximize (in } \mu) \mathbb{E}_P(\mu^*) + \mathbb{E}_{\tilde{P}}(\mu), \quad (2.29)$$

is $d_r(P, \tilde{P})^r$, but the dimension of the vector λ (or μ) in (2.29) is much smaller than the dimension of the matrix π in the primal (2.23). The problems (2.29) are unconstrained, nonlinear, and the objectives

$$\lambda \mapsto \mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\lambda^*) \quad \text{and} \quad \mu \mapsto \mathbb{E}_P(\mu^*) + \mathbb{E}_{\tilde{P}}(\mu)$$

are moreover concave. In addition a subdifferential (an element of the subgradient) of the objective with respect to λ and μ is available, as

$$\frac{\partial}{\partial \lambda_i} \mathbb{E}_P(\lambda) + \mathbb{E}_{\tilde{P}}(\lambda^*) = P_i - \tilde{P}_j \quad \text{and} \quad \frac{\partial}{\partial \mu_j} \mathbb{E}_P(\mu^*) + \mathbb{E}_{\tilde{P}}(\mu) = \tilde{P}_j - P_i,$$

where for the first equation $i \in \operatorname{argmin}_k \{d_{k,j}^r - \lambda_k\}$ and for the second one $j \in \operatorname{argmin}_k \{d_{i,k}^r - \mu_k\}$, so that equality holds in the duality equations $\lambda_j^* = d_{i,j}^r - \lambda_i$ and $\mu_i^* = d_{i,j}^r - \mu_j$. The nonlinear conjugate gradient method (cf. Ruszczyński [119]) is an appropriate choice to compute successive improvements of the unconstrained problems (2.29).

This algorithmic approach to compute $d_r(P, \tilde{P})$ notably provides the dual variables λ and μ and the distance, but not the primal solution π . However, the primal π is supported only at points (i, j) with $\lambda_i^* + \mu_j^* = d_{i,j}^r$. This can be exploited to determine the primal variable π in a dual–primal step.

Example 2.28. As an example we consider two discrete distributions on \mathbb{R}^3 , whose probability mass functions on the vectors

$$\begin{pmatrix} x_{i,0} \\ x_{i,1} \\ x_{i,2} \end{pmatrix}$$

are given by

$$P = \left[\begin{array}{cccccccccccc} 0.02 & 0.04 & 0.08 & 0.06 & 0.21 & 0.09 & 0.15 & 0.06 & 0.09 & 0.08 & 0.12 \\ \begin{pmatrix} 10 \\ 13 \\ 15 \end{pmatrix} & \begin{pmatrix} 10 \\ 13 \\ 14 \end{pmatrix} & \begin{pmatrix} 10 \\ 13 \\ 13 \end{pmatrix} & \begin{pmatrix} 10 \\ 13 \\ 11 \end{pmatrix} & \begin{pmatrix} 10 \\ 11 \\ 12 \end{pmatrix} & \begin{pmatrix} 10 \\ 11 \\ 9 \end{pmatrix} & \begin{pmatrix} 10 \\ 8 \\ 10 \end{pmatrix} & \begin{pmatrix} 10 \\ 8 \\ 8 \end{pmatrix} & \begin{pmatrix} 10 \\ 8 \\ 6 \end{pmatrix} & \begin{pmatrix} 10 \\ 7 \\ 6 \end{pmatrix} & \begin{pmatrix} 10 \\ 5 \\ 5 \end{pmatrix} \end{array} \right]$$

and

$$\tilde{P} = \left[\begin{array}{ccccccc} 0.12 & 0.18 & 0.30 & 0.16 & 0.16 & 0.08 \\ \begin{pmatrix} 10 \\ 13 \\ 14 \end{pmatrix} & \begin{pmatrix} 10 \\ 13 \\ 12 \end{pmatrix} & \begin{pmatrix} 10 \\ 11 \\ 10 \end{pmatrix} & \begin{pmatrix} 10 \\ 7 \\ 9 \end{pmatrix} & \begin{pmatrix} 10 \\ 7 \\ 8 \end{pmatrix} & \begin{pmatrix} 10 \\ 7 \\ 5 \end{pmatrix} \end{array} \right].$$

The matrix in Table 2.2 collects the distances $\mathbf{d}_{i,j} = \sum_{t=0}^2 |x_{i,t} - x_{j,t}|$. Later, these vectors will be interpreted as the values on the paths of a tree (see Fig. 2.12), but for the simple Wasserstein distance as we discuss it here, the treestructure is irrelevant.

The solutions of the Wasserstein problem (2.23) and its dual (2.25) are displayed in the Table 2.3.

Table 2.2 The distance matrix with entries $\mathbf{d}_{i,j}$ from the Example 2.28. The optimal transportation plan sits only on the 16 pairs which are italicized

Distance $\mathbf{d}_{i,j}$	1	2	3	4	5	6
1	I	3	7	12	13	16
2	0	2	6	11	12	15
3	I	I	5	10	11	14
4	3	I	3	8	9	12
5	4	2	2	7	8	11
6	7	5	I	4	5	8
7	9	7	3	2	3	6
8	11	9	5	2	I	4
9	13	11	7	4	3	2
10	14	12	8	3	2	3
11	16	14	10	5	4	I

Table 2.3 The solutions of the primal and the dual Wasserstein problem

Probabilities, $\pi_{i,j}$	0.12	0.18	0.30	0.16	0.16	0.08
0.02	0.02	0	0	0	0	0
0.04	0.04	0	0	0	0	0
0.08	0.06	0.02	0	0	0	0
0.06	0	0.06	0	0	0	0
0.21	0	0.10	0.11	0	0	0
0.09	0	0	0.09	0	0	0
0.15	0	0	0.10	0.05	0	0
0.06	0	0	0	0	0.06	0
0.09	0	0	0	0.09	0	0
0.08	0	0	0	0.02	0.06	0
0.12	0	0	0	0	0.04	0.08

(a) The transportation plan π solving the primal Wasserstein problem (2.23) for the two distributions given in Example 2.28, $d_1(P, \tilde{P}) = \sum_{i,j} \pi_{i,j} d_{i,j} = 1.91$

Dual variables	$\lambda_i + \mu_j$	μ					
		6	6	6	5	4	1
λ	-5	1	1	1	0	-1	-4
	-6	0	0	0	-1	-2	-5
	-5	1	1	1	0	-1	-4
	-5	1	1	1	0	-1	-4
	-4	2	2	2	1	0	-3
	-5	1	1	1	0	-1	-4
	-3	3	3	3	2	1	-2
	-3	3	3	3	2	1	-2
	-1	5	5	5	4	3	0
	-2	4	4	4	3	2	-1
	0	6	6	6	0	4	1

(b) The variables λ (leftmost column) and μ (upmost row) solving the dual Wasserstein problem (2.27) for the two distributions given in Example 2.28. Their sum $\lambda_i + \mu_j$ is shown as matrix elements. They satisfy $\lambda_i + \mu_j \leq d_{i,j}$, with equality in the shaded cells (cf. Table 2.2), and $\sum_i P_i \lambda_i + \sum_j \tilde{P}_j \mu_j = 1.91$

2.9 Continuity of the Dual Variables, and the Kantorovich–Rubinstein Theorem

To investigate the continuity of the dual variables define the diameter $\Delta := \sup_{\xi \in \Xi, \tilde{\xi} \in \tilde{\Xi}} d(\xi, \tilde{\xi})$ (Δ may be unbounded, but is bounded for discrete measures and even by 1 for the distances discussed in Remark 2.24).

By convexity of the function $x \mapsto x^r$ it holds that

$$d(\xi_2, \tilde{\xi})^r \geq d(\xi_1, \tilde{\xi})^r + r d(\xi_1, \tilde{\xi})^{r-1} (d(\xi_2, \tilde{\xi}) - d(\xi_1, \tilde{\xi})),$$

from which follows that

$$\begin{aligned} d(\xi_1, \tilde{\xi})^r - \mu(\tilde{\xi}) - (d(\xi_2, \tilde{\xi})^r - \mu(\tilde{\xi})) &\leq r d(\xi_1, \tilde{\xi})^{r-1} (d(\xi_1, \tilde{\xi}) - d(\xi_2, \tilde{\xi})) \\ &\leq r d(\xi_1, \tilde{\xi})^{r-1} d(\xi_1, \xi_2) \end{aligned} \quad (2.30)$$

by the triangle inequality, $d(\xi_1, \tilde{\xi}) \leq d(\xi_2, \tilde{\xi}) + d(\xi_1, \xi_2)$. As one may assume by (2.27) that $\lambda(\xi) = \inf_{\tilde{\xi}} d(\tilde{\xi}, \xi)^r - \mu(\tilde{\xi})$, it follows that

$$\lambda(\xi_1) - (\lambda(\xi_2) - (d(\xi_2, \tilde{\xi})^r - \mu(\tilde{\xi}))) \leq r \Delta^{r-1} d(\xi_1, \xi_2),$$

and thus

$$\lambda(\xi_1) - \lambda(\xi_2) \leq r \Delta^{r-1} d(\xi_1, \xi_2).$$

By interchanging the roles of ξ_1 and ξ_2 it follows that λ is continuous with Lipschitz constant $r \Delta^{r-1}$ —provided that the diameter is bounded, $\Delta < \infty$. The same reasoning as above can be repeated to verify that μ is Lipschitz continuous as well with the same Lipschitz constant.

Kantorovich–Rubinstein Theorem. A particular situation arises for the Kantorovich distance (i.e., the Wasserstein of order $r = 1$). It follows from (2.30) directly that

$$\lambda(\xi_1) - \lambda(\xi_2) \leq d(\xi_1, \xi_2),$$

that is the dual functions λ and μ are Lipschitz continuous with constant 1, irrespective of the diameter (notice as well that $r \Delta^{r-1} = \Delta^0 = 1$ whenever $r = 1$).

Moreover,

$$-\lambda(\xi) \leq \inf_{\tilde{\xi}} d(\tilde{\xi}, \xi) - \lambda(\tilde{\xi}) \leq -\lambda(\xi)$$

by Lipschitz-1 continuity and by choosing $\tilde{\xi} = \xi$, hence $\mu(\xi) = -\lambda(\xi)$. This is the content of the Kantorovich–Rubinstein Theorem.

Theorem 2.29 (Kantorovich–Rubinstein Theorem). *Let (Ξ, d) be a Polish space, then*

$$d_1(P, \tilde{P}) = \sup_{\lambda} \mathbb{E}_P \lambda - \mathbb{E}_{\tilde{P}} \lambda,$$

where the supremum is among all Lipschitz continuous functions λ , i.e.,

$$\sup_{\xi \neq \tilde{\xi}} \frac{\lambda(\xi) - \lambda(\tilde{\xi})}{d(\xi, \tilde{\xi})} \leq 1,$$

which are integrable with respect to P and \tilde{P} .

2.10 Multistage Generalization: The Nested Distance

Multistage optimization problems do not consider just one single stage, but, as its name indicates and as was outlined in the introduction, multiple and subsequent stages. In mathematical terms it is not a single random variable which has to be considered, but an entire stochastic process instead. To this end, let (Ω, \mathcal{F}, P) and $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ be two probability spaces and let $\xi : \Omega \rightarrow \Xi$ and $\tilde{\xi} : \tilde{\Omega} \rightarrow \Xi$ be two random variables with common image space $\Xi \subseteq \mathbb{R}^m$, which is endowed with a metric d . We assume that (Ξ, d) is a Polish space and the Wasserstein distance d_r on $\mathcal{P}_r(\Xi)$ is well defined. This distance will now be extended for stochastic processes ξ_t defined on a filtered probability space $(\Omega, \mathfrak{F} = (\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_T), P)^{12}$ and another process $\tilde{\xi}_t$ defined on $(\tilde{\Omega}, \tilde{\mathfrak{F}} = (\tilde{\mathcal{F}}_0, \tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_T), \tilde{P})$.

2.10.1 The Inherited Distance

Consider a stochastic process

$$\xi_t : \Omega \rightarrow (\Xi_t, d_t), \quad t = 0, 1, \dots, T,$$

with possibly different state spaces (Ξ_t, d_t) for every $t = 1, \dots, T$. The value ξ_0 is considered as being deterministic. These random variables $(\xi_t)_{t=0}^T$ can be compounded to a single random variable ξ via

$$\begin{aligned} \xi : \Omega &\rightarrow \Xi_0 \times \Xi_1 \times \dots \times \Xi_T \\ \omega &\mapsto (\xi_0(\omega), \dots, \xi_T(\omega)), \end{aligned} \tag{2.31}$$

where each ω is mapped to its path (the trajectory) in the state space $\Xi := \Xi_0 \times \Xi_1 \times \dots \times \Xi_T$. This setting generalizes the usual definition of a stochastic process as the state spaces of the partial observations

¹²Often also called a *stochastic basis*.

$$\xi_t = \text{proj}_t \circ \xi : \Omega \rightarrow \Xi_t \quad t = 0, 1, \dots, T$$

may differ at different times ($\text{proj}_t : \Xi \rightarrow \Xi_t$ is the natural projection).

For ξ a process as in (2.31), P^ξ can be considered again. $P^\xi = P \circ \xi^{-1}$ is called the *law of the process* ξ , it is a probability measure on the product $\Xi := \Xi_0 \times \Xi_1 \times \dots \times \Xi_T$.

Now note that any of the spaces Ξ_t are equipped with a distance function d_t , and there are many metrics d such that (Ξ, d) is a metric space. Given two processes ξ resp. $\tilde{\xi}$ (with the same state spaces Ξ_t) on Ω ($\tilde{\Omega}$, resp.), a (semi-)distance is inherited to $\Omega \times \tilde{\Omega}$ in an analogous way as in Definition 2.3, for example by

$$d(\omega, \tilde{\omega}) := \sum_{t=0}^T w_t d_t \left(\xi_t(\omega), \tilde{\xi}_t(\tilde{\omega}) \right), \quad (2.32)$$

the weighted ℓ^1 -distance (with weights $w_t > 0$), or

$$d(\omega, \tilde{\omega}) := \left(\sum_{t=0}^T w_t d_t \left(\xi_t(\omega), \tilde{\xi}_t(\tilde{\omega}) \right)^2 \right)^{\frac{1}{2}}, \quad (2.33)$$

the ℓ^2 -distance, or

$$d(\omega, \tilde{\omega}) := \max_{t=0, \dots, T} w_t d_t \left(\xi_t(\omega), \tilde{\xi}_t(\tilde{\omega}) \right),$$

the ℓ^∞ -distance.

For any of these choices d is a cost function or (semi-)distance on $\Omega \times \tilde{\Omega}$, and the Wasserstein distance

$$d_r \left(P^\xi, P^{\tilde{\xi}} \right) \quad (2.34)$$

of the laws is available.

The following example elaborates that this simple application of the Wasserstein distance (2.34) is not suitable yet to distinguish stochastic processes.

Example 2.30. To observe the hidden caveat for the (final) Wasserstein distance consider the example depicted in Fig. 2.4. Two processes are shown there, which have the same states. The paths of successive observations, for both processes, are $(2, 2, 3)$ or $(2, 2, 1)$. Each path has the same probability in both processes (p , and $1 - p$, resp.).

The Wasserstein distance of these processes is simply 0: indeed, the state space is

$$\Xi = \tilde{\Xi} = \{(2, 2, 1), (2, 2, 3)\}$$

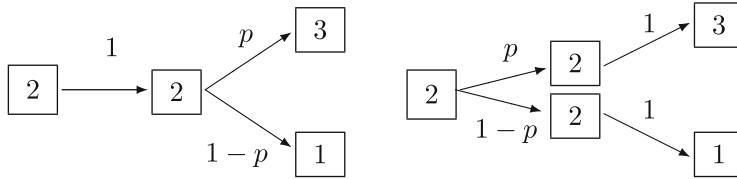


Fig. 2.4 Two processes with identical final probabilities and identical states. The second process reveals already at an earlier stage that the final observation will be 3 (1, resp.). The nested distance of the trees is $4p(1-p)$

for both processes. The distance matrix then is $d = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$ and $\pi = \begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}$ is a feasible transport plan. Hence, the Wasserstein distance of the laws P and \tilde{P} of the processes is $d_r(P, \tilde{P}) = \sum_{i,j} d_{i,j} \pi_{i,j} = 0$.

However, the processes P and \tilde{P} depicted in Fig. 2.4 are certainly not the same processes: having observed the partial path $(2, 2)$ in the second process, we already know whether the final observation will be 1 or 3. This knowledge (information) is *not* available for the first process. So as the distance was identified to be $d_r(P, \tilde{P}) = 0$, the Wasserstein distance, in its genuine form, does not qualify as a distance for filtered stochastic processes.

The reason why the Wasserstein distance does not detect this difference is because it does not take conditional probabilities into account (\mathcal{F}_t for $t = 0, 1 \dots T-1$), but only final probabilities, where the sigma algebras coincide, $\mathcal{F}_T = \tilde{\mathcal{F}}_T$. But the sigma algebras differ at stage 1 (cf. (1.21)),

$$\mathcal{F}_1 = \sigma(\{(2, 2, 1), (2, 2, 3)\}) \subsetneq \sigma(\{(2, 2, 1)\}, \{(2, 2, 3)\}) = \tilde{\mathcal{F}}_1.$$

Definition 2.31 (The Filtration Induced by the Process). The *history process* is

$$\xi_{0:t} := \text{proj}_{0:t} \circ \xi := (\text{proj}_0 \circ \xi, \dots, \text{proj}_t \circ \xi) = (\xi_0, \dots, \xi_t),$$

that is $\xi_{0:t}(\omega) := (\xi_0(\omega), \dots, \xi_t(\omega)) \in \Xi_0 \times \dots \times \Xi_t$.

The history process generates the *natural filtration* of the process ξ ,

$$\mathfrak{F}^\xi = \left(\mathcal{F}_t^\xi \right)_{t=0}^T, \quad \mathcal{F}_t^\xi := \sigma \left(\{ \xi_{0:t}^{-1}(A_0 \times \dots \times A_t) : A_s \in \mathcal{B}(\Xi_s) \} \right),$$

where $\mathcal{B}(\Xi_s)$ denotes the Borel sets on Ξ_s . Notice that the relation $\xi \triangleleft \mathfrak{F}$ implies that the filtration \mathfrak{F} is finer than \mathfrak{F}^ξ .

2.10.2 The Nested Distance

The nested distance is based on the Wasserstein distance. Extending the Wasserstein distance to stochastic processes the nested distance takes notice of all sigma algebras contained in the filtrations of the filtered probability spaces.

Definition 2.32 (The Nested Distance). The *nested distance* of order $r \geq 1$ of two filtered probability spaces $\mathbb{P} = (\Omega, (\mathcal{F}_t), P)$ and $\tilde{\mathbb{P}} = (\tilde{\Omega}, (\tilde{\mathcal{F}}_t), \tilde{P})$, for which a distance $\mathbf{d} : \Omega \times \tilde{\Omega} \rightarrow \mathbb{R}$ is defined, is the optimal value of the optimization problem

$$\begin{aligned} & \underset{\text{(in } \pi\text{)}}{\text{minimize}} \quad \left(\int \mathbf{d}(\omega, \tilde{\omega})^r \pi(d\omega, d\tilde{\omega}) \right)^{\frac{1}{r}} \\ & \text{subject to} \quad \pi(A \times \tilde{\Omega} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A \mid \mathcal{F}_t) \quad (A \in \mathcal{F}_t, t \in \mathbf{T}), \\ & \quad \pi(\Omega \times B \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \tilde{P}(B \mid \tilde{\mathcal{F}}_t) \quad (B \in \tilde{\mathcal{F}}_t, t \in \mathbf{T}), \end{aligned} \quad (2.35)$$

where the infimum in (2.35) is among all bivariate probability measures $\pi \in \mathcal{P}(\Omega \times \tilde{\Omega})$ which are defined on $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$ ¹³ and $\mathbf{T} = \{0, 1 \dots T\}$. Its optimal value, the nested distance, is denoted by

$$\mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}}).$$

Remark 2.33. The nested distance is often called *multistage distance* or *process distance* as well. A feasible measure π is called a nested transport plan.

The nested distance was initially constructed on nested distributions (cf. Definition 1.7), both were introduced by Pflug in [92]. The definition given here notably applies for continuous time, $\mathbf{T} = \{t \in \mathbb{R} : t \geq 0\}$ as well.

The Markov-constructions contained in Rüschendorf [116] can be compared with the nested distribution for two stages, such that the distance on Markov-constructions can be considered as a special case of the definition provided here.

The multistage formulation presented here is based on filtrations. The following discussion of the nested distance is adapted from [94].

Discussion of the Nested Distance. We recall first that the conditional probability is defined by the conditional expectation by $P(A \mid \mathcal{F}_t) = \mathbb{E}(\mathbb{1}_A \mid \mathcal{F}_t)$ for every $A \in \mathcal{F}_T$, it is thus a random variable itself,

$$P(A \mid \mathcal{F}_t) = \mathbb{E}(\mathbb{1}_A \mid \mathcal{F}_t) : \Omega \rightarrow [0, 1],$$

which is measurable with respect to \mathcal{F}_t . Its characterizing property is

¹³ $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$ is the smallest sigma-algebra on the product space $\Omega \times \tilde{\Omega}$, which contains all rectangles $A \times \tilde{A}$ for $A \in \mathcal{F}$, $\tilde{A} \in \tilde{\mathcal{F}}$.

$$\int_B P(A | \mathcal{F}_t) dP = \int_B P(A | \mathcal{F}_t)(\omega) P(d\omega) = P(A \cap B) \quad (A \in \mathcal{F}_T, B \in \mathcal{F}_t).$$

The identity

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t)$$

thus expresses that

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)(\omega, \tilde{\omega}) = P(A | \mathcal{F}_t)(\omega) \quad \pi \text{ almost everywhere}$$

for every $A \in \mathcal{F}_T$. The right-hand side of this equation is notably independent of $\tilde{\omega}$. It is sometimes helpful to make this independence explicit by using the notations

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t) \circ \text{id} = P(A | \mathcal{F}_t)(\text{id}),$$

where id is the projection $\text{id} : \Omega \times \tilde{\Omega} \rightarrow \Omega$, $\text{id}(\omega, \tilde{\omega}) = \omega$ and $\tilde{\text{id}}(\omega, \tilde{\omega}) = \tilde{\omega}$, respectively.

Remark 2.34. Two stochastic processes $\xi_t : \Omega \rightarrow \Xi_t$ and $\tilde{\xi}_t : \Omega \rightarrow \tilde{\Xi}_t$ on the same probability space $(\Omega, \mathcal{F}; P)$ induce the filtered probability spaces $\mathbb{P}^\xi := (\Omega, \mathfrak{F}^\xi, P^\xi)$ and $\tilde{\mathbb{P}}^{\tilde{\xi}} := (\Omega, \tilde{\mathfrak{F}}^{\tilde{\xi}}, P^{\tilde{\xi}})$, for which the nested distance is available, provided that there is a cost function

$$\mathbf{d} : (\Xi_0 \times \dots \Xi_T) \times (\tilde{\Xi}_0 \times \dots \tilde{\Xi}_T) \rightarrow \mathbb{R}. \quad (2.36)$$

This justifies the name *process distance*. In addition it should be repeated that the state spaces Ξ_t and $\tilde{\Xi}_t$ do not necessarily have to coincide. Then \mathbf{d} in (2.36) is more a cost function than a distance function. Notice that the inherited distance defined in (2.7) is rather a cost function too.

Remark 2.35 (The Initial Stage, $t = 0$). For the trivial sigma-algebra $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $P(A | \mathcal{F}_0)$ is deterministic (a constant) and satisfies $P(A | \mathcal{F}_0) = P(A)$ (almost everywhere).

Remark 2.36 (The Final Stage, $t = T$). For $A \in \mathcal{F}_T$ it holds that $P(A | \mathcal{F}_T) = \mathbb{E}(\mathbf{1}_A | \mathcal{F}_T) = \mathbf{1}_A$ and $\pi(A \times \tilde{\Omega} | \mathcal{F}_T \otimes \tilde{\mathcal{F}}_T) = \mathbf{1}_{A \times \tilde{\Omega}}$. But as $\mathbf{1}_A \circ \text{id} = \mathbf{1}_{A \times \tilde{\Omega}}$ always holds true it follows that the constraints in (2.35) are redundant for $t = T$, they can be omitted.

Lemma 2.37. Let $(\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}$ and $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}, \tilde{\xi}) \sim \tilde{\mathbb{P}}$ be nested distributions with $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\tilde{\mathcal{F}}_0 = \{\emptyset, \tilde{\Omega}\}$.¹⁴ The product measure $\pi := P \otimes \tilde{P}$ is

¹⁴If not otherwise specified, \mathbf{d} is always the distance inherited from ξ and $\tilde{\xi}$.

feasible for the multistage distance (i.e., the nested distance is well defined). It holds moreover that

$$\mathbf{d}_r(P, \tilde{P})^r \leq \mathbf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})^r \leq \mathbb{E}_{P \otimes \tilde{P}}(\mathbf{d}^r). \quad (2.37)$$

Proof. The first inequality follows in view of Remark 2.35. We shall verify that all constraints in (2.35) are satisfied for $\pi := P \otimes \tilde{P}$. For this choose $C \in \mathcal{F}_t$ and $D \in \tilde{\mathcal{F}}_t$ and observe that, for $\pi = P \otimes \tilde{P}$,

$$\begin{aligned} & \int_{C \times D} P(A | \mathcal{F}_t)(\text{id}) \cdot \tilde{P}(B | \mathcal{F}_t)(\tilde{\text{id}}) d\pi \\ &= \int_C P(A | \mathcal{F}_t)(\text{id}) dP \cdot \int_D \tilde{P}(B | \mathcal{F}_t)(\tilde{\text{id}}) d\tilde{P} \\ &= P(A \cap C) \cdot \tilde{P}(B \cap D) \\ &= \pi((A \cap C) \times (B \cap D)) = \pi((A \times B) \cap (C \times D)) \\ &= \int_{C \times D} \pi(A \times B | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) d\pi. \end{aligned}$$

It follows that the conditional probabilities $P(A | \mathcal{F}_t) \circ \text{id} \cdot \tilde{P}(B | \mathcal{F}_t) \circ \tilde{\text{id}}$ and $\pi(A \times B | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)$ are (π -almost everywhere) identical, as equality holds for any sets $C \in \mathcal{F}_t$ and $D \in \tilde{\mathcal{F}}_t$. By choosing $A = \Omega$ ($B = \tilde{\Omega}$, resp.) it follows that $\pi = P \otimes \tilde{P}$ is feasible and hence $\mathbf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})^r \leq \mathbb{E}_{P \otimes \tilde{P}}(\mathbf{d}^r)$. \square

The following example demonstrates that convergence with respect to the distance of the multivariate distributions \mathbf{d}_r is quite different from convergence of the nested distributions (i.e., with respect to the nested distance \mathbf{dl}_r).

Example 2.38 (See Heitsch et al. [55]). Consider the following nested distributions

$$\mathbb{P}_\epsilon = \left[\begin{array}{cc} \frac{0.5}{2} & \frac{0.5}{2+\epsilon} \\ \left[\begin{array}{c} 1.0 \\ 3 \end{array} \right] & \left[\begin{array}{c} 1.0 \\ 1 \end{array} \right] \end{array} \right] \quad \text{and} \quad \mathbb{P}_0 = \left[\begin{array}{c} \frac{1.0}{2} \\ \left[\begin{array}{cc} 0.5 & 0.5 \\ 3 & 1 \end{array} \right] \end{array} \right].$$

Notice that the pertaining multivariate distribution of \mathbb{P}_ϵ on \mathbb{R}^2 converges weakly to the one of \mathbb{P}_0 , if $\epsilon \rightarrow 0$. However, the nested distributions do not converge to \mathbb{P}_0 : The nested distance is $\mathbf{dl}(\mathbb{P}_\epsilon, \mathbb{P}_0) = 1 + \epsilon$ for all ϵ . The limit \mathbb{P}_ϵ as $\epsilon \rightarrow 0$ in the sense of nested distances is

$$\bar{\mathbb{P}} = \left[\begin{array}{cc} 0.5 & 0.5 \\ \hline 2 & 2 \\ \left[\begin{array}{c} 1.0 \\ \hline 3 \end{array} \right] & \left[\begin{array}{c} 1.0 \\ \hline 1 \end{array} \right] \end{array} \right]$$

which is different from \mathbb{P}_0 . To put it differently: the topology of the tree is not changed by going to the limit in the nested distance sense. The filtration of the limiting nested distribution $\bar{\mathbb{P}}$ is larger than the one generated by the scenario values. The concept of nested distributions can handle this.

The following proposition shows that the left side of inequality (2.37) can be refined by considering finer filtrations, which are between the original filtration and the full clairvoyant filtration.

Proposition 2.39. *Let $\mathbb{P} \sim (\Omega, \mathfrak{F}, P, \xi)$ and $\tilde{\mathbb{P}} \sim (\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}, \tilde{\xi})$ be filtered spaces with filtrations $\mathfrak{F} = (\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_T)$ and $\tilde{\mathfrak{F}} = (\tilde{\mathcal{F}}_0, \tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_T)$, respectively. Denote by \mathbb{P}^t the pertaining nested distribution made clairvoyant from time t onwards, that is*

$$\mathbb{P}^t \sim (\Omega, \mathfrak{F}^t, P, \xi^t) \quad \text{with} \quad \mathfrak{F}^t = (\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_{t-1}, \mathcal{F}_T, \dots, \mathcal{F}_T, \mathcal{F}_T).$$

In a similar manner we define $\tilde{\mathbb{P}}^t$. Then, for $1 \leq t \leq T$,

$$\mathbf{d}_r(P, \tilde{P}) = \mathbf{d}\mathbf{l}_r(\mathbb{P}^1, \tilde{\mathbb{P}}^1) \leq \dots \leq \mathbf{d}\mathbf{l}_r(\mathbb{P}^t, \tilde{\mathbb{P}}^t) \leq \dots \leq \mathbf{d}\mathbf{l}_r(\mathbb{P}^T, \tilde{\mathbb{P}}^T) = \mathbf{d}\mathbf{l}_r(\mathbb{P}, \tilde{\mathbb{P}}).$$

Proof. The proof follows from the fact that the multivariate distance is always not larger than the nested distance. Arguing this way for the subtrees at stage t and considering the recursive structure of the nested distance, the assertion is obvious. \square

Example 2.40. Figure 2.5 shows two trees (nested distributions) \mathbb{P} and $\tilde{\mathbb{P}}$. Their nested distance is $\mathbf{d}\mathbf{l}(\mathbb{P}, \tilde{\mathbb{P}}) = 8.75$. Figure 2.6 shows the same trees, but both are made clairvoyant from time 2 onwards. Their distance is reduced to $\mathbf{d}\mathbf{l}(\mathbb{P}^2, \tilde{\mathbb{P}}^2) = 0.5$. Finally, in Fig. 2.7, the same trees are now made totally clairvoyant. Their distance is $\mathbf{d}\mathbf{l}(\mathbb{P}^1, \tilde{\mathbb{P}}^1) = 0$, since their set of trajectories is identical.

The following lemma recovers the properties of the Wasserstein distance for the nested distance.

Lemma 2.41 (Monotonicity and Convexity).

(i) Suppose that $r_1 \leq r_2$, then

$$\mathbf{d}\mathbf{l}_{r_1}(\mathbb{P}, \tilde{\mathbb{P}}) \leq \mathbf{d}\mathbf{l}_{r_2}(\mathbb{P}, \tilde{\mathbb{P}}).$$

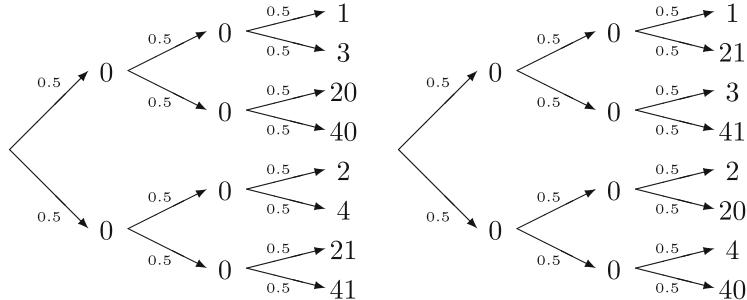


Fig. 2.5 The two original trees \mathbb{P} (*left*) and $\tilde{\mathbb{P}}$ (*right*). Their nested distance is $\text{dl}(\mathbb{P}, \tilde{\mathbb{P}}) = 8.75$

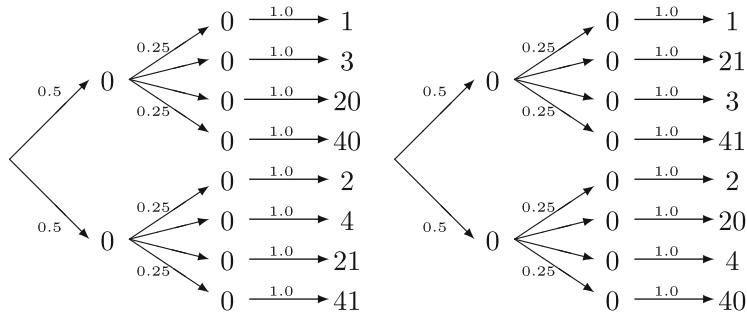


Fig. 2.6 The two trees of Fig. 2.5 have been made clairvoyant from time 2 onwards leading to the new trees \mathbb{P}^2 and $\tilde{\mathbb{P}}^2$. Their distance is $d(\mathbb{P}^2, \tilde{\mathbb{P}}^2) = 0.5$

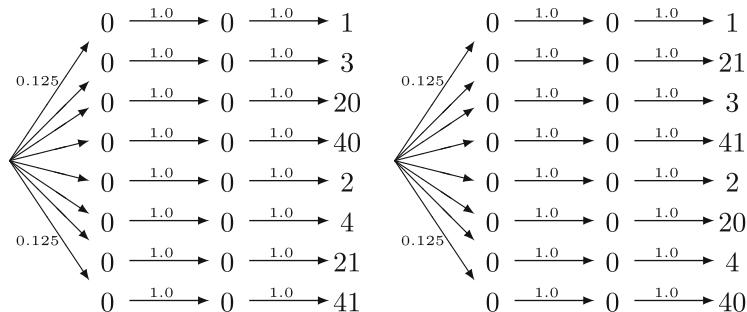


Fig. 2.7 The two trees of Fig. 2.5 have now been made further clairvoyant, namely from time 1 onwards leading to the new trees \mathbb{P}^1 and $\tilde{\mathbb{P}}^1$. Their distance is $d(\mathbb{P}^1, \tilde{\mathbb{P}}^1) = 0.0$, i.e., they are identical

(ii) The nested distance is r -convex in any of its components, that is for $0 \leq \lambda \leq 1$ it holds that

$$\mathsf{dI}_r \left(\mathbb{P}, \mathcal{C}(\tilde{\mathbb{P}}_0, \tilde{\mathbb{P}}_1, \lambda) \right)^r \leq \lambda \mathsf{dI}_r \left(\mathbb{P}, \tilde{\mathbb{P}}_0 \right)^r + (1 - \lambda) \mathsf{dI}_r \left(\mathbb{P}, \tilde{\mathbb{P}}_1 \right)^r,$$

and

$$\begin{aligned} \mathbf{dI}_r & \left(\mathbb{P}, \mathcal{C}(\tilde{\mathbb{P}}_0, \tilde{\mathbb{P}}_1, \lambda) \right) \\ & \leq \lambda^{\frac{1}{r}} \mathbf{dI}_r \left(\mathbb{P}, \tilde{\mathbb{P}}_0 \right) + (1 - \lambda)^{\frac{1}{r}} \mathbf{dI}_r \left(\mathbb{P}, \tilde{\mathbb{P}}_1 \right) \\ & \leq \max \{ \lambda, 1 - \lambda \}^{\frac{1}{r}-1} \cdot \left(\lambda \mathbf{dI}_r \left(\mathbb{P}, \tilde{\mathbb{P}}_0 \right) + (1 - \lambda) \mathbf{dI}_r \left(\mathbb{P}, \tilde{\mathbb{P}}_1 \right) \right). \end{aligned}$$

Here $\mathcal{C}(\tilde{\mathbb{P}}_0, \tilde{\mathbb{P}}_1, \lambda)$ is the compound distribution defined in (1.24) of the introduction.

- (iii) \mathbf{dI}_r satisfies the triangle inequality, $\mathbf{dI}_r \left(\mathbb{P}, \tilde{\mathbb{P}} \right) \leq \mathbf{dI}_r \left(\mathbb{P}, \tilde{\mathbb{P}} \right) + \mathbf{dI}_r \left(\tilde{\mathbb{P}}, \tilde{\mathbb{P}} \right)$.

Proof. The proof of Lemma 2.10 applies. As for the triangle inequality we refer to the proof contained in Villani [137] involving the gluing lemma. A similar proof applies here. \square

2.10.3 The Nested Distance for Trees

The Wasserstein distance between discrete probability measures can be calculated by solving the linear program (2.23). To establish the corresponding linear program for the nested distance we use trees that model the whole space and filtration. Recall that we denote by $m \prec i$ that node m is a predecessor of node i , not necessarily the immediate predecessor. Problem (2.35) reads

$$\begin{aligned} & \text{minimize}_{\substack{(\text{in } \pi)}} \quad \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \\ & \text{subject to} \quad \sum_{j \succ n} \pi(i, j | k, l) = P(i | k) \quad (k \prec i, l), \\ & \quad \sum_{i \succ m} \pi(i, j | k, l) = \tilde{P}(j | l) \quad (l \prec j, k), \\ & \quad \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1, \end{aligned} \tag{2.38}$$

where again $\pi_{i,j}$ is a matrix defined on the terminal nodes ($i \in \mathcal{N}_T$, $j \in \tilde{\mathcal{N}}_T$) and $k \in \mathcal{N}_t$, $l \in \tilde{\mathcal{N}}_t$ are arbitrary nodes on the same stage t . The conditional probabilities $\pi(i, j | k, l)$ are given by

$$\pi(i, j | k, l) = \frac{\pi_{i,j}}{\sum_{i' \succ k, j' \succ l} \pi_{i',j'}}. \tag{2.39}$$

In view of this quotient it becomes evident that the constraint $\sum_{i,j} \pi_{i,j} = 1$ is necessary in (2.38) to specify a probability measure, as otherwise every multiple of any feasible π would be feasible as well.

Formulation as a Linear Program. The constraints in (2.38) can be rewritten as

$$P(i) \cdot \sum_{i' \succ m, j' \succ n} \pi_{i',j'} = P(m) \cdot \sum_{j' \succ n} \pi_{i,j'} \quad (m \prec i, n) \quad \text{and}$$

$$\tilde{P}(j) \cdot \sum_{i' \succ m, j' \succ n} \pi_{i',j'} = \tilde{P}(n) \cdot \sum_{i' \succ m} \pi_{i',j} \quad (m, n \prec j).$$

As P and \tilde{P} are given, the latter equations show that (2.38) is equivalent to

$$\begin{aligned} & \underset{\text{(in } \pi\text{)}}{\text{minimize}} \quad \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \\ & \text{subject to} \quad P(i) \cdot \sum_{i' \succ k, j' \succ l} \pi_{i',j'} = P(k) \cdot \sum_{j' \succ l} \pi_{i,j'} \quad (k \prec i), \\ & \quad \tilde{P}(j) \cdot \sum_{i' \succ k, j' \succ l} \pi_{i',j'} = \tilde{P}(l) \cdot \sum_{i' \succ k} \pi_{i',j} \quad (l \prec j), \\ & \quad \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1, \end{aligned}$$

which is indeed a *linear* program. (This is not immediate in the formulation (2.38), as it involves quotients.)

The nested structure of the transportation plan π , which is induced by the trees, is schematically depicted in Fig. 2.8a.

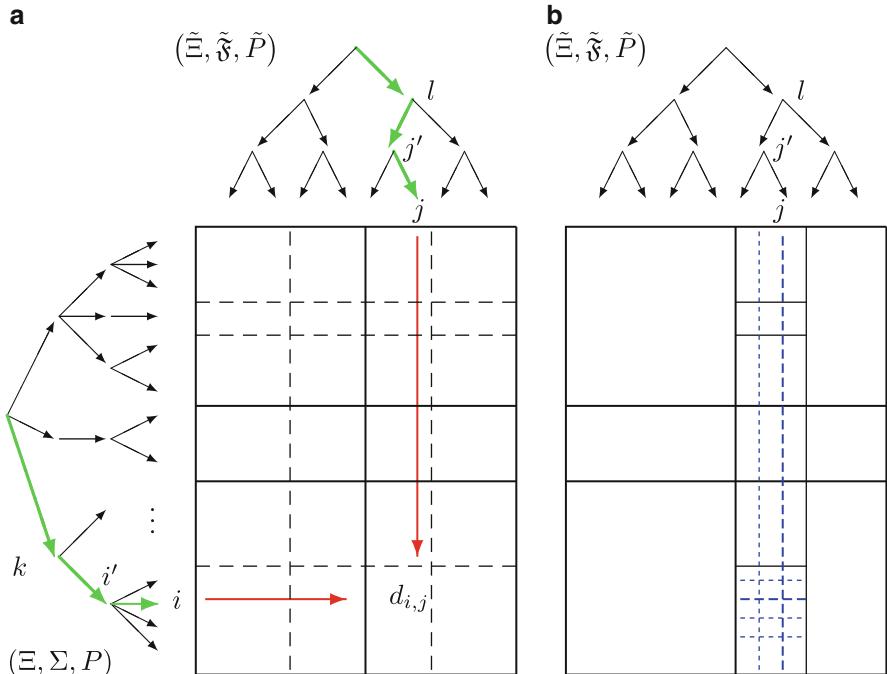


Fig. 2.8 Schematic structure of the distance matrix d and the transport matrix π , as it is imposed by the structures of the trees and the respective constraints

Remark 2.42. As is the case for the Wasserstein distance, many constraints in (2.38) are linearly dependent. For computational reasons (loss of significance during numerical evaluations, which can impact linear dependencies and the feasibility) it is advisable to remove linear dependencies. This is partially accomplished by the simpler program

$$\begin{aligned} & \text{minimize}_{(\pi)} \quad \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \\ & \text{subject to} \quad \sum_{\{j' : j' = l\}} \pi(i', j' | k, l) = Q(i') = P(i' | k) \quad (i' \in \mathcal{N} \setminus \{1\}, k = i'-, l = j'-) , \\ & \quad \sum_{\{i' : i' = k\}} \pi(i', j' | k, l) = \tilde{Q}(j') = \tilde{P}(j' | l) \quad (j' \in \tilde{\mathcal{N}} \setminus \{1\}, k = i'-, l = j'-) , \\ & \quad \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1 , \end{aligned} \tag{2.40}$$

where the conditional probabilities are

$$\begin{aligned} Q(i') &= P(i' | i'-) = P(i' | k) \\ \tilde{Q}(j') &= \tilde{P}(j' | j'-) = \tilde{P}(j' | l) \end{aligned}$$

for $k = i'-, l = j'-$. Here only one-step conditional transportation measures are required. They are defined as

$$\pi(i', j' | k, l) = \frac{\sum_{i > i', j > j'} \pi_{i,j}}{\sum_{i > k, j > l} \pi_{i,j}} . \tag{2.41}$$

Equation (2.40) is equivalent to (2.38) by the following lemma, and which can be reformulated as an LP as above (recall that $\mathcal{N} \setminus \{1\}$ denotes all nodes except the root). A computational advantage of (2.40) is given by the fact that the conditional probabilities involved are considered only at successive stages.

Further constraints can be removed from (2.40) by taking into account that $\sum_{\{i' : i' \in k\}} Q(i') = 1$. Hence, for each node k it is possible to drop one constraint out of all equations related to $\{i' : i' \in k\}$ (cf. Sect. 2.7 and Fig. 2.3 for the Wasserstein distance).

Lemma 2.43 (Tower Property). *To compute the nested distance it is enough to condition on the immediately following sigma algebra: the conditions*

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t) \text{ for all } A \in \mathcal{F}_T$$

in (2.35) may be replaced by

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t) \text{ for all } A \in \mathcal{F}_{t+1} .$$

Proof. The result follows from the tower property of the conditional expectation.

Let $A \in \mathcal{F}_T$ and observe first that

$$\begin{aligned}\mathbb{E}_\pi (\mathbf{1}_A(\text{id}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) &= \mathbb{E}_\pi (\mathbf{1}_{A \times \tilde{\Omega}} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\ &= \pi (A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A | \mathcal{F}_t)(\text{id}) = \mathbb{E}_P (\mathbf{1}_A | \mathcal{F}_t)(\text{id}),\end{aligned}$$

such that by linearity

$$\mathbb{E}_\pi (\lambda \circ \text{id} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \mathbb{E}_P (\lambda | \mathcal{F}_t) \circ \text{id}$$

for every $\lambda \triangleleft \mathcal{F}_t$. It follows then that

$$\begin{aligned}\pi (A \times \tilde{\Omega} | \mathcal{F}_{T-2} \otimes \tilde{\mathcal{F}}_{T-2}) &= \pi (\pi (A \times \tilde{\Omega} | \mathcal{F}_{T-1} \otimes \tilde{\mathcal{F}}_{T-1}) | \mathcal{F}_{T-2} \otimes \tilde{\mathcal{F}}_{T-2}) \\ &= \pi (P(A | \mathcal{F}_{T-1}) | \mathcal{F}_{T-2} \otimes \tilde{\mathcal{F}}_{T-2}) = P(A | \mathcal{F}_{T-2}),\end{aligned}$$

which is the assertion for $t = T - 2$. The assertion for general t follows by repeated application of the previous argument. \square

Notice that the nested distance may be defined not only between trees, but also between a filtered stochastic process and a tree. In the following example, we intend to approximate a simple stochastic process (with only two periods) by a discrete process sitting on a tree. It is crucial that the approximation aims at minimizing the nested distance and not the multivariate distance.

Example 2.44. Consider correlated normal variables $\xi_1 \sim N(0, 1)$ and $\xi_2 \sim N(\xi_1, 1)$, that is the joint distribution is

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \right).$$

Figure 2.9 displays the density of this distribution.

Let \mathbb{P} be the nested distribution pertaining to the two-stage process (ξ_1, ξ_2) . We approximate this distribution by a tree with 9 leaves. First, we consider a tree of height 2 with bushiness 3. To this end the first stage variable ξ_1 is approximated by a discrete distribution sitting on 3 points. It is well known that the optimal approximation of an $N(\mu, \sigma^2)$ distribution in the d_1 (Kantorovich) sense by a 3-point distribution is

$$\left[\frac{0.3035 \quad 0.3930 \quad 0.3035}{\mu - 1.029\sigma \quad \mu \quad \mu + 1.029\sigma} \right]$$

The distance is 0.3397σ . Therefore, the best approximation of ξ_1 is

$$\left[\frac{0.3035 \quad 0.3930 \quad 0.3035}{-1.029 \quad 0.0 \quad 1.029} \right]$$

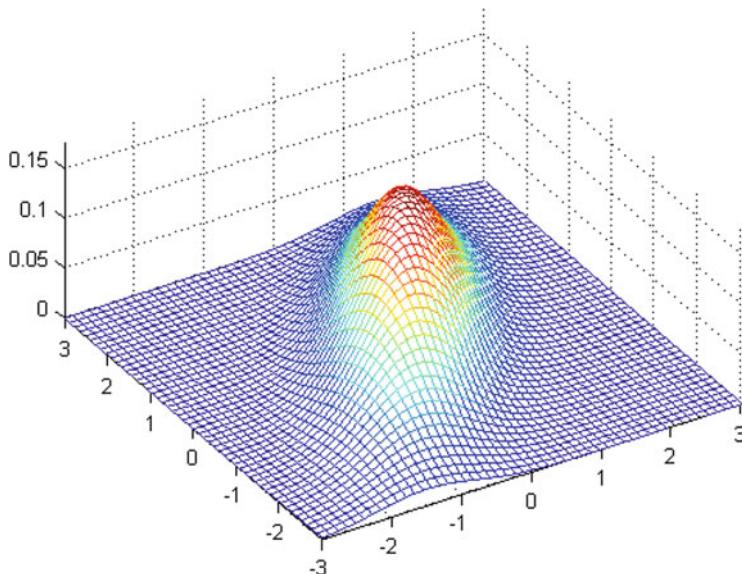


Fig. 2.9 A bivariate normal distribution

Then, conditional on the first coordinate, the second coordinate is approximated, resulting in the following nested distribution $\tilde{\mathbb{P}}^*$

$$\tilde{\mathbb{P}}^* = \left[\begin{array}{ccc} 0.3035 & 0.3930 & 0.3035 \\ -1.029 & 0.0 & 1.029 \\ \left[\begin{array}{ccc} 0.3035 & 0.393 & 0.3035 \\ -2.058 & -1.029 & 0.0 \end{array} \right] & \left[\begin{array}{ccc} 0.3035 & 0.393 & 0.3035 \\ -1.029 & 0.0 & 1.029 \end{array} \right] & \left[\begin{array}{ccc} 0.3035 & 0.393 & 0.3035 \\ 0.0 & 1.029 & 2.058 \end{array} \right] \end{array} \right].$$

The resulting nested distance is $d_{\mathbb{L}}(\mathbb{P}, \tilde{\mathbb{P}}^*) = 0.76$. The example is illustrated in Fig. 2.10.

As a comparison we have calculated the best approximation of the two-dimensional distribution (ξ_1, ξ_2) by a discrete probability sitting on 9 points. Notice that this approximation does not respect the tree structure, it can be seen as a fan with 9 leaves. The calculated approximate distribution is

$$\bar{\mathbb{P}} = \left[\begin{array}{ccccccccc} 0.114 & 0.108 & 0.152 & 0.148 & 0.078 & 0.046 & 0.188 & 0.114 & 0.052 \\ \left(\begin{array}{c} 1.205 \\ 1.205 \end{array} \right) & \left(\begin{array}{c} 0.277 \\ 1.601 \end{array} \right) & \left(\begin{array}{c} -1.068 \\ -0.855 \end{array} \right) & \left(\begin{array}{c} 0.088 \\ -0.660 \end{array} \right) & \left(\begin{array}{c} -0.577 \\ -2.074 \end{array} \right) & \left(\begin{array}{c} -1.855 \\ -2.522 \end{array} \right) & \left(\begin{array}{c} -0.412 \\ 0.397 \end{array} \right) & \left(\begin{array}{c} 0.894 \\ 0.132 \end{array} \right) & \left(\begin{array}{c} 0.052 \\ 1.673 \end{array} \right) \end{array} \right].$$

These points are shown in Fig. 2.11. While the multivariate distance is smaller than before, the nested distance is $d_{\mathbb{L}}(\mathbb{P}, \bar{\mathbb{P}}) = 1.12$ which is much larger than before because $\bar{\mathbb{P}}$ does not respect the filtration structure.

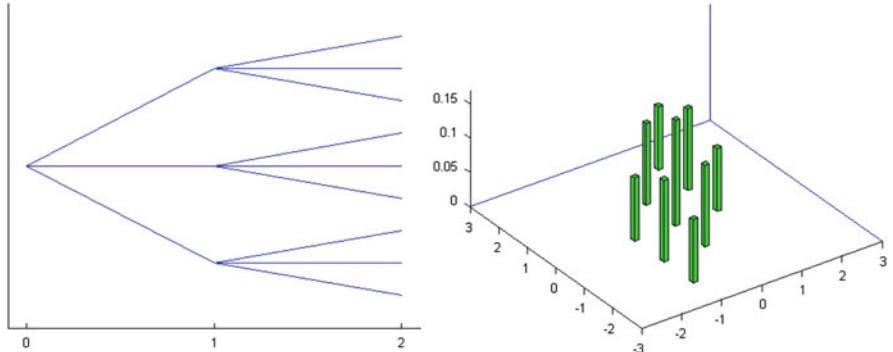


Fig. 2.10 An “optimal” discrete approximation $\tilde{\mathbb{P}}$ to the process \mathbb{P} of Fig. 2.9. *Left:* the prespecified tree structure. *Right:* the visualization of the values and probabilities

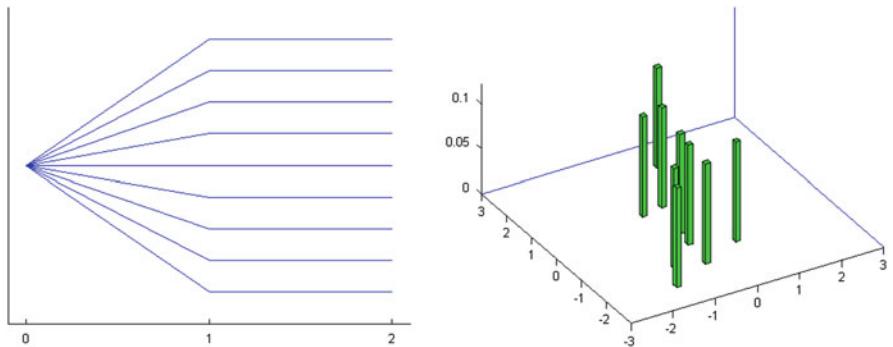


Fig. 2.11 A discrete approximation $\tilde{\mathbb{P}}$ to the bivariate distribution of Fig. 2.9. *Left:* the structure of the fan, i.e., a trivial tree. *Right:* the visualization of the values and probabilities

Rapid Computation of the Nested Distance. In view of the tower property it should be noted that instead of solving the full problem (2.40), the nested distance can be calculated in a recursive way.¹⁵ For this observe that \mathbf{d}_r is the nested distance of both trees, starting at their roots. For a rapid computation we extend the distance \mathbf{d}_r to subtrees, starting at given nodes of the trees.

For two leaves $i \in \mathcal{N}_T$, $j \in \tilde{\mathcal{N}}_T$ at the final stage of the tree define first

$$\mathbf{d}_T^r(i, j) := \mathbf{d}(\xi_i, \tilde{\xi}_j)^r.$$

Given $\mathbf{d}_{t+1}^r(i', j')$ for $i' \in \mathcal{N}_{t+1}$ and $j' \in \tilde{\mathcal{N}}_{t+1}$, set

¹⁵Notice that this recursive calculation corresponds to the way the nested distance was introduced in Sect. 1.4.1.

$$\mathbf{d}^r_t(k, l) := \sum_{i' \in k_+, j' \in l_+} \pi_t(i', j' | k, l) \cdot \mathbf{d}^r_{t+1}(i', j') \quad (k \in \mathcal{N}_t, l \in \tilde{\mathcal{N}}_t),$$

where the one-step conditional probabilities $\pi(\cdot, \cdot | k, l)$ solve the usual Wasserstein problem, conditioned on k and l , that is

$$\begin{aligned} & \text{minimize}_{\pi_t(\cdot, \cdot | k, l)} \sum_{i' \in k_+, j' \in l_+} \pi_t(i', j' | k, l) \cdot \mathbf{d}^r_{t+1}(i', j') \\ & \text{subject to} \quad \sum_{j' \in l_+} \pi_t(i', j' | k, l) = Q(i') = P(i' | k) \quad (i' \in k_+), \\ & \quad \sum_{i \in m_+} \pi_t(i', j' | k, l) = \tilde{Q}(j') = \tilde{P}(j' | l) \quad (j' \in l_+), \\ & \quad \pi_t(i', j' | k, l) \geq 0. \end{aligned}$$

The values $\mathbf{d}^r_t(k, l)$ can be interpreted as the nested distances of the subtrees starting in nodes k and l . Finally the transport plan π on the leaves is recomposed by

$$\pi(i, j) = \pi(i_1, j_1 | i_0, j_0) \cdots \pi(i_{T-1}, j_{T-1} | i_{T-2}, j_{T-2}) \cdot \pi(i, j | i_{T-1}, j_{T-1}) \quad (2.42)$$

with $i_t = \text{pred}_t(i)$, $j_t = \text{pred}_t(j)$. The nested distance is given by $\mathbf{d}^r_r(\mathbb{P}, \tilde{\mathbb{P}})^r = \mathbf{d}^r_0(1, 1)$, where $(i_0, j_0) = (1, 1)$ is the pair of root nodes of both trees.

Algorithm 2.1 summarizes this procedure in order to efficiently compute the nested distance for tree processes in a nested, recursive manner.

Example 2.45. As an example for Algorithm 2.1 to efficiently compute the nested distance we consider the nested distributions \mathbb{P} (a tree with 11 leaves) and $\tilde{\mathbb{P}}$ (a tree with 6 leaves) shown below and depicted in Fig. 2.12.

$$\begin{aligned} \mathbb{P} &= \left[\begin{array}{cccc} 0.2 & 0.3 & 0.3 & 0.2 \\ \hline 13 & 11 & 8 & 6 \\ \left[\begin{array}{cccc} 0.1 & 0.2 & 0.4 & 0.3 \\ \hline 15 & 14 & 13 & 11 \end{array} \right] & \left[\begin{array}{cc} 0.7 & 0.3 \\ \hline 12 & 9 \end{array} \right] & \left[\begin{array}{ccc} 0.5 & 0.2 & 0.3 \\ \hline 10 & 8 & 6 \end{array} \right] & \left[\begin{array}{cc} 0.4 & 0.6 \\ \hline 7 & 5 \end{array} \right] \end{array} \right], \\ \tilde{\mathbb{P}} &= \left[\begin{array}{ccc} 0.3 & 0.3 & 0.4 \\ \hline 13 & 11 & 7 \\ \left[\begin{array}{cc} 0.4 & 0.6 \\ \hline 14 & 12 \end{array} \right] & \left[\begin{array}{c} 1.0 \\ \hline 10 \end{array} \right] & \left[\begin{array}{ccc} 0.4 & 0.4 & 0.2 \\ \hline 9 & 8 & 5 \end{array} \right] \end{array} \right]. \end{aligned}$$

The pertaining multivariate distributions and their Wasserstein distances were already considered in Example 2.28.

Initialization. Table 2.2 collects the distances of two paths of the trees (the state space). Here, the ℓ^1 -distance is employed.

Algorithm 2.1 Nested computation of the nested distance $\mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})$ of two tree-processes \mathbb{P} and $\tilde{\mathbb{P}}$

Initialization ($t = T$):

For all combinations of leaf nodes $i \in \mathcal{N}_T$ and $j \in \tilde{\mathcal{N}}_T$ with predecessors $(i_0, i_1, \dots, i_{T-1}, i)$ and $(j_0, j_1, \dots, j_{T-1}, j)$ define

$$\mathbf{d}_T^r(i, j) := \mathbf{d}\left((\xi_{i_0}, \xi_{i_1}, \dots, \xi_i), (\tilde{\xi}_{j_0}, \tilde{\xi}_{j_1}, \dots, \tilde{\xi}_j)\right)^r.$$

Iteration, backwards:

For $t = T - 1$ down to 0, and

for every combination of inner nodes $k \in \mathcal{N}_t$ and $l \in \tilde{\mathcal{N}}_t$ solve the LP (cf. (2.23))

$$\begin{aligned} \mathbf{d}_t^r(k, l) := & \underset{(\text{in } \pi)}{\text{minimize}} \quad \sum_{i' \in k+, j' \in l+} \pi(i', j'|k, l) \cdot \mathbf{d}_{t+1}^r(i', j') \\ \text{subject to} \quad & \sum_{j' \in l+} \pi(i', j'|k, l) = Q(i') \quad i' \in k+ \\ & \sum_{i' \in k+} \pi(i', j'|k, l) = \tilde{Q}(j') \quad j' \in l+ \\ & \pi(i', j'|k, l) \geq 0 \end{aligned} \quad (2.43)$$

Final Assignment:

The nested distance of the trees is the distance of the trees at their roots 1,

$$\mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r = \mathbf{d}_0^r(1, 1).$$

The optimal transport plan at the leaf nodes $i \in \mathcal{N}_T$ and $j \in \tilde{\mathcal{N}}_T$ is

$$\pi(i, j) := \pi_1(i_1, j_1 | i_0, j_0) \cdots \pi_{T-1}(i, j | i_{T-1}, j_{T-1}).$$

π_t is the optimal transport plan obtained in (2.43) at stage t .

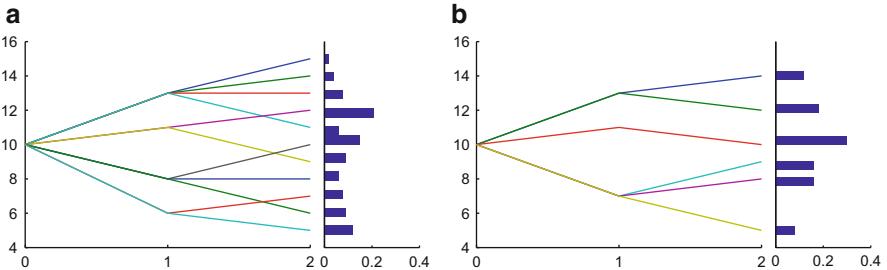


Fig. 2.12 Depicted are two trees in three stages, annotated is the histogram of their final probability distribution. (a) Initial tree \mathbb{P} . (b) Target tree $\tilde{\mathbb{P}}$

Iteration, backwards. Based on the distance matrix and the structure of the trees the respective subproblems are computed for each combination of nodes at the same stage. For stage 1, the result of the above two subtrees is displayed in Table 2.4a; the result (0.8) is the corresponding, new entry in the distance Table 2.4b at the earlier stage 0.

Table 2.4 The distance at different levels: distance \mathbf{d}_l , probabilities π , and dual variables (λ and μ ; cf. Table 2.3)

Distance of states, $\mathbf{d}_2(i, j)$	0	0	Probabilities, $\pi_2(i, j)$	0.4	0.6
1	I	3	0.1	0.1	0
0	0	2	0.2	0.2	0
1	I	I	0.4	0.1	0.3
1	3	I	0.3	0	0.3

(a) Stage 1, the first two subtrees, primal and dual solutions. The Wasserstein distance of the subtrees is $\sum_{i,j} \pi_2(i, j) \mathbf{d}_2(i, j) = 0.8$ —the first entry in Table 2.4b

Distance of subtrees, $\mathbf{d}_1(i, j)$	7.6	5.6	3	Probabilities, $\pi_1(i, j)$	0.3	0.3	0.4
-6.8	0.8	4.8	11	0.2	0.2	0	0
-3.9	3.7	1.7	7.3	0.3	0.1	0.2	0
-1	9.4	4.6	2	0.3	0	0.1	0.2
0	14	9.2	3	0.2	0	0	0.2

(b) Stage 0, the combination of all subtrees. The nested distance is $\sum_{i,j} \pi_1(i, j) \mathbf{d}_1(i, j) = 2.33$

Table 2.5 The conditional probabilities of the tree, and the probabilities of the nested distance, $\sum_{i,j} \pi_{i,j} d_{i,j} = 2.33$. The unconditional probabilities for the nested distance notably do not coincide with Wasserstein distance (Table 2.3)

Conditional probabilities		1						
		0.3		0.3	0.4			
		0.4	0.6	1	0.4	0.4	0.2	
1	0.2	0.1	0.02	0	0	0	0	
		0.2	0.04	0	0	0	0	
		0.4	0.02	0.06	0	0	0	
		0.3	0	0.06	0	0	0	
	0.3	0.7	0.04	0.03	0.14	0	0	
		0.3	0	0.03	0.06	0	0	
	0.3	0.5	0	0	0.05	0.08	0.02	
		0.2	0	0	0.02	0	0.04	
		0.3	0	0	0.03	0	0.02	
0.2	0.4	0	0	0	0.04	0.04	0	
	0.6	0	0	0	0.04	0.04	0.04	

Final Assignment. The nested distance finally is the distance of the subtrees at level 0, $\mathbf{d}_1(\mathbb{P}, \tilde{\mathbb{P}}) = 2.33$. Moreover, the transport plan π can be computed. The resulting transport plan is displayed in Table 2.5.

It should be noted that the final probabilities of any of the sub-problems can be recovered in the probability at an earlier stage. For example, the probabilities π from Table 2.4a are in the upper-left section of the matrix in Table 2.5, but multiplied with its conditional probability 10 %, which is the result from Table 2.4a.

Example 2.46. The nested distance of the trees in Fig. 1.13 (Chap. 1) is 0, a minimal transport plan is

$$\pi = \begin{pmatrix} 0.42 & 0 \\ 0 & 0.18 \\ 0.28 & 0 \\ 0 & 0.12 \end{pmatrix}.$$

This example demonstrates that the nested distance correctly identifies equivalent processes.

Example 2.47. The nested distance of the trees addressed in Example 2.30 (Fig. 2.4) is $4p(1-p) > 0$, the corresponding nested transport plan is given by

$$\pi = \begin{pmatrix} p^2 & p(1-p) \\ (1-p)p & (1-p)^2 \end{pmatrix}.$$

The nested distance thus correctly identifies two different trees in this situation (except for the degenerate cases $p = 0$ or $p = 1$, where they coincide again). This is in notable contrast to the final Wasserstein distance of these trees, which was found to be 0.

2.11 Dual Representation of the Nested Distance

The duality for the Wasserstein distance was established in Sect. 2.8 by considering LP duality for the defining optimization problem. This section provides a martingale representation of the nested distance. The martingale corresponds to successively solving Wasserstein problems on subsequent stages, the main result is Theorem 2.49 below. This (dual) representation is adapted from [94].

To prepare for the dual representation of the nested distance it is helpful to observe that one may state the dual problem (2.25) in the equivalent form

$$\begin{aligned} & \text{maximize} \\ & \quad (\text{in } M_0, \lambda, \mu) \quad M_0 \\ & \text{subject to } \mathbb{E}\lambda = 0, \tilde{\mathbb{E}}\mu = 0, \\ & \quad M_0 + \lambda(\xi) + \mu(\tilde{\xi}) \leq d(\xi, \tilde{\xi})^r \text{ for all } \xi \text{ and } \tilde{\xi}. \end{aligned}$$

The defining equations for the nested distance in Definition 2.32 involve constraints for each of the stages. As the variables in the dual program correspond to

constraints in the primal, the dual program for the nested distance involves variables for each stage.

To establish the dual representation of the nested distance it is necessary to incorporate the constraints (2.35) in the Lagrangian function. To this end we define projections, which act on one component while leaving the other unaffected, by

$$\begin{aligned}\text{proj}_t : L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T) &\rightarrow L^1(\mathcal{F}_t \otimes \tilde{\mathcal{F}}_T) \\ \lambda(\text{id}) \cdot \mu(\tilde{\text{id}}) &\mapsto \mathbb{E}(\lambda|\mathcal{F}_t)(\text{id}) \cdot \mu(\tilde{\text{id}})\end{aligned}$$

and

$$\begin{aligned}\tilde{\text{proj}}_t : L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T) &\rightarrow L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_t) \\ \lambda(\text{id}) \cdot \mu(\tilde{\text{id}}) &\mapsto \lambda(\text{id}) \cdot \mathbb{E}(\mu|\mathcal{F}_t)(\tilde{\text{id}}).\end{aligned}$$

The one-sided projections proj and $\tilde{\text{proj}}$ are well defined by linearity, because functions of the form $(x, y) \mapsto \mathbb{1}_A(x) \mathbb{1}_B(y)$ form a basis for $L^1(\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T)$.

Proposition 2.48 (Characterization of the Projection). *The measure π satisfies the marginal condition*

$$\pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = P(A|\mathcal{F}_t) \quad \text{for all } A \in \Omega \quad (2.44)$$

if and only if

$$\mathbb{E}_\pi \lambda = \mathbb{E}_\pi \text{proj}_t \lambda \quad \text{for all } \lambda \lhd \mathcal{F}_T \otimes \tilde{\mathcal{F}}_t. \quad (2.45)$$

Moreover, $\text{proj}_t(\lambda) = \mathbb{E}_\pi(\lambda | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_T)$ if π has marginal P .

Proof. Note first that the left-hand side and the right-hand side of (2.44) are probability measures, it is thus sufficient to show that

$$\int_{C \times D} \pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) d\pi = \int_{C \times D} P(A|\mathcal{F}_t)(\text{id}) d\pi$$

for all sets $C \in \mathcal{F}_t$ and $D \in \tilde{\mathcal{F}}_t$.

To this end observe that

$$\begin{aligned}\int_{C \times D} \pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) d\pi &= \mathbb{E}_\pi(\mathbb{1}_{C \times D} \mathbb{E}_\pi(\mathbb{1}_{A \times \tilde{\Omega}} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)) \\ &= \mathbb{E}_\pi \mathbb{E}_\pi(\mathbb{1}_{C \times D} \mathbb{1}_{A \times \tilde{\Omega}} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\ &= \mathbb{E}_\pi \mathbb{1}_{(C \times D) \cap (A \times \tilde{\Omega})} = \pi((A \cap C) \times D).\end{aligned}$$

Next,

$$\begin{aligned}
\int_{C \times D} P(A | \mathcal{F}_t) d\pi &= \mathbb{E}_\pi \mathbf{1}_{C \times D} \mathbb{E}_P(\mathbf{1}_A | \mathcal{F}_t) \circ \text{id} \\
&= \mathbb{E}_\pi \mathbf{1}_C(\text{id}) \cdot \mathbf{1}_D(\tilde{\text{id}}) \cdot \mathbb{E}_P(\mathbf{1}_A | \mathcal{F}_t)(\text{id}) \\
&= \mathbb{E}_\pi \mathbb{E}_P(\mathbf{1}_C \mathbf{1}_A | \mathcal{F}_t)(\text{id}) \cdot \mathbf{1}_D(\tilde{\text{id}}) \\
&= \mathbb{E}_\pi \mathbb{E}_P(\mathbf{1}_{C \cap A} | \mathcal{F}_t)(\text{id}) \cdot \mathbf{1}_D(\tilde{\text{id}}) \\
&= \mathbb{E}_\pi \text{proj}_t(\mathbf{1}_{C \cap A}(\text{id}) \cdot \mathbf{1}_D(\tilde{\text{id}})),
\end{aligned}$$

and by the assertion (2.45) thus

$$\begin{aligned}
\int_{C \times D} P(A | \mathcal{F}_t) d\pi &= \mathbb{E}_\pi \mathbf{1}_{A \cap C}(\text{id}) \cdot \mathbf{1}_D(\tilde{\text{id}}) \\
&= \mathbb{E}_\pi \mathbf{1}_{(A \cap C) \times D} = \pi((A \cap C) \times D),
\end{aligned}$$

from which the first assertion (2.44) follows.

As for the converse it is enough to prove (2.45) for functions of the form $\mu \circ \text{id} \cdot \tilde{\mu} \circ \tilde{\text{id}}$ for $\mu \in \mathcal{F}_T$ and $\tilde{\mu} \in \tilde{\mathcal{F}}_t$, as these products form a basis.

For the function $\mathbf{1}_A$ ($A \in \mathcal{F}_T$)

$$\begin{aligned}
\mathbb{E}_P(\mathbf{1}_A | \mathcal{F}_t)(\text{id}) &= P(A | \mathcal{F}_t) = \pi(A \times \tilde{\Omega} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\
&= \mathbb{E}(\mathbf{1}_{A \times \tilde{\Omega}} | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \mathbb{E}_\pi(\mathbf{1}_A(\text{id}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t),
\end{aligned}$$

and by linearity thus $\mathbb{E}_P(\mu | \mathcal{F}_t)(\text{id}) = \mathbb{E}_\pi(\mu(\text{id}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)$. With this identity it follows further that

$$\mathbb{E}_P(\mu | \mathcal{F}_t)(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}}) = \mathbb{E}_\pi(\mu(\text{id}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \cdot \tilde{\mu}(\tilde{\text{id}}) = \mathbb{E}_\pi(\mu(\text{id}) \tilde{\mu}(\tilde{\text{id}}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t).$$

Taking expectations gives that

$$\begin{aligned}
\mathbb{E}_\pi \text{proj}_t(\mu(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}})) &= \mathbb{E}_\pi \mathbb{E}_P(\mu | \mathcal{F}_t)(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}}) \\
&= \mathbb{E}_\pi \mathbb{E}_\pi(\mu(\text{id}) \tilde{\mu}(\tilde{\text{id}}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \\
&= \mathbb{E}_\pi \mu(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}}),
\end{aligned}$$

which is the assertion for the basis functions $\mu(\text{id}) \cdot \tilde{\mu}(\tilde{\text{id}})$. \square

2.11.1 Martingale Representation of the Nested Distance

Proposition 2.48 is the essential tool to describe the dual representation of the nested distance.

Theorem 2.49 (Duality for the Nested Distance). *The infimum of problem (2.35) to compute the nested distance $\mathbf{d}_r^r(\mathbb{P}, \tilde{\mathbb{P}})$ equals the supremum of all numbers M_0 such that*

$$M_T(\omega, \tilde{\omega}) \leq \mathbf{d}(\omega, \tilde{\omega})^r \quad (\omega, \tilde{\omega}) \in \Omega \times \tilde{\Omega},$$

where M_t is an \mathbb{R} -valued process on $\Omega \times \tilde{\Omega}$ of the form

$$M_t = M_0 + \sum_{s=1}^t (\lambda_s + \mu_s)$$

and the measurable functions $\lambda_t \llcorner \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}$ and $\mu_t \llcorner \mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_{t-1}$ satisfy $\text{proj}_{t-1}(\lambda_t) = 0$ and $\text{proj}_{t-1}(\mu_t) = 0$.

The process M_t , for which the supremum is attained, is a martingale with respect to the optimal measure π ,

$$M_t = \mathbb{E}(\mathbf{d}^r | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \quad \pi\text{-a.e. in } \Omega \times \tilde{\Omega}$$

and

$$\mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r = \mathbb{E}_\pi(\mathbf{d}^r) = M_0 = \mathbb{E}_\pi M_T.$$

Before we conclude this section with the proof of the theorem we give an example, which displays the martingale process.

Example 2.50 (Continuation of Example 2.45). Table 2.6 collects the final stage of the process M_T . Comparing M_T with the initial distance (Table 2.2) it becomes apparent that $M_T \leq \mathbf{d}$. Moreover it holds that $M_T = \mathbf{d}$ for all entries for which $\pi_{i,j} > 0$ (these entries are bold in the table. Cf. Table 2.5 to compare this representation by duality with the primal nested distance).

Proof of Theorem 2.49. Evoking Proposition 2.48 the constraints in (2.35) can be encoded in the Lagrangian as

$$\begin{aligned} \inf_{\pi \geq 0} \sup_{M_0, f_i, g_i} & \mathbb{E}_\pi \mathbf{d}^r + M_0 \cdot (1 - \mathbb{E}_\pi \mathbf{1}) + \\ & - \sum_{s=0}^{T-1} (\mathbb{E}_\pi f_{s+1} - \mathbb{E}_\pi \text{proj}_s(f_{s+1})) + \end{aligned}$$

Table 2.6 The final state of the dual martingale process M_T corresponding to Example 2.45

Conditional probabilities		1					
		0.3		0.3	0.4	0.4	
		0.4	0.6	1	0.4	0.4	0.2
1	0.2	0.1	I	1	1	-2.8	-1.8
		0.2	0	0	0	-3.8	-2.8
		0.4	I	I	-1	-4.8	-3.8
		0.3	1	I	-3	-6.8	-5.8
	0.3	0.7	4	2	2	-1.2	-0.2
		0.3	7	5	I	-4.2	-3.2
	0.3	0.5	6.2	4.2	3	2	3
		0.2	8.2	6.2	5	0	I
		0.3	10.2	8.2	7	2	3
	0.2	0.4	7.6	5.6	4.4	3	2
		0.6	9.6	7.6	6.4	5	4
							I

$$-\sum_{s=0}^{T-1} (\mathbb{E}_\pi g_{s+1} - \mathbb{E}_\pi \tilde{\text{proj}}_s(g_{s+1})) ,$$

the infimum being among positive measures $\pi \geq 0$, not only probability measures; the functions in the inner supremum satisfy $f_t \triangleleft \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}$ and $g_t \triangleleft \mathcal{F}_{t-1} \otimes \tilde{\mathcal{F}}_t$. According to Sion's minimax theorem (cf. Sion [132]) this saddle point has the same objective value as

$$\sup_{M_0, f_t, g_t} M_0 + \inf_{\pi \geq 0} \mathbb{E}_\pi \begin{bmatrix} \mathbf{d}^r - M_0 \cdot \mathbf{1} \\ -\sum_{s=0}^{T-1} (f_{s+1} - \text{proj}_s(f_{s+1})) \\ -\sum_{s=0}^{T-1} (g_{s+1} - \tilde{\text{proj}}_s(g_{s+1})) \end{bmatrix}. \quad (2.46)$$

Now notice that the infimum over all $\pi \geq 0$ is $-\infty$ unless the integrand is positive for every measure $\pi \geq 0$, which means that

$$M_0 + \sum_{s=0}^{T-1} (f_{s+1} - \text{proj}_s(f_{s+1})) + \sum_{s=0}^{T-1} (g_{s+1} - \tilde{\text{proj}}_s(g_{s+1})) \leq \mathbf{d}^r$$

necessarily has to hold. For a positive integrand in (2.46), the infimum over positive measures $\inf_{\pi \geq 0} \mathbb{E}_\pi$ is 0. Equation (2.46) thus can be reformulated as

$$\begin{aligned} & \text{maximize} && M_0 \\ & \text{in } M_0, f_t, g_t && \\ & \text{subject to} && M_0 + \sum_{s=0}^{T-1} (f_{s+1} - \text{proj}_s(f_{s+1})) + \sum_{s=0}^{T-1} (g_{s+1} - \tilde{\text{proj}}_s(g_{s+1})) \leq \mathbf{d}^r, \\ & && f_t \triangleleft \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}, g_t \triangleleft \mathcal{F}_t \otimes \tilde{\mathcal{F}}_{t-1}. \end{aligned}$$

Using the setting $\lambda_s := f_s - \text{proj}_{s-1}(f_s)$ and $\mu_s := g_s - \text{proj}_{s-1}(g_s)$ allows rewriting the latter equation as

$$\begin{aligned} & \text{maximize (in } M_0, \lambda_t, \mu_t) \ M_0 \\ & \text{subject to} \quad M_0 + \sum_{s=1}^T (\lambda_s + \mu_s) \leq \mathbf{d}^r \\ & \quad \text{proj}_{t-1}(\lambda_t) = 0, \ \text{proj}_{t-1}(\mu_t) = 0, \end{aligned}$$

which is the desired formulation.

To accept the martingale property observe that

$$\mathbb{E}_\pi (M_T | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = \mathbb{E}_\pi \left(M_0 + \sum_{s=0}^{T-1} \lambda_s + \mu_s | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right) = M_0 + \sum_{s=0}^t \lambda_s + \mu_s,$$

provided that $\mathbb{E}_\pi (\lambda_s + \mu_s | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = 0$ whenever $s > t$. This can be seen as follows: recall that $\lambda(\text{id}) \cdot \mathbb{1}_B(\tilde{\text{id}})$ are base functions. Now if

$$\text{proj}_t \lambda(\text{id}) \cdot \mathbb{1}_B(\tilde{\text{id}}) = \mathbb{E}(\lambda | \mathcal{F}_t)(\text{id}) \cdot \mathbb{1}_B(\tilde{\text{id}}) = 0,$$

then $\mathbb{E}(\lambda | \mathcal{F}_t) = 0$ and consequently $\mathbb{E}(\lambda(\text{id}) \tilde{\lambda}(\text{id}) | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) = 0$.

It holds moreover that

$$M_t = \mathbb{E}_\pi (M_T | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \leq \mathbb{E}_\pi (\mathbf{d}^r | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t)$$

due to the constraints. But the vanishing duality gap forces

$$M_0 = \mathbb{E}_\pi \mathbf{d}^r,$$

such that we conclude in addition that

$$M_t = \mathbb{E}_\pi (\mathbf{d}^r | \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t) \quad \pi - \text{a.e.},$$

which completes the proof. \square

Chapter 3

Risk and Utility Functionals

*How bad is bad?*¹

Risk functionals are designed to answer this question by measuring risk on a scale. They summarize all possible random outcomes in a single real number and thus allow comparing different random outcomes from a risk perspective.

Risk functionals have gained increasing attention in different fields in the past. Investigations, in particular in mathematical finance and operations research, have led to a significant conceptual progress and understanding. In addition to that risk functionals are nowadays part of the insurance industry regulations in Canada and the US, as well as in banks following the Basel directives.

In insurance, for example, all potential future claims have to be aggregated and summarized in a single premium. The premium, however, is due already at the very beginning of the contract, not knowing the future realizations of the contract. A risk functional can be employed in this situation to provide a reasonable lump-sum premium.

In this chapter, risk and utility functionals are introduced both in the unconditional and in the conditional form. A summary of the definitions is contained in Appendix A. The extension to multiperiod functionals is also presented there, although these extensions are not so relevant for the purpose of this book.

3.1 Single-Period Risk and Utility Functionals

A probability space (Ω, \mathcal{F}, P) and a vector space \mathcal{Y} of real valued, \mathcal{F} -measurable random variables is considered.

¹cf. Artzner et al. [7].

Definition 3.1. A single-stage probability functional assigns an extended real value to the elements of \mathcal{Y} . We distinguish between

- risk functionals: $\mathcal{R} : \mathcal{Y} \rightarrow \mathbb{R} \cup \{\infty\}$,
(the value $+\infty$ means unacceptable risk), and
- utility (or acceptability) functionals: $\mathcal{U} : \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty\}$
(the value $-\infty$ means absolute unacceptability).

An interpretation, which is often offered in connection with risk functionals, considers losses of a financial portfolio. Any real valued random variable $Y \in \mathcal{Y}$ represents a loss, which has a distribution specified by P . From an insurers perspective, for example, it is the aim then to keep claims (losses) as small as possible. If, conversely, $Y \in \mathcal{Y}$ represents a revenue or a profit, then $-Y$ represents a loss again, such that these concepts can be exchanged against each other. Thus by changing the sign, a loss or cost variable becomes a profit variable and vice versa. Similarly, by setting

$$\mathcal{U}(Y) = -\mathcal{R}(Y),$$

to each risk functional there corresponds an utility functional and vice versa. In this book we assume (if not otherwise specified) as standard that Y represents costs or losses, and the goal is to make its risk $\mathcal{R}(Y)$ as small as possible.

Practitioners prefer the name risk measure over the more mathematical term *risk functional*. When no confusion with the term probability measure is possible, we also use sometimes the term *risk measure*.

Important axioms on risk functionals have been outlined and described in the seminal paper [6] by Artzner et al. In insurance, respective axioms have been considered much earlier, for example by Wang [139], Denneberg [25] and by Gerber and Deprez [26].

Definition 3.2 (Axioms for Risk Functionals). A positively homogeneous *risk functional* (risk measure) is a mapping $\mathcal{R} : \mathcal{Y} \rightarrow \mathbb{R} \cup \{\infty\}$ with the following properties:

- (M) MONOTONICITY: $\mathcal{R}(Y_1) \leq \mathcal{R}(Y_2)$ whenever $Y_1 \leq Y_2$ almost surely;
- (C) CONVEXITY: $\mathcal{R}((1-\lambda)Y_0 + \lambda Y_1) \leq (1-\lambda)\mathcal{R}(Y_0) + \lambda\mathcal{R}(Y_1)$ for $0 \leq \lambda \leq 1$;
- (T) TRANSLATION EQUIVARIANCE: $\mathcal{R}(Y + c) = \mathcal{R}(Y) + c$ if $c \in \mathbb{R}$;
- (H) POSITIVE HOMOGENEITY: $\mathcal{R}(\lambda Y) = \lambda\mathcal{R}(Y)$ whenever $\lambda > 0$.

If only the properties (M), (C), and (T) are fulfilled, the functional \mathcal{R} is often simply called *risk functional* as well, while positively homogeneous risk functionals satisfying (H) are also called *coherent* risk measures. Here we shall mention explicitly, if a risk functional is not positively homogeneous.

3.2 Examples of Risk and Utility Functionals

Expectation. The simplest example of a risk functional is the expectation,

$$\mathcal{R}(Y) = \mathbb{E}(Y) = \int Y dP,$$

it satisfies all above relations (M), (C), (T), and (H).

Average Value-at-Risk. An extension of the expectation is the Average Value-at-Risk.

Definition 3.3 (The (Upper) Average Value-at-Risk). The (*upper*) Average Value-at-Risk at level α ($\alpha \in [0, 1]$) is

$$\text{AV@R}_\alpha(Y) := \frac{1}{1-\alpha} \int_\alpha^1 V@R_p(Y) dp, \quad (3.1)$$

where

$$V@R_\alpha(Y) := \inf \{y : P(Y \leq y) \geq \alpha\} \quad (3.2)$$

is the *Value-at-Risk* (the left-continuous, lower semicontinuous *quantile*, or *lower inverse cdf*) at level α . It is denoted by $G_Y^{-1}(\alpha)$ as well, with $G_Y(y) = P\{Y \leq y\}$.

The *max-risk functional* is the Average Value-at-Risk at level $\alpha = 1$, defined as

$$\text{AV@R}_1(Y) := \lim_{\alpha \nearrow 1} \text{AV@R}_\alpha(Y) = \text{ess sup}(Y).$$

The essential supremum $\text{ess sup}(Y) = \sup \{y : G_Y(y) < 1\}$ is also called the max-risk functional. The Average Value-at-Risk (cf. (1.8) in the introduction) is well defined by (3.1) for $0 \leq \alpha < 1$ and finite valued for $Y \in L^1$. To insure that $\text{AV@R}_1(Y)$ is finite valued it is frequently assumed that $Y \in L^\infty$.

Remark 3.4. The Average Value-at-Risk is called *upper* Average Value-at-Risk, because (3.1) involves the upper quantiles only, from α up to 1. The addition *upper* will usually be suppressed throughout this book.

The functional $\text{AV@R}_\alpha(\cdot)$ is a risk functional for every $\alpha \in [0, 1]$ fixed, it satisfies all axioms (M), (C), (T), and (H). Notably, the Value-at-Risk ($V@R$) fails to be convex (property (C)) and thus $V@R$ is *not* a risk functional.

The Average Value-at-Risk at level $\alpha = 0$ is the expectation, because

$$\text{AV@R}_0(Y) = \int_0^1 G_Y^{-1}(u) du = \mathbb{E}(Y).$$

The Average Value-at-Risk has many further representations. Important forms include the formula

$$\text{AV@R}_\alpha(Y) = \min_{q \in \mathbb{R}} \quad q + \frac{1}{1-\alpha} \mathbb{E}(Y - q)_+ \quad (3.3)$$

developed in Rockafellar and Uryasev [111] and in Pflug [89]. Of importance moreover is the dual representation

$$\text{AV@R}_\alpha(Y) = \max_{Z \in L^\infty} \{ \mathbb{E}(YZ) : \mathbb{E}(Z) = 1, 0 \leq Z, (1-\alpha)Z \leq \mathbb{1} \} \quad (3.4)$$

$$= \max_{Q \in \mathcal{P}(\Omega)} \left\{ \mathbb{E}_Q(Y) : \frac{dP}{dQ} \leq \frac{1}{1-\alpha} \right\}, \quad (3.5)$$

where (3.5) is referred to as change of measure, sometimes as change of numéraire. If $G_Y(G_Y^{-1}(\alpha)) = \alpha$, then

$$\text{AV@R}_\alpha(Y) = \mathbb{E}(Y | Y \geq \text{V@R}_\alpha(Y)). \quad (3.6)$$

Equation (3.6) gives rise calling the Average Value-at-Risk *Conditional Value-at-Risk*, as is done in many publications. However, formula (3.6) holds true only for continuous distributions, but not necessarily for discrete distributions.

In insurance the term *Conditional Tail Expectation* (CTE) is in more frequent use for the risk functional AV@R.

For the proofs of the representations (3.3)–(3.6) and for further discussions we refer to Pflug and Römisch [97].

The Level Parameter. In its level parameter α , the Average Value-at-Risk is a nondecreasing function; that is the mapping

$$\alpha \mapsto \text{AV@R}_\alpha(Y)$$

is nondecreasing.

Lemma 3.5 (Concavity). *For every $Y \in L^1$ the mapping (Lorentz curve)*

$$\alpha \mapsto (1-\alpha) \cdot \text{AV@R}_\alpha(Y)$$

is concave and continuous in the closed interval $[0, 1]$.

Proof. This is evident, because

$$(1-\alpha) \cdot \text{AV@R}_\alpha(Y) = \int_\alpha^1 G_Y^{-1}(u) du = \mathbb{E}(Y) - \int_0^\alpha G_Y^{-1}(u) du$$

and $G_Y^{-1}(\cdot)$ is nondecreasing.

Continuity in the open interval $(0, 1)$ follows from concavity of the function $\alpha \mapsto (1 - \alpha) \cdot \text{AV@R}_\alpha(Y)$, and continuity at both boundary points 0 and 1 follows from the fact that $Y \in L^1$. \square

Distortion Risk Functionals. *Distortion risk functionals* were introduced by Denneberg [24, 25] and further developed by Acerbi [1, 2] under the name *spectral risk functionals* or *spectral risk measures*. Distortion risk functionals generalize the Average Value-at-Risk (cf. Pflug and Römisch [91] for an introductory discussion).

Definition 3.6 (Distortion Risk Functional). For $\sigma : [0, 1] \rightarrow [0, \infty)$ a nonnegative, nondecreasing function satisfying $\int_0^1 \sigma(u) du = 1$ the functional

$$\mathcal{R}_\sigma(Y) := \int_0^1 \sigma(u) G_Y^{-1}(u) du = \int_0^1 \sigma(u) \text{V@R}_u(Y) du \quad (3.7)$$

is called *distortion risk functional*. σ is called *distortion function*, or *distortion density*.

The distortion risk functional is a risk functional. Indeed, translation equivariance (T) holds because $\int_0^1 \sigma(u) du = 1$, monotonicity (M) is ensured by the assumption $\sigma \geq 0$, positive homogeneity (H) is automatically satisfied for functionals with representation (3.7), since V@R is positively homogeneous and convexity (C) follows by assuming that σ is nondecreasing. For a comprehensive treatment of distortion risk functionals we refer to Pichler [101].

Remark 3.7. The name distortion is somehow suggestive: indeed, consider the function $G_\sigma(\alpha) := \int_0^\alpha \sigma(u) du$. As σ is a density on the unit interval $[0, 1]$, G_σ is a cdf allowing a generalized inverse $G_\sigma^{-1}(\cdot)$ according to (3.2). Then one may consider the distorted random variable

$$Y_\sigma := G_Y^{-1}(G_\sigma^{-1}(U)), \quad (3.8)$$

where U is a uniformly distributed random variable² for which $Y = G_Y^{-1}(U)$ (Y 's probability integral transform). Notice that $G_\sigma(0) = 0$ and $G_\sigma(1) = 1$. Further G_σ is convex, because σ is nondecreasing, and thus $G_\sigma(\alpha) \leq \alpha$. It follows that

$$Y_\sigma \geq Y,$$

and it holds thus that

$$\begin{aligned} \mathbb{E}(Y) &\leq \mathbb{E}(Y_\sigma) = \int_0^1 G_Y^{-1}(G_\sigma^{-1}(u)) du \\ &= \int_0^1 G_Y^{-1}(u) dG_\sigma(u) = \int_0^1 G_Y^{-1}(u) \sigma(u) du = \mathcal{R}_\sigma(Y). \end{aligned}$$

²A random variable $U : \Omega \rightarrow \mathbb{R}$ is uniformly distributed if $P(U \leq u) = u$ whenever $u \in [0, 1]$.

This demonstrates that $\mathcal{R}_\sigma(Y)$ is just the expectation of the *distorted* random variable Y_σ . The distorted random variable Y_σ —in view of (3.8)—has the same outcomes as Y , but its probabilities are distorted according to

$$G_{Y_\sigma}(y) = P(Y_\sigma \leq y) = P(U \leq G_\sigma(G_Y(y))) = G_\sigma(G_Y(y)) \leq G_Y(y). \quad (3.9)$$

This is the defining equation for the first order stochastic dominance relation (cf. the books by Stoyan and Müller [80], and by Shaked and Shanthikumar [123] on stochastic dominance, as well as the survey [49] by Gutjahr and Pichler).

Notice as well that the density g_{Y_σ} of the distorted random variable Y_σ (if it exists) satisfies

$$g_{Y_\sigma}(y) = g_Y(y) \cdot \sigma(G_Y(y)),$$

which follows by differentiating (3.9).

Representation in Terms of the Average Value-at-Risk. The Average Value-at-Risk is a distortion risk functional itself, as obviously $\mathcal{R}_{\sigma_\alpha}(Y) = \text{AV@R}_\alpha(Y)$ for the distortion density

$$\sigma_\alpha(u) = \begin{cases} 0 & \text{if } u < \alpha \\ \frac{1}{1-\alpha} & \text{if } u \geq \alpha. \end{cases} \quad (3.10)$$

A representation of general distortion risk functional in terms of the Average Value-at-Risk, that is the form

$$\mathcal{R}_\sigma(Y) = \int_0^1 \text{V@R}_\alpha(Y) \sigma(\alpha) d\alpha = \int_0^1 \text{AV@R}_\alpha(Y) \mu_\sigma(d\alpha) \quad (3.11)$$

for a probability measure μ_σ on the interval $[0, 1]$, is of elementary importance. To develop the equivalent representation (3.11) one must associate with σ the measure

$$\mu_\sigma(A) := \sigma(0) \cdot \delta_0(A) + \int_A (1-u) d\sigma(u) \quad (A \subset [0, 1]) \quad (3.12)$$

on the unit interval $[0, 1]$. By Riemann–Stieltjes integration by parts it is obvious that

$$\begin{aligned} 0 \leq \mu_\sigma(A) \leq \mu_\sigma([0, 1]) &= \sigma(0) + \int_0^1 (1-u) d\sigma(u) \\ &= \sigma(0) + (1-u)\sigma(u)|_{u=0}^1 - \int_0^1 \sigma(u) d(1-u) = \int_0^1 \sigma(u) du = 1, \end{aligned}$$

such that μ_σ is a probability measure on the interval $[0, 1]$. For this measure it holds that

$$\begin{aligned}
\int_0^1 \text{AV}@\mathbf{R}_\alpha(Y) \mu_\sigma(d\alpha) &= \sigma(0) \cdot \text{AV}@\mathbf{R}_0(Y) + \int_0^1 (1-u) \text{AV}@\mathbf{R}_u(Y) d\sigma(u) \\
&= \sigma(0) \cdot \text{AV}@\mathbf{R}_0(Y) + \int_0^1 \int_u^1 \mathbf{V}@\mathbf{R}_\alpha(Y) d\alpha d\sigma(u) \\
&= \sigma(0) \cdot \text{AV}@\mathbf{R}_0(Y) + \sigma(u) \cdot \int_u^1 \mathbf{V}@\mathbf{R}_\alpha(Y) d\alpha \Big|_{u=0}^1 \\
&\quad - \int_0^1 \sigma(u) d \int_u^1 \mathbf{V}@\mathbf{R}_\alpha(Y) d\alpha \\
&= \int_0^1 \sigma(u) \mathbf{V}@\mathbf{R}_u(Y) du = \mathcal{R}_\sigma(Y),
\end{aligned}$$

which indeed establishes the representation (3.11).

The converse relation between the measure μ and the distortion functional σ is provided by

$$\sigma_\mu(p) := \int_0^p \frac{1}{1-u} \mu(du). \quad (3.13)$$

σ_μ is obviously nonnegative, nondecreasing, and provided that $\mu(\{1\}) = 0$ it holds that

$$\begin{aligned}
\int_0^1 \sigma_\mu(p) dp &= \int_0^1 \int_0^p \frac{1}{1-u} \mu(du) dp = \int_0^1 \int_u^1 \frac{1}{1-u} dp \mu(du) \\
&= \int_0^1 \mu(du) = \mu([0, 1]) = 1.
\end{aligned}$$

The essential relation is

$$\begin{aligned}
\mathcal{R}_{\sigma_\mu}(Y) &= - \int_0^1 \sigma_\mu(u) d \int_u^1 \mathbf{V}@\mathbf{R}_\alpha(Y) d\alpha \\
&= - \sigma_\mu(u) \cdot \int_u^1 \mathbf{V}@\mathbf{R}_\alpha(Y) d\alpha \Big|_{u=0}^1 + \int_0^1 \int_u^1 \mathbf{V}@\mathbf{R}_\alpha(Y) d\alpha d\sigma_\mu(u) \\
&= \sigma_\mu(0) \cdot \int_0^1 \mathbf{V}@\mathbf{R}_\alpha(Y) d\alpha + \int_0^1 \int_u^1 \mathbf{V}@\mathbf{R}_\alpha(Y) d\alpha \cdot \frac{1}{1-u} \mu(du) \\
&= \int_0^1 \text{AV}@\mathbf{R}_u(Y) \mu(du),
\end{aligned}$$

which provides the required identity (3.11).

Monotonic Properties of Distortion Risk Functionals.

Definition 3.8 (Second Order Stochastic Dominance). A random variable Y_1 is dominated by the random variable Y_2 in the second order sense, if for all monotonically increasing and concave utility functions U

$$\mathbb{E}[U(Y_1)] \leq \mathbb{E}[U(Y_2)],$$

whenever the expectations are well defined. The symbol $Y_1 \prec_{\text{SSD}} Y_2$ is used for this relation.

Risk functionals may exhibit monotonicity properties with respect to second order stochastic dominance. For utility functionals \mathcal{U} the simple relation $Y_1 \prec_{\text{SSD}} Y_2 \implies \mathcal{U}(Y_1) \leq \mathcal{U}(Y_2)$ may hold for profit variables Y . For risk functionals \mathcal{R} applied to cost or loss variables Y , the analogous property must be expressed differently, since $-Y$ is the corresponding profit variable and $\mathcal{U} = -\mathcal{R}$ is the corresponding utility functional.

Definition 3.9. A risk functional \mathcal{R} is called *antimonotonic with respect to negative SSD*, if

$$-Y_1 \prec_{\text{SSD}} -Y_2 \text{ implies that } \mathcal{R}(Y_1) \geq \mathcal{R}(Y_2).$$

Notice that $-Y_1 \prec_{\text{SSD}} -Y_2$ is not the same as $Y_2 \prec_{\text{SSD}} Y_1$.³

Lemma 3.10. *Distortion functionals are antimonotonic with respect to negative SSD.*

Proof. In view of the representation (3.11) it suffices to prove this property for the upper AV@R. The AV@R can be written as the minimum of the expectation of the convex monotonic functions $y \mapsto V_q(y) := q + \frac{1}{1-\alpha}(y-q)_+$ by (3.3). Notice that $y \mapsto -V_q(-y)$ is concave and monotonic, which implies the assertion. \square

Further Risk Functionals. Further risk functionals can be constructed by maximization. Indeed, it is easily observed that

$$\mathcal{R}(Y) := \sup_{\mu \in \mathcal{M}} \mathcal{R}_\mu(Y)$$

is a risk functional, provided that any \mathcal{R}_μ is a risk functional.

This gives rise to the following definition.

Definition 3.11. We shall say that a risk functional \mathcal{R} is *generated by the distortions* \mathcal{S} , if

³As a counterexample, let $Y_1 \sim N(\mu, \sigma_1^2)$ and $Y_2 \sim N(\mu, \sigma_2^2)$ with $\sigma_1^2 > \sigma_2^2$. Then $-Y_1 \prec_{\text{SSD}} -Y_2$, but also $Y_1 \prec_{\text{SSD}} Y_2$.

$$\mathcal{R}(Y) = \sup_{\sigma \in \mathcal{S}} \mathcal{R}_\sigma(Y),$$

where \mathcal{S} is a set of distortion functions, that is, all \mathcal{R}_σ 's are distortion risk functionals.

3.3 Dual Representation of Risk Functionals

All risk functionals introduced in Sect. 3.2 return the same result for random variables, which share the same law. This gives rise to the notion of version independence.

Definition 3.12 (Version Independence, cf. Appendix A). A probability functional is called *version independent*,⁴ if its result $\mathcal{R}(Y)$ ($\mathcal{U}(Y)$, resp.) depends on the distribution of Y only:

$$\mathcal{R}(Y) = \mathcal{R}(Y'), \text{ if } P(Y \leq y) = P(Y' \leq y) \text{ for all } y \in \mathbb{R}.$$

A similar notion exists for multiperiod functionals: they are version independent if their value only depends on the nested distribution (see Appendix A.2).

3.3.1 Kusuoka's Representation

The main result about version independent risk functionals is Kusuoka's representation.

Theorem 3.13 (Kusuoka's Theorem). *Suppose that $\mathcal{R} : L^\infty \rightarrow \mathbb{R}$ is a version independent risk functional on a probability space without atoms. Then \mathcal{R} has the representation*

$$\mathcal{R}(Y) = \sup_{\mu \in \mathcal{M}} \int_0^1 \text{AV@R}_\alpha(Y) \mu(d\alpha), \quad (3.14)$$

where \mathcal{M} is a set of probability measures on $[0, 1]$.

Proof. Cf. Kusuoka [73], and Jouini et al. [63]. □

Corollary 3.14. *Suppose that $\mathcal{R} : L^\infty \rightarrow \mathbb{R}$ is a version independent risk functional on a probability space without atoms. Then \mathcal{R} has the representation*

⁴The terms *law invariant* or *distribution based* are in frequent use as well.

$$\mathcal{R}(Y) = \sup_{\sigma \in \mathcal{S}} \mathcal{R}_\sigma(Y),$$

where \mathcal{S} is a collection of continuous and bounded distortion densities. In view of Definition 3.11, \mathcal{R} is generated by continuous and bounded distortion densities.

Proof. Let \mathcal{M} be the set of probability measures in Kusuoka's representation. Define

$$\mathcal{M}^b := \{\mu_c : \mu \in \mathcal{M}, 0 < c < 1\}, \text{ where } \mu_c(A) := \mu(A \cap [0, c]) + \mu([c, 1]) \cdot \delta_c(A),$$

and observe that, for $Y \in L^\infty$,

$$\begin{aligned} \mathcal{R}(Y) &\geq \int_0^1 \text{AV@R}_\alpha(Y) d\mu(\alpha) \\ &\geq \int_0^1 \text{AV@R}_\alpha(Y) d\mu_c(\alpha) \xrightarrow{c \rightarrow 1} \int_0^1 \text{AV@R}_\alpha(Y) d\mu(\alpha) \end{aligned}$$

because $\alpha \mapsto \text{AV@R}_\alpha(Y)$ is increasing and $\text{AV@R}_1(Y) = \lim_{\alpha \nearrow 1} \text{AV@R}_\alpha(Y)$. It follows that

$$\mathcal{R}(Y) = \sup_{\mu \in \mathcal{M}} \int_0^1 \text{AV@R}_\alpha(Y) d\mu(\alpha) = \sup_{\mu \in \mathcal{M}^b} \int_0^1 \text{AV@R}_\alpha(Y) d\mu(\alpha).$$

Every measure $\mu \in \mathcal{M}^b$ allows a distortion function according to (3.13), as $\mu_c(\{1\}) = 0$. The associated distortion function σ , however, is possibly not continuous.

But it is possible to find a continuous distortion function σ^ε such that

$$\int_u^1 \sigma(p) dp - \varepsilon \leq \int_u^1 \sigma^\varepsilon(p) dp \leq \int_u^1 \sigma(p) dp$$

for all $u \in (0, 1)$. Then it holds that

$$\begin{aligned} \mathcal{R}_\sigma(Y) &= \int_0^1 G_Y^{-1}(u) \sigma(u) du = - \int_0^1 G_Y^{-1}(u) d \int_u^1 \sigma(p) dp \\ &= - G_Y^{-1}(u) \cdot \int_u^1 \sigma(p) dp \Big|_{u=0}^1 + \int_0^1 \int_u^1 \sigma(p) dp dG_Y^{-1}(u). \end{aligned}$$

Note now that G_Y^{-1} is an increasing function and $\int_u^1 \sigma(p) dp$ is nonnegative, such that

$$\begin{aligned} \mathcal{R}_\sigma(Y) &\leq G_Y^{-1}(0) + \int_0^1 \varepsilon d G_Y^{-1}(u) + \int_u^1 \sigma^\varepsilon(p) dp dG_Y^{-1}(u) \\ &\leq 2\varepsilon \|Y\|_\infty + G_Y^{-1}(0) + \int_0^1 \int_u^1 \sigma^\varepsilon(p) dp dG_Y^{-1}(u). \end{aligned}$$

One may repeat this computation to arrive at $\mathcal{R}_{\sigma^\varepsilon}(Y) \leq \mathcal{R}_\sigma(Y) \leq 2\varepsilon \|Y\|_\infty + \mathcal{R}_{\sigma^\varepsilon}(Y)$, which finally establishes the desired approximation by a continuous and bounded function σ^ε . \square

The following corollary offers an alternative way of evaluating the risk functional \mathcal{R} , which involves the cumulative distribution function G_Y itself instead of its inverse G_Y^{-1} , provided that Y is nonnegative (the payoff of an insurance contract, to give an example, is always nonnegative, $Y \geq 0$).

Corollary 3.15. *If $Y \geq 0$, then*

$$\mathcal{R}(Y) = \sup_{\sigma \in S} \int_0^\infty H_\sigma(G_Y(q)) dq,$$

where $H_\sigma(u) = \int_u^1 \sigma(p) dp$ and $G_Y(q) = P(Y \leq q)$ is the cdf of Y .

Proof. By Riemann–Stieltjes integration by parts and change of variable it holds that

$$\begin{aligned} \int_0^1 \sigma(u) G_Y^{-1}(u) du &= - \int_0^1 G_Y^{-1}(u) dH_\sigma(u) \\ &= -G_Y^{-1}(u) H_\sigma(u) \Big|_{u=0}^1 + \int_0^1 H_\sigma(u) dG_Y^{-1}(u) \\ &= \int_0^\infty H_\sigma(G_Y(q)) dq, \end{aligned}$$

from which the assertion is immediate. \square

3.3.2 The Dual Representation

Risk functionals are convex functionals, hence they admit a dual representation in terms of the Fenchel–Moreau transform (cf. Rockafellar [110], also known as Legendre transformation). The following statements provide the conjugate function for distortion risk functionals first, the dual representation for general risk functionals can be derived from that. This representation, as well as the alternative representation in the following Sect. 3.4, is adapted from [102].

3.3.2.1 Distortion Risk Functionals

Theorem 3.16 (Dual Representation of Distortion Risk Functionals). *Let \mathcal{R}_σ be a distortion risk functional. Then \mathcal{R}_σ has the representation*

$$\mathcal{R}_\sigma(Y) = \sup \left\{ \mathbb{E}(Y \cdot Z) \mid \begin{array}{l} \mathbb{E}(Z) = 1, \text{ and} \\ \text{AV@R}_\alpha(Z) \leq \frac{1}{1-\alpha} \int_\alpha^1 \sigma(u) du \text{ for all } \alpha \in [0, 1] \end{array} \right\}. \quad (3.15)$$

Remark 3.17. We note that \mathcal{R}_σ can be stated equally well as

$$\mathcal{R}_\sigma(Y) = \sup_Z \mathbb{E}(YZ) - \mathcal{R}_\sigma^*(Z) = \sup_{Z \preccurlyeq \sigma} \mathbb{E}(YZ),$$

where

$$\mathcal{R}_\sigma^*(Z) := \begin{cases} 0 & \mathbb{E}(Z) = 1, \text{ and } Z \preccurlyeq \sigma \\ +\infty & \text{else} \end{cases}$$

is called *dual* or *conjugate function* and the binary relation $Z \preccurlyeq \sigma$ (cf. Shapiro [126]) is defined as

$$Z \preccurlyeq \sigma : \iff \begin{cases} \mathbb{E}(Z) = 1 \text{ and} \\ \text{AV@R}_\alpha(Z) \leq \frac{1}{1-\alpha} \int_\alpha^1 \sigma(u) du \text{ for all } \alpha \in [0, 1]. \end{cases} \quad (3.16)$$

Proof. We shall employ the Fenchel–Moreau theorem which states that, for \mathcal{R}_σ lower semicontinuous,

$$\mathcal{R}_\sigma(Y) = \sup_{Z \in L^1} \mathbb{E}(Y \cdot Z) - \mathcal{R}_\sigma^*(Z),$$

where

$$\mathcal{R}_\sigma^*(Z) = \sup_{Y \in L^\infty} \mathbb{E}(Y \cdot Z) - \mathcal{R}_\sigma(Y).$$

Above all, the condition $E(Z) = 1$ is necessary because of translation equivariance (T).

Moreover it holds that

$$\begin{aligned} \mathcal{R}_\sigma^*(Z) &= \sup_{Y \in L^\infty} \mathbb{E}(YZ) - \mathcal{R}_\sigma(Y) = \sup_{Y \in L^\infty} \int_0^1 G_Y^{-1}(u) G_Z^{-1}(u) du \\ &\quad - \int_0^1 G_Y^{-1}(u) \sigma(u) du \end{aligned}$$

by Chebyshev's sum inequality (sometimes also Hardy–Littlewood or simply rearrangement inequality), and hence

$$\mathcal{R}_\sigma^*(Z) = \sup_{Y \in L^\infty} \int_0^1 G_Y^{-1}(u) (G_Z^{-1}(u) - \sigma(u)) du. \quad (3.17)$$

Now choose any measurable set B and consider $Y_B := c \cdot \mathbb{1}_{[0,1] \setminus B}$, for which $G_{Y_B}^{-1}(\cdot) = c \cdot \mathbb{1}_{[P(B),1]}(\cdot)$. It follows from (3.17) further that

$$\mathcal{R}_\sigma^*(Z) \geq \sup_{c>0} c \cdot \int_{P(B)}^1 G_Z^{-1}(u) - \sigma(u) du = \begin{cases} +\infty & \text{if } \int_{P(B)}^1 G_Z^{-1}(u) - \sigma(u) du > 0 \\ 0 & \text{if } \int_{P(B)}^1 G_Z^{-1}(u) - \sigma(u) du \leq 0. \end{cases}$$

The relation $\int_{P(B)}^1 G_Z^{-1}(u) - \sigma(u) du \leq 0$ is equivalent to $(1 - \alpha) \text{AV@R}_{P(B)}(Z) \leq \int_{P(B)}^1 \sigma(u) du$. As the set B was chosen arbitrarily it readily follows that

$$\text{AV@R}_\alpha(Z) \leq \frac{1}{1 - \alpha} \int_\alpha^1 \sigma(u) du \quad \text{for all } \alpha \in (0, 1)$$

is a necessary condition.

Notice now, by integration by parts,

$$\begin{aligned} \mathcal{R}_\sigma^*(Z) &= \sup_{Y \in L^\infty} \int_0^1 G_Y^{-1}(\alpha) (G_Z^{-1}(\alpha) - \sigma(\alpha)) d\alpha \\ &= \sup_{Y \in L^\infty} - \int_0^1 G_Y^{-1}(\alpha) d \int_\alpha^1 G_Z^{-1}(u) - \sigma(u) du \\ &= \sup_{Y \in L^\infty} G_Y^{-1}(\alpha) \left(\int_\alpha^1 G_Z^{-1}(u) - \sigma(u) \right) \Big|_{\alpha=0}^1 \\ &\quad + \int_0^1 \int_\alpha^1 G_Z^{-1}(u) - \sigma(u) du dG_Y^{-1}(\alpha) \\ &= \sup_{Y \in L^\infty} G_Y^{-1}(0) (1 - \mathbb{E}(Z)) \\ &\quad + \int_0^1 \left((1 - \alpha) \text{AV@R}_\alpha(Z) - \int_\alpha^1 \sigma(u) du \right) dG_Y^{-1}(\alpha). \end{aligned}$$

It follows from this representation that the conditions $\mathbb{E}(Z) = 1$ and $(1 - \alpha) \text{AV@R}_\alpha(Z) \leq \int_\alpha^1 \sigma(u) du$ for all $\alpha \in [0, 1]$ are sufficient as well, as G_Y^{-1} is nondecreasing. \square

The following corollary, the dual characterization of the Average Value-at-Risk, is an immediate consequence. The notable point is that the constraints can be relaxed, they have to be insured only for the limited number of levels contained in $[\alpha, 1]$.

Corollary 3.18. *The Average Value-at-Risk at level α has the equivalent representations*

$$\text{AV@R}_\alpha(Y) = \sup \left\{ \mathbb{E}(YZ) \mid \begin{array}{l} \mathbb{E}(Z) = 1, \\ \text{AV@R}_p(Z) \leq \frac{1}{1-\alpha} \text{ for all } p \in [\alpha, 1] \end{array} \right\} \quad (\alpha < 1) \quad (3.18)$$

and

$$\text{AV@R}_\alpha(Y) = \sup \left\{ \mathbb{E}(YZ) \mid \begin{array}{l} \mathbb{E}Z = 1, 0 \leq Z, Z \leq \frac{1}{1-\alpha} \end{array} \right\} \quad (\alpha \leq 1). \quad (3.19)$$

Proof. By sending $p \rightarrow 1$ in the constraint $\text{AV@R}_p(Z) \leq \frac{1}{1-\alpha}$ it follows that $Z \leq \frac{1}{1-\alpha}$. To reduce the formula to the identity (3.4) it remains to be shown that the constraints in (3.18) imply $Z \geq 0$. Indeed, suppose that $p := P(Z < 0) > 0$. Then $1 = \mathbb{E}Z \leq \int_p^1 G_Z^{-1}(u) du = (1-p) \text{AV@R}_p(Z)$, or $\text{AV@R}_p(Z) \geq \frac{1}{1-p}$. But the Average Value-at-Risk is characterized by the distortion σ_α (cf. (3.10)), and it holds that $\text{AV@R}_p(Z) \leq \int_p^1 \sigma_\alpha(u) du = \min \left\{ \frac{1}{1-p}, \frac{1}{1-\alpha} \right\}$, which is a contradiction.

For the second assertion observe first that $\frac{1}{1-\alpha} \geq \text{AV@R}_p(Z) \xrightarrow[p \rightarrow 1]{} \text{ess sup } Z$, hence $(1-\alpha)Z \leq 1$; conversely, if $0 \leq Z$ and $(1-\alpha)Z \leq 1$, then

$$\frac{1}{1-\alpha} \geq \text{ess sup } Z \geq \text{AV@R}_p(Z).$$

Hence, the constraints in (3.18) and (3.19) are equivalent whenever $\alpha < 1$. For $\alpha = 1$ the statement is obtained by Hölder $L^1 - L^\infty$ duality. \square

The following result, which is obtained in Pflug and Römisch [97, Proposition 2.65] and which is used at different places in this book, can be formulated as a corollary.

Corollary 3.19. *The distortion risk functional $\mathcal{R}_\sigma(\cdot)$ has the representation*

$$\mathcal{R}_\sigma(Y) = \max \{ \mathbb{E}(Y \cdot \sigma(U)) : U \text{ is uniformly } [0, 1] \text{ distributed} \}. \quad (3.20)$$

*The maximum is attained whenever Y and U are coupled in a co-monotone way.*⁵

Proof. Consider the random variable $\sigma(U)$ and observe that $G_{\sigma(U)}^{-1}(u) = \sigma(u)$. Indeed,

$$P(\sigma(U) \leq \sigma(u)) \geq P(U \leq u) = u,$$

such that $G_{\sigma(U)}^{-1} \geq \sigma(u)$. But $\mathbb{E}(\sigma(U)) = 1 = \int_0^1 \sigma(u) du$, from which follows that $G_{\sigma(U)}^{-1}(\cdot) = \sigma(\cdot)$ (a.e.). $\sigma(U)$ thus is feasible for (3.15), because

⁵Two random variables X and Y are coupled in a co-monotone way if $P(Y \leq y, Z \leq z) = \min(P(Y \leq y), P(Z \leq z))$. In this case $\mathbb{E}(Y \cdot Z) = \int_0^1 G_Y^{-1}(u) \cdot G_Z^{-1}(u) du$.

$$(1 - \alpha) \text{AV@R}_\alpha(\sigma(U)) = \int_\alpha^1 G_{\sigma(U)}^{-1}(u) du = \int_\alpha^1 \sigma(u) du.$$

If, finally, Y and U are coupled in a co-monotone way, then

$$\mathbb{E}(Y\sigma(U)) = \int_0^1 \sigma(u)G_Y^{-1}(u) du = \mathcal{R}_\sigma(Y),$$

which concludes the proof. \square

General Risk Functionals. The Fenchel–Moreau theorem applies for general convex functions: provided that \mathcal{R} is lower semi-continuous it holds that

$$\mathcal{R}(Y) = \sup_{Z \in \mathcal{Y}^*} \mathbb{E}(Y \cdot Z) - \mathcal{R}^*(Z),$$

where

$$\mathcal{R}^*(Z) := \sup_{Y \in \mathcal{Y}} \mathbb{E}(Y \cdot Z) - \mathcal{R}(Y) \quad (3.21)$$

is the conjugate function.

Definition 3.20. Let \mathcal{R} be a risk functional with conjugate \mathcal{R}^* .

- (i) We shall say that a dual variable Z is *feasible*, if $\mathcal{R}^*(Z) < \infty$.
- (ii) For a set \mathcal{S} of distortion densities the set

$$\begin{aligned} \mathcal{Z}_{\mathcal{S}} &:= \{Z \in L^1 : \exists \sigma \in \mathcal{S} \text{ such that } Z \preceq \sigma\} \\ &= \left\{ Z \in L^1 \left| \begin{array}{l} \mathbb{E}(Z) = 1, \text{ and } \exists \sigma \in \mathcal{S} \text{ such that} \\ \text{AV@R}_\alpha(Z) \leq \frac{1}{1-\alpha} \int_\alpha^1 \sigma(u) du \text{ for all } \alpha \in (0, 1) \end{array} \right. \right\} \end{aligned}$$

is called its *dual set*.

Theorem 3.16 identifies the conjugate function of a distortion risk functional \mathcal{R}_σ with distortion function σ as $\mathcal{R}_\sigma^*(Z) = \begin{cases} 0 & \text{if } Z \in \mathcal{Z}_{\{\sigma\}} \\ \infty & \text{else} \end{cases}$. The following corollary generalizes this observation to general risk functionals via a Kusuoka representation.

Corollary 3.21. Let \mathcal{R} be a version independent risk functional with Kusuoka representation

$$\mathcal{R}(Y) = \sup_{\sigma \in \mathcal{S}} \mathcal{R}_\sigma(Y).$$

Then the representation

$$\mathcal{R}(Y) = \sup_{Z \in \mathcal{Z}_S} \mathbb{E}(YZ) = \sup_{\sigma \in \mathcal{S}^*} \mathcal{R}_\sigma(Y)$$

in terms of support functions holds true, where $\mathcal{S}^* := \{G_Z^{-1}(\cdot) : Z \in \mathcal{Z}_S\} \supset \mathcal{S}$ is a set of distortion functionals.

Proof. It follows from Theorem 3.16 that

$$\mathcal{R}(Y) = \sup_{\sigma \in \mathcal{S}} \mathcal{R}_\sigma(Y) = \sup_{\sigma \in \mathcal{S}} \sup_{Z \in \mathcal{Z}_{\{\sigma\}}} \mathbb{E}(YZ) = \sup_{Z \in \mathcal{Z}_S} \mathbb{E}(YZ),$$

as \mathcal{R}_σ is lower semi-continuous and the maximum of lower semi-continuous functions is lower semi-continuous. \square

3.4 An Alternative Description of Distortion Risk Functionals

As was already emphasized in the introduction risk functionals are important in stochastic optimization. Suppose thus one intends to solve a stochastic optimization problem, which is given in the form

$$\begin{aligned} & \text{minimize} \\ & \quad (\text{in } Y) \quad \mathcal{R}_\sigma(Y) \\ & \text{subject to } Y \in \mathcal{Y} \end{aligned} \tag{3.22}$$

with a distortion risk functional \mathcal{R}_σ in the objective. Then one may employ the representation

$$\mathcal{R}_\sigma(Y) = \int_0^1 G_Y^{-1}(u)\sigma(u) du$$

from the definition of the distortion risk functional. In this representation the outcomes of Y have to be sorted in (3.22), which is cumbersome and slow in a real application.

Employing the dual representation

$$\mathcal{R}_\sigma(Y) = \sup \{\mathbb{E}(YZ) : 0 \leq Z, \mathbb{E}(Z) = 1, Z \preceq \sigma\}$$

in computing (3.22) leads to the problem

$$\begin{aligned}
& \underset{\text{(in } Y\text{)}}{\text{minimize}} \quad \sup \mathbb{E}(YZ) \\
& \text{subject to} \quad Y \in \mathcal{Y}, \\
& \quad 0 \leq Z, \mathbb{E}(Z) = 1, Z \leq \sigma,
\end{aligned}$$

which is a minimax problem, which cannot be solved fast neither.

For the Average Value-at-Risk, however, formula (3.3) provides a welcome and widely used alternative, because it expresses the Average Value-at-Risk as an infimum instead of a supremum. Indeed, problem (3.22) rewrites as

$$\begin{aligned}
& \underset{\text{(in } Y \text{ and } q\text{)}}{\text{minimize}} \quad q + \frac{1}{1-\alpha} \mathbb{E}(Y - q)_+ \\
& \text{subject to} \quad Y \in \mathcal{Y}, \\
& \quad q \in \mathbb{R},
\end{aligned} \tag{3.23}$$

which transforms the initial optimization problem in a simple minimization with a single, additional parameter $q \in \mathbb{R}$. It is known that the optimal value q in (3.23) is the Value-at-Risk of the optimal random variable Y , $q = V@R_\alpha(Y)$.

The following theorem generalizes the helpful formula (3.3) for distortion risk functionals and thus explains how distortion risk functionals can be employed to solve problems of type (3.22).

Theorem 3.22. *The distortion risk functional \mathcal{R}_σ has the representation*

$$\mathcal{R}_\sigma(Y) = \inf \left\{ \mathbb{E}(h(Y)) : \int_0^1 h^*(\sigma(u)) du \leq 0 \right\}, \tag{3.24}$$

where the infimum is among all measurable functions $h : \mathbb{R} \rightarrow \mathbb{R}$.⁶

The initial problem (3.22) thus can be reformulated as

$$\begin{aligned}
& \underset{\text{(in } Y \text{ and } h\text{)}}{\text{minimize}} \quad \mathbb{E}[h(Y)] \\
& \text{subject to} \quad Y \in \mathcal{Y}, \\
& \quad \int_0^1 h^*(\sigma(u)) du \leq 0,
\end{aligned}$$

where a function h appears in addition in the optimization procedure.

To prove the statement we shall address the following corollary first.

Corollary 3.23. *The distortion risk functional \mathcal{R}_σ has the representation*

⁶The conjugate dual function of h is $h^*(\sigma) = \sup_{y \in \mathbb{R}} \sigma \cdot y - h(y)$.

$$\mathcal{R}_\sigma(Y) = \inf \mathbb{E}[h(Y)] + \int_0^1 h^*(\sigma(u)), \quad (3.25)$$

where the infimum is among all measurable functions $h : \mathbb{R} \rightarrow \mathbb{R}$.

Proof. From the definition of the conjugate function it follows that $\sigma \cdot y \leq h(y) + h^*(\sigma)$ (often referred to as Fenchel's inequality, or Fenchel–Young inequality). For a uniform random variable U thus

$$\sigma(U) \cdot Y \leq h(Y) + h^*(\sigma(U)),$$

and by taking expectations it follows that

$$\mathbb{E}[\sigma(U) \cdot Y] \leq \mathbb{E}[h(Y)] + \mathbb{E}[h^*(\sigma(U))] = \mathbb{E}[h(Y)] + \int_0^1 h^*(\sigma(u)) du.$$

The left side is independent from h , and the right side is independent from U , it thus follows by (3.20) that

$$\mathcal{R}_\sigma(Y) = \sup_{U \text{ uniform}} \mathbb{E}[\sigma(U) Y] \leq \inf_h \mathbb{E}[h(Y)] + \int_0^1 h^*(\sigma(u)) du. \quad (3.26)$$

$$= \inf_h \left\{ \mathbb{E}[h(Y)] : \int_0^1 h^*(\sigma(u)) du \leq 0 \right\}. \quad (3.27)$$

To accept the latter equality consider the function $h_c(y) := h(y) + c$. Its conjugate is

$$h_c^*(\sigma) = \sup_y \sigma \cdot y - h_c(y) = \sup_y \sigma \cdot y - h(y) - c = h^*(\sigma) - c,$$

such that (3.26) is not affected by adding a constant $c \in \mathbb{R}$ to the function h . The choice $c := \int_0^1 h^*(\sigma(u)) du$ establishes equality in (3.27).

To establish equality in (3.26) define the function

$$h_\sigma(y) := \sigma(0) \cdot y + \int_0^1 (1-\alpha) G_Y^{-1}(\alpha) + (y - G_Y^{-1}(\alpha))_+ d\sigma(\alpha).$$

Notice that h_σ is increasing and convex, as $y \mapsto (1-\alpha)c + (y-c)_+$ is convex and increasing for every c , and because σ is an increasing function.

Recall next that $\text{AV@R}_\alpha(Y) = \inf_{q \in \mathbb{R}} q + \frac{1}{1-\alpha} \mathbb{E}(Y - q)_+$ and the fact that the infimum is attained at $q = G_Y^{-1}(\alpha)$ (cf. Pflug [89]), such that $\text{AV@R}_\alpha(Y) = G_Y^{-1}(\alpha) + \frac{1}{1-\alpha} \mathbb{E}(Y - G_Y^{-1}(\alpha))_+$. Employing (3.13) (note that $d\sigma(\alpha) = \frac{1}{1-\alpha} d\mu_\sigma(\alpha)$ whenever $\alpha > 0$) it follows that

$$\begin{aligned}
\mathcal{R}_\sigma(Y) &= \int_0^1 \text{AV@R}_\alpha(Y) d\mu_\sigma(\alpha) \\
&= \int_0^1 G_Y^{-1}(\alpha) + \frac{1}{1-\alpha} \mathbb{E}(Y - G_Y^{-1}(\alpha))_+ d\mu_\sigma(\alpha) \\
&= \mathbb{E} \int_{0^-}^1 (1-\alpha) G_Y^{-1}(\alpha) + (Y - G_Y^{-1}(\alpha))_+ d\sigma(\alpha) \\
&= \mathbb{E} h_\sigma(Y),
\end{aligned} \tag{3.28}$$

where we have used that

$$(1-\alpha) G_Y^{-1}(\alpha) + (y - G_Y^{-1}(\alpha))_+ \xrightarrow[\alpha \rightarrow 0]{} \max \{y, G_Y^{-1}(0)\}$$

(even if $G_Y^{-1}(0) = -\infty$).

To establish the assertion it is enough to show that $\int_0^1 h_\sigma^*(\sigma(u)) du = 0$. For this observe that h_σ is almost everywhere differentiable (as it is convex) with derivative

$$h'_\sigma(y) = \sigma(0) + \int_{\{G_Y^{-1}(\cdot) \leq y\}} d\sigma(u) = \sigma(0) + \int_0^{G_Y(y)} d\sigma(u) = \sigma(G_Y(y)). \tag{3.29}$$

Next observe that $h_\sigma^*(\sigma(u)) = \sup_y \sigma(u)y - h_\sigma(y)$, the supremum is attained at y satisfying $\sigma(u) = h'_\sigma(y) = \sigma(G_Y(y))$, that is at $y = G_Y^{-1}(u)$, such that $h_\sigma^*(\sigma(u)) = \sigma(u)G_Y^{-1}(u) - h_\sigma(G_Y^{-1}(u))$, and it follows that

$$\int_0^1 h_\sigma^*(\sigma(u)) du = \int_0^1 \sigma(u)G_Y^{-1}(u) du - \int_0^1 h_\sigma(G_Y^{-1}(u)) du = \mathcal{R}_\sigma(Y) - \mathbb{E} h_\sigma(Y).$$

By (3.28) it follows thus that $\int_0^1 h_\sigma^*(\sigma(u)) du = 0$, which finally establishes the assertion. \square

Remark 3.24. In case that σ is bounded, then, according to (3.29), h_σ is Lipschitz continuous with constant

$$L(h_\sigma) = \sup_{y \in \mathbb{R}} h'_\sigma(y) = \sup_{y \in \mathbb{R}} \sigma(G_Y(y)) = \|\sigma\|_\infty < \infty.$$

It follows thus that the function h in (3.24)—without loss of generality—can be assumed to be Lipschitz continuous with constant $L(h) \leq \|\sigma\|_\infty$.

3.5 The Impact of the Probability Measure on Risk Functionals

The risk functionals considered above are all version independent in the sense of Definition 3.12, they just depend on the distribution function of the random variables. However, in many situations the underlying probability measure is not known precisely, as is the case, for example, for the empirical measure. In this section we investigate how risk functionals depend on the underlying probability measure.

In Definition 3.2, the probability space (Ω, \mathcal{F}, P) is fixed and only the random variables $Y \in \mathcal{Y}$ vary. To consider also the dependency of the risk functional on the underlying probability measure we write explicitly $\mathcal{R}_P(Y)$.

3.5.1 Compound Concavity and Convex-Concavity

The following definition relates to the mapping $P \mapsto \mathcal{R}_P(Y)$.

Definition 3.25 (Compound Concavity). A version-independent functional is called *compound concave*, if the mapping $P \mapsto \mathcal{R}_P(Y)$ is concave.

(CC) COMPOUND CONCAVITY:⁷ $\mathcal{R}_{\lambda P_1 + (1-\lambda)P_0}(Y) \geq \lambda \mathcal{R}_{P_1}(Y) + (1-\lambda) \mathcal{R}_{P_0}(Y)$ for $0 \leq \lambda \leq 1$, all real random variables Y and probability measures P_0 and P_1 (we set $\mathcal{R}(Y) := \infty$ if Y is not in the domain of \mathcal{R}).

One may also consider the joint mapping $(Y, P) \mapsto \mathcal{R}_P(Y)$ and combine the two properties convexity (C) and compound concavity (CC) to the notion of convex-concavity of risk functionals.

Definition 3.26 (Convex-Concavity). A risk functional \mathcal{R} is

(C-CC) CONVEX-CONCAVE, if the mapping $(Y, P) \mapsto \mathcal{R}_P(Y) = \mathcal{R}(P^Y)$ is convex in Y (property (C) in Definition 3.2) and compound concave in P ((CC) in Definition 3.25; cf. also Appendix A).

Theorem 3.27. Any distortion risk functional \mathcal{R} is compound concave, that is,

$$P \mapsto \mathcal{R}_P(Y)$$

is concave for every Y fixed. Since $Y \mapsto \mathcal{R}_P(Y)$ is convex in Y it is convex-concave, \mathcal{R} is convex-concave in the sense of Definition 3.26.

In particular, the Average Value-at-Risk is convex-concave.

⁷The name *compound concave* is motivated from the fact that $\lambda P_1 + (1 - \lambda) P_0$ is called the *compound* of the two probabilities P_0 and P_1 . It can be seen as the result, when P_1 is chosen with probability λ and P_0 is chosen with probability $1 - \lambda$.

Proof. Observe first that

$$\begin{aligned}
\text{AV@R}_{\alpha;P_\lambda}(Y) &= \min_q q + \frac{1}{1-\alpha} \mathbb{E}_{P_\lambda} (Y - q)_+ \\
&= \min_q q + \frac{1}{1-\alpha} \int (Y - q)_+ d(1-\lambda) P_0 + \lambda P_1 \\
&= \min_q (1-\lambda) \left(q + \frac{1}{1-\alpha} \int (Y - q)_+ dP_0 \right) \\
&\quad + \lambda \left(q + \frac{1}{1-\alpha} \int (Y - q)_+ dP_1 \right) \\
&\geq \min_{q_0, q_1} (1-\lambda) \left(q_0 + \frac{1}{1-\alpha} \int (Y - q_0)_+ dP_0 \right) \\
&\quad + \lambda \left(q_1 + \frac{1}{1-\alpha} \int (Y - q_1)_+ dP_1 \right) \\
&= (1-\lambda) \text{AV@R}_{\alpha;P_0}(Y) + \lambda \cdot \text{AV@R}_{\alpha;P_1}(Y),
\end{aligned}$$

and this is the assertion for the Average Value-at-Risk. The assertion for the distortion risk can be derived from the representation $\mathcal{R}_\sigma(Y) = \int_0^1 \text{AV@R}_\alpha(Y) d\mu_\sigma(\alpha)$ (cf. (3.11)) for a distortion risk functional.

It demonstrates the strength of representation (3.25) that the assertion for general distortion functionals follows directly from

$$\begin{aligned}
\mathcal{R}_{\sigma;P_\lambda}(Y) &= \min_h \mathbb{E}_{P_\lambda} h(Y) + \int_0^1 h^*(\sigma(u)) du \\
&= \min_h (1-\lambda) \int h(Y) dP_0 + \lambda \int h(Y) dP_1 + \int_0^1 h^*(\sigma(u)) du \\
&= \min_h (1-\lambda) \left(\int h(Y) dP_0 + \int_0^1 h^*(\sigma(u)) du \right) \\
&\quad + \lambda \left(\int h(Y) dP_1 + \int_0^1 h^*(\sigma(u)) du \right) \\
&\geq \min_{h_0, h_1} (1-\lambda) \left(\int h_0(Y) dP_0 + \int_0^1 h_0^*(\sigma(u)) du \right) \\
&\quad + \lambda \left(\int h_1(Y) dP_1 + \int_0^1 h_1^*(\sigma(u)) du \right) \\
&= (1-\lambda) \mathcal{R}_{\sigma;P_0}(Y) + \lambda \cdot \mathcal{R}_{\sigma;P_1}(Y).
\end{aligned}$$

□

Example 3.28. The assumption that \mathcal{R} is a distortion functional is crucial: other convex risk functionals need not be convex-concave. As a counterexample to concavity in P consider the convex functional

$$\mathcal{R}(Y) = \max \left\{ \frac{1}{2} \mathbb{E}(Y) + \frac{1}{2} \text{AV@R}_{0.9}(Y), \text{AV@R}_{0.7}(Y) \right\}. \quad (3.30)$$

For the two measures

$$P_1 = \begin{bmatrix} 0.50 & 0.35 & 0.15 \\ 0 & 40 & 80 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 0.05 & 0.85 & 0.10 \\ 0 & 40 & 80 \end{bmatrix}$$

and the identity $Y = \text{id}$ one calculates $\mathcal{R}_{P_1}(Y) = 60$ and $\mathcal{R}_{P_2}(Y) = 61$, while for the convex combination

$$\mathcal{R}_{\frac{1}{2}P_1 + \frac{1}{2}P_2}(Y) = 57 < 60 = \min \{\mathcal{R}_{P_1}(Y), \mathcal{R}_{P_2}(Y)\}.$$

Thus \mathcal{R} does not satisfy condition (CC), it is not even quasiconcave in the probability measure.

A consequence of the non-compound concavity is that employing this risk functional, randomized decisions may outperform non-randomized ones. An example of this sort was already given in Example 1.9 for value-at-risk minimization. Here we give another example which uses the convex risk functional (3.30). Consider Fig. 3.1. At the first node, the decision has to be made whether the upper or the

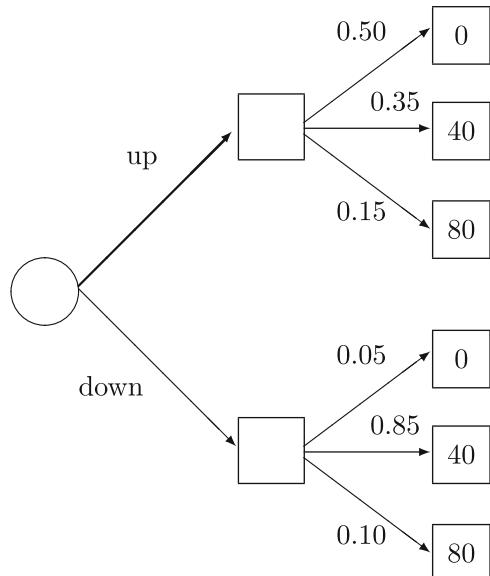


Fig. 3.1 Functionals, which are not compound concave, favorize randomized decisions: the functional defined in (3.30) applied to the upper subtree (decision *up*) gives 61 and to the lower subtree (decision *down*) gives 60. However, to decide with probability 1/2 for *up* und with probability 1/2 for *down* gives a smaller risk of 57

lower random cost variable is chosen. The choice *up* leads to the risk value 60, while the choice *down* leads to risk 61. The optimal decision is *up*. However, the randomized decision choosing *up* with probability 1/2 and *down* with probability 1/2 has a smaller risk of 57. For compound concave risk functionals randomizations increase the risk and cannot be risk-optimal.

3.5.2 Continuity with Respect to the Probability Measure

In many applications the probability measure is often not known precisely, just an empirical measure is available or an approximation of the true measure. In this situation it is essential that the risk functional will not vary too much neither, if the probability measure differs by some controllable amount.

It turns out that the Wasserstein distance is adapted to capture this problem. The following lemma outlines the close relationship between the Wasserstein distance and risk functionals. Indeed, the distortion risk functional can be expressed explicitly, that is with equality, by the quadratic Wasserstein distance.

Lemma 3.29. *Let $Y \in L^2(\mathbb{R})$ be a random variable on \mathbb{R} with second moment and σ a distortion function with $\|\sigma\|_2^2 = \int_0^1 \sigma(u)^2 du < \infty$. Then*

$$2 \cdot \mathcal{R}_\sigma(Y) = \|Y\|_2^2 + \|\sigma\|_2^2 - d_2(P^Y, P^{\sigma(U)})^2,$$

where U is a uniformly distributed random variable.

Proof. Recall that G_Y^{-1} is the quantile function of the image measure P^Y and the quantile function of $\sigma(U)$ is $\sigma(\cdot) = G_{\sigma(U)}^{-1}(\cdot)$ (cf. the proof of Corollary 3.19). By Theorem 2.15 thus

$$\begin{aligned} d_2(P^Y, P^\sigma)^2 &= \int_0^1 \left(G_Y^{-1}(\alpha) - G_{\sigma(U)}^{-1}(\alpha) \right)^2 d\alpha \\ &= \int_0^1 G_Y^{-1}(\alpha)^2 d\alpha - 2 \int_0^1 G_Y^{-1}(\alpha) \sigma(\alpha) d\alpha + \int_0^1 \sigma(\alpha)^2 d\alpha \\ &= \mathbb{E}(Y^2) - 2\mathcal{R}_\sigma(Y) + \int_0^1 \sigma(\alpha)^2 d\alpha, \end{aligned}$$

which provides the assertion. \square

Remark 3.30. The statement of the latter lemma is perhaps less of surprise by comparing the alternative representation (3.25) for the distortion risk functional with the dual (2.27) of the Wasserstein distance.

A general relationship on continuity is provided by the following statement, which is adapted from [100] and [99].

Theorem 3.31 (Continuity with Respect to Changing the Measure). *The distortion risk functional is robust in the following sense: it satisfies*

$$|\mathcal{R}_{\sigma;P}(Y) - \mathcal{R}_{\sigma;\tilde{P}}(Y)| \leq \|\sigma\|_{r_\beta^*} \cdot H_\beta(Y) \cdot d_r(P, \tilde{P})^\beta,$$

where $r_\beta^* \geq \frac{r}{r-\beta}$ and $H_\beta(Y)$ is the Hölder constant for which $|Y(x) - Y(y)| \leq H_\beta(Y) \cdot d(x, y)^\beta$.

Proof. Define $G_Y(y) := P(Y \leq y)$ and $\tilde{G}_Y(y) := \tilde{P}(Y \leq y)$ and set $r_\beta := \frac{r}{\beta}$. Let r_β^* be the conjugate exponent, defined by $\frac{1}{r_\beta} + \frac{1}{r_\beta^*} = 1$. By Hölder's inequality it holds that

$$\begin{aligned} \mathcal{R}_{\sigma;P}(Y) - \mathcal{R}_{\sigma;\tilde{P}}(Y) &= \int_0^1 G_Y^{-1}(u) - \tilde{G}_Y^{-1}(u) \sigma(u) du \\ &\leq \left(\int_0^1 |G_Y^{-1}(u) - \tilde{G}_Y^{-1}(u)|^{r_\beta} du \right)^{\frac{1}{r_\beta}} \cdot \left(\int_0^1 \sigma(u)^{r_\beta^*} du \right)^{\frac{1}{r_\beta^*}}. \end{aligned} \quad (3.31)$$

Now note that

$$\begin{aligned} d_{r_\beta}(P^Y, \tilde{P}^Y)^{r_\beta} &= \int |x - y|^{r_\beta} \pi(Y^{-1}(dx), Y^{-1}(dy)) \\ &= \int |Y(x) - Y(y)|^{r_\beta} \pi(dx, dy) \\ &\leq H_\beta(Y)^{r_\beta} \cdot \int |x - y|^{r_\beta \cdot \beta} \pi(dx, dy), \end{aligned}$$

where π has the marginals P and \tilde{P} , such that $d_{r_\beta}(P^Y, \tilde{P}^Y)^{r_\beta} \leq H_\beta(Y)^{r_\beta} \cdot d_r(P, \tilde{P})^r$. It follows from (3.31) that

$$\mathcal{R}_{\sigma;P}(Y) - \mathcal{R}_{\sigma;\tilde{P}}(Y) \leq H_\beta(Y) \cdot \|\sigma\|_{r_\beta^*} \cdot d_r(P, \tilde{P})^\beta.$$

The assertion of the theorem finally follows by interchanging the measures P and \tilde{P} . \square

Example 3.32. As an example consider the Average Value-at-Risk, which has distortion function $\sigma_\alpha = \frac{1}{1-\alpha} \mathbb{1}_{[\alpha, 1]}$. It is easily observed that

$$\|\sigma_\alpha\|_{r_\beta^*} = \left(\int_\alpha^1 \left(\frac{1}{1-\alpha} \right)^{r_\beta^*} \right)^{1/r_\beta^*} = (1-\alpha)^{\frac{1-r_\beta^*}{r_\beta^*}} = (1-\alpha)^{-\frac{r}{\beta}}.$$

For the important case $r = \beta = 1$ Hölder continuity is simply Lipschitz continuity, and $\|\sigma_\alpha\|_\infty = \frac{1}{1-\alpha}$. It follows thus that

$$|\text{AV@R}_{\alpha;P}(Y) - \text{AV@R}_{\alpha;\tilde{P}}(Y)| \leq \frac{1}{1-\alpha} \cdot L(Y) \cdot d_1(P, \tilde{P}),$$

as was already outlined in Pflug and Wozabal [98].

The continuity result of the previous Theorem 3.31 can be extended from distortion risk functionals to more general risk functionals. This is the content of the subsequent corollary.

Corollary 3.33 (Continuity with Respect to Changing the Measure). *Let $\mathcal{R}(\cdot) = \sup_{\sigma \in S} \mathcal{R}_\sigma(\cdot)$ be a version independent risk functional. Provided that $\sup_{\sigma \in S} \|\sigma\|_{r_\beta^*} < \infty$ for some $r_\beta^* \geq \frac{r}{r-\beta}$ the risk functional \mathcal{R} is continuous with respect to changing the measure, it satisfies*

$$|\mathcal{R}_P(Y) - \mathcal{R}_{\tilde{P}}(Y)| \leq H_\beta(Y) \cdot d_r(P, \tilde{P}) \cdot \sup_{\sigma \in S} \|\sigma\|_{r_\beta^*}.$$

where H_β is Y 's Hölder constant for the exponent β .

Proof. For $\varepsilon > 0$, choose $\sigma_0 \in S$ such that $\sup_{\sigma \in S} \mathcal{R}_{\sigma;P}(Y) < \mathcal{R}_{\sigma_0;P}(Y) + \varepsilon$. Then

$$\begin{aligned} \sup_{\sigma} \mathcal{R}_{\sigma;P}(Y) - \sup_{\sigma} \mathcal{R}_{\sigma;\tilde{P}}(Y) &\leq \mathcal{R}_{\sigma_0;P}(Y) + \varepsilon - \sup_{\sigma} \mathcal{R}_{\sigma;\tilde{P}}(Y) \\ &\leq \mathcal{R}_{\sigma_0;P}(Y) - \mathcal{R}_{\sigma_0;\tilde{P}}(Y) + \varepsilon \\ &\leq H_\beta(Y) \cdot d_r(P, \tilde{P}) \cdot \|\sigma_0\|_{r_\beta^*} + \varepsilon \\ &\leq H_\beta(Y) \cdot d_r(P, \tilde{P}) \cdot \sup_{\sigma \in S} \|\sigma\|_{r_\beta^*} + \varepsilon \end{aligned}$$

by Theorem 3.31. As $\varepsilon > 0$ was chosen arbitrarily the assertion follows by interchanging the roles of P and \tilde{P} . \square

3.6 Conditional Risk Functionals

In multistage optimization risk functionals can be considered in the objective and in constraints. As constraints they may appear on the root node, but on every other node equally well. Constraints, for example expressed via the Average Value-at-Risk, can be considered at every node, each with a different level of risk. In addition, the risk level may even be considered random, that is, dependent on realizations from the root up to the actual node, which has already been uncovered. In line with the

conditional expectation these risk functionals are called *conditional risk functionals*, or risk functionals on conditional basis (cf. Werner et al. [141]).

In the basic form, the conditional form of a version independent risk functional is just the functional applied to the family of conditional distributions. This definition has been adopted, among others, by Pflug and Römisch in [97]. The Average Value-at-Risk, as a special case, is addressed, amongst others, in [17, 18, Section 2.3.1] by Cheridito et al. and in Kovacevic and Pflug [71].

In this section we generalize the concept and define conditional risk functionals for general risk functionals and at random level. It turns out that conditional risk functionals inherit elementary properties of risk functionals. The presentation follows [96] and [95].

Definition 3.34 (Conditional Risk Functional). Let \mathcal{R} be a risk functional with conjugate \mathcal{R}^* (see (3.21)) and $Z_t \in \mathcal{Z}$ be feasible according to Definition 3.20 with $Z_t \triangleleft \mathcal{F}_t$. The *conditional risk functional at random level* Z_t is defined as

$$\mathcal{R}_{Z_t}(Y|\mathcal{F}_t) := \text{ess sup } \{\mathbb{E}(YZ'|\mathcal{F}_t) \mid \mathbb{E}(Z'|\mathcal{F}_t) = \mathbf{1} \text{ and } \mathcal{R}^*(Z_t Z') = 0\}. \quad (3.32)$$

For the definition of the essential supremum of a family of random variables we refer to Dunford and Schwartz [31], or to Karatzas and Shreve [65, Appendix A].

The supremum in (3.32) is not over the empty set, as $Z' = \mathbf{1}$ satisfies the constraints. The following lemma ensures that the risk functional can be considered at random level $\mathbb{E}(Z|\mathcal{F}_t)$ as well, whenever Z is a feasible dual variable.

Lemma 3.35. *Let \mathcal{R} be a risk functional with Kusuoka representation (see (3.14)), and $\mathcal{F}_t \subset \mathcal{F}$. Then it holds that*

$$\mathcal{R}(\mathbb{E}(Z|\mathcal{F}_t)) \leq \mathcal{R}(Z), \quad (3.33)$$

and in particular

$$\text{AV@R}_\alpha(\mathbb{E}(Z|\mathcal{F}_t)) \leq \text{AV@R}_\alpha(Z).$$

If moreover $\mathcal{R}^*(Z) < \infty$ is finite (i.e., Z is feasible), then $\mathcal{R}^*(\mathbb{E}(Z|\mathcal{F}_t)) < \infty$ is finite as well (and $\mathbb{E}(Z|\mathcal{F}_t)$ is feasible).

Proof. For the convex function $z \mapsto (z - q)_+$ it holds by the conditional Jensen inequality (cf. Williams [142, Section 34]) that

$$(\mathbb{E}(Z|\mathcal{F}_t) - q)_+ \leq \mathbb{E}((Z - q)_+ | \mathcal{F}_t)$$

for every $q \in \mathbb{R}$. Hence, by (3.3),

$$\begin{aligned}
\text{AV@R}_\alpha(\mathbb{E}(Z|\mathcal{F}_t)) &= \min_{q \in \mathbb{R}} q + \frac{1}{1-\alpha} \mathbb{E}(\mathbb{E}(Z|\mathcal{F}_t) - q)_+ \\
&\leq \min_{q \in \mathbb{R}} q + \frac{1}{1-\alpha} \mathbb{E}\mathbb{E}((Z-q)_+|\mathcal{F}_t) \\
&= \min_{q \in \mathbb{R}} q + \frac{1}{1-\alpha} \mathbb{E}(Z-q)_+ = \text{AV@R}_\alpha(Z).
\end{aligned}$$

By integration with respect to α the assertion follows for distortion risk functionals. The assertion follows for general risk functionals by employing Kusuoka's representation.

If Z is further feasible, then, by Theorem 3.16, $\text{AV@R}_\alpha(\mathbb{E}(Z|\mathcal{F}_t)) \leq \text{AV@R}_\alpha(Z) \leq \frac{1}{1-\alpha} \int_\alpha^1 \sigma(u) du$ for some distortion density σ of the corresponding Kusuoka representation. This verifies the second statement. \square

Conditional Average Value-at-Risk. For the Average Value-at-Risk it is natural and convenient to define a conditional version by involving a random risk level $\alpha_t \triangleleft \mathcal{F}_t$ instead of the random variable $Z_t \triangleleft \mathcal{F}_t$.

Definition 3.36 (Conditional Average Value-at-Risk at Random Level). A \mathcal{F}_t -measurable function α_t with values in $[0, 1]$ is called *random risk level*. The *Average Value-at-Risk at random risk level $\alpha_t \triangleleft \mathcal{F}_t$* is

$$\text{AV@R}_{\alpha_t}(Y|\mathcal{F}_t) := \text{ess sup} \left\{ \mathbb{E}(YZ'|\mathcal{F}_t) \middle| \begin{array}{l} \mathbb{E}(Z'|\mathcal{F}_t) = 1, Z' \geq 0, \text{ and} \\ (1-\alpha_t)Z' \leq 1 \end{array} \right\}. \quad (3.34)$$

The relation of the Average Value-at-Risk at random risk level and the conditional version of AV@R_α is established by the level $\alpha_t := 1 - (1-\alpha)Z_t$, as

$$\text{AV@R}_{1-(1-\alpha)Z_t}(Y|\mathcal{F}_t) = \text{AV@R}_{\alpha;Z_t}(Y|\mathcal{F}_t)$$

in view of (3.34) and the fact that the dual set of AV@R_α is given by $\mathcal{Z} = \{Z \geq 0 : \mathbb{E}Z = 1, (1-\alpha)Z \leq 1\}$.

The conditional risk functional, as it is defined in (3.32), is given as an essential supremum of risk functionals. The essential supremum itself is defined as the Radon–Nikodým derivative of an associated measure in the references given above. This is the motivation of the following characterization, which is available for the conditional Average Value-at-Risk.

Proposition 3.37 (Characterization). *The conditional Average Value-at-Risk at random level α_t , $\text{AV@R}_{\alpha_t}(Y|\mathcal{F}_t)$, is characterized by*

$$\mathbb{E}[\mathbb{1}_B \text{AV@R}_{\alpha_t}(Y|\mathcal{F}_t)] = \sup \{\mathbb{E}(YZ') : \mathbb{E}(Z'|\mathcal{F}_t) = 1_B, Z' \geq 0, (1-\alpha_t)Z' \leq 1\},$$

where $B \in \mathcal{F}_t$ is an arbitrary measurable set.

Proof. By the defining characterization provided in Karatzas and Shreve [65, Appendix A] the essential supremum in $\text{AV@R}_{\alpha_t}(Y|\mathcal{F}_t)$ satisfies

$$\begin{aligned} & \mathbb{E}[\mathbb{1}_B \text{AV@R}_{\alpha_t}(Y|\mathcal{F}_t)] \\ &= \sup \left\{ \sum_{k=1}^K \mathbb{E}[\mathbb{1}_{B_k} Y Z'_k] : 0 \leq Z_k, \mathbb{E}(Z'_k|\mathcal{F}_t) = \mathbb{1}, (1 - \alpha_t) Z'_k \leq \mathbb{1} \right\}, \end{aligned}$$

where $B_k \in \mathcal{F}_t$ are pairwise disjoint sets, $B_k \cap B_l = \emptyset$ whenever $k \neq l$, with $B = \bigcup_{k=1}^K B_k$ (it is said that the sets $(B_k)_{k=1}^K$ form a tessellation of B). Define now $Z' := \sum_{k=1}^K \mathbb{1}_{B_k} Z'_k$ and observe that

$$\mathbb{E}(Z'|\mathcal{F}_t) = \sum_{k=1}^K \mathbb{1}_{B_k} \mathbb{E}(Z'_k|\mathcal{F}_t) = \sum_{k=1}^K \mathbb{1}_{B_k} = \mathbb{1}_B.$$

Moreover $\sum_{k=1}^K \mathbb{E}(\mathbb{1}_{B_k} Y Z'_k) = \mathbb{E}(YZ')$ and clearly $(1 - \alpha_t) Z' \leq \mathbb{1}$, which establishes the assertion. \square

3.6.1 Properties of Conditional Risk Functionals

Conditional risk functionals are defined based on risk measures. In what follows we demonstrate that the properties of the initial risk measure are maintained by conditional risk functionals.

Theorem 3.38. *The conditional risk functional satisfies the following properties:*

- (i) PREDICTABILITY: $\mathcal{R}_{Z_t}(Y|\mathcal{F}_t) = Y$ if $Y \triangleleft \mathcal{F}_t$;
- (ii) MONOTONICITY: $\mathcal{R}_{Z_t}(Y_1|\mathcal{F}_t) \leq \mathcal{R}_{Z_t}(Y_2|\mathcal{F}_t)$ whenever $Y_1 \leq Y_2$;
- (iii) CONVEXITY: The mapping $Y \mapsto \mathcal{R}_{Z_t}(Y|\mathcal{F}_t)$ is convex, more specifically

$$\mathcal{R}_{Z_t}((1 - \lambda)Y_0 + \lambda Y_1|\mathcal{F}_t) \leq (1 - \lambda)\mathcal{R}_{Z_t}(Y_0|\mathcal{F}_t) + \lambda\mathcal{R}_{Z_t}(Y_1|\mathcal{F}_t)$$

for $\lambda \triangleleft \mathcal{F}_t$ and $0 \leq \lambda \leq 1$, almost surely;

- (iv) TRANSLATION EQUIVARIENCE: $\mathcal{R}_{Z_t}(Y|\mathcal{F}_t) = \mathcal{R}_{Z_t}(Y|\mathcal{F}_t) + c$ if $c \triangleleft \mathcal{F}_t$;
- (v) POSITIVE HOMOGENEITY: $\mathcal{R}_{Z_t}(\lambda Y|\mathcal{F}_t) = \lambda\mathcal{R}_{Z_t}(Y|\mathcal{F}_t)$ whenever $\lambda \geq 0$ is bounded and $\lambda \triangleleft \mathcal{F}_t$;
- (vi) CONCAVITY: the mapping $Z_t \mapsto Z_t \cdot \mathcal{R}_{Z_t}(Y|\mathcal{F}_t)$ is concave; more specifically

$$Z_\lambda \cdot \mathcal{R}_{Z_\lambda}(Y|\mathcal{F}_t) \geq (1 - \lambda)Z_0 \cdot \mathcal{R}_{Z_0}(Y|\mathcal{F}_t) + \lambda Z_1 \cdot \mathcal{R}_{Z_1}(Y|\mathcal{F}_t)$$

almost everywhere, where $Z_\lambda = (1 - \lambda)Z_0 + \lambda Z_1$ ($\lambda \in [0, 1]$) and $Z_0, Z_1 \triangleleft \mathcal{F}_t$.

Proof. Predictability follows from predictability of the expected value, as $\mathbb{E}(YZ'|\mathcal{F}_t) = Y \cdot \mathbb{E}(Z'|\mathcal{F}_t) = Y$ whenever $Y \triangleleft \mathcal{F}_t$. Translation equivariance follows from $\mathbb{E}((Y + c)Z'|\mathcal{F}_t) = \mathbb{E}(YZ'|\mathcal{F}_t) + c \cdot \mathbb{E}(Z'|\mathcal{F}_t) = \mathbb{E}(YZ'|\mathcal{F}_t) + c$, and positive homogeneity from $\text{ess sup } \mathbb{E}(\lambda YZ'|\mathcal{F}_t) = \lambda \text{ess sup } \mathbb{E}(Z'|\mathcal{F}_t)$, as $0 \leq \lambda \triangleleft \mathcal{F}_t$.

Monotonicity is inherited from the conditional expected value, as $\mathbb{E}(Y_1 Z'|\mathcal{F}_t) \leq \mathbb{E}(Y_2 Z'|\mathcal{F}_t)$ whenever $Y_1 \leq Y_2$ and $Z' \geq 0$.

As for convexity observe that

$$\begin{aligned} (1 - \lambda) \mathcal{R}_{Z_t}(Y_0|\mathcal{F}_t) + \lambda \mathcal{R}_{Z_t}(Y_1|\mathcal{F}_t) \\ = (1 - \lambda) \underset{\mathcal{R}^*(Z_t Z'_0)=0}{\text{ess sup}} \mathbb{E}(Y_0 Z'_0|\mathcal{F}_t) + \lambda \underset{\mathcal{R}^*(Z_t Z'_1)=0}{\text{ess sup}} \mathbb{E}(Y_1 Z'_1|\mathcal{F}_t) \\ \geq \underset{\mathcal{R}^*(Z_t Z')=0}{\text{ess sup}} (1 - \lambda) \mathbb{E}(Y_0 Z'|\mathcal{F}_t) + \lambda \mathbb{E}(Y_1 Z'|\mathcal{F}_t) \\ = \mathcal{R}_{Z_t}((1 - \lambda) Y_0 + \lambda Y_1|\mathcal{F}_t). \end{aligned}$$

To accept concavity let $Z_0, Z_1 \triangleleft \mathcal{F}_t$ be feasible and Z'_0 and Z'_1 be chosen such that $\mathcal{R}^*(Z_0 Z'_0) < \infty$ and $\mathcal{R}^*(Z_1 Z'_1) < \infty$. Define $Z'_\lambda := \frac{1}{Z_\lambda} ((1 - \lambda) Z_0 Z'_0 + \lambda Z_1 Z'_1)$ and observe that $\mathbb{E}(Z'_\lambda|\mathcal{F}_t) = \frac{(1-\lambda)Z_0+\lambda Z_1}{Z_\lambda} = 1$. Then, by convexity of \mathcal{Z}_{S^*} , $Z_\lambda := (1 - \lambda) Z_0 + \lambda Z_1 \triangleleft \mathcal{F}_t$ is feasible as well, and $\mathcal{R}^*(Z_\lambda Z'_\lambda) \leq (1 - \lambda) \mathcal{R}^*(Z_0 Z'_0) + \lambda \mathcal{R}^*(Z_1 Z'_1) < \infty$, such that $Z_\lambda Z'_\lambda$ is feasible. It follows that

$$Z_\lambda \cdot \mathcal{R}_{Z_\lambda}(Y|\mathcal{F}_t) \geq Z_\lambda \cdot \mathbb{E}(Y|Z'_\lambda|\mathcal{F}_t) = (1 - \lambda) Z_0 \mathbb{E}(Y|Z'_0|\mathcal{F}_t) + \lambda Z_1 \mathbb{E}(Y|Z'_1|\mathcal{F}_t)$$

Taking the essential supremum (with respect to Z'_0 and Z'_1) reveals that

$$Z_\lambda \cdot \mathcal{R}_{Z_\lambda}(Y|\mathcal{F}_t) \geq (1 - \lambda) Z_0 \cdot \mathcal{R}_{Z_0}(Y|\mathcal{F}_t) + \lambda Z_1 \cdot \mathcal{R}_{Z_1}(Y|\mathcal{F}_t),$$

which insures concavity. \square

Chapter 4

From Data to Models

Multistage decision problems are designed to solve real world problems. It is especially important to model the reality in an appropriate way so that optimal solutions of the model can be used as decisions for the real problem at hand. Recall that we consider always the mathematical optimization problem as an approximation (see Fig. 1.2 in the introduction) and are aware of possible model errors. A crucial part of modeling is to find an appropriate stochastic model and—as a further step—an appropriate tree representation of the multistage scenario process. Typically, there is a sample of past data available, but not more.

It needs two steps to come from the sample of observations to the scenario model (see Fig. 4.1):

- (i) In the first step a *probability model* is identified, i.e., the description of the uncertainties as random variables or random processes by identifying the probability distribution. This step is based on statistical methods of model selection and parameter estimation. If several probability measures represent the data equally well, one speaks of *ambiguity*. In the non-ambiguous situation, one and only one probability model is selected and this model is the basis for the next step.
- (ii) In the following scenario generation step, a *scenario model* is found, which is an approximation of (i) by a *finite model* of lower complexity than the probability model.

Let the original problem be

$$(Opt) \quad \min \{F(x) = \mathcal{R}_P[Q(x, \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}\}. \quad (4.1)$$

The *scenario model* differs from the original model (4.1) insofar, as P is replaced by \tilde{P} , and the filtration \mathfrak{F} is replaced by a finite filtration $\tilde{\mathfrak{F}}$:

$$\widetilde{(Opt)} \quad \min \{\tilde{F}(\tilde{x}) = \mathcal{R}_{\tilde{P}}[Q(\tilde{x}, \xi)] : \tilde{x} \in \mathbb{X}, \tilde{x} \triangleleft \tilde{\mathfrak{F}}\}.$$

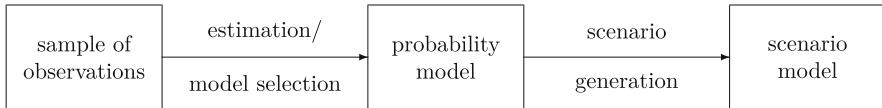


Fig. 4.1 From the observed data to the tree model

It is essential that the scenario model is finite (i.e., only finitely many values of the uncertainties are possible), but also that it is a good approximation to the original model. The finiteness is essential to keep the computational complexity low, and the approximate character is needed to allow the results to be carried over to the original model.

Consider Fig. 1.2 of the Introduction: the scenario model is the approximate problem which serves as a proxy for the original problem. Especially for multiperiod problems, the solution of the approximate problem is not directly applicable to the original problem, but an extension function is needed, which extends the solution of the approximate problem to a solution of the original problem. Of course, the extended solution of the approximate problem is typically suboptimal for the original problem. The respective gap is called the approximation error. It is the important goal of scenario generation to make this gap acceptably small.

The quality of a scenario approximation depends on the distance between the original model \mathbb{P} (written as a nested distribution) and the generated scenario model $\tilde{\mathbb{P}}$. Well-known quantitative stability theorems (see, e.g., Dupačová [34], Rachev and Römisch [106] or Heitsch et al. [55]) establish the relation between distances of probability models on one side and distances between optimal values or solutions on the other side. In Chap. 2 several distances were presented and studied. In this chapter we focus on approximation by Wasserstein distance and its multiperiod generalization, the nested distance.

4.1 Approximations of Single-Period Probability Distributions

In this section we consider the problem of approximating a given probability measure P on \mathbb{R}^m by a discrete probability \tilde{P} . In principle, there are three quite widely used methods.

- The Monte Carlo method, where the mass points are generated using pseudo-random numbers,
- the quasi-Monte Carlo method, where the mass points are generated by a low-discrepancy recursion,
- the optimal quantization method, where the mass points are found by solving an optimal facility location problem.

The methods are ordered not only in increasing approximation quality but also in increasing computational complexity: pseudo random numbers are easy to generate, but the approximate discrete distributions have an inherent randomness and have—with some probability—a quite low approximation quality. In contrast, to calculate optimal discretizations is a quite complex (in fact NP hard) problem, but they exhibit low approximation error. We discuss the Monte Carlo method in Sect. 4.1.1 and the optimal quantization methods in Sect. 4.1.3 and the following. For the quasi-Monte Carlo method we refer to the vast literature (books by Harald Niederreiter [83], Christiane Lemieux [74] or Josef Dick and Friedrich Pillichshammer [27]).

4.1.1 Approximation Quality of the Monte Carlo Generation Method

All information about probability distributions comes ultimately from samples, and our knowledge about the distribution of scenario processes comes from past observation series. The *empirical measure* assigns equal probabilities to all past observations and is the best nonparametric estimate of the underlying distribution. However, it is often necessary to generate more points than were observed. In this case, a probability model has to be estimated and artificial data may be randomly sampled from the estimated distribution by the Monte Carlo method. Whether the data come from the original observations or they are sampled from a theoretical distribution, the analysis of the approximation properties is the same and will be discussed below. In particular, it is important to understand how many sample points are necessary to ensure a certain approximation quality measured by the Wasserstein distance.

Let (ξ_1, \dots, ξ_n) be an independent, identically distributed (i.i.d.) sample from a probability distribution P with distribution function G on \mathbb{R}^m . The pertaining *empirical measure* is

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i},$$

and the empirical distribution function is

$$\hat{G}_n(u) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, u]}(\xi_i),$$

where $\mathbb{1}_A$ is the indicator function of the set A and $(-\infty, u] = (-\infty, u_1] \times \dots \times (-\infty, u_m]$ is a multidimensional interval. The famous Glivenko–Cantelli theorem (cf. van der Vaart [134, Chapter 19]) states that

$$\sup_u |\hat{G}_n(u) - G(u)| \xrightarrow{n \rightarrow \infty} 0 \quad \text{with probability 1,}$$

which can be reformulated as

$$\int h(u) \hat{P}_n(du) \xrightarrow{n \rightarrow \infty} \int h(u) P(du) \quad \text{with probability 1} \quad (4.2)$$

for every bounded, continuous functions h . If P has a finite first moment, i.e., if $\int \|u\| P(du) < \infty$ (if the Euclidean norm is used) or $\int d(u, u_0) P(du) < \infty$ for some u_0 (if the distance d is used), then (4.2) holds for all Lipschitz functions h and it follows that

$$d_1(\hat{P}_n, P) \xrightarrow{n \rightarrow \infty} 0 \quad \text{with probability 1.}$$

Lemma 4.1. *If P has finite r -th moment, then $d_r(\hat{P}_n, P) \rightarrow 0$ for $n \rightarrow \infty$ almost surely.*

Proof. We consider for simplicity only the case of the Euclidean distance on \mathbb{R}^m . We may find a constant C such that $\int_{\{\|u\|>C\}} \|u\|^r P(du) < \epsilon$. Since $\int_{\{\|u\|>C\}} \|u\|^r \hat{P}_n(du) \rightarrow \int_{\{\|u\|>C\}} \|u\|^r P(du)$ a.s. by the law of large numbers, $\sup_{m \geq n} \int_{\{\|u\|>C\}} \|u\|^r \hat{P}_m(du) < 2\epsilon$ for large enough n with arbitrary high probability. By weak convergence, there are measures $\hat{\pi}_n$ with marginals P and \hat{P}_n such that $\int \|u - v\| \hat{\pi}_n(du, dv) \rightarrow 0$ w. pr. 1. Now,

$$\begin{aligned} \int \|u - v\|^r \hat{\pi}_n(du, dv) &\leq \int_{\{\|u\|\leq C, \|v\|\leq C\}} \|u - v\|^r \hat{\pi}_n(du, dv) \\ &\quad + \int_{\{\|u\|>C \text{ and/or } \|v\|>C\}} \|u - v\|^r \hat{\pi}_n(du, dv) \\ &\leq (2C)^{r-1} \int \|u - v\| \hat{\pi}_n(du, dv) \\ &\quad + 2^{r-1} \left(\int_{\{\|u\|>C\}} \|u\|^r P(du) + \int_{\{\|v\|>C\}} \|v\|^r \hat{P}_n(dv) \right) \end{aligned}$$

which can be made arbitrarily small by choosing first C and then n large enough. \square

We have seen that the Wasserstein distance is not only appropriate for approximating stochastic optimization problems, but also has its importance in statistics, since it is the right distance for the convergence of the empirical distribution. In particular, natural confidence sets are formed by Wasserstein neighborhoods and these sets are good candidates for ambiguity sets, see Chap. 7.

For a precise statement about the quality of the approximation of the empirical distribution, the speed of convergence of $d_r(\hat{P}_n, P)$ to 0 is crucial. The relatively old result

$$\mathbb{E} \left[d_1 \left(\hat{P}_n, P \right) \right] \leq K \cdot n^{-1/m}$$

for some constant K is due to Dudley [30], and hence, by Chebysev's inequality,

$$P \left\{ d_1 \left(\hat{P}_n, P \right) \geq \epsilon \right\} \leq \frac{K}{\epsilon} \cdot n^{-1/m}.$$

The following more refined estimate was derived by Kersting (cf. [66]) under some smoothness assumptions on the distribution G of P for the dimension $m = 1$.

Theorem 4.2. *Suppose that G has density g such that for all $\lambda > 0$, $g(G^{-1}(\lambda t))/g(G^{-1}(t))$ converges to a positive number as $t \rightarrow 0$.*

- (i) *If G has bounded support, then $n^{1/2} \cdot d_1 \left(\hat{P}_n, P \right)$ is bounded in probability, if n tends to infinity.*
- (ii) *If the density is symmetric around 0, is twice differentiable, monotonic on $(-\infty, 0]$, and positive on the whole \mathbb{R} , then $n^{1/2} \alpha_n^{-1} d_1 \left(\hat{P}_n, P \right)$ is bounded in probability, where $\alpha_n \rightarrow \infty$ is the solution of $G(-\alpha_n^2) = n^{1/2} \alpha_n$.*

Proof. See Kersting [66]. □

In general, the convergence rate is smaller than $n^{-1/2}$ if P has unbounded support. For dimension m , the rate is typically $n^{-1/m}$. The following result shows the boundedness in probability of $n^{1/m} \cdot d_1 \left(\hat{P}_n, P \right)$.

Theorem 4.3. *If G is a distribution function in \mathbb{R}^m with density g , then*

$$\lim_{n \rightarrow \infty} P \left\{ n^{1/m} d_1 \left(\hat{P}_n, P \right) \geq t \right\} = \int_{\mathbb{R}^m} (1 - \exp(-t^m b_m g(u))) g(u) du,$$

where $b_m = \frac{2 \cdot \pi^{m/2}}{m \cdot \Gamma(m/2)}$ is the volume of the Euclidean unit ball in \mathbb{R}^m .

Proof. See Graf and Luschgy [48, Theorem 9.2]. □

Other types of inequalities can be derived from large deviations results like Sanov's theorem and concentration inequalities of the Talagrand type. These results are based on assumptions about higher moments of P .

In what follows we cite three theorems of this type to provide upper bounds of the deviation probability for d_1 and d_r , which do not hold only asymptotically, but already from a fixed sample size n_0 , which is specified by the theorems.

Theorem 4.4. *Let P be a probability measure on \mathbb{R}^m endowed with metric d . Suppose that $\int \exp(\alpha d(x, y)^2) P(dx) < \infty$ for some $\alpha > 0$. Then there exists a $\lambda > 0$ such that for all $m' > m$ and $\varepsilon > 0$*

$$P \left\{ d_1 \left(\hat{P}_n, P \right) > \varepsilon \right\} \leq \exp \left(-\frac{\lambda}{2} n \varepsilon^2 \right),$$

where $n \geq n_0 \cdot \max \left\{ \varepsilon^{-m'-2}, 1 \right\}$.

Proof. See Bolley, Guillin, and Villani [14, Theorem 2.1]. \square

Theorem 4.5. Let P be a probability measure on \mathbb{R}^m endowed with metric $\|\cdot\|$. Suppose that $\int \exp(\alpha \|x\|) P(dx) < \infty$. Then for $m' > m$, there exist constants K and n_0 such that

$$P \left\{ d_r \left(\hat{P}_n, P \right) > \varepsilon \right\} \leq \exp \left(-K n^{1/r} \min \left\{ \varepsilon, \varepsilon^2 \right\} \right)$$

for $\varepsilon > 0$ and $n \geq n_0 \cdot \max \left\{ \varepsilon^{-(2r+m')}, 1 \right\}$.

Proof. See Bolley, Guillin, and Villani [14, Theorem 2.8 (i)]. \square

Theorem 4.6. Let P be a probability measure on \mathbb{R}^m endowed with metric $\|\cdot\|$ such that $\int \|x\|^q P(dx) < \infty$. Then

- (i) For any $r \in [1, \frac{q}{2}]$, $\delta \in (0, \frac{q}{r} - 2)$ and all $m' > m$, there exist a constant n_0 such that

$$P \left\{ d_r \left(\hat{P}_n, P \right) > \varepsilon \right\} \leq \varepsilon^{-q} n^{-\frac{q}{2r} + \frac{\delta}{2}}$$

for $\varepsilon > 0$ and $n \geq n_0 \cdot \max \left\{ \varepsilon^{-q \frac{2r+m'}{q-r}}, \varepsilon^{m'-m} \right\}$.

- (ii) For any $r \in [\frac{q}{2}, q]$, $\delta \in (0, \frac{q}{r} - 1)$ and all $m' > m$, there exist a constant n_0 such that

$$P \left\{ d_r \left(\hat{P}_n, P \right) > \varepsilon \right\} \leq \varepsilon^{-q} n^{1-\frac{q}{r}+\delta}$$

for $\varepsilon > 0$ and $n \geq n_0 \cdot \max \left\{ \varepsilon^{-q \frac{2r+m'}{q-r}}, \varepsilon^{m'-m} \right\}$.

Proof. See Bolley, Guillin, and Villani [14, Theorem 2.7]. \square

4.1.2 Quasi-Monte Carlo Approximations

Quasi-Monte Carlo sequences mimic some, but not all of the properties of Monte Carlo sequences from the uniform distribution in the m -dimensional hypercube $U[0, 1]^m$. To a sequence $\tilde{\xi}_n$ we associate the *empirical measure* $\tilde{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\tilde{\xi}_i}$. The sequence $\tilde{\xi}_n$ is called a Quasi Monte Carlo sequence in $[0, 1]^m$, if its star-discrepancy D^* satisfies

$$D^*(\tilde{P}_n, U[0, 1]^m) \leq C \frac{(\log n)^m}{n}$$

for all n and some constant C . Here, the star-discrepancy is

$$D^*(\tilde{P}_n, U[0, 1]^m) = \sup \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{1}_R(\tilde{\xi}_i) - \lambda(R) : R \in \mathfrak{R}_0 \right\},$$

where the supremum is over the set \mathfrak{R}_0 of all rectangles in $[0, 1]^m$, which have the origin $(0, \dots, 0)$ as one cornerpoint. $\lambda(R)$ is the Lebesgue-measure (area/volume/ m -dimensional volume) of the rectangle R . The star-discrepancy D^* equals

$$D^*(\tilde{P}_n, U[0, 1]^m) = \sup_u |\tilde{G}_n(u) - G(u)|,$$

where \tilde{G}_n is the empirical distribution of \tilde{P}_n and $G(u) = \prod_{i=1}^m u_i$ is the distribution of the uniform distribution. It is related to the discrepancy D defined as

$$D(\tilde{P}_n, U[0, 1]^m) = \sup \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{1}_R(\tilde{\xi}_i) - \lambda(R) : R \in \mathfrak{R} \right\},$$

where the supremum is over the set \mathfrak{R} of all rectangles in $[0, 1]^m$ by

$$\sup_u |\tilde{G}_n(u) - G(u)| \leq D(\tilde{P}_n, U[0, 1]^m) \leq 2^m \sup_u |\tilde{G}_n(u) - G(u)|.$$

In general, there is no relation between the star-discrepancy and the Wasserstein distance. However, in the case of a compact support, e.g. the unit cube, some relations can be found: for the univariate situation ($m = 1$), the following bound is obvious:

$$\mathbf{d}_1(\tilde{P}_n, U[0, 1]) = \int_0^1 |\tilde{G}_n(u) - G(u)| du \leq \sup_u |\tilde{G}_n(u) - G(u)| = D^*(\tilde{P}_n, U[0, 1]).$$

In the general case one may find a relation in the following way: let ϵ be the side length of the largest cube $\times_{i=1}^m (a_i, a_i + \epsilon)$, which does not contain a mass point of \tilde{P}_n . Obviously $\epsilon^m \leq D^*(\tilde{P}_n, U[0, 1]^m)$. This implies that the largest distance of any point in $[0, 1]^m$ to the next mass point of \tilde{P}_n is bounded by the diameter $\epsilon\sqrt{m}$. Therefore also the Wasserstein distance can be bounded by

$$\mathbf{d}_r(\tilde{P}_n, U[0, 1])^r \leq (\sqrt{m}\epsilon)^r \leq m^{r/2} \cdot D^*(\tilde{P}_n, U[0, 1]^m)^{r/m}.$$

It follows that a low discrepancy sequence has also a small Wasserstein distance to the uniform distribution (cf. Glasserman [46]).

4.1.3 Optimal and Nearly Optimal Single-Period Discretizations

Let P be a probability measure on \mathbb{R}^m endowed with some distance d . To this distance one may associate the pertaining Wasserstein (transportation-) distance denoted by the same symbol d (d_r , resp.). The problem of optimal discretization consists in finding a discrete probability \tilde{P} sitting on (not more than) s points, such that its distance to P ,

$$\min \{d_r(P, \tilde{P}) : \tilde{P} \text{ sits on } s \text{ points}\}, \quad (4.3)$$

is minimal. The minimal value of (4.3) is called the *quantization error*.

In the previous section the supporting points of the approximating probability measure have been sampled by the Monte Carlo method, and every of these s sampled points has been assigned a mass (or probability) of $1/s$. In this section we intend to find the solution of the quantization problem (4.3), or at least an approximate solution.

It was already mentioned (Theorem 2.25) that discrete measures are dense in the space (\mathcal{P}_r, d_r) , that is to say for every $P \in \mathcal{P}_r$ there is a measure $\tilde{P} = \sum_{i=1}^n P_i \cdot \delta_{z^{(i)}} \in \mathcal{P}_r$ with finite support $Z = (z^{(1)}, \dots, z^{(s)})$ and masses P_i such that $d_r(P, \tilde{P}) < \varepsilon$ for every arbitrary $\varepsilon > 0$.

We address the problem of finding good approximations \tilde{P} in two steps. First, given the locations $z^{(i)}$, $i = 1, \dots, s$, we find the best probabilities P_i such that $d_r(P, \sum_{i=1}^n P_i \cdot \delta_{z^{(i)}})$ is as small as possible. This problem has an explicit solution. Next we address the question, how many supporting points $z^{(i)}$ are necessary in order to achieve a desired approximation quality in terms of the Wasserstein distance, and where they should be located. This approximation is based on iterative algorithms, which are presented in Sect. 4.1.4.

In some rare cases the optimal quantization can be found in an analytic manner, as the following two examples illustrate.

Example 4.7 (Laplace Distribution in \mathbb{R}^1 , cf. Graf and Luschgy [48]). For the Euclidean distance on \mathbb{R} and the Laplace distribution with density $g(x) = \frac{1}{2} \exp(-|x|)$, the optimal supporting points for even $s = 2k$ are

$$z^{(i)} = \begin{cases} 2 \log \left(\frac{i}{\sqrt{k^2+k}} \right) & \text{if } 1 \leq i \leq k, \\ 2 \log \left(\frac{\sqrt{k^2+k}}{s+1-i} \right) & \text{if } k+1 \leq i \leq s. \end{cases}$$

The quantization error is given by

$$\begin{cases} \log \left(1 + \frac{2}{s} \right) & \text{if } s \text{ is even,} \\ \frac{2}{s+1} & \text{if } s \text{ is odd,} \end{cases}$$

such that an approximation quality of ε can be obtained by $\sim 2/\varepsilon$ points.

Example 4.8 (The Exponential Distribution in \mathbb{R}^1). For the Euclidean distance on \mathbb{R} and the exponential distribution with density $g(x) = \exp(-x) \mathbf{1}_{\{x \geq 0\}}$, the optimal supporting points are

$$z^{(i)} = 2 \log \left(\frac{\sqrt{s^2 + s}}{s + 1 - i} \right); \quad i = 1, \dots, s.$$

The quantization error is $\log \left(1 + \frac{1}{s} \right)$. It follows that s points can approximate this measure with a quality of $\sim \frac{1}{s}$.

4.1.3.1 Optimal Probabilities

For a finite set Z of possible supporting points we consider probability measures \tilde{P} , with support contained in Z , i.e., measures of the form $\tilde{P} = \sum_{z \in Z} P_z \cdot \delta_z$, where $P_z \geq 0$ and $\sum_{z \in Z} P_z = 1$ (i.e., they satisfy $\tilde{P}(Z) = 1$).

We intend to identify the probability measures \tilde{P} with support contained in Z , which approximate a given measure P as good as possible. It turns out that the best probability masses P_z do not depend on the order r of the Wasserstein distance d_r . For this reason we give the best approximations just for the basic distance d .

In order to characterize the probability measure located on Z and approximating P in the best possible way consider the pushforward probability measure $P^T := P \circ T^{-1}$, where T is the *transport map*, also called *quantizer*

$$T : \mathbb{R}^m \rightarrow Z \text{ with } T(\xi) \in \operatorname{argmin}_{z \in Z} d(\xi, z), \quad (4.4)$$

assigning to $\xi \in \mathbb{R}^m$ its nearest point in Z in a measurable way. To be more precise, one could define

$$T(\xi) = z^{(i)} \quad \text{on the set } \left\{ \xi \in \mathbb{R}^m \mid \begin{array}{l} d(\xi, z^{(i)}) = \min_j d(\xi, z^{(j)}) \text{ and} \\ d(\xi, z^{(k)}) > \min_j d(\xi, z^{(j)}) \text{ for } k < i \end{array} \right\} \quad (4.5)$$

to ensure measurability. The following lemma clarifies that P^T is the probability measure located on Z approximating P in the best possible way, i.e., P^T solves the problem

$$\min d_r(P, \tilde{P}) \quad \text{subject to } \tilde{P}(Z) = 1.$$

Lemma 4.9 (Lower Bounds and Best Probabilities). *Let P be a probability measure and T the transport map defined in (4.5).*

- (i) *For any measure \tilde{P} with $\tilde{P}(Z) = 1$ and any π having marginals P and \tilde{P} it holds that*

$$\iint \mathbf{d}^r d\pi \geq \int \min_{z \in Z} \mathbf{d}(\xi, z)^r P(d\xi). \quad (4.6)$$

(ii) The measure $\tilde{P} = P^T$ minimizes (4.6) and satisfies^{1,2}

$$\mathbf{d}_r(P, P^T)^r = \min_{\pi} \iint \mathbf{d}^r d\pi = \int \min_{z \in Z} \mathbf{d}(\xi, z)^r P(d\xi) = \mathbb{E}_P [\mathbf{d}(\text{id}, T(\text{id}))^r] \quad (4.7)$$

where the min is over all π with marginals P and P^T .

Proof. Let π have the marginals of P and \tilde{P} . Then

$$\begin{aligned} \iint_{\mathbb{R}^m \times Z} \mathbf{d}(\xi, z)^r \pi(d\xi, dz) &\geq \int_{\mathbb{R}^m} \int_Z \min_{z \in Z} \mathbf{d}(\xi, z)^r \pi(d\xi, dz) \\ &= \int_Z \min_{z \in Z} \mathbf{d}(\xi, z)^r P(d\xi), \end{aligned}$$

because \tilde{P} is located on Z , and thus (4.6) holds.

Employing the transport map T , define the transport plan $\pi := P^{\text{id} \times T}$, where id is the identity on \mathbb{R}^m , that is

$$\pi(A \times B) = P(\{\xi : (\xi, T(\xi)) \in A \times B\}) = P(A \cap T^{-1}(B)). \quad (4.8)$$

π is feasible, it has the marginals $\pi(A \times Z) = P(\{\xi : \xi \in A, T(\xi) \in Z\}) = P(A)$ and $\pi(\mathbb{R}^m \times B) = P(\{\xi : T(\xi) \in B\}) = P^T(B)$. Thus

$$\iint_{\mathbb{R}^m \times Z} \mathbf{d}(\xi, z)^r \pi(d\xi, dz) = \int_{\mathbb{R}^m} \mathbf{d}(\xi, T(\xi))^r P(d\xi) = \int_{\mathbb{R}^m} \min_{z \in Z} \mathbf{d}(\xi, z)^r P(d\xi),$$

which is (4.7). \square

Remark 4.10. Lemma 4.9 identifies P^T as the measure closest to P and supported just by Z . Moreover (4.8) is the optimal transport plan. The corresponding optimal dual variables are

$$\lambda(\xi) = \min_{z \in Z} \mathbf{d}(\xi, z)^r = \mathbf{d}(\xi, T(\xi))^r \text{ and } \mu(z) = 0,$$

so that with the choice (4.7), the corresponding dual problem (2.27) is solved.

It should be noted as well that the transport map T and therefore the optimal P^T does *not depend* on the order r .

¹See also [36, Theorem 2].

² id is the identity.

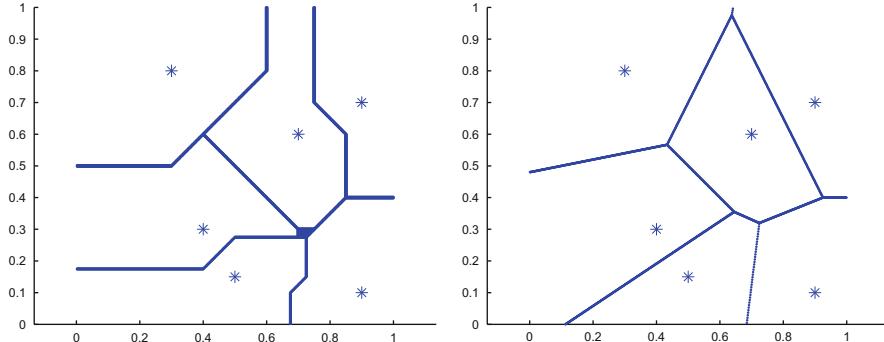


Fig. 4.2 Voronoi tessellations for the 1-distance $d(u, v) = |u_1 - v_1| + |u_2 - v_2|$ (left) and the Euclidean distance $d(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2}$ (right)

Given a probability measure P , Lemma 4.9 allows us to find the best approximation with support $Z = (z^{(1)}, \dots, z^{(s)})$. Based on the set Z (which can be seen as a $m \times s$ matrix) introduce the *Voronoi tessellation* $\mathcal{V}_Z = \{V_Z^{(i)} : i = 1, \dots, s\}$ of \mathbb{R}^m , where $V_Z^{(i)} = \{\mathbf{T}(\xi) = z^{(i)}\}$. According to the rule (4.5) the Voronoi sets, i.e.,

$$V_Z^{(i)} = \left\{ u \in \mathbb{R}^m \left| \begin{array}{l} d(u, z^{(i)}) = \min_j d(u, z^{(j)}) \text{ and} \\ d(u, z^{(k)}) > \min_j d(u, z^{(j)}) \text{ for } k < i \end{array} \right. \right\}, \quad (4.9)$$

form a partition (for a comprehensive treatment see Graf and Luschgy [48] and the œuvre of G. Pagès, e.g. [8]). Notice that the Voronoi partition depends on the chosen distance. For an illustration, see Fig. 4.2.

Denote by P_Z the optimal quantization of P with support Z , i.e.,

$$P_Z = \sum P(V_Z^{(i)}) \cdot \delta_{z^{(i)}}.$$

Calculating $P(V_Z^{(i)})$ requires a multidimensional integration over complicated, typically polyhedral sets in general. If, however, the basic P has finite support, i.e., $P = \sum_j P_j \delta_{\xi_j}$, then P_Z takes the explicit form

$$P_Z = \sum_i P_Z^{(i)} \delta_{z_j^{(i)}}, \quad (4.10)$$

where

$$P(V_Z^{(i)}) = P_Z^{(i)} = \sum_{\{j : \xi_j \in V_Z^{(i)}\}} P_j, \quad (4.11)$$

which is computationally easy.

4.1.3.2 Optimal Location of Supporting Points

Having identified the best probability weights, it is of subsequent interest to find good locations for the support of the approximating probability measure. This problem is often referred to as *facility location* problem. Finding locations $\{z^{(1)}, \dots, z^{(s)}\}$ minimizing the distance to P is, in view of (4.7), minimizing the map

$$Z = \{z^{(1)}, \dots, z^{(s)}\} \mapsto D(Z) = \int \min_i d(\xi, z^{(i)})^r P(d\xi) \quad (4.12)$$

globally over $(z^{(1)}, \dots, z^{(s)}) \in [\mathbb{R}^m]^s$. Optimal points of (4.12) are sometimes called *principal points* or *representative points*.

This problem is not convex and typically hard to solve. However, given a set of locations $\{z^{(1)}, \dots, z^{(s)}\}$, the following lemma reveals an improvement which can be exploited to successively compute improved locations in an algorithm.

Lemma 4.11 (Improved Locations). *Let P be a probability measure and \mathbf{T} the transport map defined in (4.4). Then*

$$d_r(P, P^{\tilde{\mathbf{T}}}) \leq d_r(P, P^{\mathbf{T}}),$$

where the map $\tilde{\mathbf{T}}$ is the conditional expectation, defined as $\tilde{\mathbf{T}} = \mathbb{E}_P(\text{id} | \mathbf{T}) \circ \mathbf{T}$ (i.e., $\tilde{\mathbf{T}}(\xi) := \mathbb{E}_P(\tilde{\xi} | \mathbf{T}(\tilde{\xi})) = \mathbf{T}(\tilde{\xi})$).

Proof. For the last assertion apply the conditional Jensen's inequality to the convex function $\varphi_\xi: y \mapsto d(\xi, y)^r$ and obtain $\varphi_\xi \circ \mathbb{E}(\text{id} | \mathbf{T}) \leq \mathbb{E}(\varphi_\xi \circ \text{id} | \mathbf{T})$, from which follows (cf. Fig. 4.3) that

$$d(\xi, \mathbb{E}(\text{id} | \mathbf{T}) \circ \mathbf{T})^r \leq \mathbb{E}(d_r(\xi, \text{id})^r | \mathbf{T}) \circ \mathbf{T}.$$

The measure $\tilde{\pi}(A \times B) := P(A \cap \tilde{\mathbf{T}}^{-1}(B))$ has marginals P and $P^{\tilde{\mathbf{T}}}$. Hence

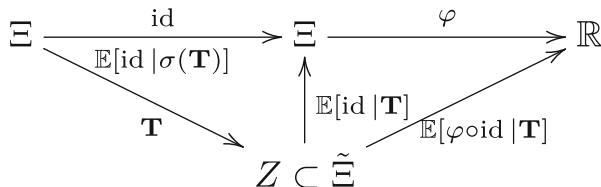


Fig. 4.3 The diagram displays the domain and the codomain for conditional expectation. It does not commute, but it holds that $\mathbb{E}[\text{id} | \mathbf{T}] \circ \mathbf{T} = \mathbb{E}[\text{id} | \sigma(\mathbf{T})]$ and $\mathbb{E}[\varphi \circ \text{id} | \mathbf{T}] \circ \mathbf{T} = \mathbb{E}[\varphi \circ \text{id} | \sigma(\mathbf{T})]$

Algorithm 4.1

A typical hierarchical cluster algorithm (complete linkage), clustering n points into s clusters

- (i) **INITIALIZATION.** Suppose that n points $(\eta^{(1)}, \dots, \eta^{(n)})$ in \mathbb{R}^m (endowed with the metric d) are given. The set $\{\eta^{(i)} : i = 1, \dots, n\}$ is iteratively partitioned into disjoint clusters, such that their number decreases from step to step. At the beginning every point is a cluster of itself.
- (ii) **ITERATION.** Suppose that the current partition of the set is $Z = \bigcup_j C_j$. Find the pair of clusters C_j, C_k , such that

$$\sup \{d(\eta, \eta') : \eta \in C_j, \eta' \in C_k\}$$

is minimal. Create a new cluster by merging C_j and C_k .

- (iii) **STOPPING CRITERION.** If the number of clusters has decreased to the desired number s , then stop. Otherwise goto (ii).

$$\begin{aligned} d_r(P, P^{\tilde{T}})^r &\leq \int d(\xi, \tilde{T}(\xi))^r P(d\xi) = \int d(\xi, \mathbb{E}(\text{id} | T) \circ T(\xi))^r P(d\xi) \\ &\leq \int \mathbb{E}(d(\xi, \text{id})^r | T)(T(\xi)) P(d\xi) = \int d(\xi, T(\xi))^r P(d\xi) \\ &= d_r(P, P^T)^r, \end{aligned}$$

which is the desired assertion. \square

While the conditional expectation leads to an improvement of the approximation distance, it is the *optimal* improvement, if the distance is the Euclidean norm, $d(u, v) = \|u - v\|_2$. This follows from the fact that the conditional expectation $Z^* := \mathbb{E}[\xi | \mathcal{F}]$ minimizes the function

$$Z \mapsto \mathbb{E}(\xi - Z)^2 = \mathbb{E}\|\xi - Z\|_2^2 \quad (4.13)$$

among all \mathcal{F} -measurable random variables Z . If the underlying distance is not the two-norm, one has to consider the mapping $\tilde{P} \mapsto \operatorname{argmin} \{\int d(z, \xi)^r \tilde{P}(d\xi) : z \in \mathbb{R}^m\}$ applied for the probabilities P conditioned on the Voronoi sets $V^{(i)}$. This leads to the Algorithm 4.2. This algorithm is of iterative nature and needs a starting point set Z . To get such a starting point configuration a cluster algorithm (for example, Algorithm 4.1) can be used (there are many variants of this cluster algorithms (single linkage, average linkage etc.), we refer to general books on cluster algorithms here, for example Hartigan [51]).

Algorithm 4.2

Optimal discretization of the probability measure P by a discrete probability sitting on s points: a deterministic, but numerically difficult algorithm

-
- (i) **INITIALIZATION.** Generate n random points $\{\eta^{(i)} : 1 \leq i \leq n\}$ from the distribution P , where n is much larger (e.g. 10 times larger) than s . Use a cluster algorithm like Algorithm 4.1 to define s clusters C_1, \dots, C_s . Find the representations of the clusters as

$$z^{(i)} \in \operatorname{argmin}_z \left\{ \sum_{\eta \in C_i} d(z, \eta)^r \right\}.$$

Form the initial point set $Z(0) := \{z^{(i)} : 1 \leq i \leq s\}$ and set $k = 0$.

- (ii) **VORONOI PARTITION.** Find the Voronoi sets $V_{Z(k)}^{(i)}$ for $1 \leq i \leq s$ according to (4.9).
 (iii) **OPTIMIZATION STEP.** For all i find the *center of order* r of each Voronoi set, that is, let

$$z^{(i)}(k+1) \in \operatorname{argmin}_z \left\{ \int_{V_{Z(k)}^{(i)}} d(\xi, z)^r P(d\xi) \right\}.$$

Form the new set $Z(k+1) = \{z^{(i)}(k+1) : 1 \leq i \leq s\}$.

- (iv) **INTEGRATION STEP.** Calculate $D(Z(k+1))$ according to (4.12). Stop if $|D(Z(k+1)) - D(Z(k))|$ is smaller than a tolerance. Otherwise set $k := k + 1$ and goto (ii).
-

4.1.3.3 The Deterministic Iteration

Recall that our goal is to find optimal quantizers, i.e., global minimizer of $Z \mapsto D(Z)$, with

$$D(Z) = \int \min_{z \in Z} d(\xi, z)^r P(d\xi) \quad (4.14)$$

among all sets Z of cardinality s . Since this problem is nonconvex and any permutation of the minimum is a minimum as well, also locally optimal solutions are of interest.

Computing (4.14) requires evaluating integrals with respect to P , as well as nonlinear optimizations to be carried out numerically. This is typically a difficult task, especially for higher dimensions. However, we assume in this section that integration is possible and present an algorithm (Algorithm 4.2) to obtain the minimum of (4.12). The next section provides a stochastic algorithm, which does not require evaluating the integrals explicitly.

Before introducing the algorithms we mention the necessary conditions of optimality, and thus the differentiability properties of the mapping $Z \mapsto D(Z)$. Suppose that $z \mapsto d(\xi, z)$ is almost everywhere (with respect to the Lebesgue measure) differentiable with derivative $\nabla_z d(\xi, z)$, and let $\nabla_{z^{(i)}} D(Z)$ be the column vector consisting of the entries

$$\int_{V_Z^{(i)}} r d(\xi, z^{(i)})^{r-1} \cdot \nabla_z d(\xi, z^{(i)}) P(d\xi), \quad i = 1, \dots, s$$

$(V_Z^{(i)})$ are the Voronoi partitions defined in (4.9)). The collection of all these vectors is denoted by $\nabla_Z D(Z)$, it is an $m \times s$ matrix. Then $Z \mapsto D(Z)$ is as well differentiable with derivative $\nabla_Z D(Z)$ (cf. Pflug [88, Corollary 3.52, page 184]).

Example 4.12 (Derivative of Selected Distances). Typical examples for distances in \mathbb{R}^m are

- (i) the weighted ℓ^1 -distance $d(u, v) = \sum_{j=1}^m w_j |u_j - v_j|$, where w_i are some positive weights. The role of weights is to make different components of the scenario vector comparable, especially, if they represent different dimensions in different units, like prices, quantities, demands etc.;
- (ii) the weighted ℓ^2 -distance $d(u, v) = \sqrt{\sum_{j=1}^m w_j (u_j - v_j)^2}$;
- (iii) the weighted ℓ^p -distance $d(u, v) = \left(\sum_{j=1}^m w_j |u_j - v_j|^p \right)^{1/p}$, a generalization of the two previous cases.

The corresponding derivatives are

- (i) $\nabla_{z_i} d(u, z) = w_i \text{ sign}(z_i - u_i)$,
- (ii) $\nabla_{z_i} d(u, z) = w_i \frac{z_i - u_i}{d(u, z)}$ and
- (iii) $\nabla_{z_i} d(u, z) = w_i \frac{|z_i - u_i|^{p-1}}{d(u, z)^{p-1}} \text{ sign}(z_i - u_i)$.

Proposition 4.13. If $Z(k)$ is the sequence of point sets generated by the deterministic iteration algorithm (Algorithm 4.2), then

$$D(Z(k+1)) \leq D(Z(k)).$$

If $D(Z(k^* + 1)) = D(Z(k^*))$ for some k^* , then $D(Z(k)) = D(Z(k^*))$ for all $k \geq k^*$ and

$$\nabla_{z^{(i)}} D(Z(k^*)) = 0 \text{ for all } i.$$

Proof. Notice that

$$\begin{aligned} D(Z(k)) &= \int \min_{z \in Z(k)} d(\xi, z)^r P(d\xi) = \sum_{i=1}^s \int_{V_{Z(k)}^{(i)}} d(\xi, z^{(i)}(k))^r P(d\xi) \\ &\geq \sum_{i=1}^s \int_{V_{Z(k)}^{(i)}} d(\xi, z^{(i)}(k+1))^r P(d\xi) \\ &\geq \int \min_i d(\xi, z^{(i)}(k+1)) P(d\xi) = D(Z(k+1)). \end{aligned}$$

If $D(Z(k^* + 1)) = D(Z(k^*))$, then necessarily, for all i and all j ,

$$z_j^{(i)}(k) \in \operatorname{argmin}_y \left\{ \int_{V_{Z(k)}^{(i)}} d(\xi, y)^r P(d\xi) \right\},$$

Algorithm 4.3

Iterative search for optimal discretizations in \mathbb{R} for the Wasserstein distance of order $r = 1$

- **INITIALIZATION.** Start with some initial values $z^{(1)}, \dots, z^{(s)}$
- **ITERATION STEP.** Find the pertaining breakpoints b_i , the intervals I_i and the probabilities p_i according to (4.16) and (4.17). Define the new points $z_{new}^{(i)}$ as the conditional medians within I_i

$$z_{new}^{(i)} = G^{-1} \left(\frac{G(b_{i-1}) + G(b_i)}{2} \right)$$

- **STOPPING CRITERION.** Iterate until convergence or until the change of the configuration is smaller than some threshold.

which is equivalent to

$$\int_{V_{Z(k)}^{(i)}} r d(\xi, z^{(i)}(k))^{r-1} \cdot \nabla_{z^{(i)}(k)} d(\xi, z^{(i)}(k)) P(d\xi) = 0,$$

i.e., $\nabla_Z D(Z(k^*)) = 0$ and evidently, the iteration has reached a fixpoint. \square

We remark here that this method is related to the k-means method of cluster analysis (see, e.g., McQueen [75]), sometimes Lloyd's algorithm or Voronoi-iteration.

Remark 4.14. The most problematic step in Algorithm 4.2 is the **OPTIMIZATION STEP**. It requires not only a multidimensional integration over a polyhedral set but also an optimization. Some special cases are, however, easy:

- If P has finite support, i.e., if $P = \sum_{i=1}^N P_i \delta_{\xi_i}$, $r = 2$ and $d(u, v) = \|u - v\|_2$, then

$$z^{(i)}(k+1) := \frac{\sum_{\xi_j \in V_{Z(k)}^{(i)}} P(\xi_j) \xi_j}{\sum_{\xi_j \in V_{Z(k)}^{(i)}} P(\xi_j)}$$

can be written as an expression with finite sums.

- If P has finite support as before, $r = 1$ and $d(u, v) = \|u - v\|_1$, then $z^{(i)}(k+1)$ can be found as the solution of the linear program

$$\min_z \frac{\sum_{\xi_j \in V_{Z(k)}^{(i)}} P(\xi_j) \|z - \xi_j\|_1}{\sum_{\xi_j \in V_{Z(k)}^{(i)}} P(\xi_j)}.$$

- If the dimension is $m = 1$, then these steps can be analytically performed for any probability P , see Algorithms 4.3 and 4.4.

Algorithm 4.4

Iterative search for optimal discretizations in \mathbb{R} with for the Wasserstein distance of order $r = 2$

- **INITIALIZATION.** Start with some initial values $z^{(1)}, \dots, z^{(s)}$
- **ITERATION STEP.** Find the breakpoints b_i , the intervals I_i and the probabilities p_i according to (4.16) and (4.17). Define the new points as the conditional means

$$z_{new}^{(i)} = \frac{1}{p_i} \int_{I_i} u dG(u)$$

- **STOPPING CRITERION.** Iterate until convergence or until the change of the configuration is smaller than some threshold.

Optimal One-Dimensional Discretization. We specialize the previous algorithms to the one-dimensional case. Let the probability measure on \mathbb{R} have the distribution function G . We aim at finding a discrete probability \tilde{P} sitting on (not more than) s points $z^{(1)}, \dots, z^{(s)}$ with probabilities p_1, \dots, p_s such that the Wasserstein distance $d_r(P, \tilde{P})$ is minimal. For the case $d(u, v) = |u - v|$ and $r = 1$ the problem can be written as

$$\text{minimize } D(z^{(1)}, \dots, z^{(s)}) = \int \min_i |u - z^{(i)}| dG(u). \quad (4.15)$$

For every given configuration $z^{(1)} < z^{(2)} < \dots < z^{(s)}$ define the pertaining breakpoints as

$$b_i = \frac{1}{2} (z^{(i)} + z^{(i+1)}) \text{ for } i = 1, \dots, s-1 \text{ and } b_0 = -\infty, b_s = \infty, \quad (4.16)$$

as well as the intervals $I_i = [b_{i-1}, b_i]$ for $i = 1, \dots, s$ and the probabilities $p_i = G(b_i) - G(b_{i-1})$. In order that a configuration is at least a local minimizer of (4.15), $z^{(i)}$ must be the conditional median of G within I_i , i.e., the following equations must hold:

$$z^{(i)} = G^{-1} \left(p_1 + \dots + p_{i-1} + \frac{1}{2} p_i \right) = G^{-1} \left(\frac{G(b_{i-1}) + G(b_i)}{2} \right). \quad (4.17)$$

Algorithm 4.3 finds at least a locally optimal solution of (4.15).

A similar algorithm (Algorithm 4.4) works for the case $r = 2$, i.e., the minimization of

$$D(z^{(1)}, \dots, z^{(s)}) = \int \min_i (u - z_i)^2 dG(u).$$

Algorithm 4.5

Optimal discretization of the probability measure P by a discrete probability sitting on s points for the distance d : a stochastic approximation algorithm

- (i) **INITIALIZATION.** Sample n random variates from distribution P , where n is much larger than s . Use a cluster algorithm, e.g., Algorithm 4.1 to find s clusters. Let $Z(0) = (z^{(1)}(0), \dots, z^{(s)}(0))$ be the cluster medians. Choose the sequence $a_k > 0$ with $\sum_{k=0}^{\infty} a_k = \infty$ and $\sum_{k=0}^{\infty} a_k^2 < \infty$ (according to Theorem 4.17) and set $k = 0$.
- (ii) **ITERATION.** Use a new independent sample $\xi(k)$ for the following stochastic optimization step: find the index $i \in \{1, \dots, s\}$ such that

$$d(\xi(k), z^{(i)}(k)) = \min_{\ell} d(\xi(k), z^{(\ell)}(k)).$$

Set

$$z^{(i)}(k+1) = z^{(i)}(k) - a_k \cdot r d(\xi(k), z^{(i)})^{r-1} \cdot \nabla_z d(\xi(k), z^{(i)}),$$

and leave all other points unchanged to form the new point set $Z(k+1)$.

- (iii) **STOPPING CRITERION.** Set $k = k + 1$ and goto (ii). Stop, if either the predetermined number of iterations are performed or if the relative change of the point set Z is below some threshold ϵ .
- (iv) **DETERMINATION OF THE PROBABILITIES.** After having fixed the final point set Z , generate another sample $\xi(1), \dots, \xi(n)$ and find the probabilities

$$p_i = \frac{1}{n} \# \left\{ \ell : d(\xi(\ell), z^{(i)}) = \min_k d(\xi(\ell), z^{(k)}) \right\}.$$

The final approximate distribution is $\tilde{P} = \sum_{i=1}^s p_i \cdot \delta_{z^{(i)}}$.

4.1.4 The Stochastic Approximation Algorithms for Multidimensional Quantization

As was remarked, the facility location problem (4.12) requires multidimensional integration. However, there is an algorithm of the stochastic approximation type which does not require the knowledge of the distribution G , but needs only a random number generator for G . This algorithm further avoids the optimization and integration steps of Algorithm 4.2 by using *stochastic approximation*. It requires that we can sample independent and identically distributed sequences of vectors of arbitrary length n ,

$$\xi(1), \dots, \xi(n),$$

each $\xi(i)$ distributed according to P . The validity of Algorithm 4.5, i.e., its convergence to a local optimum is stated in Theorem 4.17, its proof needs Propositions 4.15 and 4.16.

Proposition 4.15. *Suppose that $\mathfrak{F} = (\mathcal{F}_1, \mathcal{F}_2, \dots)$ is a filtration and (Y_k) is a sequence of random variables, which are uniformly bounded from below and*

adapted to \mathfrak{F} . In addition, let (A_k) and (B_k) be sequences of nonnegative random variables also adapted to \mathfrak{F} . If the recursion

$$\mathbb{E}[Y_{k+1}|\mathcal{F}_k] \leq Y_k - A_k + B_k \quad (4.18)$$

is satisfied and $\sum_k B_k < \infty$ a.s., then Y_k converges and $\sum_k A_k < \infty$ a.s.

Proof. Let $S_k = \sum_{\ell=1}^k B_\ell$ and $T_k = \sum_{\ell=1}^k A_\ell$. Then (4.18) implies that

$$\mathbb{E}[Y_{k+1} - S_k | \mathcal{F}_k] = \mathbb{E}[Y_{k+1} | \mathcal{F}_k] - S_{k-1} - B_k \leq Y_k - A_k - S_{k-1} \leq Y_k - S_{k-1}.$$

Hence $Y_{k+1} - S_k$ is a supermartingale, which is bounded from below and therefore converges a.s. Since S_k converges by assumption, it follows that Y_k converges a.s. Similarly,

$$\mathbb{E}[Y_{k+1} - S_k + T_k | \mathcal{F}_k] \leq \mathbb{E}[Y_{k+1} | \mathcal{F}_k] - S_{k-1} - B_k + T_{k-1} + A_k \leq Y_k - S_{k-1} + T_{k-1}$$

and $Y_{k+1} - S_k + T_k$ converges a.s., which implies that T_k converges a.s. \square

Proposition 4.16. *Let $F(\cdot)$ be a real function defined on \mathbb{R}^d , which has a Lipschitz-continuous derivative $f(\cdot)$. Consider a recursion of the form*

$$X_{k+1} = X_k - a_k f(X_k) + a_k R_{k+1} \quad (4.19)$$

with some starting point X_0 , where $\mathbb{E}[R_{k+1} | R_1, \dots, R_k] = 0$.

If $a_k \geq 0$, $\sum_k a_k = \infty$ and $\sum_k a_k^2 \|R_{k+1}\|^2 < \infty$ a.s., then $F(X_k)$ converges. If further $\sum_k a_k R_{k+1}$ converges a.s., then $f(X_k)$ converges to zero a.s.

Proof. Let $Y_k := F(X_k)$ and let K be the Lipschitz constant of f . Using the recursion (4.19) and the mean value theorem, there is a $\theta \in [0, 1]$ such that

$$\begin{aligned} F(X_{k+1}) &= F(X_k) + f(X_k + \theta(-a_k f(X_k) + a_k R_{k+1}))^\top \\ &\quad \cdot (-a_k f(X_k) + a_k R_{k+1}) \\ &\leq F(X_k) + f(X_k)^\top \cdot (-a_k f(X_k) \\ &\quad + a_k R_{k+1}) + K \| -a_k f(X_k) + a_k R_{k+1} \|^2 \\ &\leq F(X_k) - a_k \|f(X_k)\|^2 + a_k f(X_k) R_{k+1} + 2K a_k^2 \|f(X_k)\|^2 \\ &\quad + 2K a_k^2 \|R_{k+1}\|^2. \end{aligned}$$

Taking the conditional expectation with respect to R_1, \dots, R_k one gets

$$\begin{aligned} \mathbb{E}[F(X_{k+1}) | R_1, \dots, R_k] &\leq F(X_k) - a_k \|f(X_k)\|^2 + 2K a_k^2 \|f(X_k)\|^2 \\ &\quad + 2K a_k^2 \|R_{k+1}\|^2 \\ &\leq F(X_k) - \frac{a_k}{2} \|f(X_k)\|^2 + 2K a_k^2 \|R_{k+1}\|^2 \end{aligned}$$

for k large enough. Proposition 4.15, applied for $Y_k = F(X_k)$, $A_k = \frac{a_k}{2} \|f(X_k)\|^2$ and $B_k = 2Ka_k^2 \|R_{k+1}\|^2$, implies now that $F(X_k)$ converges and

$$\sum_k a_k \|f(X_k)\|^2 < \infty \quad (4.20)$$

a.s. It remains to be shown that $f(X_k) \rightarrow 0$ a.s. Since $\sum_k a_k = \infty$, it follows from (4.20) that $\liminf_k \|f(X_k)\| = 0$ a.s. We argue now pointwise on the set of probability 1, where $\sum_k a_k \|f(X_k)\|^2 < \infty$, $\liminf_k \|f(X_k)\| = 0$ and $\sum_k a_k R_{k+1}$ converges. Suppose that $\limsup_k \|f(X_k)\|^2 > 2\epsilon$. Let $m_\ell < n_\ell < m_{\ell+1}$ be chosen such that

$$\begin{aligned} \|f(X_k)\|^2 &> \epsilon \text{ for } m_\ell < k \leq n_\ell \text{ and} \\ \|f(X_k)\|^2 &\leq \epsilon \text{ for } n_\ell < k \leq m_{\ell+1}. \end{aligned} \quad (4.21)$$

Let ℓ_0 be such large that

$$\sum_{k=m_{\ell_0}}^{\infty} a_k \|f(X_k)\|^2 \leq \frac{\epsilon^2}{2K} \quad \text{and} \quad \left\| \sum_{k=s}^t a_k R_{k+1} \right\| < \frac{\epsilon}{2} \quad \text{for all } s, t \geq m_{\ell_0}.$$

Then, for $\ell \geq \ell_0$ and $m_\ell \leq k \leq n_\ell$, by the recursion (4.19) and (4.21), as well as the Lipschitz property of f ,

$$\begin{aligned} \|f(X_{i+1}) - f(X_{m_\ell})\| &\leq K \|X_{i+1} - X_{m_\ell}\| = K \left\| \sum_{k=m_\ell}^i a_k f(X_k) + a_k R_{k+1} \right\| \\ &\leq K \sum_{k=m_\ell}^i a_k \|f(X_k)\| + K \left\| \sum_{k=m_\ell}^i a_k R_{k+1} \right\| \\ &\leq \frac{K}{\epsilon} \sum_{k=m_\ell}^i a_k \|f(X_k)\|^2 + \frac{\epsilon}{2} < \epsilon. \end{aligned}$$

Since $\|f(X_{m_\ell})\| \leq \epsilon$ it follows that $\limsup_k \|f(X_k)\| \leq 2\epsilon$ for every ϵ and this contradiction establishes the result. \square

Theorem 4.17. *Suppose that the step lengths a_k in Algorithm 4.5 satisfy $a_k \geq 0$, $\sum_k a_k = \infty$ and $\sum_k a_k^2 < \infty$. Suppose further that the assumptions of Proposition 4.16 are fulfilled. If $Z(k)$ is the sequence of point sets generated by the stochastic approximation Algorithm 4.5, then $D(Z(k))$ converges a.s. and*

$$\nabla_Z D(Z(k)) \rightarrow 0 \quad \text{a.s.}$$

as $k \rightarrow \infty$. In particular, if $D(Z)$ has a unique local minimizer Z^* , then

$$Z(k) \rightarrow Z^* \quad a.s.$$

Proof. The matrices $Z(k)$ satisfy the recursion

$$Z(k+1) = Z(k) - a_k \nabla_Z D(Z(k)) - a_k W(k)$$

with

$$\begin{aligned} W(k) &= \sum_{i=1}^s r \mathbf{d}(\xi(k), z^{(i)})^{r-1} \cdot \nabla_z \mathbf{d}(\xi(k), z^{(i)}) \cdot \mathbb{1}_{\xi(k) \in A_{Z(k)}^{(i)}} \\ &\quad - \int_{A_{Z(k)}^{(i)}} r \mathbf{d}(\xi, z^{(i)})^{r-1} \cdot \nabla_z \mathbf{d}(\xi, z^{(i)}) P(d\xi). \end{aligned}$$

Notice that the $W(k)$'s are independent and bounded, $\mathbb{E}[W(k)] = 0$ and $\sum_i a_i W(i)$ is convergent almost surely. Proposition 4.16 applied for $X_k = Z(k)$, $F(\cdot) = D(\cdot)$, $f(\cdot) = \nabla_Z D(\cdot)$ and $R_k = W(k)$ leads to the assertion. \square

Example. A good choice for the stepsizes a_k is

$$a_k = \frac{C}{(k+30)^{3/4}}.$$

These stepsizes satisfy the requirements $a_k \rightarrow 0$, $\sum_k a_k = \infty$ and $\sum_k a_k^2 < \infty$.

There are also algorithms which find the globally optimal discretization by the branch and bound method. However, these algorithms are such complex that only very small problems, say to find two or three optimal points in \mathbb{R}^2 or \mathbb{R}^3 can be effectively handled. In addition, the probability measure P must have bounded support, see, for instance, Algorithm 4.6.

4.1.5 Asymptotic Distribution of Optimal Quantizers

Even though Algorithm 4.5 provides successively improved approximations, it remains to quantify the quality of these approximations as the number of points increases to infinity. To this end we define the n -th quantization error.

Definition 4.18. The n -th quantization error of a probability measure P is

$$\mathbf{d}_{r;n}(P) := \inf \left\{ \mathbf{d}_r \left(P, \sum_{i=1}^n P_i \delta_{z_i} \right) : z_i \in \mathbb{R}^m, P_i \geq 0, \sum_{i=1}^n P_i = 1 \right\}.$$

Algorithm 4.6

Optimal discretization of probability P by a probability \tilde{P} sitting on s points:
A global optimization algorithm

- Suppose that the optimal configuration of s points in a bounded set, for simplicity in the unit cube $[0, 1]^m$ in \mathbb{R}^m , is to be found. The optimal configuration is an element of $[0, 1]^{sm}$. At stage ℓ the unit cube is dissected into smaller cubes, say $[0, 1]^m = \bigcup C_j$. By considering all selections $C_{j_1} \times C_{j_2} \times \cdots \times C_{j_s}$ a dissection of the search space is defined. The “local” problem finds a stochastic lower and a stochastic upper bound for

$$\min_{z^{(i)} \in C_{j_i}} \int \min_i d(u, z^{(i)}) P(du).$$

- BOUNDING STEP.** Configurations which have a lower bound larger than the upper bound of another configuration are excluded and not investigated further.
- BRANCHING STEP.** The best configuration will be refined by dissecting the pertaining cubes into smaller cubes.
- STOPPING CRITERION.** If the gap between the upper bound and the lower bound is small enough, then stop.

The n -th quantization error of a measure P is the Wasserstein distance of its best possible approximation supported by not more than n points (quantizers). It is thus a lower bound of any approximation supported by up to n points. To understand the quality of an approximation it is thus of essential interest to understand the quantization error.

The quantization error can be described—asymptotically—on the finite-dimensional space \mathbb{R}^m sufficiently precise. This is the content of the following theorem. The theorem is due to Zador [144], who stated it under additional assumptions; cf. also Na and Neuhoff [81].

Theorem 4.19 (Asymptotic Quantization Error). *Suppose that P is a probability measure on \mathbb{R}^m and $P \in \mathcal{P}_{r+\delta}(\mathbb{R}^m)$ for some $\delta > 0$. Then*

$$\lim_{n \rightarrow \infty} n^{\frac{1}{m}} \cdot d_{r;n}(P) = \left\| \frac{dP_a}{d\lambda^m} \right\|_{\frac{m}{m+r}}^{\frac{1}{r}} \cdot \inf_{n \geq 1} n^{\frac{1}{m}} \cdot d_{r;n}(U[0, 1]^m), \quad (4.22)$$

where $\frac{dP_a}{d\lambda^m}$ is the density of the absolutely continuous part of P with respect to the Lebesgue-measure λ^m on \mathbb{R}^m and $U[0, 1]^m$ is the uniform distribution on the unit cube $[0, 1]^m$. Moreover, the constant

$$\inf_{n \geq 1} n^{\frac{1}{m}} \cdot d_{r;n}(U[0, 1]^m) > 0$$

is strictly positive.

Remark 4.20. The constants $\inf_{n \geq 1} n^{\frac{1}{m}} \cdot d_{r;n}(U[0, 1]^m)$ depend on the norm used on \mathbb{R}^m . For the norm $\|x\|_\infty = \max |x_i|$ it is known (cf. Graf and Luschgy [48]) that

$$\inf_{n \geq 1} n^{\frac{1}{m}} \cdot d_{r;n}(U[0,1]^m) = \frac{m}{2^r(m+r)}.$$

This constant notably depends on r , while the order of convergence in (4.22) does not depend on r .

The following corollary formulates a result, which is important for our considerations.

Corollary 4.21. *For $P \in \mathcal{P}_{r+\delta}(\mathbb{R}^m)$ there are approximations \tilde{P}_n^* , located on no more than n points, such that*

$$d_r(P, \tilde{P}_n^*) = \mathcal{O}\left(n^{-\frac{1}{m}}\right). \quad (4.23)$$

Proof of Theorem 4.19 and Corollary 4.21. For the lengthy proof we refer to [48, Theorem 6.2]. \square

Remark 4.22. It should be noted that $\frac{m}{m+r} < 1$, so that $\|\cdot\|_{\frac{m}{m+r}}$ in (4.22) is not a norm; it is, however, a convenient abbreviation for $(\int g^p d\lambda^m)^{1/p} =: \|g\|_p$, even for $0 < p < 1$.

The condition $P \in \mathcal{P}_{r+\delta}(\mathbb{R}^m)$ ensures that $\left\| \frac{dP_a}{d\lambda^m} \right\|_{\frac{m}{m+r}} < \infty$ (cf. the discussion in Graf and Luschgy [48] subsequent to Theorem 6.2).

Remark 4.23. In order to obtain an approximating measure with distance no more than ε , a total of at least

$$n = \mathcal{O}(\varepsilon^{-m})$$

supporting points are necessary (cf. (4.23) and Nemirovski and Shapiro [131], as well as Shapiro [124] for related results).

Remark 4.24. For every choice $\{z^{(1)}, \dots, z^{(n)}\} \subset [0,1]^m$ the quantity

$$\int_0^1 \dots \int_0^1 \min_{z \in \{z^{(1)}, \dots, z^{(n)}\}} \|z - u\|^r du_1 \dots du_n$$

is an upper bound for $d_{r;n}(U[0,1]^m)$, so that upper bounds for the constant in (4.22) are easily available. Good choices typically take notice of the geometry of the norm and exploit clever arrangements in the unit cube $[0,1]^m$. Useful arrangements, however, depend on the norm $\|\cdot\|$.

4.1.5.1 Asymptotically Best Approximations

The discrete measures $P_n = \sum_{i=1}^n P_i^n \delta_{z_i^n}$, where the weights and locations are chosen such that

$$n^{\frac{1}{m}} \cdot d_r(P_n, P) \rightarrow \lim_n n^{\frac{1}{m}} \cdot d_{r;n}(P),$$

converge weakly to P ($P_n \rightarrow P$) by Theorem 2.23. To investigate the distribution of the quantization points z_i^n themselves one should investigate the measure

$$\sum_{i=1}^n \frac{1}{n} \delta_{z_i^n}$$

with equal weights $\frac{1}{n}$ instead of P_n . This is elaborated in [48, Theorem 7.5], and it holds that

$$\frac{1}{n} \sum_{i=1}^n \delta_{z_i^n} \rightarrow P_r,$$

where the measure P_r has density

$$g_r(x) = \frac{g(x)^{\frac{m}{m+r}}}{\int_{\mathbb{R}^m} g(x)^{\frac{m}{m+r}} dx},$$

provided that P has density g . This result thus exposes the optimal locations: they should be chosen, asymptotically, with density proportional to $g^{m/m+r}$.

Remark 4.25 (The Special Case $m = 1$). Let the points z_i^n be chosen such that

$$\int_{-\infty}^{z_i^n} g^{\frac{1}{1+r}}(x) dx = \frac{2i-1}{2n} \int_{-\infty}^{\infty} g^{\frac{1}{1+r}}(x) dx \quad i = 1, \dots, n;$$

z_i^n is the $\frac{i-\frac{1}{2}}{n}$ quantile of a distribution with density $\sim g^{\frac{1}{1+r}}$. Then the supporting points z_i^n are asymptotically optimal, and the measures $P_n = \sum_{i=1}^n P_i^n \delta_{z_i^n}$ converge,

$$P_n = \sum_{i=1}^n P_i^n \delta_{z_i^n} \rightarrow P,$$

if P_i^n is the probability of the Voronoi interval associated with z_i^n , that is

$$P_i^n = \int_{\frac{1}{2}(z_{i-1}^n + z_i^n)}^{\frac{1}{2}(z_i^n + z_{i+1}^n)} g(x) dx.$$

It is thus reasonable to sample initial points in Algorithm 4.5 from P_r with density $g^{\frac{m}{m+r}}$, and then improve the locations by continuing the iteration with these points. Rejection sampling is a convenient algorithm to sample from $g^{\frac{m}{m+r}}$.

4.2 Approximations of Multiperiod Distributions

A multiperiod distribution can be dissected into the chain of conditional distributions. Typically, the total distribution is approximated by approximating all conditional distributions.

Conditional Distributions. Suppose that a probability measure P_1 on Ξ_1 is given, as well as conditional probability measures $P_2(\cdot|\xi_1)$ on Ξ_2 , for every ξ_1 . Then the total measure on the product $\Xi_1 \times \Xi_2$ is obtained by

$$P(A_1 \times A_2) = \int_{A_1} \int_{A_2} P_2(d\xi_2|\xi_1) P_1(d\xi_1), \quad (4.24)$$

(denoted also $P = P_1 \circ P_2$).

The disintegration theorem (see, e.g., Durrett [38, Chapter 4, Theorem 1.6]) states that the converse is true as well. That is, a multivariate probability measure P on a product $\Xi_1 \times \Xi_2$ can be dissected according to (4.24), where

- (i) $P_1(A_1) = P(A_1 \times \Xi_2)$ for all $A_1 \in \mathcal{F}_1$,
- (ii) $\xi_1 \mapsto P_2(A|\xi_1)$ is (Borel) measurable for all $A \in \mathcal{F}_2$, and
- (iii) $A \mapsto P_2(A|\xi_1)$ is a probability measure on \mathcal{F}_2 for every $\xi_1 \in \Xi_1$ (it is occasionally said that the transition probability $P_2(\cdot|\xi_1)$ lives on the fiber $\{\xi_1\} \times \Xi_2$).

This process can be repeated in higher dimensions or products as $\Xi = \Xi_1 \times \dots \times \Xi_T$, such that a measure P on Ξ can be disintegrated with

$$P(A_1 \times A_2 \times \dots \times A_T) = \int_{A_1} \int_{A_2} \dots \int_{A_T} P_T(d\xi_T|\xi_{1:T-1}) \dots P_2(d\xi_2|\xi_1) P_1(d\xi_1)$$

(that is, $P = P_1 \circ P_2 \circ \dots \circ P_T$). The total probability measure P is the law of a process, which sequentially reveals the components of ξ . This process is notably not necessarily Markovian, as the transition probability to the next stage, $P_{t+1}(\cdot|\xi_{1:t})$, may depend on the full history $\xi_{1:t} = (\xi_1, \dots, \xi_t)$ and not just on the last stage ξ_t (cf. Mirkov and Pfugl [78]).

In this setting it is possible to derive an upper bound for the nested distance of two filtered spaces, just by controlling the transition probabilities. For this we equip the space $\Xi = \tilde{\Xi} = \Xi_0 \times \dots \times \Xi_T$ again with a distance \mathbf{d} according to (2.32) or (2.33), that is

$$\mathbf{d}(\xi, \tilde{\xi}) := \left(\sum_{t=0}^T w_t \cdot \mathbf{d}_t(\xi_t, \tilde{\xi}_t)^p \right)^{1/p} \quad (\xi, \tilde{\xi} \in \Xi),$$

where $w_t > 0$ are appropriate weights and $p \geq 1$. We moreover make the distance available for fractional paths (i.e., paths, which are not continued up to the final

stage T) by setting

$$\mathbf{d}(\xi_{0:t}, \tilde{\xi}_{0:t}) := \left(\sum_{s=0}^t w_s \cdot \mathbf{d}_s(\xi_s, \tilde{\xi}_s)^p \right)^{1/p},$$

where we employ the same symbol \mathbf{d} for the distance of paths starting at the origin. (For a relation of these distances for different exponents than p we refer to Lemma C.1 in Appendix C.)

Proposition 4.26. *Let \mathbb{P} and $\tilde{\mathbb{P}}$ be filtered probability spaces and consider the distance $\mathbf{d}(\xi, \tilde{\xi}) = \left(\sum_{s=0}^t w_s \cdot \mathbf{d}_s(\xi_s, \tilde{\xi}_s)^r \right)^{1/r}$ for paths ξ and $\tilde{\xi}$. Moreover assume that the transition probabilities satisfy*

$$\mathbf{d}_r(P_{t+1}(\cdot | \xi_{0:t}), \tilde{P}_{t+1}(\cdot | \tilde{\xi}_{0:t})) \leq \epsilon_{t+1} + \kappa_{t+1} \cdot \mathbf{d}(\xi_{0:t}, \tilde{\xi}_{0:t}), \quad (4.25)$$

where $\xi_{0:t} \in \Xi_{0:t}$, $\tilde{\xi}_{0:t} \in \tilde{\Xi}_{0:t}$ and $\epsilon_0 := \mathbf{d}(\xi_0, \tilde{\xi}_0)$.³ Then the nested distance is bounded by

$$\begin{aligned} \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r &\leq \epsilon_T^r w'_T + \epsilon_{T-1}^r w'_{T-1} (1 + w'_T \kappa_T^r) + \dots \\ &= \sum_{t=0}^T \epsilon_t^r w'_t \prod_{s=t+1}^T (1 + w'_s \kappa_s^r), \end{aligned} \quad (4.26)$$

where the modified weights are $w'_t = 2^{r-1} w_t$.

Proof. Let π be a transport plan, which is feasible for the nested distance. The last stage $\Xi_T \times \tilde{\Xi}_T$ can be dissected from the transport plan. It holds then that

$$\begin{aligned} \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r &\leq \iint \mathbf{d}(\xi, \tilde{\xi})^r \pi(d\xi, d\tilde{\xi}) \\ &= \iint (\mathbf{d}(\xi_{0:T-1}, \tilde{\xi}_{0:T-1})^r + w_T \mathbf{d}_T(\xi_T, \tilde{\xi}_T)^r) \\ &\quad \pi_T(d\xi_T, d\tilde{\xi}_T | \xi_{0:T-1}, \tilde{\xi}_{0:T-1}) \pi(d\xi_{0:T-1}, d\tilde{\xi}_{0:T-1}) \\ &= \int (\mathbf{d}(\xi_{0:T-1}, \tilde{\xi}_{0:T-1})^r + w_T \int \mathbf{d}_T(\xi_T, \tilde{\xi}_T)^r \\ &\quad \pi_T(d\xi_T, d\tilde{\xi}_T | \xi_{0:T-1}, \tilde{\xi}_{0:T-1})) \pi(d\xi_{0:T-1}, d\tilde{\xi}_{0:T-1}) \end{aligned}$$

³Usually both trees start at the same state $\xi_0 = \tilde{\xi}_0$, such that $\epsilon_0 = 0$ in this case.

by conditioning on stage $T - 1$, where π_T is the conditional transport plan. The inner integral itself exactly represents the Wasserstein distance given the histories up to $T - 1$. Due to the assumption (4.25) this is bounded by

$$\begin{aligned} \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r &\leq \iint \left(\mathbf{d}\left(\xi_{0:T-1}, \tilde{\xi}_{0:T-1}\right)^r + w_T \mathbf{d}_r\left(P_T(\cdot | \xi_{0:T-1}), \tilde{P}_T(\cdot | \tilde{\xi}_{0:T-1})\right)^r \right) \\ &\quad \pi(d\xi_{0:T-1}, d\tilde{\xi}_{0:T-1}) \\ &\leq \int \mathbf{d}\left(\xi_{0:T-1}, \tilde{\xi}_{0:T-1}\right)^r \\ &\quad + w_T \left(\epsilon_T + \kappa_T \cdot \mathbf{d}\left(\xi_{0:T-1}, \tilde{\xi}_{0:T-1}\right) \right)^r \pi(d\xi_{0:T-1}, d\tilde{\xi}_{0:T-1}) \\ &\leq \iint \mathbf{d}\left(\xi_{0:T-1}, \tilde{\xi}_{0:T-1}\right)^r \\ &\quad + w_T 2^{r-1} \left(\epsilon_T^r + \kappa_T^r \cdot \mathbf{d}\left(\xi_{0:T-1}, \tilde{\xi}_{0:T-1}\right)^r \right) \pi(d\xi_{0:T-1}, d\tilde{\xi}_{0:T-1}) \\ &= w_T 2^{r-1} \epsilon_T^r \\ &\quad + (1 + w_T 2^{r-1} \kappa_T^r) \cdot \iint \mathbf{d}\left(\xi_{0:T-1}, \tilde{\xi}_{0:T}\right)^r \pi(d\xi_{0:T-1}, d\tilde{\xi}_{0:T-1}). \end{aligned}$$

It is evident that the measure π is feasible for the problem truncated at $T - 1$, such that the computation can be repeated for the next stage $T - 2$. It follows hence by induction that

$$\begin{aligned} \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r &\leq \epsilon_T^r w'_T + \epsilon_{T-1}^r w'_{T-1} (1 + w'_T \kappa_T^r) + \dots \\ &= \sum_{t=0}^T \epsilon_t^r w'_t \prod_{s=t+1}^T (1 + w'_s \kappa_s^r). \end{aligned}$$

which is the assertion. \square

Lemma 4.27. *The assertion of Proposition 4.26 is satisfied in particular, if one process satisfies a Lipschitz condition*

$$\mathbf{d}_r\left(P_{t+1}(\cdot | \xi_{0:t}), P_{t+1}\left(\cdot | \tilde{\xi}_{0:t}\right)\right) \leq \kappa_{t+1} \cdot \mathbf{d}\left(\xi_{0:t}, \tilde{\xi}_{0:t}\right) \quad (4.27)$$

for all $\xi, \tilde{\xi}$,⁴ and if the conditional distributions conditioned on the same $\xi_{0:t}$ are close

$$\mathbf{d}_r(P_{t+1}(\cdot | \xi_{0:t}), \tilde{P}_{t+1}(\cdot | \xi_{0:t})) \leq \epsilon_{t+1}$$

for all ξ .

Proof. It is immediate from the triangle inequality that

$$\begin{aligned} & \mathbf{d}_r(P_{t+1}(\cdot | \xi_{0:t}), \tilde{P}_{t+1}(\cdot | \tilde{\xi}_{0:t})) \\ & \leq \mathbf{d}_r(P_{t+1}(\cdot | \xi_{0:t}), P_{t+1}(\cdot | \tilde{\xi}_{0:t})) + \mathbf{d}_r(P_{t+1}(\cdot | \tilde{\xi}_{0:t}), \tilde{P}_{t+1}(\cdot | \tilde{\xi}_{0:t})) \\ & \leq \kappa_{t+1} \cdot \mathbf{d}(\xi_{0:t}, \tilde{\xi}_{0:t}) + \epsilon_{t+1}, \end{aligned}$$

what is the condition for Proposition 4.26. \square

Example (Gaussian Processes Have the Lipschitz Property (4.27), cf. Mirkov and Pflug [78]). Consider a normal distribution P on $\mathbb{R}^{m_1+m_2}$, i.e.,

$$P = N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

The conditional distribution, given $u \in \mathbb{R}^{m_1}$, is

$$P_2(\cdot | u) = N\left(\mu_1 - \Sigma_{12}\Sigma_{11}^{-1}(u - \mu_1), \Sigma_{22} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{21}\right)$$

and it holds that

$$\mathbf{d}_r(P_2(\cdot | u), P_2(\cdot | v)) \leq K \|u - v\|.$$

with $K = \|\Sigma_{12}\Sigma_{11}^{-1}\|$.

Indeed, let $Y \sim \mathcal{N}(0, \Sigma)$ and let $Y_1 = a_1 + Y$ and $Y_2 = a_2 + Y$. Then $Y_1 \sim \mathcal{N}(a_1, \Sigma)$, $Y_2 \sim \mathcal{N}(a_2, \Sigma)$ and a transportation plan is to transport u to $u - a_1 + a_2$. Therefore

$$\mathbf{d}_r(N(a_1, \Sigma), N(a_2, \Sigma))^r \leq \|a_1 - a_2\|^r.$$

⁴Notice that if $r = 1$, the factor 2^{r-1} disappears. The condition (4.27) is related to the *ergodic coefficient* introduced in by Dobrushin [28] and studied in Pflug [88]; the ergodic coefficient of a Markov chain is

$$\kappa = \sup_{u \neq v} \frac{\mathbf{d}(P(\cdot | u), P(\cdot | v))}{\mathbf{d}(u, v)}.$$

Setting $a_1 := \mu_1 - \Sigma_{12}\Sigma_{11}^{-1}(u - \mu_1)$, $a_2 := \mu_1 - \Sigma_{12}\Sigma_{11}^{-1}(v - \mu_1)$ and $\Sigma := \Sigma_{22} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{12}$ one gets

$$\mathbf{d}_r(P_2(\cdot|u), P_2(\cdot|v))^r \leq \|\Sigma_{12}\Sigma_{11}^{-1}(u - v)\|^r \leq \|\Sigma_{12}\Sigma_{11}^{-1}\|^r \|u - v\|^r$$

and the assertion is verified.

The following example outlines that the regularity condition (4.25) in Proposition 4.26 is indeed necessary and essential:

Example 4.28. Let $P = P_1 \circ P_2$ be the measure on \mathbb{R}^2 , such that

$$P_1 = \mathcal{U}[0, 1] \text{ and}$$

$$P_2(\cdot|u) = \begin{cases} \mathcal{U}[0, 1] & \text{if } u \in \mathbb{Q}, \\ \mathcal{U}[1, 2] & \text{if } u \in \mathbb{R} \setminus \mathbb{Q}, \end{cases}$$

where $\mathcal{U}[a, b]$ denotes the uniform distribution on $[a, b]$ and \mathbb{Q} is the set of rational numbers. Obviously, P is the uniform distribution on $[0, 1] \times [1, 2]$, since the rational numbers have probability zero. P_2 does not have the Lipschitz property (4.25). Now, let $\tilde{P}^{(n)} = \tilde{P}_1^{(n)} \circ \tilde{P}_2^{(n)}$, where

$$\begin{aligned} \tilde{P}_1^{(n)} &= \sum_{k=1}^n \frac{1}{n} \delta_{k/n} \\ \tilde{P}_2^{(n)}(\cdot|u) &= \begin{cases} \sum_{k=1}^n \frac{1}{n} \delta_{k/n} & \text{if } u \in \mathbb{Q}, \\ \sum_{k=1}^n \frac{1}{n} \delta_{1+k/n} & \text{if } u \in \mathbb{R} \setminus \mathbb{Q}. \end{cases} \end{aligned}$$

Notice that $\tilde{P}^{(n)} = \tilde{P}_1^{(n)} \circ \tilde{P}_2^{(n)}$ converges weakly to the uniform distribution on $[0, 1] \times [0, 1]$ because \tilde{P}_1 takes only rational values. In particular, $\tilde{P}^{(n)}$ does not converge to P and the nested distance is $\mathbf{d}_1(\mathbb{P}, \tilde{\mathbb{P}}^{(n)}) = 1$. However, for $n \rightarrow \infty$,

$$\mathbf{d}_1(P_1, \tilde{P}_1^{(n)}) = \frac{1}{n} \rightarrow 0 \quad \text{and} \quad \mathbf{d}_1(P_2(\cdot| \cdot), \tilde{P}_2^{(n)}(\cdot| \cdot)) = \frac{1}{n} \rightarrow 0.$$

The Branching Structure of Practically Useful Trees. The results of the previous lemma and Proposition 4.26 can be used in the following way to develop a branching structure for multistage optimization trees.

Suppose the Lipschitz constants κ_t satisfying (4.27) are known. Consider the natural numbers

$$N_t \sim (w_t (1 + \kappa_{t+1} w_{t+1}) \cdot \dots \cdot (1 + \kappa_T w_T) \cdot N)^{m_t}, \quad (4.28)$$

where T is the height of the desired tree and m_t the dimension at stage t . According to Corollary 4.21 there are discrete approximations with N_t supporting points such that

$$\begin{aligned} \mathbf{d}_r(P_{t+1}(\cdot|\cdot), \tilde{P}_{t+1}(\cdot|\cdot)) \\ \sim N_{t+1}^{-1/m_t} \sim \frac{1}{w_{t+1} (1 + \kappa_{t+2} w_{t+2}) \dots (1 + \kappa_T w_T)} \cdot \frac{1}{N} =: \epsilon_{t+1}^r. \end{aligned}$$

Then, according to (4.26), the approximation of the nested distance is bounded by

$$\mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^r \leq \sum_{t=1}^T \frac{1}{N} = \frac{T}{N}.$$

The number of leaves of this tree, however, is

$$\prod_{t=1}^T N_t \sim N^{\sum_{t=1}^T m_t},$$

which is a huge number to obtain an approximation quality of order $\frac{1}{N}$, often too big for numerical computations.

Remark 4.29. The identity (4.28) reveals that if w_t decrease fast enough, then worse approximations are acceptable towards the end of the tree. This corresponds to the intuition that bad approximations towards the end have less impact onto the overall approximation, as they are of less weight for the entire problem. Note, however, that the constants in (4.28) are of minor importance compared to the dimensions m_t , which dominate the size of the tree necessary to achieve a small nested distance.

4.3 Construction of Scenario Trees

Let a stochastic scenario process (ξ_t) , $t = 1, \dots, T$ be given and let \mathfrak{F} be the filtration generated by (ξ_t) . If it is necessary to model a larger filtration (if additional information is available for the decisions), one may always assume that the filtration is generated by an extended process $((\xi_t, \eta_t))$, where only the values (ξ_t) enter the objective and the constraints, while the additional process (η_t) describes the additional available information, which is not contained in (ξ_t) (cf. Remark 1.3).

We denote the nested distribution of the original or the extended process by \mathbb{P} . On the other hand, consider a finite tree process (v_t) , $t = 1, \dots, T$ which generates the filtration $\tilde{\mathfrak{F}}$ on which a finite scenario process $(\tilde{\xi}_t)$, $t = 1, \dots, T$ is sitting, such that the pertaining nested distribution is $\tilde{\mathbb{P}}$. With a slight abuse of language, we identify the tree with its nested distribution and call $\tilde{\mathbb{P}}$ a (*valuated*) tree. The main problems in scenario generation are

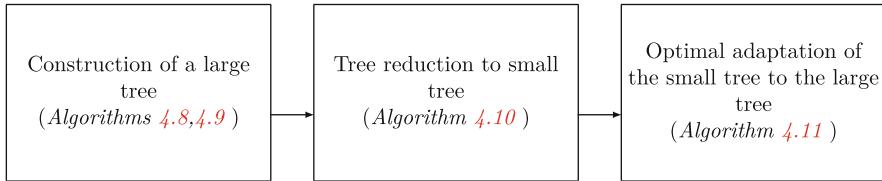


Fig. 4.4 The general structure of scenario tree generation using Algorithms 4.8, 4.10, and 4.11

- **Distance calculation.** Given a finite scenario tree $\tilde{\mathbb{P}}$, calculate the nested distance $d(\mathbb{P}, \tilde{\mathbb{P}})$.

Ideally, one would want to minimize $d(\mathbb{P}, \tilde{\mathbb{P}})$, among all trees $\tilde{\mathbb{P}}$ with a given size, say number of nodes or number of leaves, but there are no feasible algorithms, which can do this hypercomplex optimization. For this reason, the task is practically done in three steps: (i) a large tree model is estimated, (ii) then reduced to smaller size (tree reduction), and (iii) then adapted for better fitting to the original process (tree adaptation).

- **Large tree construction.** Find a large tree $\bar{\mathbb{P}}$, such that $\bar{\mathbb{P}}$ is close to \mathbb{P} .
- **Tree reduction.** Given a larger finite tree model $\bar{\mathbb{P}}$, find a finite tree model $\tilde{\mathbb{P}}$, which is much smaller than $\bar{\mathbb{P}}$, but represents $\bar{\mathbb{P}}$ well.
- **Tree adaptation.** Suppose that the smaller tree $\tilde{\mathbb{P}}$ found in the previous step has tree structure $\bar{\mathbb{T}}$. Find a tree $\tilde{\mathbb{P}}$ which minimizes the distance

$$d\left(\tilde{\mathbb{P}}, \tilde{\mathbb{P}}\right) \rightarrow \min$$

among all trees with the same tree structure $\bar{\mathbb{T}}$.

The typical sequence of algorithms for finding scenario trees based on the nested distance is illustrated in Fig. 4.4. Starting with a large tree $\bar{\mathbb{P}}$, which is constructed on the basis of a simulation program for the scenario process (ξ_t) , the tree is reduced to a smaller tree $\tilde{\mathbb{P}}$ and finally adapted to the larger tree $\bar{\mathbb{P}}$ resulting in the final tree $\tilde{\mathbb{P}}$.

Different solution techniques for continuous, as well as for discrete approximation are outlined in Pflug [90] and Hochreiter and Pflug [59].

4.3.1 Distance Calculation

The nested distance between two processes, finite or not, is always bounded by the transportation effort of some admissible transportation plan. Even if this plan is not the optimal one, typically the bound is very good, if it is chosen in a clever way.

Let a stochastic process $(\xi_t), t = 1, \dots, T$ with values in \mathbb{R}^m be given as well as a discrete tree with node sets $\tilde{\mathcal{N}} = \{1\} \cup \mathcal{N}_1 \cup \dots \cup \mathcal{N}_T$ and values $\tilde{\xi}(n)$ sitting on its nodes. Since we are interested in trees which are close to the process (ξ_t) in nested distance, we assume that the probabilities on the tree are left open and determined such that the distance to the original process is minimal. Assume that $\xi_0 = \tilde{\xi}(1)$ (the root node). In order to construct a transportation plan between the two processes, define a Voronoi tessellation $\{V_i : i \in \tilde{\mathcal{N}}_T\}$ on \mathbb{R}^{mT} by

$$\begin{aligned} V_i &= \left\{ (z_1, \dots, z_T) \in \mathbb{R}^{mT} : d(z_1, \dots, z_T, \tilde{\xi}(i_1), \dots, \tilde{\xi}(i)) \right. \\ &\quad \left. = \min_j d(z_1, \dots, z_T, \tilde{\xi}(j_1), \dots, \tilde{\xi}(j)) \right\} \end{aligned}$$

where $i_t = \text{pred}_t(i)$. If there are several indices j , to which $z = (z_1, \dots, z_T)$ has the minimal distance, choose the smallest of them to guarantee uniqueness. The distance d is typically defined as

$$d(z, z') = \sum_{t=1}^T w_t \|z_t - z'_t\|.$$

A natural transportation mapping between $z \in \mathbb{R}^{mT}$ and $i \in \tilde{\mathcal{N}}_T$ is given by

$$\mathbf{T}(z) = i \quad \text{if } z \in V_i.$$

and for this choice the optimal leaf probabilities are

$$\tilde{P}(i) = P\{\mathbf{T}(\xi) = i\} = P\{(\xi_1, \dots, \xi_T) \in V_i\}.$$

The upper estimate of the nested distance is then given by

$$d\mathbb{I}_r^r(\mathbb{P}, \tilde{\mathbb{P}}) \leq \sum_{i \in \tilde{\mathcal{N}}_T} \mathbb{E}[d^r((\xi_1, \dots, \xi_T), (\tilde{\xi}(i_1), \dots, \tilde{\xi}(i))) \cdot \mathbb{1}_{\xi \in V_i}] \cdot \tilde{P}(i). \quad (4.29)$$

The integrals in (4.29) are very difficult to carry out numerically. The geometry of Voronoi sets may be very complicated (cf. Fig. 4.2). For this reason, the best method is to use Monte Carlo integration, where a large number of samples $\xi^{(k)} = (\xi_1^{(k)}, \dots, \xi_T^{(k)})$, $k = 1, \dots, K$ are drawn, the corresponding $\mathbf{T}(\xi)$ is found and the distance $d\mathbb{I}_r^r(\mathbb{P}, \tilde{\mathbb{P}})$ is estimated by

$$\frac{1}{K} \sum_{k=1}^K d^r((\xi_1^{(k)}, \dots, \xi_T^{(k)}), (\tilde{\xi}(i_1), \dots, \tilde{\xi}(i))) \cdot \mathbb{1}_{\mathbf{T}(\xi)=i}$$

Only in the univariate case ($m = 1$) and if $r = 1$ there is some hope to avoid Monte Carlo integration and to use analytics. To this end, the following functions are needed

$$\begin{aligned} E_t(z_t; u_1, v_1, \dots, u_{t-1}, v_{t-1}, u_t, v_t) \\ := \mathbb{E} \left(|\xi_t - z_t| \mathbb{1}_{u_t \leq \xi_t \leq v_t} \middle| \begin{array}{l} u_1 < \xi_1 \leq v_1, u_2 < \xi_2 \leq v_2, \dots \\ u_{t-1} < \xi_{t-1} \leq v_{t-1} \end{array} \right). \end{aligned} \quad (4.30)$$

It is convenient to use different numbering for the nodes by assigning to each node n the pair $(t(n), j(n))$ where $t(n)$ is the stage of node n and $j(n)$ is its index within all nodes in stage $t(n)$. Without loss of generality one may assume that the values $\tilde{\xi}(n) = \tilde{\xi}(t(n), j(n))$ are ordered with respect to i , i.e., $\tilde{\xi}(t, j) < \tilde{\xi}(t, j')$ for $j < j'$. Suppose that $s = s(t)$ is the number of nodes at stage t and let $b_{t,0} = -\infty$, $b_{t,s} = \infty$ and $b_{t,j} = (\tilde{\xi}_{t,j} + \tilde{\xi}_{t,j+1})/2$ for $j = 1, \dots, s-1$. For a node $i \in \tilde{\mathcal{N}}_T$, let $j(i_t)$ be the index of the predecessor of i at stage t . An upper estimate of the nested distance is then given by

$$\mathbf{d}_1(\mathbb{P}, \tilde{\mathbb{P}}) \leq \sum_{i \in \tilde{\mathcal{N}}_T} \sum_{t=1}^T w_t E_t(\tilde{\xi}(i_t); b_{1,j(i_1)-1}, b_{1,j(i_1)}, \dots, b_{t,j(i_t)-1}, b_{t,j(i_t)})$$

when the basic distance is $\mathbf{d}(z, z') = \sum_{t=1}^T w_t |z_t - z'_t|$.

4.3.2 The Construction of Large Trees

Suppose that a scenario process (ξ_t) has been estimated, i.e., its (nested) distribution \mathbb{P} has been identified. The goal is now to construct a valued probability tree $\tilde{\mathbb{P}}$ (the scenario tree), which represents ξ_t in the best possible way. Recall from Sect. 1.4 that scenario trees are represented by a triple consisting of the tree structure (i.e., the predecessor relations), the values of the process sitting on the nodes, and the probabilities sitting on the arcs of the tree. To be more precise, a scenario tree $\tilde{\mathbb{P}} = \tilde{\mathbb{P}}(n, \text{pred}, \tilde{\xi}, Q)$ is characterized by

- the number n of nodes,
- a function pred mapping $\{1, 2, \dots, n\}$ to $\{0, 1, \dots, n\}$; $\text{pred}(k) = \ell$ (also written as $k- = \ell$) means that node ℓ is a direct predecessor of node k ; the root is node 1 and its predecessor is encoded as 0,
- a valuation of each node $\tilde{\xi}(i)$, $i = 1, \dots, n$ lying in \mathbb{R}^m ,
- the conditional probability $\tilde{Q}(i)$ of reaching node i from its predecessor; for the root, we have $\tilde{Q}(1) = 1$.

It is always assumed that these model structures are consistent, i.e., that they form a tree of height T meaning that all leaves of the tree are at the same level T . The distance of each node to the root is the *stage* of the node. The root is at stage 0 and the leaves of the tree are at stage T . $\tilde{\mathcal{N}}_t$ is the set of all nodes at stage t . Let $\tilde{\Omega} = \tilde{\mathcal{N}}_T$ be the set of all leaf nodes, which can be seen as a probability space carrying the unconditional probabilities $P(n)$ to reach the leaf node n from the root. If $\text{pred}_t(n)$ denotes the predecessor of node n at stage t , then these mappings induce a filtration $\tilde{\mathfrak{F}} = (\tilde{\mathcal{F}}_0, \dots, \tilde{\mathcal{F}}_T)$, with $\tilde{\mathcal{F}}_0$ being the trivial sigma-algebra and $\tilde{\mathcal{F}}_T$ the power set of $\tilde{\Omega}$. The process (ξ_t) takes the values $\tilde{\xi}(i)$ for all nodes i with probability $\tilde{P}(i) = \tilde{Q}(i) \cdot \prod_{j \prec i} \tilde{Q}(j)$, where \prec denotes the predecessor relation.

On the other hand, the basic stochastic process (ξ_t) also induces a nested distribution on the filtered probability space $\underbrace{\mathbb{R}^m \times \dots \times \mathbb{R}^m}_{T+1}$ with filtration $\mathfrak{F} = (\mathcal{F}_0, \dots, \mathcal{F}_T)$. To measure the quality of approximation of the nested distribution \mathbb{P} by the finite tree $\bar{\mathbb{P}}$, we use the nested distance. However, in view of Proposition 4.26 and Lemma 4.27 it is often easier to minimize stagewise conditional distances than the full nested distance, as it is, e.g., the argument in the next theorem.

It is assumed that the process (ξ_t) satisfies condition (4.25).

Theorem 4.30. *Suppose that the process (ξ_t) satisfies condition (4.25) and that the conditional distributions of ξ_t given the past ξ_1, \dots, ξ_{t-1} exhibit the following property: the absolute continuous part of their densities with respect to the Lebesgue measure have uniformly bounded $^{m/m+r}$ moments. Then there is a sequence of finite trees $\bar{\mathbb{P}}_n$, such that*

$$\mathbf{d}_r(\mathbb{P}, \bar{\mathbb{P}}_n) \rightarrow 0 \quad \text{as } n \text{ tends to } \infty. \quad (4.31)$$

Proof. By Proposition 4.26, the conditional distances $\mathbf{d}_r(P_{t+1}(\cdot|u^t), \tilde{P}_{t+1}(\cdot|u^t))$ have to be made uniformly small in order to guarantee (4.31). By assumption and Theorem 4.19 this can be achieved by a finite tree. \square

Of course, in order to get the nested distance small, the approximating tree must become bushier. The bushiness of the tree, i.e., the number of successors of each internal node is crucial for a good approximation. This obvious fact is illustrated by the following example.

Example 4.31. Consider a lognormal stationary process of the following form: let

$$\eta_0 \sim N(\mu_0, \sigma_0^2)$$

$$\eta_t = b \eta_{t-1} + \epsilon_t$$

with $\epsilon_t \sim N(\mu, \sigma^2)$ (independent), with $\mu = \mu_0(1-b)$ and $\sigma^2 = \sigma_0^2(1-b^2)$. Then (η_t) is a stationary Gaussian process. Let $\xi_t = \exp(\eta_t)$ be the lognormal process. The parameters were chosen as $\mu_0 = 4$, $\sigma_0^2 = 0.36$, $b = 0.8$. The process ξ was approximated by optimal discretization (see Algorithm 4.7 below) and finite trees with different heights T and different bushiness (i.e., the number of descendants in

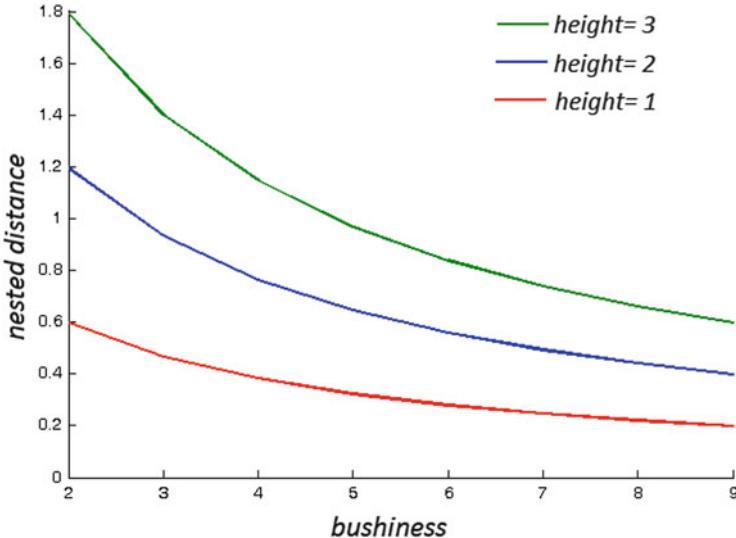


Fig. 4.5 The optimal nested distance decreases with increasing bushiness, but increases with the height T of the tree

every node) were constructed. Bushiness 2 corresponds to a binary tree, bushiness 3 to a ternary tree and so on. Figure 4.5 demonstrates the nested distance between the original process and the optimal finite tree. As expected, the distance decreases with increasing bushiness, but the marginal distance decrease gets smaller for larger bushiness. For processes with T periods, the approximation gets more difficult if T is increased.

Algorithms for Scenario Tree Construction. Let $(\xi_t), t = 0, \dots, T$ be a stochastic process with discrete time and with (typically continuous) values in \mathbb{R}^m . Assume that the distribution of this process is known. To be more precise, let ξ_0 be deterministic and let

$$G_t(z_t | z_{t-1}, \dots, z_0) \quad (4.32)$$

be the conditional distribution functions of ξ_t given the past $\xi_0 = z_0, \dots, \xi_{t-1} = z_{t-1}$.

Example 4.32. A typical example is given by a k th-order Markovian process, where the conditional distribution depends on the past k observations: $G_t(z_t | z_{t-1}, \dots, z_{t-k})$. It is clear that in this case one has to know also some past values z_{-1}, \dots, z_{-k+1} in order to describe the conditional distributions at the first k stages and one may replace (4.32) by $G_t(z_t | z_{t-1}, \dots, z_{t-k})$ also for $t < k$.

We construct trees by approximating the conditional distributions and evoking Proposition 4.26 and Lemma 4.27. These results tell that we can achieve a small nested distance (but typically not the best one) also by minimizing the

Algorithm 4.7

General structure of the tree construction by discretization of the conditional probabilities based on Proposition 4.26

- **INITIAL STEP.** Set $\tilde{\xi}(1) = \xi_0$ and $\tilde{Q}(1) = 1$. Find the optimal discretization of the distribution G_1 . Suppose that the optimal points are $\tilde{\xi}(2), \dots, \tilde{\xi}_1(k)$ with probabilities $\tilde{Q}(2), \dots, \tilde{Q}(k)$.
- **RECURSIVE STEP.** Given that the values are determined up to stage t . For the given past values $\tilde{\xi}_1(i_1), \tilde{\xi}_2(i_2), \dots, \tilde{\xi}_t(i_t)$, find the optimal discretizations of the conditional distribution $G_{t+1}(\cdot | \tilde{\xi}_1(i_1), \tilde{\xi}_2(i_2), \dots, \tilde{\xi}_t(i_t))$ as the successors of the path $\tilde{\xi}(1), \tilde{\xi}_1(i_1), \tilde{\xi}_2(i_2), \dots, \tilde{\xi}_t(i_t)$.

Wasserstein distances of the conditional distribution.⁵ The basic algorithm for large tree construction is done stagewise in ascending order using optimal discretizations, see Algorithm 4.7.

The main part of the algorithm is to find the optimal discretization of a probability distribution. If the distribution is one-dimensional and the distribution function is numerically available, one may use Algorithms 4.3 or 4.4. Otherwise, a specialized multidimensional deterministic discretization algorithm may be used or (which is advised here) the stochastic approximation algorithm (Algorithm 4.5). If a predetermined error d_t must not be exceeded, the needed bushiness at each node has to be determined as it is described in Algorithm 4.9. Otherwise the bushiness can be determined beforehand (Algorithm 4.8), but in this case no guarantee for a small distance can be given.

Example 4.33 (Continuation of Example 2.44). The best approximation to the process (ξ_1, ξ_2) with $\xi_1 \sim N(0, 1)$ and $\xi_2 \sim N(\xi_1, 1)$ was shown to be

$$\tilde{\mathbb{P}}^* = \left[\begin{array}{ccc} 0.3035 & 0.3930 & 0.3035 \\ -1.029 & 0.0 & 1.029 \\ \left[\begin{array}{ccc} 0.3035 & 0.3930 & 0.3035 \\ -2.058 & -1.029 & 0.0 \end{array} \right] & \left[\begin{array}{ccc} 0.3035 & 0.3930 & 0.3035 \\ -1.029 & 0.0 & 1.029 \end{array} \right] & \left[\begin{array}{ccc} 0.3035 & 0.3930 & 0.3035 \\ 0.0 & 1.029 & 2.058 \end{array} \right] \end{array} \right]$$

We have used the Algorithm 4.9 to generate a discrete approximation with bushiness $(3, 3)$. After 500 steps, the algorithm produced the following nested distribution

$$\tilde{\mathbb{P}} = \left[\begin{array}{ccc} 0.275 & 0.4 & 0.325 \\ -1.039 & -0.048 & 0.938 \\ \left[\begin{array}{ccc} 0.305 & 0.360 & 0.324 \\ -2.062 & -0.99 & 0.085 \end{array} \right] & \left[\begin{array}{ccc} 0.335 & 0.345 & 0.325 \\ -1.124 & -0.061 & 0.955 \end{array} \right] & \left[\begin{array}{ccc} 0.325 & 0.410 & 0.265 \\ -0.031 & 1.083 & 2.094 \end{array} \right] \end{array} \right],$$

which is sufficiently close to the optimal \mathbb{P}^* .

⁵Anna Timonina introduces in her dissertation at the University of Vienna an algorithm, which improves the nested distance over the one obtained by stagewise Wasserstein minimization.

Algorithm 4.8

Tree generation with fixed bushiness

PARAMETERS. Let T be the desired height of the tree and let (b_1, \dots, b_T) be the given bushiness parameters per stage.

- **DETERMINING THE ROOT.** The value at the root is $\tilde{\xi}(1) = \xi_0$. Its stage is 0. Set the root as the current open node.
 - **SUCCESSOR GENERATION.** Enumerate the tree stagewise from the root to the leaves.
 1. Let k be the node to be considered next and let $t < T$ be its stage. Let $\tilde{\xi}(k_{t-1}), \tilde{\xi}(k_{t-2}), \dots, \tilde{\xi}(1)$ be the already fixed values at node k and all its predecessors. Call the stochastic approximation algorithm (Algorithm 4.5) to generate b_t points $z^{(1)}, \dots, z^{(b_t)}$ out of the probability distribution $G_{t+1}(\cdot | \tilde{\xi}(k_{t-1}), \tilde{\xi}(k_{t-2}), \dots, \tilde{\xi}(1))$ and find the corresponding conditional probabilities $p(z^{(i)})$.
 2. Store the b_t successor nodes of node k with node numbers (n_1, \dots, n_{b_t}) and assign to them the values $\tilde{\xi}(n_1) = z^{(1)}, \dots, \tilde{\xi}(n_{b_t}) = z^{(b_t)}$, as well as their conditional probabilities $\tilde{Q}(n_i) = p(z^{(i)})$.
 - **STOPPING CRITERION.** If all nodes at stage $T - 1$ have been considered as parent nodes, the generation of the tree is finished. One may then calculate the unconditional probabilities out of the conditional probabilities.
-

Algorithm 4.9

Dynamic tree generation with flexible bushiness

- **PARAMETERS.** Let T be the desired height of the tree and let (b_1, \dots, b_T) be the minimal bushiness values and (d_1, \dots, d_T) the maximal transportation distances. These two vectors have to be fixed in advance.
- **DETERMINING THE ROOT.** The value at the root is $\tilde{\xi}(1) = \xi_0$. Its stage is 0. Set the root as the current open node.
- **WHILE** there are open nodes **DO**
 - (i) Let k be the next open node and let $t < T$ be its stage. Let $\tilde{\xi}(k_{t-1}), \tilde{\xi}(k_{t-2}), \dots, \tilde{\xi}(1)$ be the already fixed values at node k and at its predecessors. Set the initial number of successors of k to $s = b_{t+1}$.
 - (ii) Call the stochastic approximation algorithm (Algorithm 4.5) to generate s points $z = (z^{(1)}, \dots, z^{(s)})$ out of the distribution $G_{t+1}(\cdot | \tilde{\xi}(k_{t-1}), \tilde{\xi}(k_{t-2}), \dots, \tilde{\xi}(1))$ and calculate the distance $d = D(z | G_{t+1}(\cdot | \tilde{\xi}(k_{t-1}), \tilde{\xi}(k_{t-2}), \dots, \tilde{\xi}(1)))$ (cf. (4.14)).
 - (iii) If the distance d is larger than d_{t+1} , then increase b by one and go back to (ii). Otherwise goto (iv).
 - (iv) Store the b successor nodes of node k using the optimal values $z^{(1)}, \dots, z^{(b)}$ as well as their optimal conditional probabilities $\tilde{Q}(z^{(i)})$ and mark the new successors as open.
 - **STOPPING CRITERION.** If all nodes at stage $T - 1$ have been considered as parent nodes, the generation of the tree is finished.

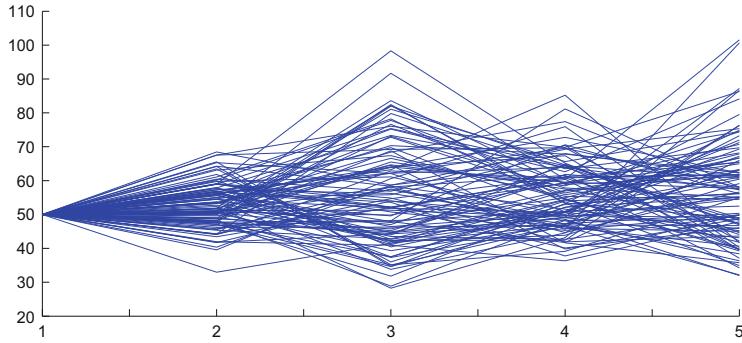


Fig. 4.6 100 trajectories of the process (4.33)

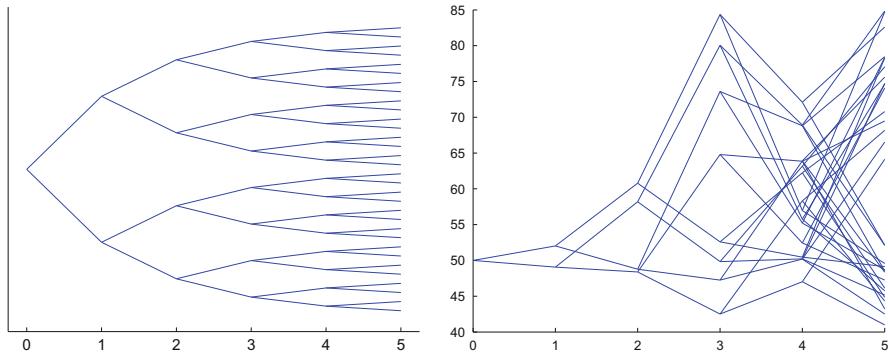


Fig. 4.7 A binary tree generated on the basis of the process (4.33) using Algorithm 4.8. The tree has 63 nodes and 32 leaves. *Left:* the tree structure; *right:* the values

A larger tree is generated in the following example.

Example 4.34. Consider the Markovian process

$$\xi_{t+1} = \xi_t^{0.8/(1+\sqrt{t})} \cdot \eta_t^{1-0.8/(1+\sqrt{t})}, \quad t = 0, \dots, 4 \quad (4.33)$$

with starting value $\xi_0 = 50$. Here, η_t is an i.i.d. sequence with distribution

$$\log \eta(t) \sim \begin{cases} N(4, 0.4) & \text{if } t \text{ is even,} \\ N(4, 0.2) & \text{if } t \text{ is odd.} \end{cases}$$

Notice that the conditional variance heavily depends on t . Figure 4.6 shows 100 sampled trajectories of this process. Algorithm 4.8 is used to generate a binary tree approximating the process $\xi(\cdot)$, this tree is shown in Fig. 4.7. A larger tree was then generated employing Algorithm 4.9 with a predetermined accuracy for every stage

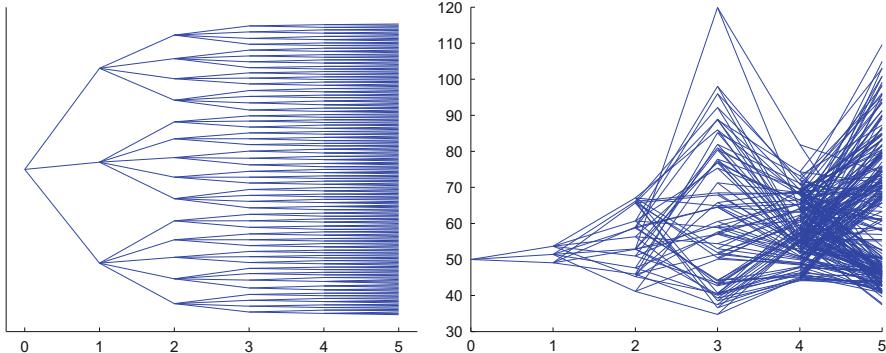


Fig. 4.8 A dynamically generated tree on the basis of the process (4.33) using Algorithm 4.9. The stagewise maximal distances were 0.9, 2, 6, 8, 9. The tree has 390 nodes and 224 leaves. *Left:* the tree structure; *right:* the values

($d = (0.9, 2, 6, 8, 9)$). The structure and the values of the latter tree are displayed in Fig. 4.8.

4.4 Scenario Tree Reduction

If a large tree has been estimated, one may use a reduction algorithm, which reduces the tree to an acceptable size. To do so, one may collapse two different subtrees to a single one, if (i) these subtrees are close to each other and (ii) the resulting subtree is a good compromise between the two former subtrees.

To begin with, suppose that two discrete distributions $P_1 = \sum_{i=1}^k P_i^{(1)}\delta_{\xi_i}$ and $P_2 = \sum_{j=1}^\ell P_j^{(2)}\delta_{\eta_j}$ have to be merged into one. ξ_i and η_j may be vector-valued. To find the “compromise” distribution between P_1 and P_2 consider the optimal transportation plan π , which transports P_1 to P_2 . In order to introduce a parameter for size reduction, let p be the percentage of total probability, which will be retained in the new distribution. One may arrange the elements of π in ascending order, say

$$\pi_{i_1 j_1} \geq \pi_{i_2, j_2} \geq \cdots \geq \pi_{i_m j_m},$$

and choose the smallest m such that $\pi_{i_1 j_1} + \pi_{i_2, j_2} + \cdots + \pi_{i_m j_m} \geq p$. The new “averaged” distribution $P_{12}(p)$ sits on the m points

$$\begin{aligned}
& (\xi_{i_1} + \eta_{j_1})/2 \text{ with probability } \pi_{i_1, j_1}/s \\
& (\xi_{i_2} + \eta_{j_2})/2 \text{ with probability } \pi_{i_2, j_2}/s \\
& \vdots \\
& (\xi_{i_m} + \eta_{j_m})/2 \text{ with probability } \pi_{i_m, j_m}/s.
\end{aligned}$$

with $s = \pi_{i_1, j_1} + \pi_{i_2, j_2} + \dots + \pi_{i_m, j_m}$. Notice that the number m of mass points of $P_{12}(p)$ can be controlled by the choice of p . The maximal number is obtained by choosing $p = 1$, the number of mass points can be considerably reduced by choosing p small.

A quite similar algorithm can be applied for merging subtrees. It is a recursive version of the just presented algorithm.

Remark 4.35. Suppose that the trees \mathbb{P}_1 and \mathbb{P}_2 are collapsed to one tree $\mathbb{P}_{12}(p)$ using the retention rate parameter p . If $p = 100\%$, then

$$\mathbf{dl}(\mathbb{P}_1, \mathbb{P}_{12}(1)) + \mathbf{dl}(\mathbb{P}_2, \mathbb{P}_{12}(1)) = \mathbf{dl}(\mathbb{P}_1, \mathbb{P}_2).$$

In this sense, one may say that $\mathbb{P}_{12}(1)$ is a true “mean” between \mathbb{P}_1 and \mathbb{P}_2 . If however $p < 1$, then typically

$$\begin{aligned}
\mathbf{dl}(\mathbb{P}_1, \mathbb{P}_{12}(p)) &< \mathbf{dl}(\mathbb{P}_1, \mathbb{P}_2) \text{ and} \\
\mathbf{dl}(\mathbb{P}_2, \mathbb{P}_{12}(p)) &< \mathbf{dl}(\mathbb{P}_1, \mathbb{P}_2),
\end{aligned}$$

as well as

$$\mathbf{dl}(\mathbb{P}_1, \mathbb{P}_{12}(p)) + \mathbf{dl}(\mathbb{P}_2, \mathbb{P}_{12}(p)) > \mathbf{dl}(\mathbb{P}_1, \mathbb{P}_2).$$

In this case, the collapsed tree \mathbb{P}_{12} is no longer a “mean” tree, but still a good compromise between \mathbb{P}_1 and \mathbb{P}_2 .

Example 4.36. The tree reduction algorithm selects two close subtrees and merges them into one. In this example, we assume that the two subtrees are \mathbb{P}_1 and \mathbb{P}_2 shown in Fig. 4.9. These two trees are merged into one (sub)tree using the Algorithm 4.10 with different values of p .

Table 4.1 shows the number of nodes and scenarios of the merged trees. It demonstrates that the choice $p = 1$ may lead to large number of nodes, even larger than the sum of nodes of the two initial trees. However, choosing a smaller p may lead to considerable reductions. Figure 4.10 displays the merged tree $\mathbb{P}_{12}(0.5)$, which has less scenarios.

Algorithm 4.10

Recursive tree reduction algorithm

- **STEP 1—CHOICE OF THE SUBTREES TO BE MERGED.** Let a tree \mathbb{P} be given. At each level t the nested distance between all subtrees is calculated. Let \mathbb{P}_1 and \mathbb{P}_2 be the two subtrees at stage t which are closest to each other and should be merged into one. To do so, we use the algorithm MERGING TREES.
- **STEP 2—MERGING TREES.**
 1. For merging two trees into one, the new value ξ_1 at the new root is the mean of the two values of the two old roots.
 2. For the successors of the two roots the averaging algorithm with parameter p is used. Suppose that the selected pairs of nodes are

$$(i_1, j_1), \dots (i_m, j_m).$$

Then, in a recursive step, the subtrees with roots i_1 and j_1 have to be merged, as well as all other pairs i_2 and j_2 up to i_m and j_m .

- **STOP OR CONTINUE.** If the new tree is small enough, stop. Otherwise choose another level t and another pair of close subtrees to be merged into one by going to STEP 1.

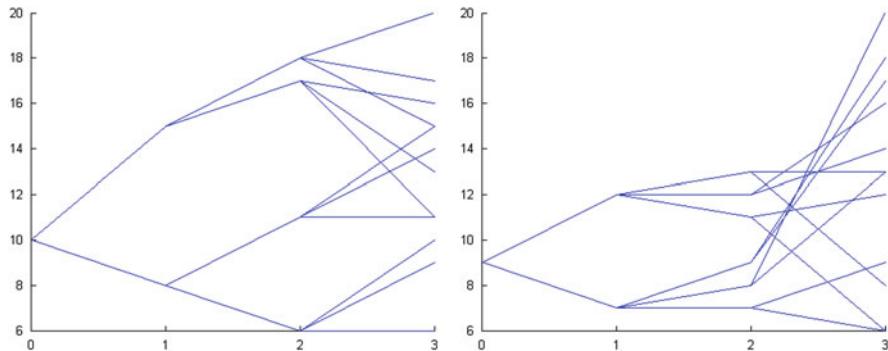


Fig. 4.9 Left: \mathbb{P}_1 (tree1), Right: \mathbb{P}_2 (tree2). Their distance is $d(\mathbb{P}_1, \mathbb{P}_2) = 9.97$

Table 4.1 The sizes of the merged trees

	Number of nodes	Number of scenarios
\mathbb{P}_1	19	12
\mathbb{P}_2	21	12
$\mathbb{P}_{12}(1)$	83	66
$\mathbb{P}_{12}(0.8)$	32	13
$\mathbb{P}_{12}(0.5)$	17	10

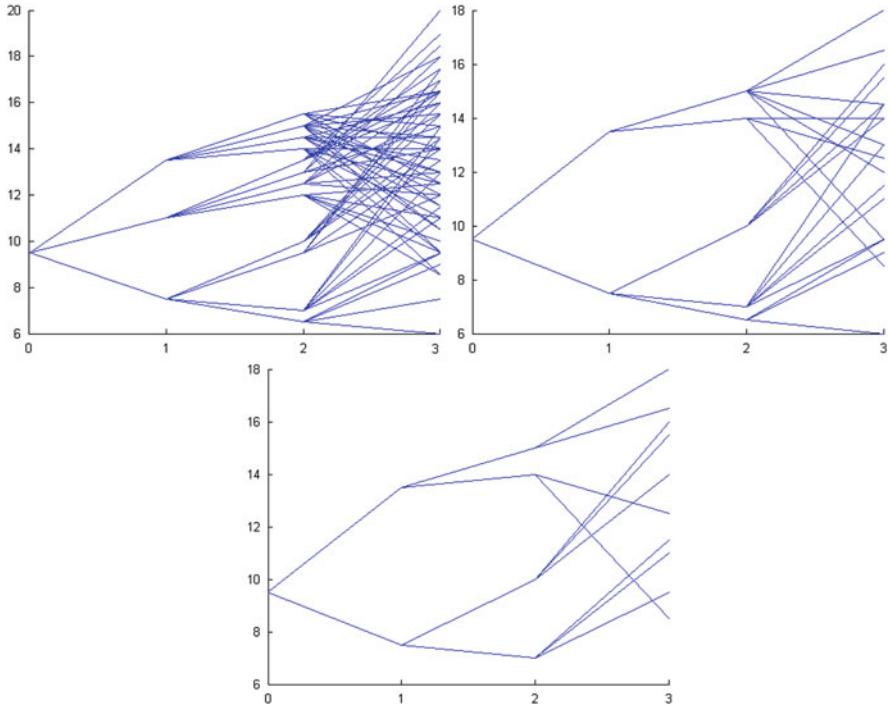


Fig. 4.10 These trees have been constructed using Algorithm 4.10. The first tree $\mathbb{P}_{12}(1)$ has a retention probability of $p = 1$. It results in $d(\mathbb{P}_1, \mathbb{P}_{12}(1)) = 4.98$ and $d(\mathbb{P}_2, \mathbb{P}_{12}(1)) = 4.98$. The second tree is $\mathbb{P}_{12}(0.8)$, i.e., it has retention probability of 0.8. The distances are $d(\mathbb{P}_1, \mathbb{P}_{12}(0.8)) = 4.95$ and $d(\mathbb{P}_2, \mathbb{P}_{12}(0.8)) = 5.28$. Finally, the third tree is $\mathbb{P}_{12}(0.5)$ with distances $d(\mathbb{P}_1, \mathbb{P}_{12}(0.5)) = 5.46$ and $d(\mathbb{P}_2, \mathbb{P}_{12}(0.5)) = 5.66$

4.5 Improvement of Approximating Trees

This section presents an algorithm, which allows improving the approximation quality of a smaller tree $\tilde{\mathbb{P}}$ with respect to a larger tree \mathbb{P} . Successively, the tree $\tilde{\mathbb{P}}$ will be modified to decrease its distance to \mathbb{P} , but only by adapting the scenario values and probabilities. The tree topology, i.e., the filtration structure will not be changed, relying on the fact that the filtration structure was well estimated in previous steps and need not to be improved.

We proceed in an analogous manner as for improving the approximation quality of distribution with respect to the Wasserstein distance: given a tree and an approximating tree, we shall outline in the sequel how to impose better probabilities on the approximating tree, and next, where to better locate the paths of the approximating process in order to obtain improved approximations of the initial tree in terms of the nested distance (the third step in Fig. 4.4).

The results for stochastic processes, however, differ in comparison to probability measures with some extent. Whereas computing the optimal probability masses is

immediate for probability measures (cf. (4.10) and (4.11)), a linear program has to be solved for the nested distance. But the situation is similar for improving the paths: with a natural modification, the paths of the process can easily be adapted, making the same strategy available as for the k-means algorithm for probability measures.

4.5.1 Improvement of the Probability Measure

In view of (2.40) it is obvious that the transport plan π , which respects the marginals imposed by the measure P and the tree structures determined by $(\mathcal{F}_t)_t$ and $(\tilde{\mathcal{F}}_t)_t$, but which is not specified by a probability function \tilde{P} , solves the problem

$$\begin{aligned} & \text{minimize}_{\pi} \quad \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r \\ & \text{subject to} \quad \sum_{j' \in l+} \pi(i', j'|k, l) = P(i'|k) \quad (k \in \mathcal{N} \setminus \mathcal{N}_0, k = i'-, l), \\ & \quad \sum_{i' \in k+} \pi(i', j'|k, l) = \sum_{i' \in k'+} \pi(i', j'|k', l) \quad (k, k' \in \mathcal{N} \setminus \mathcal{N}_0, l = j'-) \\ & \quad \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1. \end{aligned} \tag{4.34}$$

The probability measure \tilde{P} then can be assembled by $\tilde{P}(j) = \sum_{i \in \mathcal{N}_T} \pi_{i,j}$ in the usual way. The constraints $\sum_{i' \in k+} \pi(i', j'|k, l) = \sum_{i' \in k'+} \pi(i', j'|k', l)$ reflect the fact that π is a transport plan respecting both marginals, $(\mathcal{F}_t)_t$ and $(\tilde{\mathcal{F}}_t)_t$. They ensure that

$$\sum_{i' \in k+} \pi(i', j'|k, j'-) = \tilde{P}(j'|j'-)$$

is well defined, irrespective of the choice of k (cf. Fig. 2.8).

The constraint (4.34) is *not* a linear constraint, as the conditional probabilities $\pi(i', j'|k, l)$ represent divisions. Consequently, the problem (4.34) is not an LP.

The recursive structure of the nested distance can be used to iteratively improve approximations in order to solve problem (4.34). For this consider, for fixed $l \in \tilde{\mathcal{N}}_t$, the problem (cf. Fig. 2.8b)

$$\begin{aligned} & \text{minimize}_{\bar{\pi}_t(\cdot|k, l)} \quad \sum_{k \in \mathcal{N}_t} \pi(k, l) \cdot \sum_{i' \in k+, j' \in l+} \bar{\pi}_t(i', j'|k, l) \cdot \mathbf{d}_r(i', j')^r \\ & \text{subject to} \quad \sum_{j' \in l+} \bar{\pi}_t(i', j'|k, l) = P(i'|k) \quad (i' \in k+), \\ & \quad \sum_{i' \in k+} \bar{\pi}_t(i', j'|k, l) = \sum_{i' \in k'+} \bar{\pi}_t(i', j'|k', l) \quad (j' \in l+) \\ & \quad \bar{\pi}_t(i', j'|k, l) \geq 0, \end{aligned} \tag{4.35}$$

where π is fixed. Here, $\pi(k, l)$ is the total mass transported between nodes $k \in \mathcal{N}_t$ and $l \in \tilde{\mathcal{N}}_t$,

$$\pi(k, l) = \sum_{\substack{i > k \\ j > l}} \pi_{i,j}.$$

Let $\bar{\pi}_t$ be the solution of the linear problem (4.35). Defining

$$\bar{\pi}_{i,j} := \bar{\pi}_{T-1}(i, j | i_{T-1}, j_{T-1}) \cdots \bar{\pi}_{t-1}(i_t, j_t | i_{t-1}, j_{t-1}) \cdots \bar{\pi}_0(i_1, j_1 | 1, 1) \quad (4.36)$$

in the usual way it is obvious that

$$\sum_{i,j} \bar{\pi}_{i,j} d_{i,j}^r \leq \sum_{i,j} \pi_{i,j} d_{i,j}^r,$$

where $\pi_{i,j}$ is defined as in (4.36), but $\pi_t(\cdot|k, l)$ replaced by $\bar{\pi}_t(\cdot|k, l)$. In contrast to (4.34) the problem (4.35) is linear and can be employed at every stage. This gives a sequential improvement of probability masses, which can be exploited algorithmically.

4.5.2 Improvement of the Paths

In Sect. 4.1.3.2 it was elaborated how to construct improved locations in order to obtain better approximations of a given probability measure. This approach is repeated here for a stochastic process instead of a random variable. For this purpose it is advantageous to consider the paths of the process: every path ξ is a sequence

$$\xi = (\xi(1), \xi(i_1) \dots \xi(i_{T-1}), \xi(i)) \in \Xi_0 \times \Xi_1 \cdots \times \Xi_T,$$

where $i \in \mathcal{N}_T$ is a leaf node and $1, i_1, \dots, i_{T-1}$ are its predecessors. Describing the process as the respective collection of all paths gives a collection of $|\mathcal{N}_T|$ matrices $(\xi(1), \dots, \xi(i_{T-1}), \xi(i))_{i \in \mathcal{N}_T}$, each of them has dimension $m \times (T + 1)$.

Now suppose that a transport plan π is given, such that the nested distance satisfies

$$d_r(\mathbb{P}, \tilde{\mathbb{P}})^r = \sum_{i,j} \pi_{i,j} \cdot d_{i,j}^r,$$

where $d_{i,j}$ is the distance of the entire path $(\xi(1), \dots, \xi(i_{T-1}), \xi(i))$ and $(\tilde{\xi}(1), \dots, \tilde{\xi}(j_{T-1}), \tilde{\xi}(j))$, that is,

$$d_{i,j} = d((\xi(1), \dots, \xi(i_{T-1}), \xi(i)), (\tilde{\xi}(1), \dots, \tilde{\xi}(j_{T-1}), \tilde{\xi}(j))).$$

Note that the path $(\xi(1), \dots, \xi(i_{T-1}), \xi(i))$ is fixed, whereas $(\tilde{\xi}(1), \dots, \tilde{\xi}(j_{T-1}), \tilde{\xi}(j))$ can be changed, so that one can look up the infimum

$$\begin{aligned} & \underset{\text{in } \tilde{\xi}}{\text{minimize}} \quad \sum_{i,j} \pi_{i,j} \cdot d\left((\xi(1), \dots, \xi(i_{T-1}), \xi(i)), (\tilde{\xi}(1), \dots, \tilde{\xi}(j_{T-1}), \tilde{\xi}(j))\right)^r \\ & \text{subject to } (\tilde{\xi}(1), \dots, \tilde{\xi}(j_{t-1}), \tilde{\xi}(j_t)) = (\tilde{\xi}(1), \dots, \tilde{\xi}(j'_{t-1}), \tilde{\xi}(j'_t)), \text{ whenever } j_t = j'_t. \end{aligned} \quad (4.37)$$

The constraints in (4.37) ensure that the history $\tilde{\xi}(1), \dots, \tilde{\xi}(j'_{t-1}), \tilde{\xi}(j'_t)$ of $\tilde{\xi}$ up to the node j'_t at stage t is kept unchanged. However, one may alternatively consider an unconstrained optimization of the same objective

$$\left(\tilde{\xi}(n)\right)_{n \in \mathcal{N}} \mapsto \sum_{i,j} \pi_{i,j} \cdot d\left((\xi(1), \dots, \xi(i_{T-1}), \xi(i)), (\tilde{\xi}(1), \dots, \tilde{\xi}(j_{T-1}), \tilde{\xi}(j))\right)^r \quad (4.38)$$

for minimization, where all states $\tilde{\xi}(j_t)$ at all stages ($j_t \in \mathcal{N}_t$, $t = 0, \dots, T$) can be changed simultaneously in the optimization procedure. This is justified, as the formulation (4.38) respects the filtration $(\tilde{\mathcal{F}}_t)_{t=0}^T$ for fixed transportation plan $\pi_{i,j}$.

As for the Wasserstein distance of order 2, and the distance for the paths in \mathbb{R}^{Tm} given by $d^2 = \sum_{t=0}^T w_t \|\xi(i_t) - \tilde{\xi}(j_t)\|_2^2$ the best optimal value of the problem (4.37) can be given explicitly:

Proposition 4.37. *For the nested distance of order 2 and the weighted Euclidean distance the optimal values in (4.37) are given by*

$$\tilde{\xi}^*(j_t) := \frac{\sum_{i_t \in \mathcal{N}_t} \pi(i_t, j_t) \cdot \xi(i_t)}{\sum_{i_t \in \mathcal{N}_t} \pi(i_t, j_t)}.$$

Proof. The explicit decomposition of the nested distance allows the rearrangement

$$\begin{aligned} d\ell_2\left(\mathbb{P}, \tilde{\mathbb{P}}\right)^2 &= \sum_{i,j} \pi_{i,j} \cdot d\left((\xi(1), \dots, \xi(i_{T-1}), \xi(i)), (\tilde{\xi}(1), \dots, \tilde{\xi}(j_{T-1}), \tilde{\xi}(j))\right)^2 \\ &= \sum_{i,j} \pi_{i,j} \sum_{t=0}^T w_t \cdot \left\| \xi(i_t) - \tilde{\xi}(j_t) \right\|_2^2 \\ &= \sum_{t=0}^T w_t \cdot \sum_{j_t \in \mathcal{N}_t} \left(\sum_{i_t \in \mathcal{N}_t} \pi(i_t, j_t) \left\| \xi(i_t) - \tilde{\xi}(j_t) \right\|_2^2 \right). \end{aligned}$$

By the same reasoning as for (4.13) the assertion follows for every $j_t \in \mathcal{N}_t$ by considering and minimizing every map

$$q \mapsto \sum_{i_t \in \mathcal{N}_t} \pi(i_t, j_t) \cdot \|\xi(i_t) - q\|_2^2$$

separately, which has its minimum in $\tilde{\xi}^*(j_t) := \frac{\sum_{i_t \in \mathcal{N}_t} \pi(i_t, j_t) \cdot \xi(i_t)}{\sum_{i_t \in \mathcal{N}_t} \pi(i_t, j_t)}$. \square

Algorithm 4.11 summarizes the content of this Sect. 4.5, and its convergence is proved in the following Proposition 4.38. An exemplary computation is depicted in Fig. 4.11.

Proposition 4.38. *Provided that the minimization (4.37) can be done exactly—as is the case for the quadratic nested distance—Algorithm 4.11 terminates at a stationary $\text{dl}_r(\mathbb{P}, \tilde{\mathbb{P}}^{k^*})$ after finitely many iterations k^* .*

Proof. It is possible—although very inadvisable for computational purposes—to rewrite the computation of dl_r^{k+1} in Algorithm 4.11 as a single linear program of the form

$$\begin{aligned} & \text{minimize} \\ & \text{in } \pi^{k+1} \quad c(\pi^{k+1} | \pi^k) \\ & \text{subject to} \quad A\pi^{k+1} = b, \\ & \quad \pi^{k+1} \geq 0, \end{aligned}$$

where the matrix A and the vector b collect all linear conditions from (4.35), and $\pi \mapsto c(\pi | \tilde{\pi})$ is multilinear. Note that the constraint set $\Pi := \{\pi : A\pi = b, \pi \geq 0\}$ is a convex polytope, which is independent of the iterate π^k . Without loss of generality one may assume that π^k is an edge of the polytope Π . Because Π has finitely many edges and each edge $\pi \in \Pi$ can be associated with a unique scenario $\xi(\pi)$ by assumption, it is clear that the decreasing sequence

$$\text{dl}_r^{k+2}(\mathbb{P}, \tilde{\mathbb{P}}^{k+2}) = c(\pi^{k+2} | \pi^{k+1}) \leq c(\pi^{k+1} | \pi^k) = \text{dl}_r^{k+1}(\mathbb{P}, \tilde{\mathbb{P}}^{k+1})$$

cannot improve further whenever the stationary point is met. \square

Algorithm 4.11

Sequential improvement of the nested distributions $\tilde{\mathbb{P}}^k$ to approximate a given nested distribution \mathbb{P} of same length T with respect to the nested distance. The filtrations $(\mathcal{F}_t)_{t \in \{0, \dots, T\}}$ ($(\tilde{\mathcal{F}}_t)_{t \in \{0, \dots, T\}}$, resp.) are kept fixed, only the scenario values and the probabilities of $\tilde{\mathbb{P}}^k$ are changed

STEP 1—INITIALIZATION

Set $k \leftarrow 0$, and let $\tilde{\xi}^0 \triangleleft \tilde{\mathfrak{F}}$ be the values of an initial approximating tree $\tilde{\mathbb{P}}^0$. Moreover, let π^0 be a transport plan, specified by the marginals of $\tilde{\mathbb{P}}^0$ for all sigma algebras \mathcal{F}_t and $\tilde{\mathcal{F}}_t$, $t = 0, \dots, T$ as well as the processes scenario i of the original \mathbb{P} -tree and scenario j^0 of the approximating tree $\tilde{\mathbb{P}}^0$.

STEP 2—IMPROVE THE LOCATIONS

Find improved quantizers $\tilde{\xi}^{k+1}(j_t)$ for every $j_t \in \tilde{\mathcal{N}}_t$:

- In case of the quadratic Wasserstein distance (Euclidean distance and Wasserstein of order $r = 2$) set

$$\tilde{\xi}^{k+1}(j_t) := \frac{\sum_{i_t \in \mathcal{N}_t} \pi^k(i_t, j_t) \cdot \tilde{\xi}^k(i_t)}{\sum_{i_t \in \mathcal{N}_t, j_t \in \tilde{\mathcal{N}}_t} \pi^k(i_t, j_t)},$$

- or solve (4.37), for example by applying the steepest descent method, or the limited memory BFGS method.⁶

STEP 3—IMPROVE THE PROBABILITIES

Solve (4.35) with $\bar{\pi}$ replaced by π^k , and denote the solution π^{k+1} according to (4.36). The distance is

$$[\mathbf{d}_r^{k+1}]^r := \sum_{i,j} \pi_{i,j}^{k+1} \cdot \mathbf{d}\left((\tilde{\xi}(1), \dots, \tilde{\xi}(i_{T-1}), \tilde{\xi}(i)), (\tilde{\xi}^{k+1}(1), \dots, \tilde{\xi}^{k+1}(j_{T-1}), \tilde{\xi}^{k+1}(j))\right)^r.$$

STEP 4—FINAL ASSIGNMENT

Set $k \leftarrow k + 1$ and continue with **STEP 2** if

$$\mathbf{d}_r^{k+1} < \mathbf{d}_r^k - \varepsilon,$$

where $\varepsilon > 0$ is the desired improvement in each cycle k .

Otherwise, define the final approximating tree $\tilde{\mathbb{P}}^*$ such that its scenario values are $\tilde{\xi}^{k+1}(\cdot)$ and the unconditional probabilities sitting on the leaves $j \in \tilde{\mathcal{N}}_T$ are $\sum_i \pi_{i,j}^{k+1}$. Then $\mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}}^*) = \mathbf{d}_r^{k+1}$ and the procedure stops.

Remark. In case of the quadratic nested distance ($r = 2$) and the weighted Euclidean distance, the choice $\varepsilon = 0$ is possible.

⁶If necessary, additional moment conditions may be required here.

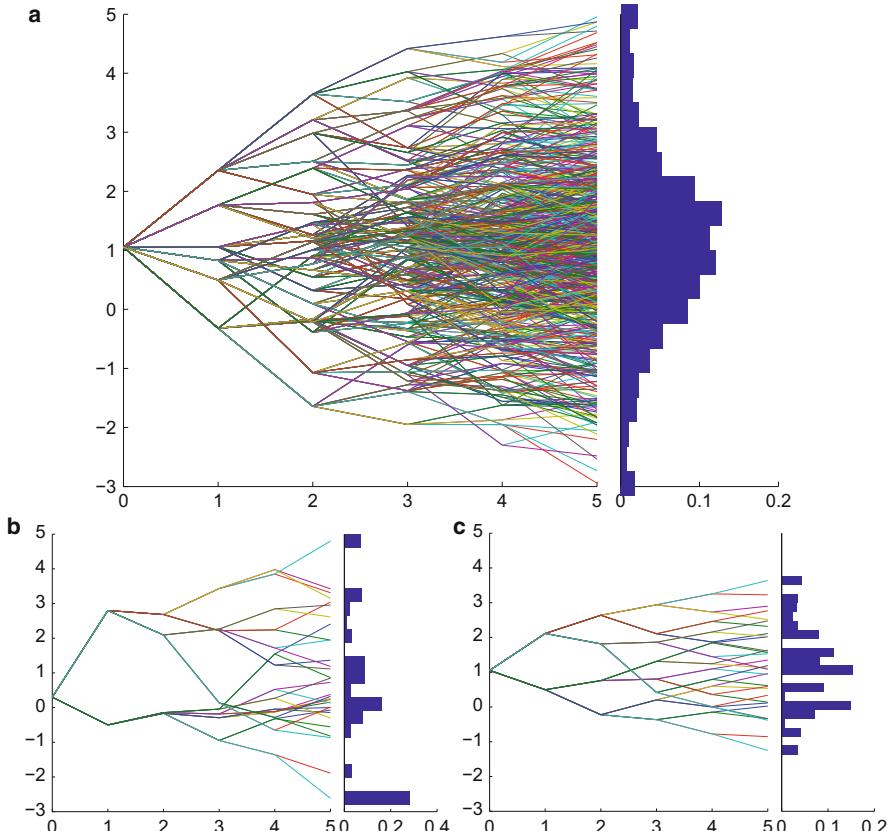


Fig. 4.11 An initial tree with 1237 nodes and two approximations with 63 nodes. Apparently, the tree Fig. 4.11c is a better approximation of the tree Fig. 4.11a than Fig. 4.11b. (a) This initial tree process branches 6 times at the first stage, 5 times at the second, etc. (b) Starting (binary) tree for Algorithm 4.11. The nested distance to the process in Fig. 4.11a is 6.23. (c) The Algorithm 4.11 modifies the starting approximation (Fig. 4.11b) resulting in a process with nested distance 1.1 to the initial tree (Fig. 4.11a)

4.6 An Alternative View on Approximations

In order to measure the distances between two discretizations on the same probability space one could first measure the distance of the respective filtrations. This concept of filtration distance for scenario generation has been developed by Heitsch and Römisch, see [52–54].

Let $(\Omega, \tilde{\mathcal{F}}, P)$ be a probability space and let \mathcal{F} and $\tilde{\mathcal{F}}$ be two sub-sigma algebras of $\tilde{\mathcal{F}}$. Then the *Kudō* distance (cf. [72]) between the sigma algebras \mathcal{F} and $\tilde{\mathcal{F}}$ is defined as

$$\mathfrak{D}(\mathcal{F}, \tilde{\mathcal{F}})$$

$$= \max \left(\sup \left\{ \|\mathbb{1}_A - \mathbb{E}(\mathbb{1}_A | \tilde{\mathcal{F}})\|_p : A \in \mathcal{F} \right\}, \sup \left\{ \|\mathbb{1}_B - \mathbb{E}(\mathbb{1}_B | \mathcal{F})\|_p : A \in \tilde{\mathcal{F}} \right\} \right). \quad (4.39)$$

The most important case is the distance between a sigma-algebra and a sub-sigma-algebra of it: if $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, then (4.39) reduces to

$$\mathfrak{D}(\mathcal{F}, \tilde{\mathcal{F}}) = \sup \left\{ \|\mathbb{1}_A - \mathbb{E}(\mathbb{1}_A | \tilde{\mathcal{F}})\|_p : A \in \mathcal{F} \right\}.$$

If $\mathfrak{F} = (\mathcal{F}_1, \mathcal{F}_2, \dots)$ is an infinite filtration, one may ask, whether this filtration converges to a limit, i.e., whether there is a \mathcal{F}_∞ , such that

$$\mathfrak{D}(\mathcal{F}_t, \mathcal{F}_\infty) \rightarrow 0, \text{ for } t \rightarrow \infty.$$

If \mathcal{F}_∞ is the smallest σ -field, which contains all \mathcal{F}_t , then $\mathfrak{D}(\mathcal{F}_t, \mathcal{F}_\infty) \rightarrow 0$. However, the following example shows that not every discrete, increasing filtration converges to the Borel σ -field.

Example 4.39. Let \mathcal{F} be the Borel-sigma-algebra on $[0,1)$ and let $\tilde{\mathcal{F}}_n$ be the sigma-algebra generated by the sets $\left[\frac{k}{2^n}, \frac{k+1}{2^n} \right), k = 0, \dots, 2^n - 1$. Moreover let $A_n = \bigcup_{k=0}^{2^n-1} \left[\frac{k}{2^n}, \frac{2k+1}{2^{n+1}} \right)$. Then $\mathbb{E}(\mathbb{1}_{A_n} | \tilde{\mathcal{F}}_n) = \frac{1}{2}$ and $\|\mathbb{1}_{A_n} - \mathbb{E}(\mathbb{1}_{A_n} | \tilde{\mathcal{F}}_n)\|_p = \frac{1}{2}$ for all n . While one has the intuitive feeling that $\tilde{\mathcal{F}}_n$ approaches \mathcal{F} , the Kudō distance is always $\frac{1}{2}$.

The Heitsch–Römisich reduction method requires that there is only one probability space $(\Omega, \tilde{\mathcal{F}}, P)$ on which both processes ξ and $\tilde{\xi}_t$, as well as both filtrations \mathfrak{F} and $\tilde{\mathfrak{F}}$ are defined.

The distance between the processes ξ_t on \mathfrak{F} and $\tilde{\xi}_t$ on $\tilde{\mathfrak{F}}$ is given as

$$\mathfrak{D}(\xi, \mathfrak{F}; \tilde{\xi}, \tilde{\mathfrak{F}}) = w_1 \left(\sum_{t=1}^T \int \|\xi_t - \tilde{\xi}_t\|^r P(d\omega) \right)^{1/r} + w_2 \sum_{t=1}^T \sup_{\|x_t\|_\infty \leq 1} \|x_t - \mathbb{E}[x_t | \tilde{\mathcal{F}}_t]\|_{r'},$$

where w_1 and w_2 are two appropriate weights. Notice that a filtration distance appears as the second summand in this definition. By some upper bounding techniques, the final rule for merging two adjacent nodes with node values $\xi_t(i)$ and $\xi_t(j)$ and with conditional probabilities $P(i| -)$ ($P(j| -)$, resp.) depends on the following quantity

$$w_1 P(i| -)^{1/r} \|\xi_t(i) - \xi_t(j)\| + w_2 \frac{\left(P(i| -) \cdot P(j| -)^{r'} + P(i| -)^{r'} P(j| -) \right)^{1/r'}}{P(i| -) + P(j| -)}.$$

Details can be found in the cited papers of Heitsch and Römisich.

Chapter 5

Time Consistency

In a multistage problem decisions x_t have to be made at several stages, say at times $t = 0, 1, \dots, T$. The solution of the problem at time 0 consists of a complete plan for all future decisions at later times. If it turns out that it is preferable to change the initial plan at later stages, then the decision problem is called *inconsistent in time*. As will be shown, time inconsistency may appear quite naturally in risk-averse stochastic multistage decision problems.

For deterministic problems such a phenomenon may not occur: suppose that an overall cost function $Q(x_0, \dots, x_T)$ is to be minimized within the feasible set \mathbb{X} and let x^* be an overall solution of the problem

$$x^* = (x_0^*, \dots, x_T^*) \in \operatorname{argmin} \{Q(x_0, x_1, \dots, x_T) : (x_0, \dots, x_T) \in \mathbb{X}\}.$$

At time t , the decisions x_0^*, \dots, x_{t-1}^* are already made and fixed, while the remaining decisions x_t, \dots, x_T may be revised, if necessary. However,

$$(x_t^*, \dots, x_T^*) \in \operatorname{argmin} \left\{ Q(x_0^*, \dots, x_{t-1}^*, x_t, \dots, x_T) \middle| \begin{array}{l} (x_t, \dots, x_T) \text{ such that} \\ (x_0^*, \dots, x_{t-1}^*, x_t, \dots, x_T) \in \mathbb{X} \end{array} \right\},$$

i.e., the decisions are time consistent. Thus for deterministic problems the following principle is valid (cf. also Carpentier et al. [16, page 249] and Shapiro [127]).

The principle of time consistency

If the optimal decision sequence is implemented, but only up to time t , and at time t the problem is resolved for the remaining times keeping the already made decisions as fixed, then the optimal solution of this subproblem coincides with that of the original problem.¹

¹More precisely: among all solutions of the problem for remaining times is also the solution of the original problem.

Time consistency holds also for controlled deterministic dynamic systems: let a dynamic system be given with state z_t and control x_t . Starting with an initial state z_0 , the state evolves as

$$z_{t+1} = g_t(z_t, x_t), \quad t = 0, \dots, T-1, \quad (5.1)$$

where the feasible set for the controls is

$$x_t \in \mathbb{X}_t(z_t). \quad (5.2)$$

The objective is

$$\min \{Q(z_0, \dots, z_T) : \text{under (5.1) and (5.2)}\},$$

for which time consistency holds as well by the same reasoning as for the deterministic case. This fact is known under the name of *Bellman principle*.

5.1 Time Consistency in Stochastic Decision Problems

For general multistage stochastic optimization problems time consistency is defined as follows:

Let a basic problem of the form

$$Opt(\mathbb{P}) : \min \{\mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : x \in \mathbb{X}(\xi), x \triangleleft \mathfrak{F}, (\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}\}$$

be given. Assume that the filtration \mathfrak{F} is generated by a tree process v_t ($t = 0, \dots, T$), i.e., $\mathcal{F}_t = \sigma(v_t)$. For any stage t and some possible value u^2 of v_t there corresponds a conditional nested distribution $\mathbb{P}(\cdot | v_t = u)$, written as $\mathbb{P}^{v_t=u}$. For finite trees, this means that the node n corresponding to $v_t = u$ and all its predecessors get probability 1, while all successors $m > n$ get the new probabilities $P(m)/P(n)$. The filtration and the scenario process remain the same (but the values are irrelevant on nodes which have conditional probability 0). The conditional decision problem is

$$Opt(\mathbb{P}^{v_t=u}) : \min \{\mathcal{R}_{\mathbb{P}^{v_t=u}}[Q(x, \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}, (\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}\}.$$

Of course, $Opt(\mathbb{P})$ and $Opt(\mathbb{P}^{v_t=u})$ may have different optimal solutions and different optimal values. But under time consistency, inserting the solution of $Opt(\mathbb{P})$ until stage t and finding the solution for the residual stages of $Opt(\mathbb{P}^{v_t=u})$

²In the finite case, u can be just the node number.

should lead to the same solutions for the remaining times. In mathematical notation, the time consistency principle can be reformulated as follows:

Definition 5.1 (Time Consistency). The optimization problem $Opt(\mathbb{P})$ is called *time consistent*, if for each solution x^* of $Opt(\mathbb{P})$ and each conditioned problem $Opt(\mathbb{P}^{v_t=u})$, the subsolution $x_{t:T}^*$ of x^* satisfies³

$$x_{t:T}^* \in \operatorname{argmin} \{ \mathcal{R}_{\mathbb{P}^{v_t=u}} [Q((x_{0:t-1}^*, x_{t:T}), \xi)] : (x_{0:t-1}^*, x_{t:T}) \in \mathbb{X}, x \triangleleft \mathfrak{F}, (\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P} \}.$$

For a time consistent problem, the initial decision plan never has to be revised. As illustration, look at Fig. 5.1. The left side shows a full problem and the right side shows a conditional problem, conditioned on node 3.

A property closely related to time consistency is dynamic decomposability. Decomposable multistage decision problems can be written as structures of nested optimization problems. To allow decomposability, assume that the constraint set $\mathbb{X}(\xi)$ is of the specific form

$$\{x \in \mathbb{X}\} \iff \{x_0 \in \mathbb{X}_0, x_1 \in \mathbb{X}_1(x_0, \xi_1), \dots, x_{T-1} \in \mathbb{X}_{T-1}(x_{0:T-2}, \xi_{1:T-1})\}.$$

Here is the definition of decomposability.

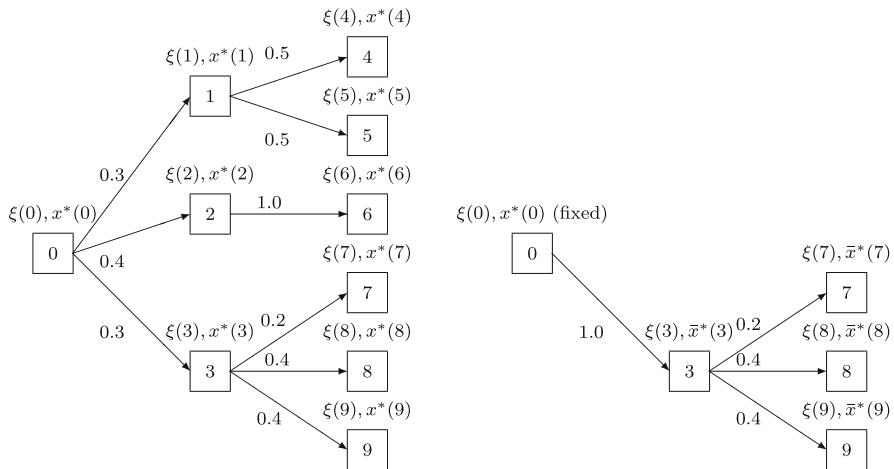


Fig. 5.1 Left: a full problem with scenario process ξ and optimal solution x^* .

Right: the conditional problem, conditioned on node 3. The optimal solution of this problem is called \bar{x}^* . The decision problem is time consistent, if $x^*(n) = \bar{x}^*(n)$ for all nodes n , which are in the subtree of the conditioning node (here node 3)

³ $(x_{0:t-1}^*, x_{t:T})$ is the concatenated vector $(x_{0:t-1}^*, \dots, x_{t-1}^*, x_t, \dots, x_T)$.

Definition 5.2 (Dynamic Decomposition). The problem

$$Opt(\mathbb{P}) : \min \{\mathcal{R}[Q(x_{0:T-1}, \xi_{1:T})] : x_t \triangleleft \mathcal{F}_t, x_t \in \mathbb{X}_t(x_{0:t-1}, \xi_{1:t})\} \quad (5.3)$$

is *dynamically decomposable*, if there exist conditional functionals \mathcal{R}_t such that (5.3) is equivalent to

$$\min_{x_0 \in \mathbb{X}_0} \mathcal{R}_0 \left(\min_{\substack{x_1 \in \mathbb{X}(x_0, \xi_1) \\ x_1 \triangleleft \mathcal{F}_1}} \mathcal{R}_1 \left(\dots \min_{\substack{x_{T-1} \in \mathbb{X}(x_{0:T-2}, \xi_{1:T-1}) \\ x_{T-1} \triangleleft \mathcal{F}_{T-1}}} \mathcal{R}_{T-1} \left(Q(x_{0:T-1}, \xi_{0:T}) \right) \right) \right). \quad (5.4)$$

Remark 5.3. If the probability functional is the expectation $\mathcal{R} = \mathbb{E}$, then time decomposability holds. This can be seen by choosing the functionals \mathcal{R}_t as $\mathcal{R}_t(\cdot) := \mathbb{E}(\cdot | \mathcal{F}_t)$:

$$\begin{aligned} & \min_{x_0 \in \mathbb{X}_0} \min_{\substack{x_1 \in \mathbb{X}(x_0, \xi_1) \\ x_1 \triangleleft \mathcal{F}_1}} \dots \min_{\substack{x_{T-1} \in \mathbb{X}(x_{0:T-2}, \xi_{1:T-1}) \\ x_{T-1} \triangleleft \mathcal{F}_{T-1}}} \mathbb{E}[Q(x, \xi)] \\ &= \min_{x_0 \in \mathbb{X}_0} \min_{\substack{x_1 \in \mathbb{X}(x_0, \xi_1) \\ x_1 \triangleleft \mathcal{F}_1}} \dots \min_{\substack{x_{T-1} \in \mathbb{X}(x_{0:T-2}, \xi_{1:T-1}) \\ x_{T-1} \triangleleft \mathcal{F}_{T-1}}} \mathbb{E}\left[\mathbb{E}_1\left[\dots \mathbb{E}_{T-1}[Q(x, \xi)]\dots\right]\right] \\ &= \min_{x_0 \in \mathbb{X}_0} \mathbb{E}\left[\min_{\substack{x_1 \in \mathbb{X}(x_0, \xi_1) \\ x_1 \triangleleft \mathcal{F}_1}} \mathbb{E}_1\left[\dots \min_{\substack{x_{T-1} \in \mathbb{X}(x_{0:T-2}, \xi_{1:T-1}) \\ x_{T-1} \triangleleft \mathcal{F}_{T-1}}} \mathbb{E}_{T-1}[Q(x_{0:T-1}, \xi_{0:T})] \right] \right], \end{aligned}$$

where $\mathbb{E}_t = \mathbb{E}(\cdot | \mathcal{F}_t)$.

Lemma 5.4. *If a stochastic problem is decomposable in time, then time consistency holds and the solution can be found by backward induction as in deterministic dynamic programming.*

Proof. See Shapiro [125]. □

Time decomposability may not hold, if

- the functional is not the expectation or
- other than the standard nonanticipativity constraints are in place (e.g., $x_t \triangleleft \mathcal{F}_s$ for $s < t$, that is decisions implemented at time t must be already been taken at time $s < t$).

Time decomposability thus depends on the choice of the functional and the form of the constraints.

5.2 Time Consistent Risk Functionals

Many conditional risk functionals repeat an initial, genuine risk functional as, for example, the expectation or the Average Value-at-Risk at some fixed level (cf. (5.4)). However, there are situations where sticking to the same, genuine risk functional on a conditional basis may lead to inconsistent risk assessments.

This section addresses and investigates several properties of conditional risk functionals. As above we consider a probability space (Ω, \mathcal{F}, P) and a sigma-algebra $\mathcal{F}_1 \subseteq \mathcal{F}$. Let $\mathcal{R}_1(\cdot|\mathcal{F}_1)$ be a conditional risk mapping (see Sect. 3.6) and let $\mathcal{R}_0(\cdot)$ be an unconditional risk functional. Typically, but not necessarily, \mathcal{R}_0 is the unconditional counterpart of $\mathcal{R}_1(\cdot|\mathcal{F}_1)$.

Definition 5.5 (Artzner et al. [7], Penner [86]). The pair $\mathcal{R}_0(\cdot), \mathcal{R}_1(\cdot|\mathcal{F}_1)$ is called *time consistent*, if for all $Y, \tilde{Y} \in \mathcal{Y}$ the implication

$$\mathcal{R}_1(Y|\mathcal{F}_1) \leq \mathcal{R}_1(\tilde{Y}|\mathcal{F}_1) \text{ a.s.} \implies \mathcal{R}_0(Y) \leq \mathcal{R}_0(\tilde{Y}) \quad (5.5)$$

holds. If \mathcal{R}_1 is the conditional version of $\mathcal{R} = \mathcal{R}_0$, we say that \mathcal{R} is time consistent.

The pair consisting of expectation and conditional expectation is time consistent, and so is the essential supremum. However, important risk functionals fail to be time consistent in the sense of the definition.

Example 5.6. Risk functionals, which are not time consistent in the sense of Definition 5.5.

- (i) The simple Average Value-at-Risk and the conditional Average Value-at-Risk at the same level $\alpha > 0$ is not a time consistent pair. Figure 5.2a exhibits this—perhaps unexpected—behavior for the level $\alpha = 90\%$. Indeed, consider the two random variables Y and \tilde{Y} and the filtration displayed. Notice that

$$\text{AV@R}_{0.9}(Y|\mathcal{F}_1) = \binom{5}{1} < \binom{6}{2} = \text{AV@R}_{0.9}(\tilde{Y}|\mathcal{F}_1),$$

while

$$\text{AV@R}_{0.9}(Y) = 4.1 > 3.3 = \text{AV@R}_{0.9}(\tilde{Y}),$$

such that the Average Value-at-Risk is not time consistent in the sense of Definition 5.5.

- (ii) The risk functional

$$\mathcal{R}(Y) := (1 - \mu) \cdot \mathbb{E}Y + \mu \text{ess sup } Y$$

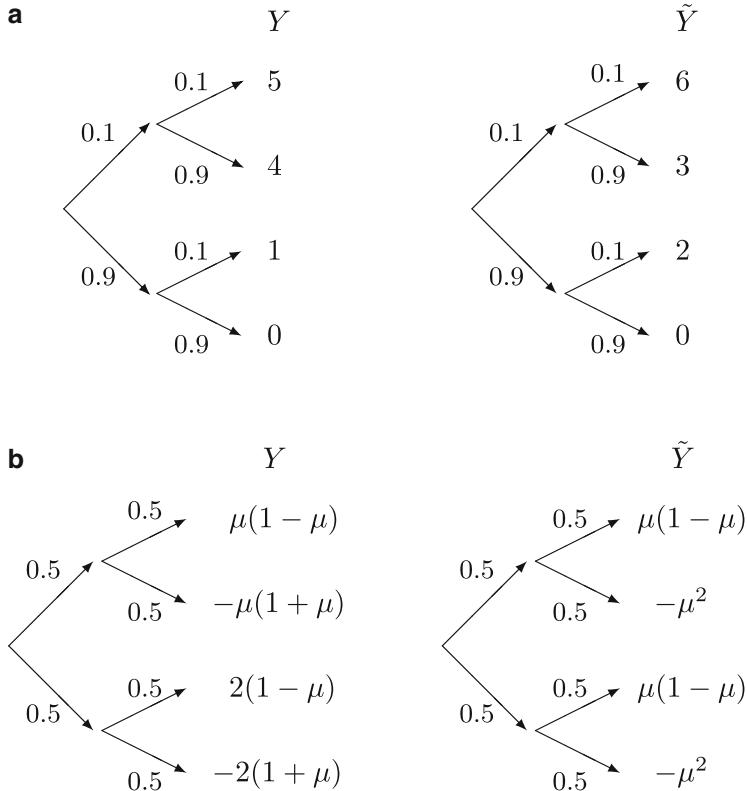


Fig. 5.2 Time inconsistency of selected risk functionals

is a convex combination of time consistent risk functionals. However, it is not time consistent whenever $0 < \mu < 1$, as Fig. 5.2b shows: here,

$$\mathcal{R}(Y|\mathcal{F}_1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} < \begin{pmatrix} \frac{1}{2}\mu(1-\mu) \\ \frac{1}{2}\mu(1-\mu) \end{pmatrix} = \mathcal{R}(\tilde{Y}|\mathcal{F}_1),$$

whereas

$$\mathcal{R}(Y) = \frac{1}{2}\mu(1-\mu) \cdot (2-\mu) > \frac{1}{2}\mu(1-\mu) = \mathcal{R}(\tilde{Y})$$

for $0 < \mu < 1$, as a quick computation shows.

- (iii) As a matter of fact all risk functionals, which repeat the initial risk functional on conditional level as above, are not time consistent—except for the expectation and the essential supremum.

Consistency can be split into acceptance and rejection consistency: a random loss Y is accepted, if its risk is below a limit ρ , say, and rejected if it is above this limit ρ .

Definition 5.7. A pair $\mathcal{R}_0(\cdot), \mathcal{R}_1(\cdot|\mathcal{F}_1)$ is called *acceptance consistent*, if for all $Y \in \mathcal{Y}$ the relation

$$\mathcal{R}_0(Y) \leq \text{ess sup } \mathcal{R}_1(Y|\mathcal{F}_1) \quad (5.6)$$

holds. It is called *rejection consistent*, if

$$\mathcal{R}_0(Y) \geq \text{ess inf } \mathcal{R}_1(Y|\mathcal{F}_1)$$

(see, e.g., Weber [140]).

For an acceptance consistent pair the a.s. conditional acceptance $\text{ess sup } \mathcal{R}_1(Y|\mathcal{F}_1) \leq \rho$ implies the unconditional acceptance $\mathcal{R}_0(Y) \leq \rho$. Likewise for a rejection consistent pair the a.s. conditional rejection $\text{ess inf } \mathcal{R}_1(Y|\mathcal{F}_1) \geq \rho$ implies the unconditional rejection $\mathcal{R}_0(Y) \geq \rho$.

Proposition 5.8. If $\mathcal{R}_0(0) = 0, \mathcal{R}_1(0|\mathcal{F}_1) = 0$ a.s. and both $\mathcal{R}_0(\cdot), \mathcal{R}_1(\cdot|\mathcal{F}_1)$ are translation equivariant, then time consistency (in the sense of Definition 5.5) of the pair implies acceptance and rejection consistency.

Proof. The translation equivariance implies that for each constant c , $\mathcal{R}_0(c) = \mathcal{R}_0(c+0) = c + \mathcal{R}_0(0) = c$ and similarly $\mathcal{R}_1(c|\mathcal{F}_1) = c$. Let $c := \text{ess sup } \mathcal{R}_1(Y|\mathcal{F}_1)$. It holds that

$$\mathcal{R}_1(Y|\mathcal{F}_1) \leq c = \mathcal{R}_1(c|\mathcal{F}_1),$$

and by time consistency ((5.5) in Definition 5.5) thus

$$\mathcal{R}_0(Y) \leq \mathcal{R}_0(c) = c = \text{ess sup } \mathcal{R}_1(Y|\mathcal{F}_1),$$

which is the assertion for acceptance consistency. The proof for rejection consistency is similar. \square

We consider now the case that $\mathcal{R}(Y) = \mathcal{R}_0(Y) = \mathcal{R}(Y|\mathcal{F}_0)$, where \mathcal{F}_0 is the trivial sigma-algebra. Recall the Definition of compound concavity ($P \mapsto \mathcal{R}_P(\cdot)$ is concave), see (CC) in Appendix A and its counterpart, compound convexity ($P \mapsto \mathcal{R}_P(\cdot)$ is convex). Notice the following lemma.

Lemma 5.9. A version-independent functional \mathcal{R} is

- (i) compound convex, if and only if $\mathcal{R}(Y) \leq \mathbb{E}(\mathcal{R}(Y|\mathcal{F}_1))$ for all $Y \in \mathcal{Y}$
- (ii) compound concave, if and only if $\mathcal{R}(Y) \geq \mathbb{E}(\mathcal{R}(Y|\mathcal{F}_1))$ for all $Y \in \mathcal{Y}$

and all sigma-algebras \mathcal{F}_1 .

Proposition 5.10. For a pair $\mathcal{R}(\cdot), \mathcal{R}(\cdot|\mathcal{F}_1)$ the following hold true:

- (i) compound convexity implies acceptance consistency.
- (ii) compound concavity implies rejection consistency.

Proof. (i) follows from the fact that

$$\mathcal{R}(Y) \leq \mathbb{E}(\mathcal{R}(Y | \mathcal{F}_1)) \leq \text{ess sup } \mathcal{R}(Y | \mathcal{F}_1),$$

which is the defining Eq. (5.6) in Definition 5.7. The equation

$$\mathcal{R}(Y) \geq \mathbb{E}(\mathcal{R}(Y | \mathcal{F}_1)) \geq \text{ess inf } \mathcal{R}(Y | \mathcal{F}_1)$$

provides (ii), i.e., rejection consistency. \square

Remark 5.11. AV@R (as a distortion measure) is compound concave, hence rejection consistent, but AV@R is not acceptance consistent. A counterexample to acceptance consistency is shown in Fig. 5.3, since

$$\text{AV@R}_{1/3}(Y) = 2, \text{ but } \text{AV@R}_{1/3}(Y | \mathcal{F}_1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Definition 5.12 (Cf. Schachermayer et al. [122] and Jobert et al. [62]). A pair $\mathcal{R}_0(\cdot), \mathcal{R}_1(\cdot | \mathcal{F}_1)$ is called *recursive*, if for all $Y \in \mathcal{Y}$ the equation

$$\mathcal{R}_0(Y) = \mathcal{R}_0(\mathcal{R}_1(Y | \mathcal{F}_1))$$

holds.

Of special interest are version-independent conditional functionals \mathcal{R} , which are auto-recursive (i.e., for which $\mathcal{R}_0 = \mathcal{R}_1 = \mathcal{R}$).

Example 5.13. Auto-recursive property of exemplary risk functionals.

- (i) expected-conditional functionals (i.e., functionals of the form $\mathbb{E}[\mathcal{R}(Y | \mathcal{F}_1)]$) are recursive.

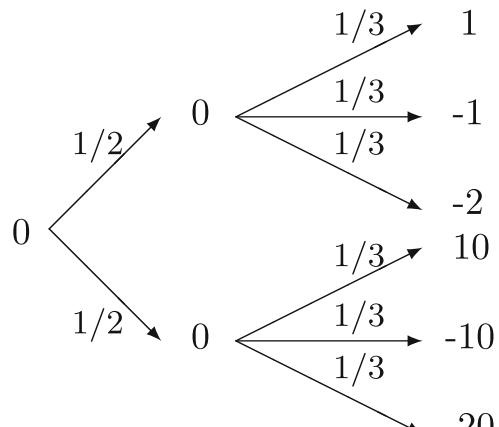


Fig. 5.3 AV@R is not acceptance consistent

(ii) The entropic functional ($\gamma > 0$)

$$-\frac{1}{\gamma} \log \mathbb{E}[e^{-\gamma Y} | \mathcal{F}_1]$$

is auto-recursive, as

$$\begin{aligned} -\frac{1}{\gamma} \log \mathbb{E}\left[\exp\left(-\gamma\left(-\frac{1}{\gamma} \log \mathbb{E}[\exp(-\gamma Y) | \mathcal{F}_1]\right)\right)\right] \\ = -\frac{1}{\gamma} \log \mathbb{E}[\mathbb{E}[\exp(-\gamma Y) | \mathcal{F}_1]] = -\frac{1}{\gamma} \log \mathbb{E}[\exp(-\gamma Y)]. \end{aligned}$$

Note that the entropic functional is not positively homogeneous, it does not satisfy (H).

(iii) The AV@R is not auto-recursive (cf. Fig. 5.3).

Theorem 5.14 (cf. Artzner et al. [7] or Kovacevic et al. [71, Theorem 5.1]). A pair $\mathcal{R}_0(\cdot), \mathcal{R}_1(\cdot | \mathcal{F}_1)$ with monotonic \mathcal{R}_0 and \mathcal{R}_1 enjoying the idempotency property $\mathcal{R}_1(\mathcal{R}_1(\cdot | \mathcal{F}_1) | \mathcal{F}_1) = \mathcal{R}_1(\cdot | \mathcal{F}_1)$ is time consistent if and only if it is recursive.

Proof. Let the pair be recursive and let $\mathcal{R}_1(Y | \mathcal{F}_1) \leq \mathcal{R}_1(\tilde{Y} | \mathcal{F}_1)$. Then, by recursivity and monotonicity

$$\mathcal{R}_0(Y) = \mathcal{R}_0(\mathcal{R}_1(Y | \mathcal{F}_1)) \leq \mathcal{R}_0(\mathcal{R}_1(\tilde{Y} | \mathcal{F}_1)) = \mathcal{R}_0(\tilde{Y}),$$

which is the assertion for time consistency.

Conversely, let the pair be time consistent. By assumption we have that

$$\mathcal{R}_1(\mathcal{R}_1(Y | \mathcal{F}_1) | \mathcal{F}_1) = \mathcal{R}_1(Y | \mathcal{F}_1).$$

Setting $\tilde{Y} := \mathcal{R}_1(Y | \mathcal{F}_1)$ it is evident that

$$\begin{aligned} \mathcal{R}_1(\tilde{Y} | \mathcal{F}_1) &\leq \mathcal{R}_1(Y | \mathcal{F}_1) \text{ and} \\ \mathcal{R}_1(Y | \mathcal{F}_1) &\leq \mathcal{R}_1(\tilde{Y} | \mathcal{F}_1), \end{aligned}$$

from which we conclude that

$$\begin{aligned} \mathcal{R}_0(\tilde{Y}) &\leq \mathcal{R}_0(Y) \text{ and} \\ \mathcal{R}_0(Y) &\leq \mathcal{R}_0(\tilde{Y}) \end{aligned}$$

due to time consistency. Hence

$$\mathcal{R}_0(Y) = \mathcal{R}_0(\tilde{Y}) = \mathcal{R}_0(\mathcal{R}_1(Y | \mathcal{F}_1)),$$

which is the equation of recursivity. \square

Example 5.15 (Ruszczynski and Shapiro [118]). Consider a probability space (Ω, \mathcal{F}, P) and a filtration $\mathfrak{F} = (\mathcal{F}_0, \dots, \mathcal{F}_T)$ of σ -fields \mathcal{F}_t with $\mathcal{F}_T = \mathcal{F}$. Define $\mathcal{Y}_t := L_p(\mathcal{F}_t)$ for $t = 1, \dots, T$ and some $p \in [1, +\infty)$.

Let, for each $t = 0, \dots, T - 1$, conditional risk mappings $\mathcal{R}_t(\cdot | \mathcal{F}_t)$ from \mathcal{Y}_t to \mathcal{Y}_t be given. Introduce a multiperiod probability functional \mathcal{R} on $\mathcal{Y} := \times_{t=1}^T \mathcal{Y}_t$ by compositions of the conditional acceptability mappings \mathcal{R}_t , $t = 0, \dots, T - 1$, namely,

$$\begin{aligned}\mathcal{R}(Y; \mathcal{F}) &:= \mathcal{R}_0[Y_1 + \dots + \mathcal{R}_{T-2}[Y_{T-1} + \mathcal{R}_{T-1}(Y_T)] \dots] \\ &= \mathcal{R}_0 \circ \mathcal{R}_1 \circ \dots \circ \mathcal{R}_{T-1} \left(\sum_{t=1}^T Y_t \right)\end{aligned}$$

for $Y_t \in \mathcal{Y}_t$ by translation equivariance. Then these functionals are recursive in a trivial way.

Example 5.16. Consider the conditional Average Value-at-Risk (of level $\alpha \in (0, 1]$) as conditional risk functional

$$\mathcal{R}_t(Y_{t+1}) := \text{AV@R}_\alpha(\cdot | \mathcal{F}_t)$$

for every $t = 1, \dots, T$. Then the multiperiod risk functional

$$n\text{AV@R}_\alpha(Y; \mathcal{F}) := \text{AV@R}_\alpha(\cdot | \mathcal{F}_0) \circ \dots \circ \text{AV@R}_\alpha(\cdot | \mathcal{F}_{T-1}) \left(\sum_{t=1}^T Y_t \right)$$

is recursive and hence time consistent. It is called the *nested Average Value-at-Risk*.

Time inconsistency appears in a natural way in stochastic risk-adverse optimality problems as the following example shows.

Example 5.17. Consider Fig. 5.4a. Here the circle nodes are decision nodes at which one may decide to follow the upper or the lower arc. The square nodes are the usual probability splitting nodes. Notice that there are 4 possible decisions: *up1-up2*, *up1-down2*, *down1-up2*, and *down1-down2*, i.e., the feasible set consists only of four points. Suppose we want to find the decision which minimizes the convex-concave functional

$$\mathbb{E}(Y) + 0.5 \cdot \text{AV@R}_{0.95}(Y).$$

A simple calculation shows that the optimal decision at the root is *up1-up2*. If, however, the lower node has realized and the problem is reconsidered (Fig. 5.4b), then the optimal decision is *down2*.

This evokes a partly philosophical question: when planning future decisions it may happen that a certain strategy is optimal from today's standpoint, but one knows already right at the beginning that at a later time, if the problem is reconsidered,

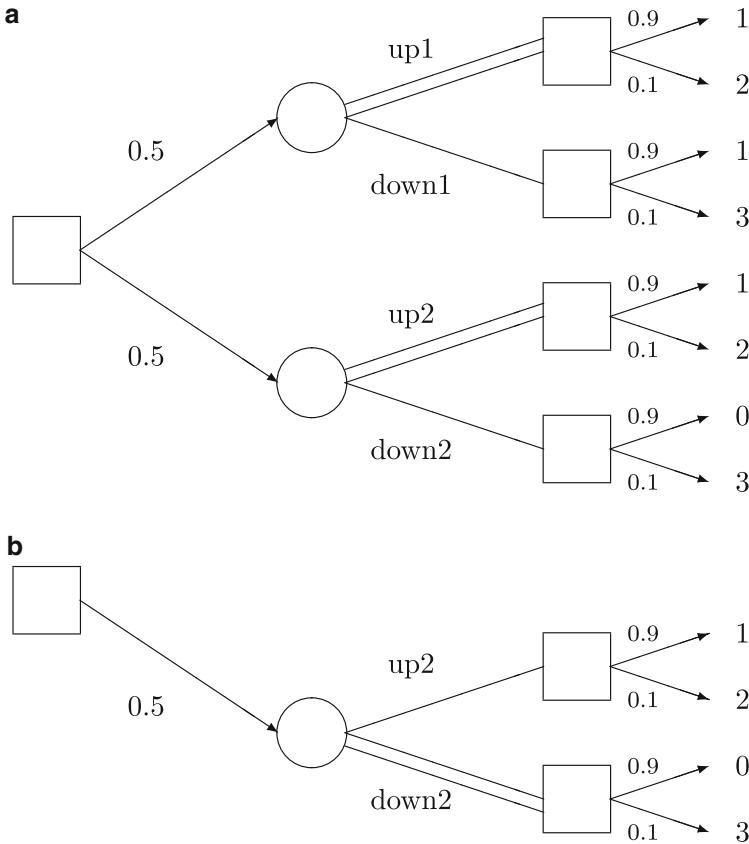


Fig. 5.4 The decision tree for Example 5.17 (above) and the decision after the observation. *Doubled lines* indicate the optimal decision. The optimal decision at the lower node has to be changed from *up2* to *down2*, i.e., the problem is not time consistent

the plan will be changed. So isn't it reasonable to change the plan already today? Or does this mean that the problem was ill-posed? Such a paradox situation can only happen for risk-averse decision problems. Risk neutral problems minimize the expected costs and are always time consistent. We argue that in risk-averse situations one has to accept one of the two possibilities:

- either to reconsider the identical problem at later stages and possibly change the decisions (or work with the changed decisions right from the beginning, being aware that they are suboptimal for the unconditional problem),
- or to change the optimality criterion at later stages in such a way, that time-consistency holds. This idea of conditionally changing the risk functionals will be pursued in Sect. 5.3.

Example 5.18. Here is another example for a time-inconsistent solution of stochastic programs. We reconsider the flowergirl problem of Example 1.4 of the introduction, but change the objective from the maximization of the expected profit to the minimization of the Average Value-at-Risk of the negative profit. For

$$\text{AV@R}_\alpha(Y) = \frac{1}{1-\alpha} \int_\alpha^1 \text{V@R}_p(Y) dp = \min \left\{ q + \frac{1}{1-\alpha} \mathbb{E}[Y - q]_+ : q \in \mathbb{R} \right\}$$

we consider

$$\text{minimize} \left\{ \text{AV@R}_\alpha \left[- \sum_{t=1}^T \xi_t + \sum_{t=0}^{T-1} b_t x_t + \sum_{t=1}^T u_t \cdot [\xi_t]_- - \ell_T \cdot [\xi_T]_+ \right] \right\}$$

under the constraints (1.14)–(1.16). The demand process ξ was chosen as in Example 1.4, the other parameters are

$$\ell = (0.91, 0.92, 0.93), \quad b = (0.95, 0.95, 0.95), \quad u = (1.0, 1.1, 1.2).$$

The optimal decisions x^* of this problem for $\alpha = 0.4$ are shown in Fig. 5.5. Notice that choosing $\alpha = 0$ would lead to the risk-neutral case, i.e., the expectation maximization case, which is time consistent. The choice $\alpha > 0$, however, refers to the risk-averse case and time-inconsistency may happen. Typically, the optimal order sizes in the risk-averse case are smaller than in the risk neutral case.

The solution x^* is the multistage solution calculated at time 0 with no information, i.e., the full tree. Call this solution the *solve-and-keep* solution. We may compare it with the *rolling horizon solution*, where at each node reached, the conditional problem is solved. We denote the rolling horizon solution by x^{**} and use the following notation:

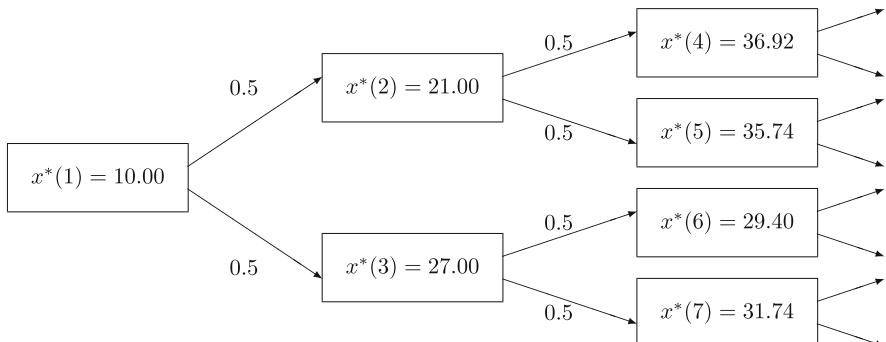


Fig. 5.5 The optimal solution of the risk-adverse flowergirl problem ($\alpha = 0.4$). The optimal value (i.e. the maximal risk-adjusted profit) is $m^* = 3.11$

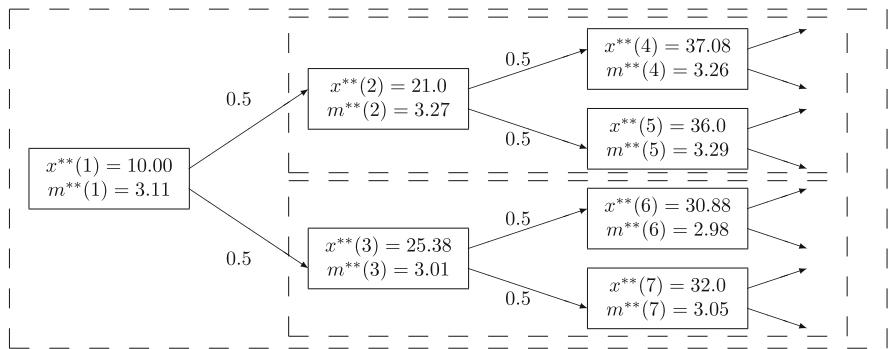


Fig. 5.6 The rolling horizon solution x^{**} of the risk-adverse problem ($\alpha = 0.4$). The optimal value of the full problem, when the solution x^{**} is inserted is $m^{***} = 3.07$. This is a worse objective than $m^* = 3.11$, which is obtained by implementing the solve-and-keep solution x^*

- $x^{**}(1)$ is the decision at time 0 for the full risk-averse problem. It coincides with $x^*(1)$.
- $x^{**}(n)$ is the decision at node n for the conditional problem conditioned at node n . For this solution, the values at the predecessors of node n are kept at the values, which were obtained previously, i.e., when solving the conditional problems at earlier stages.

The values of the rolling horizon solution can be seen in Fig. 5.6. Apparently they differ from the results in Fig. 5.5.

5.3 Time Consistency and Decomposition

5.3.1 Composition of Risk Functionals

As was seen in the previous section, time consistency of risk functionals can be enforced by considering compositions of conditional functionals as introduced Ruszczyński and Shapiro [118] (cf. Example 5.15). One such example is the nested Average Value-at-Risk

$$Y \mapsto \text{AV@R}_\alpha (\text{AV@R}_{\alpha_t} (Y | \mathcal{F}_t)). \quad (5.7)$$

Various important results are known about the composition of risk functionals. Shapiro demonstrates in [127, Theorem 2.1] that a composition of version independent risk functionals of type (5.7) is version independent in the ordinary sense, if it

is either the expectation (AV@R_0) or the max-risk functional ($\text{AV@R}_1 = \text{essup}$).⁴ Moreover recursive risk functionals are time consistent for infinite, increasing sequences of sigma-algebras \mathcal{F}_t (cf. Schachermayer et al. [122] for this result) if and only if they are of the specific form

$$Y \mapsto u^{-1}(\mathbb{E}(u(Y)|\mathcal{F}_t))$$

for some (dis)utility function u . These functionals are translation equivariant only if they are the entropic risk functional described in Example 5.13 or the expectation; they are positively homogeneous just in the case of the expectation and the max-risk functional. Kovacevic and Pflug [70] show that composed risk functionals are information monotone, if and only if again they are either the expectation of the max-risk functionals, see Appendix A.2.

These results suggest that the only interesting compositions of risk functionals are the expectation and the max-risk functional. Moreover both functionals have a useful, intuitive interpretation, whereas the composition of risk functionals often lacks an interpretation: it is not evident, e.g., how Eq. (5.7), the Average Value-at-Risk of the Average Value-at-Risk could be interpreted in an economically meaningful way.

This situation is somewhat unsatisfactory, because it is intuitive to employ a risk functional in the objective at each node separately in order to achieve the original goal, which is formulated in terms of a risk functional as well. Moreover, from computational perspective, a risk functional in the objective is desirable, as it is easy to implement and easy to handle recursively.

In contrast to the results mentioned it is possible to write down dynamic programming equations, even in the situation of a time inconsistent stochastic program. To this end we develop a decomposition theorem (Theorem 5.21 in the following section), which recomposes conditional risk measures to recover an initial risk functional. It is based on changing the measure (change of numéraire) instead of composing risk functionals, and it requires changing the risk aversion in an adaptive way over time.

5.3.2 Multistage Decomposition of Risk Functionals: The Decomposition Theorem

For every risk functional applied to a sequence (Y_t) of random variables adapted to a filtration, there is a sort of time-decomposition. This decomposition, however, depends on the problem and the random variables Y_t , the decomposition in particular

⁴Notice that ordinary version-independence means that the functional depends only on the distribution of the random variable. However, the composed functionals are all dependent only on the *nested* distribution and in this extended sense, they are version independent.

is not unique for a pair consisting of a risk functional and a predefined tree, or tree structure.

The decomposition of a risk functional with respect to incomplete information is accomplished by the conditional risk functionals \mathcal{R}_Z introduced in Definition 3.34. Theorem 5.21 recovers the initial risk functional from conditional dissections introduced in Proposition 5.19 by applying an appropriate change-of-measure through some density variable Z .

The statement of this Theorem, its proof and the subsequent discussion essentially follow our papers [95] and [96].

The following type of expected-conditional risk functional obtained from a basic functional \mathcal{R} through its conditional version \mathcal{R}_Z (see Definition 3.34) plays an important role in decomposition results.

Proposition 5.19. *For $Z \triangleleft \mathcal{F}_t$ with $Z \geq 0$ and $\mathbb{E}(Z) = 1$ the mapping*

$$Y \mapsto \mathbb{E}[Z \cdot \mathcal{R}_Z(Y | \mathcal{F}_t)] \quad (5.8)$$

is a risk functional.

Proof. The properties (M)—(H) of Definition 3.2 are immediate from Theorem 3.38. \square

Remark 5.20. In general, the risk functional (5.8) is not version independent, even if \mathcal{R} is version independent.

Theorem 5.21 (Decomposition of a Risk Functional). *Let $\mathcal{R}(\cdot) = \sup_{\sigma \in \mathcal{S}} \mathcal{R}_\sigma(\cdot)$ be a risk functional generated by the distortion functions $\sigma \in \mathcal{S}$ (cf. Definition 3.11) and $\mathcal{F}_t \subset \mathcal{F}$. Then it holds that*

$$\mathcal{R}(Y) = \sup \mathbb{E}[Z_t \cdot \mathcal{R}_{Z_t}(Y | \mathcal{F}_t)], \quad (5.9)$$

where the supremum is among all feasible random variables $Z_t \triangleleft \mathcal{F}_t$, i.e., Z_t is measurable with respect to \mathcal{F}_t and Z_t satisfies $Z_t \leq \sigma$ for some $\sigma \in \mathcal{S}$ (recall (3.16) for the relation $Z_t \leq \sigma$).

Moreover, let $\mathcal{F}_t \subset \mathcal{F}_\tau$. Then the conditional risk functional obeys the nested decomposition

$$\mathcal{R}(Y | \mathcal{F}_t) = \text{ess sup } \mathbb{E}\left[Z_\tau \cdot \mathcal{R}_{Z_\tau}(Y | \mathcal{F}_\tau) \mid \mathcal{F}_t\right], \quad (5.10)$$

the essential supremum is taken among all feasible dual random variables $Z_\tau \triangleleft \mathcal{F}_\tau$.

Suppose further that Y is independent from \mathcal{F}_t .⁵ Then the infimum is attained at $Z_t = \mathbb{1}$, and the optimal conditional risk functionals are the original risk functional,

$$\mathcal{R}_\mathbb{1}(\cdot | \mathcal{F}_t) \equiv \mathcal{R}(\cdot).$$

⁵Cf. Williams [142] for independence of two sigma algebras.

Remark 5.22. The decomposition (5.9) outlines how a decision maker has to change her or his preference after each stage in order to achieve the initial goal. This situation can be compared to a risk manager, who relaxes the risk constraints after having observed indicators, which make the success of the entire project more likely. And conversely, the risk manager will undertake additional efforts after having observed partial outcomes, which make the future, undesired outcomes less likely.

For the Average Value-at-Risk the statement can be given by involving the conditional Average Value-at-Risk at random risk level as follows.

Corollary 5.23 (Decomposition of the Average Value-at-Risk). *For every $\alpha \in [0, 1]$ the Average Value-at-Risk has the decomposition*

$$\text{AV@R}_\alpha(Y) = \sup \mathbb{E} [Z_t \cdot \text{AV@R}_{1-(1-\alpha)Z_t}(Y | \mathcal{F}_t)], \quad (5.11)$$

where the supremum is among all random variables $Z_t \triangleleft \mathcal{F}_t$ satisfying $Z_t \geq 0$, $(1-\alpha)Z_t \leq 1$ and $\mathbb{E} Z_t = 1$. The supremum is attained, if $\alpha < 1$.

Moreover, if Y is independent from \mathcal{F}_t , then the supremum in (5.11) is attained at $Z_t = 1$ for the constant level $\alpha = \alpha \cdot 1$, and

$$\text{AV@R}_\alpha(Y | \mathcal{F}_t) = \text{AV@R}_\alpha(Y) \quad (5.12)$$

(that is, $\text{AV@R}_\alpha(Y | \mathcal{F}_t)$ is constant with outcome $\text{AV@R}_\alpha(Y)$).

Remark 5.24. For $\alpha = 0$, only the random variable $Z_t = 1$ satisfies the conditions $Z_t \leq 1$ and $\mathbb{E} Z_t = 1$, such that the corollary asserts the well-known identity $\mathbb{E} Y = \mathbb{E}[\mathbb{E}(Y | \mathcal{F}_t)]$.

Remark 5.25. Equation (5.12) can be interpreted by saying that the increments, after having observed \mathcal{F}_t , are independent from this observation, such that there is no difference when measuring the risk directly, or subject to the observation \mathcal{F}_t .

In order to prove Theorem 5.21 we start with proving the corollary for the Average Value-at-Risk first, and extend the assertion to general risk functionals in a second step.

Proof of Corollary 5.23. For $\alpha = 1$ the assertion is

$$\text{ess sup } Y = \sup \mathbb{E} [Z_t \cdot \text{ess sup } (Y | \mathcal{F}_t)],$$

where $Z_t \triangleleft \mathcal{F}_t$ and $\mathbb{E} Z_t = 1$, which is true by duality of $L^1(\mathcal{F}_t)$ and $L^\infty(\mathcal{F}_t)$.

We shall assume thus $\alpha < 1$. Let $Z_t \triangleleft \mathcal{F}_t$ be feasible and fixed, satisfying $Z_t \geq 0$ and $(1-\alpha)Z_t \leq 1$ and $\mathbb{E} Z_t = 1$. In view of the characterization (Proposition 3.37) consider a finite tessellation of the entire space with sets $B_k \in \mathcal{F}_t$, that is $B_k \cap B_j = \emptyset$ whenever $k \neq j$ and $\sum_{k=1}^K \mathbf{1}_{B_k} = 1$. Moreover let Z'_k be feasible for the conditional $\text{AV@R}_{1-(1-\alpha)Z_t}(Y | \mathcal{F}_t)$, that is $Z'_k \geq 0$, $\mathbb{E}(Z'_k | \mathcal{F}_t) = 1$ and $(1-\alpha)Z_t Z'_k \leq 1$.

The combined random variable $Z' := \sum_{k=1}^K \mathbb{1}_{B_k} Z'_k$ satisfies $(1 - \alpha) Z_t Z' \leq \mathbb{1}$ and $Z_t Z' \geq 0$, and it holds that

$$\mathbb{E}(Z' | \mathcal{F}_t) = \mathbb{E}\left(\sum_{k=1}^K \mathbb{1}_{B_k} Z'_k \mid \mathcal{F}_t\right) = \sum_{k=1}^K \mathbb{1}_{B_k} \mathbb{E}(Z'_k | \mathcal{F}_t) = \sum_{k=1}^K \mathbb{1}_{B_k} = \mathbb{1},$$

and further

$$\mathbb{E} Z_t Z' = \mathbb{E} Z_t \mathbb{E}(Z' | \mathcal{F}_t) = \mathbb{E} Z_t = 1.$$

The random variable $Z_t Z'$ is thus feasible for the Average Value-at-Risk (cf. (3.4)) and it follows that

$$\text{AV@R}_\alpha(Y) \geq \mathbb{E} Y Z_t Z' = \mathbb{E} Z_t \cdot \mathbb{E}(Y Z' | \mathcal{F}_t).$$

Finally observe that the tessellation is chosen to represent the essential supremum necessary in the definition of the conditional Average Value-at-Risk at level $1 - (1 - \alpha) Z_t$, (3.34). Thus

$$\text{AV@R}_\alpha(Y) \geq \mathbb{E} Z_t \cdot \text{AV@R}_{1-(1-\alpha)Z_t}(Y | \mathcal{F}_t),$$

establishing an initial relation.

For the converse relation consider the random variable Z^* maximizing (3.4), that is, $Z^* \geq 0$, $(1 - \alpha) Z^* \leq \mathbb{1}$ and $\mathbb{E} Z^* = 1$, and Z^* moreover satisfies $\text{AV@R}_\alpha(Y) = \mathbb{E} Y Z^*$. The conditional random variable

$$Z_t^* := \mathbb{E}(Z^* | \mathcal{F}_t) \quad (5.13)$$

is feasible as well for the Average Value-at-Risk (cf. (3.15)), as

$$\text{AV@R}_{\alpha'}(Z_t^*) = \text{AV@R}_{\alpha'}(\mathbb{E}(Z^* | \mathcal{F}_t)) \leq \text{AV@R}_{\alpha'}(Z^*) \leq \frac{1}{1 - \alpha'} \int_{\alpha'}^1 \mathbb{1}_{[\alpha,1]}(u) du$$

by (3.33).

The random variable $Z' := \frac{Z^*}{Z_t^*}$ is nonnegative as well ($Z' \geq 0$), and moreover

$$(1 - (1 - (1 - \alpha) Z_t^*)) Z' = (1 - \alpha) Z_t^* Z' = (1 - \alpha) Z^* \leq \mathbb{1},$$

such that Z' is feasible for the *conditional* Average Value-at-Risk, (3.34).

Hence,

$$\begin{aligned} \text{AV@R}_\alpha(Y) &= \mathbb{E} Y Z^* = \mathbb{E} Y Z_t^* Z' = \mathbb{E} Z_t^* \mathbb{E}(Y Z' | \mathcal{F}_t) \\ &\leq \mathbb{E} Z_t^* \text{AV@R}_{1-(1-\alpha)Z_t^*}(Y | \mathcal{F}_t), \end{aligned}$$

which is the remaining relation.

In case that Y is independent from \mathcal{F}_t , then one may assume that $Y = F_Y^{-1}(U)$ for some uniformly distributed random variable U independent from \mathcal{F}_t . Recall now that the supremum for the Average Value-at-Risk is attained at $Z^* = \sigma_\alpha(U)$ (i.e., $\text{AV@R}_\alpha(Y) = \mathbb{E} Y Z^*$), and Z^* thus is independent from \mathcal{F}_t as well, that is, $\mathbb{E}(Z^* | \mathcal{F}_t) = \mathbb{E} Z^* = 1$. It follows that Z^* is feasible for Eq. (3.34), such that

$$\text{AV@R}_\alpha(Y | \mathcal{F}_t) \geq \mathbb{E}(YZ^* | \mathcal{F}_t) = \mathbb{E}(YZ^*) = \text{AV@R}_\alpha(Y),$$

because $Y \cdot Z^*$ is independent from \mathcal{F}_t as well. The converse inequality is obvious, as the requirement $\mathbb{E}(Z | \mathcal{F}_t) = 1$ is stronger than $\mathbb{E}(Z) = 1$. \square

The proof of the general decomposition theorem (Theorem 5.21) is derived from the decomposition of the Average Value-at-Risk, Corollary 5.23.

Proof of Theorem 5.21. Let Z_t be fixed, and choose Z' such that $Z_t Z'$ is feasible, that is $\mathbb{E}(Z' | \mathcal{F}_t) = 1$. In accordance to Corollary 3.21 one may assume that there is a distortion $\sigma \in \mathcal{S}$ such that $\text{AV@R}_\alpha(Z_t Z') = \frac{1}{1-\alpha} \int_\alpha^1 \sigma(u) du$ and further, that there is a uniformly distributed random variable U , coupled in a co-monotone way with $Z_t Z'$, such that $Z_t Z' = \sigma(U)$.

Define now $Z_\alpha := \mathbb{E}(\sigma_\alpha(U) | \mathcal{F}_t)$ and $Z'_\alpha := \frac{\sigma_\alpha(U)}{Z_\alpha}$, where σ_α is as in (3.10). It holds that $(1 - \alpha) Z_\alpha Z'_\alpha \leq 1$ and $\mathbb{E}(Z' | \mathcal{F}_t) = 1$. By (5.11) it follows that

$$\text{AV@R}_\alpha(Y) \geq \mathbb{E} Z_\alpha \text{AV@R}_{1-(1-\alpha)Z_\alpha}(Y | \mathcal{F}_t)$$

for every α , and by integrating with respect to the distribution function μ_σ introduced by (3.12),

$$\begin{aligned} \mathcal{R}(Y) &\geq \int_{0^-}^1 \mathbb{E} Z_\alpha \text{AV@R}_{1-(1-\alpha)Z_\alpha}(Y | \mathcal{F}_t) d\mu_\sigma(\alpha) \\ &= \mathbb{E} \int_{0^-}^1 Z_\alpha \text{AV@R}_{1-(1-\alpha)Z_\alpha}(Y | \mathcal{F}_t) d\mu_\sigma(\alpha). \end{aligned} \quad (5.14)$$

By the reverse Fatou lemma

$$\begin{aligned} \mathcal{R}(Y) &\geq \mathbb{E} \int_{0^-}^1 Z_\alpha \text{ess sup} \left\{ \mathbb{E}(YZ'_\alpha | \mathcal{F}_t) : 0 \leq (1 - \alpha) Z_\alpha Z'_\alpha \leq 1, \mathbb{E}(Z'_\alpha | \mathcal{F}_t) \leq 1 \right\} \\ &\quad d\mu_\sigma(\alpha) \\ &= \mathbb{E} \int_{0^-}^1 \text{ess sup} \left\{ \mathbb{E}(YZ_\alpha Z'_\alpha | \mathcal{F}_t) : 0 \leq (1 - \alpha) Z_\alpha Z'_\alpha \leq 1, \mathbb{E}(Z'_\alpha | \mathcal{F}_t) \leq 1 \right\} \\ &\quad d\mu_\sigma(\alpha) \\ &\geq \mathbb{E} \text{ess sup} \left\{ \int_{0^-}^1 \mathbb{E}(YZ_\alpha Z'_\alpha | \mathcal{F}_t) : 0 \leq (1 - \alpha) Z_\alpha Z'_\alpha \leq 1, \mathbb{E}(Z'_\alpha | \mathcal{F}_t) \leq 1 \right\} \\ &\quad d\mu_\sigma(\alpha) \end{aligned}$$

$$= \mathbb{E} \operatorname{ess\,sup} \left\{ \mathbb{E} \left(Y \int_{0^-}^1 Z_\alpha Z'_\alpha d\mu_\sigma(\alpha) \middle| \mathcal{F}_t \right) : \right. \\ \left. 0 \leq (1 - \alpha) Z_\alpha Z'_\alpha \leq 1, \mathbb{E}(Z'_\alpha | \mathcal{F}_t) \leq 1 \right\}.$$

Note that

$$\int_{0^-}^1 \sigma_\alpha(u) d\mu_\sigma(\alpha) = \int_{0^-}^u \frac{1}{1 - \alpha} d\mu_\sigma(\alpha) = \sigma(u) \quad (5.15)$$

by (3.13), and thus

$$\int_0^1 Z_\alpha Z'_\alpha d\mu_\sigma(\alpha) = \int_0^1 \sigma_\alpha(U) d\mu_\sigma(\alpha) = \sigma(U) = Z_t Z'$$

by (5.15). Moreover

$$\int_0^1 Z_\alpha d\mu_\sigma(\alpha) = \int_0^1 \mathbb{E}(\sigma_\alpha(U) | \mathcal{F}_t) d\mu_\sigma(\alpha) = \mathbb{E} \left(\int_0^1 \sigma_\alpha(U) d\mu_\sigma(\alpha) \middle| \mathcal{F}_t \right) \\ = \mathbb{E}(\sigma(U) | \mathcal{F}_t) = \mathbb{E}(Z_t Z' | \mathcal{F}_t) = Z_t. \quad (5.16)$$

It thus holds that

$$\begin{aligned} \mathcal{R}(Y) &\geq \mathbb{E} \operatorname{ess\,sup} \{ \mathbb{E}(YZ_t Z' | \mathcal{F}_t) : \mathbb{E}(Z' | \mathcal{F}_t) \leq 1, Z_t Z' \leq \sigma \} \\ &= \mathbb{E} Z_t \operatorname{ess\,sup} \{ \mathbb{E}(YZ' | \mathcal{F}_t) : \mathbb{E}(Z' | \mathcal{F}_t) \leq 1, Z_t Z' \leq \sigma \} \end{aligned}$$

and finally

$$\mathcal{R}(Y) \geq \mathbb{E} Z_t \mathcal{R}_{Z_t}(Y | \mathcal{F}_t).$$

For the converse inequality choose a distortion density $\sigma \in \mathcal{S}$ and a uniform random variable U such that

$$\mathcal{R}(Y) - \varepsilon < \mathcal{R}_\sigma(Y) = \mathbb{E} Y\sigma(U)$$

for $\varepsilon > 0$. Define $Z := \sigma(U)$, $Z_t := \mathbb{E}(Z | \mathcal{F}_t)$, and note that Z_t is feasible for \mathcal{R} because

$$\text{AV@R}_\alpha(Z_t) \leq \text{AV@R}_\alpha(Z) \leq \frac{1}{1 - \alpha} \int_\alpha^1 \sigma(u) du$$

by (3.33). Moreover

$$Z' := \frac{Z}{Z_t} \quad (5.17)$$

is feasible for \mathcal{R}_{Z_t} . It follows that

$$\mathcal{R}(Y) - \varepsilon < \mathbb{E} Y Z = \mathbb{E} Z_t \mathbb{E}(YZ' | \mathcal{F}_t) \leq \mathbb{E} Z_t \mathcal{R}_{Z_t}(Y | \mathcal{F}_t).$$

As $\varepsilon > 0$ was chosen arbitrarily the assertion follows.

The above proof applies for the nested decomposition (Eq. (5.10)) as well, by considering the statement conditionally on \mathcal{F}_t and by replacing \mathcal{F}_t by the intermediate sigma algebra \mathcal{F}_τ . \square

5.3.2.1 Examples of Decompositions

The following two examples address the risk functionals $\mathcal{R} = \text{AV@R}$ and the combination $\mathcal{R}(\cdot) = (1 - \mu)\mathbb{E}(\cdot) + \mu \cdot \text{AV@R}_\alpha(\cdot)$ of two simple risk functionals.

The first example outlines the *random level* of the conditional Average Value-at-Risk, which deviates from the initial level. The second example moreover demonstrates that it is not enough to adapt the level of the Average Value-at-Risk, it is necessary to change the weights of the respective AV@R measures as well. That is, the Kusuoka representation—conditionally—has different levels *and* weights. Both examples show as well that the conditional risk functionals is not independent of Y .

Example 5.26 (Decomposition of the Average Value-at-Risk). Figure 5.7a addresses the Average Value-at-risk at level $\alpha = 50\%$ for the random variable Y depicted.

The optimal dual variable Z^* is provided by (3.4), and $Z_t = \mathbb{E}(Z^* | \mathcal{F}_t)$ evaluates to $Z_t(\text{up}) = \frac{22}{15}$ and $Z_t(\text{down}) = \frac{4}{5}$ (cf. (5.13)). The optimal random level $\alpha_t = 1 - (1 - \alpha)Z_t$ deviates from the initial $\alpha = 50\%$, the random levels are $\alpha(\cdot | \text{up}) = \frac{4}{15}$ and $\alpha(\cdot | \text{down}) = \frac{3}{5}$ in this example. The individual evaluations of the conditional risk functional at random level are $\text{AV@R}_{\frac{4}{15}}(Y | \text{up}) = 193$ and $\text{AV@R}_{\frac{3}{5}}(Y | \text{down}) = 143$. Finally it holds that $\mathbb{E} Z_t \cdot \text{AV@R}_{1-(1-\alpha)Z_t}(Y | \mathcal{F}_t) = 165$, which is in line with $\text{AV@R}_{50\%}(Y) = 165$.

Remark 5.27 (Conditional Kusuoka Representation). The essential observation in the previous proof is that

$$\mathcal{R}_{Z_t}(Y | \mathcal{F}_t) = \frac{\int_0^1 Z_\alpha \text{AV@R}_{1-(1-\alpha)Z_\alpha}(Y | \mathcal{F}_t) \mu_\sigma(d\alpha)}{\int_0^1 Z_\alpha \mu_\sigma(d\alpha)} \quad (5.18)$$

by the Eqs. (5.14) and (5.17), where

$$Z_t = \int_0^1 Z_\alpha \mu_\sigma(d\alpha) \quad (5.19)$$

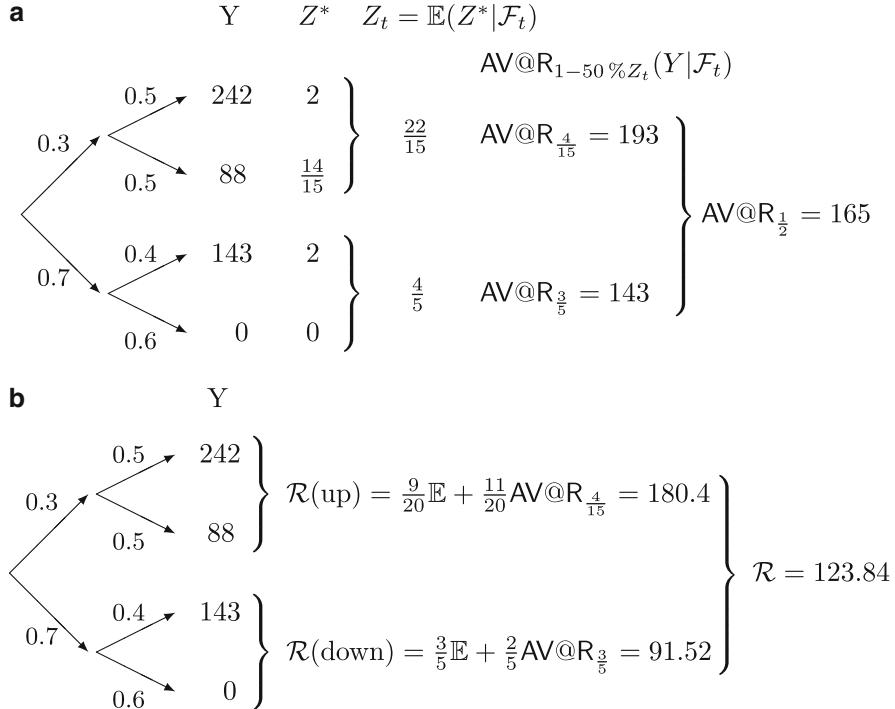


Fig. 5.7 Decomposition of risk functionals

according to (5.16). The family $(Z_\alpha)_{\alpha \in [0,1]}$ of random variables is given by $Z_\alpha := \mathbb{E}(Z_\alpha^* | \mathcal{F}_t)$, where Z_α^* is the optimal dual variable for the Average Value-at-Risk at level α (that is, $\mathbb{E}YZ_\alpha^* = \mathbb{AV} @ R_\alpha(Y)$ and $0 \leq Z_\alpha^* \leq \frac{1}{1-\alpha}$). The identity (5.18), which provides the conditional risk functionals in an explicit way, can be considered as *conditional Kusuoka representation*. The random variables in the nominator and denominator are both Bochner integrals. Note that the family $(Z_\alpha)_{\alpha \in [0,1]}$ depends on Y , consequently Z_t (Eq. (5.19)) depends on Y and the optimal decomposition (5.18) in particular depends on Y .

Example 5.28 (Continuation of Example 5.26). Consider the risk functional $\mathcal{R}(\cdot) = \frac{6}{11}\mathbb{E}(\cdot) + \frac{5}{11}\mathbb{AV} @ R_{50\%}(\cdot)$ (depicted in Fig. 5.7a) on the same tree as the previous example. As the dual variable for the expected value is 1 , (5.18) determines the conditional risk functionals as

$$\begin{aligned}
 \mathcal{R}_{\frac{40}{33}}(\cdot | \text{up}) &= \frac{\frac{6}{11} \cdot 1 \cdot \mathbb{E}(\cdot) + \frac{5}{11} \cdot \frac{22}{15} \cdot \mathbb{AV} @ R_{\frac{4}{15}}(\cdot | \text{up})}{\frac{6}{11} \cdot 1 + \frac{5}{11} \cdot \frac{22}{15}} \\
 &= \frac{9}{20}\mathbb{E}(\cdot | \text{up}) + \frac{11}{20}\mathbb{AV} @ R_{\frac{4}{15}}(\cdot | \text{up})
 \end{aligned}$$

and

$$\begin{aligned}\mathcal{R}_{\frac{10}{11}}(\cdot \mid \text{down}) &= \frac{\frac{6}{11} \cdot 1 \cdot \mathbb{E}(\cdot) + \frac{5}{11} \cdot \frac{4}{5} \cdot \mathbf{AV@R}_{\frac{3}{5}}(\cdot \mid \text{down})}{\frac{6}{11} \cdot 1 + \frac{5}{11} \cdot \frac{4}{5}} \\ &= \frac{3}{5} \mathbb{E}(\cdot \mid \text{down}) + \frac{2}{5} \mathbf{AV@R}_{\frac{3}{5}}(\cdot \mid \text{down}).\end{aligned}$$

Evaluated for the random variable Y depicted the results are $\mathcal{R}_{\frac{40}{33}}(Y \mid \text{up}) = \frac{9}{20} \cdot 165 + \frac{11}{20} \cdot 193 = 180.4$ and $\mathcal{R}_{\frac{10}{11}}(Y \mid \text{down}) = \frac{3}{5} \cdot \frac{286}{5} + \frac{2}{5} \cdot 143 = 91.52$. The decomposition theorem states that $\mathcal{R}(Y) = \mathbb{E} Z_t \cdot \mathcal{R}_{Z_t}(Y \mid \mathcal{F}_t) = \frac{3}{10} \cdot \frac{40}{33} \cdot 180.4 + \frac{7}{10} \cdot \frac{10}{11} \cdot 91.52 = 123.84$. The direct computation $\mathcal{R}(Y) = \frac{6}{11} \cdot 89.54 + \frac{5}{11} \cdot 165 = 123.84$ provides the same result.

5.4 Martingale Formulations of Time Inconsistent Stochastic Programs⁶

The genuine problem of multistage stochastic optimization, as it was addressed in the introduction, is the problem

$$\begin{aligned}&\text{minimize } \mathcal{R}[Q(x, \xi)] \\ &\text{subject to } x \in \mathbb{X}, \\ &\quad x \triangleleft \mathfrak{F},\end{aligned}\tag{5.20}$$

where the measurability constraint $x \triangleleft \mathfrak{F}$ (i.e., $x_t \triangleleft \mathcal{F}_t$ for all $t \in \{0, 1, \dots, T\}$) ensures that the decision x_t at stage t is based on information available up to time t . As the previous section presents a decomposition of the positively homogeneous risk functional \mathcal{R} , it is desirable to decompose the entire problem (5.20) in subsequent instances, and to solve the subproblems recursively.

This attempt is in line with the *dynamic programming principle*, which was developed in the 1950s to solve optimal control problems with a deterministic (i.e., nonrandom) objective function (cf. Bellman [9]). These solution techniques have been developed further to solve optimal control problems, where the objective is an expectation and where decisions depend on the current status of the system, not the entire history (cf. Fleming and Soner [44] for Markov-type problems).

The problem considered here, (5.20), is more general, as the risk functional \mathcal{R} in the objective is not of this simple type of an expectation, and the decision x_t is based on the entire history $(\xi_0, \xi_1, \dots, \xi_t)$, not just on ξ_t . Moreover we have seen

⁶This section adapts [96].

already that \mathcal{R} —in general—is not time consistent, which considerably complicates solution techniques.

We generalize the techniques available via dynamic programming to solve the problem (5.20). That is, we introduce a value function first, and characterize the optimality of a solution by respective verification theorems. It turns out that the optimal solution, together with the optimal dual solution form a martingale process, and the verification theorems involve sub- and supermartingales. Interestingly, martingales have been considered very early by Rockafellar and Wets [112, 113], although in a context which is slightly different.

To keep the presentation simple we elaborate the results for the risk functional $\mathcal{R}(\cdot) = \mathbb{E}(\cdot) + \gamma \cdot \text{AV@R}_\alpha(\cdot)$ for two fixed parameters $\gamma \in [0, 1]$ and $\alpha \in (0, 1)$ instead of a general risk functional; more precisely, we consider the problem

$$\begin{aligned} & \text{minimize } \mathbb{E}(Q(x, \xi)) + \gamma \cdot \text{AV@R}_\alpha(Q(x, \xi)) \\ & \text{subject to } x \in \mathbb{X}, \\ & \quad x \triangleleft \mathfrak{F}. \end{aligned} \tag{5.21}$$

It is evident how the results have to be restated for general distortion risk functionals \mathcal{R} . We shall assume throughout the following presentation that the minimum in (5.21) is attained.

Definition 5.29 (The Value Function). The value function of the problem (5.21) is

$$v_t(x_{0:t-1}, \alpha, \mu) := \underset{(x_{0:t-1}, x_{t:T}) \in \mathbb{X}}{\text{ess inf}} \mathbb{E}[Q(x_{0:T}) | \mathcal{F}_t] + \gamma \cdot \text{AV@R}_\alpha(Q(x_{0:T}) | \mathcal{F}_t), \tag{5.22}$$

where $x_{0:t-1}$ is the history of decisions, and $(x_{0:t-1}, x_{t:T})$ is the concatenated vector of previous and future decision. $\alpha \triangleleft \mathcal{F}_t$ and $\mu \triangleleft \mathcal{F}_t$ are \mathcal{F}_t -measurable, \mathbb{R} -valued random variables.

Remark 5.30. It should be noted that the essential infimum in (5.22) is among possible *future* decision, $x_{t:T}$, such that the concatenated decisions $(x_{0:t-1}, x_{t:T})$ are feasible, i.e., $(x_{0:t-1}, x_{t:T}) \in \mathbb{X}$. This is occasionally written as $x_{t:T} \in \mathbb{X}(x_{0:t-1})$, where $\mathbb{X}(x_{0:t-1}) = \{x_{t:T} : (x_{0:t-1}, x_{t:T}) \in \mathbb{X}\}$. In the frequent case that $\mathbb{X} = \mathbb{X}_0 \times \mathbb{X}_1 \times \dots \mathbb{X}_T$ it thus holds that $\mathbb{X}(x_{0:t-1}) = \mathbb{X}_t \times \dots \mathbb{X}_T$.

Relation of the Value Function and the Stochastic Multistage Problem (5.21). Recall that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ is the trivial sigma algebra. At time $t = 0$ there is no history, such that the value function evaluates to

$$\begin{aligned} v_0((), \alpha, \mu) &= \underset{(x_{0:T}) \in \mathbb{X}}{\text{ess inf}} \mathbb{E}[Q(x_{0:T}) | \mathcal{F}_0] + \gamma \cdot \text{AV@R}_\alpha(Q(x_{0:T}) | \mathcal{F}_0) \\ &= \sup_{x \in \mathbb{X}} \mathbb{E} Q(x) + \gamma \cdot \text{AV@R}_\alpha(Q(x)). \end{aligned}$$

This is the objective value of the initial problem (5.21), such that the value function v_0 indeed provides the objective of the initial problem.

Theorem 5.31 (Dynamic Programming Principle). *Assume that Q is lower semicontinuous with respect to x , and the values of ξ lie in some convex, compact subset of \mathbb{R}^n .*

(i) *The value function at terminal time T is given by deterministic optimization as*

$$v_T(x_{0:T-1}, \alpha, \gamma) = (1 + \gamma) \cdot \underset{x_T \in \mathbb{X}(x_{0:T-1})}{\text{ess inf}} Q(x_{0:T}).$$

(ii) *For any $t < \tau$, ($t, \tau \in \{0, 1, \dots, T\}$) the recursive relation*

$$v_t(x_{0:t-1}, \alpha, \gamma) = \underset{x_{t:\tau-1}}{\text{ess inf}} \underset{Z_{t:\tau}}{\text{ess sup}} \mathbb{E}[v_\tau(x_{0:\tau-1}, 1 - (1 - \alpha)Z_{t:\tau}, \gamma \cdot Z_{t:\tau}) | \mathcal{F}_t], \quad (5.23)$$

holds true, where $Z_{t:\tau} \triangleleft \mathcal{F}_\tau$, $0 \leq Z_{t:\tau}$, $(1 - \alpha)Z_{t:\tau} \leq 1$ and $\mathbb{E}(Z_{t:\tau} | \mathcal{F}_t) = 1$.

Proof. By definition,

$$\begin{aligned} v_T(x_{0:T-1}, \alpha, \mu) &= \underset{(x_{0:T-1}, x_{T:T}) \in \mathbb{X}}{\text{ess inf}} \mathbb{E}[Q(x_{0:T}) | \mathcal{F}_T] + \gamma \cdot \text{AV@R}_\alpha(Q(x_{0:T}) | \mathcal{F}_T) \\ &= \underset{(x_{0:T-1}, x_{T:T}) \in \mathbb{X}}{\text{ess inf}} Q(x_{0:T}) + \gamma \cdot Q(x_{0:T}) \\ &= \underset{x_T \in \mathbb{X}(x_{0:T-1})}{\text{ess inf}} (1 + \gamma) \cdot Q(x_{0:T}), \end{aligned}$$

because the random variable Q is constant conditionally on \mathcal{F}_T , and the essential infimum is deterministic and thus can be replaced by an infimum.

For an earlier time $t < T$ the nested decomposition of the conditional Average Value-at-Risk (Corollary 5.23 and (5.10)) implies that

$$\begin{aligned} v_t(x_{0:t-1}, \alpha, \gamma) &= \underset{x_{t:T}}{\text{ess inf}} \mathbb{E}[Q(x_{0:T}) | \mathcal{F}_t] + \gamma \cdot \text{AV@R}_\alpha(Q(x_{0:T}) | \mathcal{F}_t) \\ &= \underset{x_{t:T}}{\text{ess inf}} \underset{Z_{t:t+1}}{\text{ess sup}} \mathbb{E} \left[\begin{array}{l} \mathbb{E}[Q(x_{0:T}) | \mathcal{F}_{t+1}] \\ + \gamma \cdot Z_{t:t+1} \cdot \text{AV@R}_{1-(1-\alpha)Z_{t:t+1}}(Q(x_{0:T}) | \mathcal{F}_{t+1}) \end{array} \middle| \mathcal{F}_t \right], \end{aligned} \quad (5.24)$$

where the random variable $Z_{t:t+1}$ is measurable, $Z_{t:t+1} \triangleleft \mathcal{F}_t$, and satisfies $\mathbb{E}(Z_{t:t+1} | \mathcal{F}_t) \equiv 1$ with $(1 - \alpha)Z_{t:t+1} \leq 1$.

Recall now that the mapping

$$(Y, Z) \mapsto \text{AV@R}_{1-(1-\alpha)Z}(Y)$$

is convex in Y and concave in Z (Theorem 3.38 (iii) and (vi)), where Z is chosen from the $\sigma(L^\infty, L^1)$ compact set $\{Z \in L^\infty : 0 \leq Z \leq \frac{1}{1-\alpha}\}$. By Sion's minimax theorem (cf. Sion [132] and Komiya [69]) the supremum and infimum in (5.24) thus may be interchanged, such that

$$v_t(x_{0:t-1}, \alpha, \gamma)$$

$$= \operatorname{ess\,inf}_{x_t} \operatorname{ess\,sup}_{Z_{t:t+1}} \operatorname{ess\,inf}_{x_{t+1:T}} \mathbb{E} \left[\begin{array}{l} \mathbb{E}[Q(x_{0:T}) | \mathcal{F}_{t+1}] \\ + \gamma \cdot Z_{t:t+1} \cdot \text{AV@R}_{1-(1-\alpha)Z_{t:t+1}}(Q(x_{0:T}) | \mathcal{F}_{t+1}) \end{array} \middle| \mathcal{F}_t \right],$$

where we have separated first the following decision x_t from the further decision $x_{t+1:T}$.

The function Q is lower semicontinuous by assumption, thus the interchangeability principle (cf. Rockafellar and Wets [114, Theorem 14.60], or Shapiro et al. [129, p. 405]) applies and

$$v_t(x_{0:t-1}, \alpha, \gamma)$$

$$\begin{aligned} &= \operatorname{ess\,inf}_{x_t} \operatorname{ess\,sup}_{Z_{t:t+1}} \mathbb{E} \left[\begin{array}{l} \operatorname{ess\,inf}_{x_{t+1:T}} \mathbb{E}[Q(x_{0:T}) | \mathcal{F}_{t+1}] \\ + \gamma \cdot Z_{t:t+1} \cdot \text{AV@R}_{1-(1-\alpha)Z_{t:t+1}}(Q(x_{0:T}) | \mathcal{F}_{t+1}) \end{array} \middle| \mathcal{F}_t \right] \\ &= \operatorname{ess\,inf}_{x_t} \operatorname{ess\,sup}_{Z_{t:t+1}} \mathbb{E}[v_{t+1}(x_{0:t}, 1 - (1 - \alpha)Z_{t:t+1}, \gamma \cdot Z_{t:t+1}) | \mathcal{F}_t]. \end{aligned}$$

This is the assertion for $\tau = t + 1$. The above computation can be repeated at subsequent stages of time, which finally reveals the general result. \square

5.4.1 Verification Theorems

Verification theorems are used in dynamic optimization to verify the optimality of a solution. Although the multistage stochastic optimization problem (5.20) is time inconsistent, the concept of verification theorems can be maintained in a generalized form.

To elaborate the verifications for the sequence $x_{0:T}$ of optimal decisions and the optimal dual variable Z^* we recall that it is assumed that the infimum in (5.21) is attained at some $x_{0:T}^*$. By Corollary 5.23, the dual variable $Z_{t:t+1}^*$ is attained as well in (5.23). Every time step t thus reveals a density $Z_{t:t+1}^*$, which satisfies $\mathbb{E}(Z_{t:t+1}^* | \mathcal{F}_t) = 1$. We consider the products

$$Z_{\tau:t}^* := Z_{\tau:\tau+1}^* \cdot Z_{\tau+1:\tau+2}^* \cdot \dots \cdot Z_{t-1:t}^* \quad (\text{and } Z_{t:t}^* := 1)$$

and in particular the process $Z_t^* := Z_{0:t}^*$ ($t \in \{0, 1, \dots, T\}$). This process is a martingale, as

$$Z_{t-1:t} \triangleleft \mathcal{F}_t \text{ and } Z_{\tau:t} \triangleleft \mathcal{F}_t,$$

such that $\mathbb{E}(Z_t^* | \mathcal{F}_t) = \mathbb{E}(Z_t^* | \mathcal{F}_t) = Z_t^*$ ($\tau \geq t$).

The previous dynamic programming principle (Theorem 5.31) suggest to consider the process

$$m_t(x, Z) := v_t(x_{0:t-1}, 1 - (1 - \alpha) \cdot Z_t, \gamma \cdot Z_t) \quad (t \in \{0, 1 \dots T\}).$$

Theorem 5.32 (Martingale Property of Optimal Solutions). *If the processes x^* and Z^* are optimal, then the process $m_t(x_{0:t}^*, Z_t^*)$ is a martingale with respect to the filtration \mathcal{F}_t . Conversely, if $m_t(x_{0:t}, Z_t)$ is a martingale for the pair of processes $(x_{0:t}, Z_t)$, then the processes $x_{0:t}$ and Z_t are optimal.*

Proof. We employ the dynamic programming principle (Theorem 5.31). Hence,

$$\begin{aligned} m_t(x_{0:t}^*, Z_t^*) &= v_t(x_{0:t-1}^*, 1 - (1 - \alpha) \cdot Z_t^*, \gamma \cdot Z_t^*) \\ &= \text{ess inf}_{x_{t+1}} \text{ess sup}_{Z_{t+1}} \mathbb{E}[v_{t+1}((x_{0:t-1}^*, x_t), 1 - (1 - \alpha) \cdot Z_t^* \cdot Z_{t+1}, \gamma \cdot Z_{t+1} \cdot Z_{t+1}) | \mathcal{F}_t] \\ &= \text{ess inf}_{x_{t+1}} \mathbb{E}[v_{t+1}((x_{0:t-1}^*, x_t), 1 - (1 - \alpha) \cdot Z_{t+1}^*, \gamma \cdot Z_{t+1}^*) | \mathcal{F}_t] \\ &= \mathbb{E}[v_{t+1}(x_{0:t}^*, 1 - (1 - \alpha) \cdot Z_{t+1}^*, \gamma \cdot Z_{t+1}^*) | \mathcal{F}_t] \\ &= \mathbb{E}[m_{t+1}(x_{0:t+1}^*, Z_{t+1}^*) | \mathcal{F}_t], \end{aligned}$$

and $m_t(x_{0:t}^*, Z_t^*)$ is a martingale.

The converse follows from the following verification theorem. \square

In dynamic programming, optimal solutions are often characterized by verification theorems. In what follows we extend this concept of verification theorems to the more general, time inconsistent stochastic optimization problem, which has a risk functional in its objective and which involves the entire history.

Theorem 5.33 (Verification Theorem). *Let x be feasible, and let Z be a martingale satisfying feasible.*

(i) *Suppose that the process w_t satisfies*

$$\begin{aligned} w_T(x_{0:T-1}, 1 - (1 - \alpha)Z_T, \gamma Z_T) &\geq (1 + \gamma Z_T) Q(x_{0:T}) \quad \text{and} \\ w_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t) &\geq \text{ess sup}_{Z_{t+1}} \mathbb{E}[w_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t], \end{aligned}$$

then the process $w_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t)$ is a supermartingale dominating $v_t(x_{0:t-1}, \alpha Z_t, \gamma Z_t)$, $w \geq v$.

(ii) Suppose that the process u_t satisfies

$$\begin{aligned} u_T(x_{0:T-1}, 1 - (1 - \alpha)Z_T, \gamma Z_T) &\leq (1 + \gamma Z_T) Q(x_T) \quad \text{and} \\ u_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t) \\ &\leq \operatorname{ess\,inf}_{x_t} \mathbb{E}[u_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t], \end{aligned}$$

then the process $u_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t)$ is a submartingale dominated by $v_t(x_{0:t-1}, \alpha Z_t, \gamma Z_t)$, $u \leq v$.

Proof. Suppose the first assertions of (i) and (ii) hold true, then it follows readily from Theorem 5.31 (i) that

$$u_T \leq v_T \leq w_T.$$

The other stages are derived by backwards induction as follows.

$$\begin{aligned} w_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t) \\ \geq \operatorname{ess\,sup}_{Z_{t+1}} \mathbb{E}[w_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \\ \geq \operatorname{ess\,sup}_{Z_{t+1}} \mathbb{E}[v_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \\ \geq \operatorname{ess\,inf}_{x_t} \operatorname{ess\,sup}_{Z_{t+1}} \mathbb{E}[v_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \\ = v_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t), \end{aligned}$$

where we have employed the second assertion in (i), the induction hypothesis and finally the statement (ii) of the dynamic programming principle, Theorem 5.31. This ensures the $w \geq v$.

w_t is moreover a supermartingale, because

$$\begin{aligned} w_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t) \\ \geq \operatorname{ess\,sup}_{Z_{t+1}} \mathbb{E}[w_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \\ \geq \mathbb{E}[w_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t]. \end{aligned}$$

As for the second statement note that

$$\begin{aligned} u_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t) \\ \leq \operatorname{ess\,inf}_{x_t} \mathbb{E}[u_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \\ \leq \operatorname{ess\,inf}_{x_t} \mathbb{E}[v_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \end{aligned}$$

$$\begin{aligned}
&\leq \text{ess sup}_{Z_{t+1}} \text{ess inf}_{x_t} \mathbb{E} [v_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \\
&\leq \text{ess inf}_{x_t} \text{ess sup}_{Z_{t+1}} \mathbb{E} [v_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \\
&= v_t(x_{0:t-1}, 1 - (1 - \alpha) \cdot Z_t, \gamma \cdot Z_t),
\end{aligned}$$

where we have employed additionally the max–min inequality $\sup_z \inf_x L(x, z) \leq \inf_x \sup_Z L(x, z)$. u is finally a submartingale, as

$$\begin{aligned}
&u_t(x_{0:t-1}, 1 - (1 - \alpha)Z_t, \gamma Z_t) \\
&\leq \text{ess inf}_{x_t} \mathbb{E} [u_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t] \\
&\leq \mathbb{E} [u_{t+1}(x_{0:t}, 1 - (1 - \alpha) \cdot Z_{t+1}, \gamma \cdot Z_{t+1}) | \mathcal{F}_t].
\end{aligned}$$

□

5.4.2 An Algorithm for Sequential Improvement

Theorem 5.32 and the subsequent verification theorem allow to verify if a given policy x and dual Z are optimal to solve the multistage problem (5.21). If x is given, then the verification theorem allows to compute Z by looking up the respective argmax sets, and conversely it is possible to find the optimal x to given Z by computing the respective argmin sets.

Moreover, given a policy x and a dual density Z it is possible to improve on these choices by taking the respective maximizers from the verification theorem. Iterating the optimization procedures in an alternating way may not necessarily give a sequence converging to the saddle point, but an improvement can always be achieved (Cf. also Proposition B.6 in the Appendix).

The representation (3.3) is useful to solve the problem (5.21), as it rewrites as

$$\begin{aligned}
&\text{minimize}_{x,q} \mathbb{E} [Q(x) + \gamma q + \frac{\gamma}{1-\alpha} (Q(x) - q)_+] \\
&\text{subject to } x \triangleleft \mathfrak{F} \\
&\quad x \in \mathbb{X},
\end{aligned}$$

and this problem just involves minimization, which can often be implemented in a straight forward way.

The decomposition theorem can be applied to specify a local problem, typically leading to a considerable acceleration. For this recall that

$$\begin{aligned} & \mathbb{E} Q(x) + \gamma \text{AV@R}_\alpha(Q(x)) \\ &= \inf_{Z_t} \mathbb{E} [\mathbb{E}(Q(x)|\mathcal{F}_t) + \gamma Z_t \cdot \text{AV@R}_{1-(1-\alpha)Z_t}(Q(x)|\mathcal{F}_t)]; \end{aligned}$$

given the optimizing random variable Z_t this suggests to improve x locally, that is to choose

$$\begin{aligned} x_t &\in \operatorname{argmin}_{x_t} \mathbb{E}[Q(x)|\mathcal{F}_t] + \gamma Z_t \text{AV@R}_{1-(1-\alpha)Z_t}(Q(x)|\mathcal{F}_t) \\ &= \operatorname{argmin}_{q \triangleleft \mathcal{F}_t, x_t} \mathbb{E}(Q(x)|\mathcal{F}_t) + \gamma Z_t q - \frac{\gamma}{1-\alpha} \mathbb{E}((q - Q(x))_+ | \mathcal{F}_t) \end{aligned}$$

or to find at least a local improvement.

This strategy is indicated in Algorithm 5.1 and indeed gives an improvement in many situations:

Properties of an Implementation of Algorithm 5.1.

- For high dimensional problems Algorithm 5.1 is considerably faster in finding a suitable solution of the problem (cf. the exemplary comparison below).
- The quick implementation of the straightforward method needs a Hessian to perform quickly. This is memory demanding, to compute its inverse is time consuming. Algorithm 5.1, in contrast, can finish its job by applying a simple line search at each node. For this reason the algorithm can be significantly accelerated by selecting the successive nodes in a clever way, and by improving the accuracy goal in subsequent steps.
- Above all Algorithm 5.1 can be implemented in a *parallel* way, as optimizations can be done independently and simultaneously for neighboring nodes. This decreases the run-time by a factor given by the degree of parallelization.

5.4.3 Numerical Experiments

To illustrate Algorithm 5.1 we have chosen the flowergirl problem again, which was already addressed in Example 1.4 before (cf. Dupačová [37]). This multistage extension of the newsboy problem has the objective

$$\mathbb{E} Q(\xi, x(\xi)) + \gamma \text{AV@R}_\alpha(Q(\xi, x(\xi)))$$

for a suitable choice of α and γ . The problem considered thus is

$$\begin{aligned} & \text{minimize } \mathbb{E} Q(x) + \gamma \cdot \text{AV@R}_\alpha(Q(x)) \\ & \text{subject to } x \triangleleft \mathfrak{F}, \\ & \quad x \geq 0. \end{aligned} \tag{5.26}$$

Algorithm 5.1 Sequential improvement and verification of the strategy $x_{0:T}$ by exploiting the nested structure of the risk functional. A modification according to Proposition B.6 can improve convergence

STEP 0—INITIALIZATION

Let $x_{0:T}^0$ be any feasible, initial solution of the problem 5.21,

Set $k \leftarrow 0$.

STEP 1—SELECT A DUAL

Find Z^k , such that $0 \leq Z^k \leq \frac{1}{1-\alpha}$, $\mathbb{E} Z^k = 1$ and define

$$Z_t^k := \mathbb{E}(Z^k | \mathcal{F}_t). \quad (5.25)$$

A good initial choice is often Z^k satisfying

$$\mathbb{E} Z^k Q(x_{0:T}^k) = \text{AV@R}_\alpha(Q(x_{0:T}^k)).$$

STEP 2—CHECK FOR LOCAL IMPROVEMENT

Choose

$$\begin{aligned} x_t^{k+1} &\in \underset{x_t \triangleleft \mathcal{F}_t}{\operatorname{argmin}} \mathbb{E}[Q(x_{0:T}^k) | \mathcal{F}_t] + \gamma Z_t^k \text{AV@R}_{1-(1-\alpha)Z_t^k}(Q(x_{0:T}^k) | \mathcal{F}_t) \\ &= \underset{q, x_t \triangleleft \mathcal{F}_t}{\operatorname{argmin}} \mathbb{E}\left[Q(x_{0:T}^k) + \gamma Z_t^k q - \frac{\gamma}{1-\alpha} (q - Q(x_{0:T}^k))_+ \mid \mathcal{F}_t\right] \end{aligned}$$

at any arbitrary stage $t \in \mathbf{T}$ and node specified by \mathcal{F}_t ($x_{0:T} = (x_{0,t-1}, x_t, x_{t+1:T})$ is the concatenation).

STEP 3—VERIFICATION

Accept $x_{0:t}^{k+1}$ if

$$u(x_{0:T}^k) \leq \mathbb{E} Q(x_{0:T}^{k+1}) + \gamma \text{AV@R}_\alpha(Q(x_{0:T}^{k+1})),$$

else try another feasible Z^k (for example, $Z^k \leftarrow \frac{1}{2}(1 + Z^k)$, $Z^k \leftarrow (1+\alpha)\mathbb{1} - \alpha Z^k$ or $Z^k = \mathbb{1}_B$ ($P(B) \geq \alpha$)) and repeat **STEP 2**. If no direction Z^k can be found providing an improvement in (ref{eq:U}), then $x_{0:T}$ is already optimal.

Set

$$u(x_{0:T}^{k+1}) := \mathbb{E} Q(x_{0:T}^{k+1}) + \gamma \text{AV@R}_\alpha(Q(x_{0:T}^{k+1})),$$

increase $k \leftarrow k + 1$ and continue with **STEP 1** unless

$$u(x_{0:T}^{k+1}) - u(x_{0:T}^k) < \varepsilon,$$

where $\varepsilon > 0$ is the desired improvement in each cycle k .

Table 5.1 Run-times (CPU seconds) to solve problem (5.26) for the tree processes ξ with stages and leaves as indicated. Implementation of Algorithm 5.1 in its sequential, non-parallel implementation

Stages	3	4	5	6	6	7	7	8	9	10
Leaves	30	150	81	32	273	64	216	128	256	512
Straightforward implementation/s	14	639	262	58	1.970	264	3.390	1.203	5.231	26.157
Algorithm 5.1/s	20	214	69	21	637	124	725	425	1.320	4.697

The stages and trees, which were chosen to solve problem (5.26), are collected in Table 5.1. We compare a straightforward implementation with the performance of Algorithm 5.1. For this two different programs to solve problem (5.26) have been implemented in MATLAB R2009a:

- (i) Straightforward implementation of problem (5.26) by employing the function `fminunc` provided by Matlab.
- (ii) Implementation of Algorithm 5.1 to generate the solution of the problem with the same precision goal, by employing the same function `fminunc` at each node.

The run-times (in CPU seconds) on a customary laptop for different trees are compared in Table 5.1. The time indicated for Algorithm 5.1 is the time for a single cycle. It was found that successively increasing the tolerance ε to the finally desired tolerance is a time-saving alternative, such that finally 3 to 5 cycles are sufficient to obtain a solution at a desirable precision.⁷

5.5 Dualization of Nonanticipativity Constraints

The nonanticipativity constraints $x \triangleleft \mathfrak{F}$ are inherent to all multistage stochastic optimization problems of the genuine form

$$\begin{aligned} & \text{minimize } \mathcal{R}(Q(x, \xi)) \\ & \text{subject to } x \in \mathbb{X}, \\ & \quad x \triangleleft \mathfrak{F}, \end{aligned} \tag{5.27}$$

they model the fact that decisions cannot anticipate the future.

Many solution techniques involve the Lagrangian of the problem (5.27), and thus naturally the question arises how to incorporate the nonanticipativity constraint in the Lagrangian function.

⁷We are grateful to Jochen Gönsch and Michael Hassler (University Augsburg) for pointing out a shortcoming in a previous version of the algorithm.

To this end (cf. Ruszczyński [19], or Ruszczyński et al. [129, Chapter 2]) one may understand the function x_t as a function on all leaf nodes $n \in \mathcal{N}_T$,

$$\begin{aligned} x : \mathcal{N}_T \times \{0, \dots, T\} &\rightarrow \mathbb{X}_t, \\ (n, t) &\mapsto x_t(n) \end{aligned}$$

as depicted in Fig. 5.8. The nonanticipativity constraint $x \triangleleft \mathfrak{F}$ then states that x_t returns the same result for all nodes, which have the same ancestor at level t , that is

$$x_t(n) = x_t(n') \quad \text{provided that } \text{pred}_t(n) = \text{pred}_t(n').$$

These values, which are all equal at a given node, thus can be replaced by their expected value, which is again the same value. The nonanticipativity condition $x \triangleleft \mathfrak{F}$ is thus equivalent to writing

$$x_t = \mathbb{E}(x_t | \mathcal{F}_t) =: \mathfrak{P}_t x_t \quad t \in \{0, 1, \dots, T\}, \quad (5.28)$$

where $\mathfrak{P}_t(\cdot) = \mathbb{E}(\cdot | \mathcal{F}_t)$ is a linear operator, and (5.28) rewrites explicitly as

$$x_t(n) = \sum_{n_t \prec n'} P(n' | n_t) x_t(n') = (\mathfrak{P}_t x_t)(n) \quad (n \in \mathcal{N}_T)$$

for a discrete probability measure. The linear mapping $\mathfrak{P}_t x = \mathbb{E}(x | \mathcal{F}_t)$ satisfies $\mathfrak{P}_t^2 = \mathfrak{P}_t$, which is the defining property of a projection. \mathfrak{P}_t is moreover self-adjoint for the inner product $\langle \lambda, x \rangle = \mathbb{E} \lambda x$, that is,

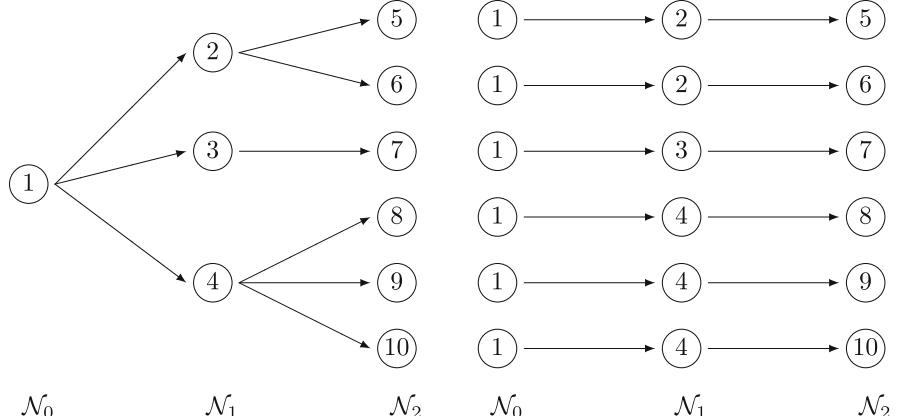


Fig. 5.8 A tree and the pertaining scenario-splitting forest. The tree of Fig. 1.10 (left) is replaced by the clairvoyant forest (right). The non-anticipativity constraints can be expressed as linking constraints: decisions at nodes with the same number have to be identical

$$\mathbb{E} \lambda \cdot (\mathfrak{P}_t x) = \mathbb{E} [\lambda \cdot \mathbb{E} (x | \mathcal{F}_t)] = \mathbb{E} [\mathbb{E} (\lambda | \mathcal{F}_t) \cdot x] = \mathbb{E} (\mathfrak{P}_t \lambda) \cdot x.$$

Hence $1 - \mathfrak{P}_t$ is a self-adjoint projection as well and we have that

$$\mathbb{E} (\lambda - \mathfrak{P}_t \lambda) x = \mathbb{E} \lambda (x - \mathfrak{P}_t x).$$

The nonanticipativity constraints $x_t \triangleleft \mathfrak{F}_t$ ($t = 0, \dots, T$) can be incorporated in a Lagrangian function as

$$L(x, \lambda) := \mathcal{R}(Q(\xi; x)) - \sum_{t=0}^T \mathbb{E} \lambda_t \cdot (x_t - \mathbb{E}(x_t | \mathcal{F}_t))$$

by employing (5.27).

The concave Lagrange dual function of problem (5.27) thus is

$$D(\lambda) := \inf_{x \in \mathbb{X}} \mathcal{R}(Q(\xi; x)) - \sum_{t=0}^T \mathbb{E} (\lambda_t - \mathbb{E} (\lambda_t | \mathcal{F}_t)) \cdot x_t. \quad (5.29)$$

If $\mathbb{E} (\lambda_t - \lambda'_t | \mathcal{F}_t) = 0$, then $D(\lambda_t) = D(\lambda'_t)$ such that one may assume $\mathbb{E} (\lambda_t | \mathcal{F}_t) = 0$ without loss of generality. With this restriction and specializing \mathcal{R} to the expectation $\mathcal{R} = \mathbb{E}$, the Lagrange dual function (5.29) rewrites

$$D(\lambda) = \inf_{x \in \mathbb{X}} \mathbb{E} Q(\xi; x) - \sum_{t=0}^T \mathbb{E} (\lambda_t \cdot x_t) = \inf_{x \in \mathbb{X}} \mathbb{E} \left(Q(\xi; x) - \sum_{t=0}^T \lambda_t \cdot x_t \right).$$

In explicit terms, for a discrete measure, the dual function is

$$\begin{aligned} D(\lambda) &= \inf_{x \in \mathbb{X}} \sum_{n \in \mathcal{N}_T} p(n) \left(Q(\xi(n); x(n)) - \sum_{t=0}^T \lambda_t(n) \cdot x_t(n) \right) \\ &= \sum_{n \in \mathcal{N}_T} p(n) \cdot \inf_{x(n) \in \mathbb{X}} \left(Q(\xi(n); x(n)) - \sum_{t=0}^T \lambda_t(n) \cdot x_t(n) \right) \quad (5.30) \\ &= -\mathbb{E} \sup_{x(n) \in \mathbb{X}} \left(\sum_{t=0}^T \lambda_t(n) \cdot x_t(n) - Q(\xi(n); x(n)) \right) \\ &= -\mathbb{E} Q^*(\xi; \lambda), \end{aligned}$$

where

$$Q^*(\xi; \lambda) := \sup_{x \in \mathbb{X}} \sum_{t=0}^T \lambda_t \cdot x_t - Q(\xi; x) = \sup_{x \in \mathbb{X}} \lambda^\top x - Q(\xi; x)$$

is the usual conjugate function. Equation (5.30) is justified by the interchangeability principle, cf. Rockafellar and Wets [114].

The essential observation is that the infimum can be computed for every path $n \in \mathcal{N}_T$ separately: for every scenario $n \in \mathcal{N}_T$ and every fixed choice of $\lambda = (\lambda_0, \dots, \lambda_T)$ there are separate decision $x = (x_0, \dots, x_T)$ maximizing $Q^*(\xi(n); \lambda)$.

By the max–min-inequality ($\sup_{\lambda} \inf_x L(x, \lambda) \leq \inf_x \sup_{\lambda} L(x, \lambda)$) it always holds that

$$\max_{\mathbb{E}(\lambda_t | \mathcal{F}_t) = 0} -\mathbb{E} Q^*(\xi; \lambda) \leq \min_{x \triangleleft \mathfrak{F}} \mathbb{E} Q(\xi; x),$$

but provided a vanishing duality gap equality holds,

$$\min_{x \triangleleft \mathfrak{F}} \mathbb{E} Q(\xi; x) = \max_{\mathbb{E}(\lambda_t | \mathcal{F}_t) = 0} -\mathbb{E} Q^*(\xi; \lambda).$$

Chapter 6

Approximations and Bounds

As was already said in the introduction, stochastic optimization problems and in particular multistage stochastic optimization problems can only be solved by approximation. Coming back to Fig. 1.2 in the Introduction we consider here in detail the quality of approximation of a complex problem by a simpler one.

While approximation techniques aim at finding sequences of simpler problems, such that the optimal values and the optimal solutions converge to the corresponding values of the basic problem, bounding techniques find simpler problems for which the optimal value is larger (smaller, resp.) than the optimal value of the basic problem. If the lower and the upper bound are close enough to each other, then the decision maker may accept these solutions as good enough and abstain from the time-consuming full numerical solution of the basic problem.

Approximations and bounds are presented here for two-stage problems first, and then extended to multistage problems. The approximations involve Lipschitz functions, such that Wasserstein and nested distances can be employed naturally. Further, bounds are presented.

6.1 Two-Stage Problems, and Approximation in the Wasserstein Distance

The Kantorovich–Rubinstein theorem (Theorem 2.29) is a first theorem, which involves Lipschitz constants to compare probability measures. We have presented a continuity result (Theorem 3.31), which compares the evaluation of risk measures by involving Lipschitz, or Hölder continuity of the random variables addressed. Here, we extend these results and make them available for stochastic optimization. Two-stage stochastic problems are addressed first, the results are then extended to multistage stochastic problems.

Theorem 6.1 (Second and First Stage, Complete Recourse). Suppose that $\xi \mapsto Q(x_0, \xi, x_1)$ is uniformly Hölder, that is

$$\left| Q(x_0, \xi, x_1) - Q\left(x_0, \tilde{\xi}, x_1\right) \right| \leq H_\beta \cdot d(\xi, \tilde{\xi})^\beta$$

for all $x_0 \in \mathbb{X}_0$, $x_1 \in \mathbb{X}_1$ and $\xi, \tilde{\xi} \in \Xi$ and some $\beta \leq 1$. For a version independent risk measure \mathcal{R} with Kusuoka representation $\mathcal{R}(\cdot) = \sup_{\sigma \in S} \mathcal{R}_\sigma(\cdot)$ according to Corollary 3.14 it holds that

$$\begin{aligned} & \left| \inf_{x_0 \in \mathbb{X}_0} \mathcal{R}_P \left(\inf_{x_1 \in \mathbb{X}_1} Q(x_0, \xi, x_1) \right) - \inf_{x_0 \in \mathbb{X}_0} \mathcal{R}_{\tilde{P}} \left(\inf_{x_1 \in \mathbb{X}_1} Q(x_0, \xi, x_1) \right) \right| \\ & \leq H_\beta \cdot d_r(P, \tilde{P})^\beta \cdot \sup_{\sigma \in S} \|\sigma\|_{r_\beta^*}, \end{aligned}$$

where d_r is the Wasserstein distance of order r ($r \geq 1$) and $r_\beta^* \geq \frac{r}{r-\beta}$.

Remark 6.2. It should be noted that the sets \mathbb{X}_0 and \mathbb{X}_1 are arbitrary sets, they are not necessarily convex or connected. In particular, $\mathbb{X}_0 = \mathbb{Z}^m$ and $\mathbb{X}_1 = \mathbb{Z}^n$, or $\mathbb{X}_0 = \{0, 1\}^m$ and $\mathbb{X}_1 = \{0, 1\}^n$ are possible choices, such that the corollary applies for stochastic integer optimization equally well.

To verify the assertion we recall the following lemma.

Lemma 6.3. Let f_α be Hölder continuous functions with uniform exponent $\beta \leq 1$ and constant H_β for every $\alpha \in A$. Then $\sup_{\alpha \in A} f_\alpha$ and $\inf_{\alpha \in A} f_\alpha$ (provided they are finite at a single point) are Hölder continuous with constant H_β for the exponent β .

Proof. For x in the domain and $\varepsilon > 0$, choose α_0 such that $\sup_{\alpha \in A} f_\alpha(x) < f_{\alpha_0}(x) + \varepsilon$. Then

$$\begin{aligned} \sup_{\alpha \in A} f_\alpha(x) - \sup_{\alpha \in A} f_\alpha(y) & \leq f_{\alpha_0}(x) + \varepsilon - \sup_{\alpha \in A} f_\alpha(y) \\ & \leq f_{\alpha_0}(x) + \varepsilon - f_{\alpha_0}(y) \\ & \leq H_\beta \cdot d(x, y)^\beta + \varepsilon. \end{aligned}$$

Interchanging the role of x and y reveals the assertion, as $\varepsilon > 0$ was chosen arbitrarily. The assertion for the infimum follows by considering $-f_\alpha$ instead of f_α . \square

Proof of Corollary 6.1. By the previous lemma the mapping $\xi \mapsto \inf_{x_1 \in \mathbb{X}_1} Q(x_0, \xi, x_1)$ is Hölder continuous. Hence, by Corollary 3.33,

$$\mathcal{R}_P \left(\inf_{x_1 \in \mathbb{X}_1} Q(x_0, \xi, x_1) \right) - \mathcal{R}_{\tilde{P}} \left(\inf_{x_1 \in \mathbb{X}_1} Q(x_0, \xi, x_1) \right) \leq H_\beta \cdot d_r(P, \tilde{P})^\beta \cdot \sup_{\sigma \in S} \|\sigma\|_{r_\beta^*}$$

for every $x_0 \in \mathbb{X}_0$.

Choose $\tilde{x}_0 \in \mathbb{X}_0$, such that

$$\mathcal{R}_{\tilde{P}} \left(\inf_{x_1 \in \mathbb{X}_1} Q(\tilde{x}_0, \xi, x_1) \right) \leq \inf_{x_0 \in \mathbb{X}_0} \mathcal{R}_{\tilde{P}} \left(\inf_{x_1 \in \mathbb{X}_1} Q(x_0, \xi, x_1) \right) + \varepsilon.$$

Then

$$\begin{aligned} & \inf_{x_0 \in \mathbb{X}_0} \mathcal{R}_P \left(\inf_{x_1 \in \mathbb{X}_1} Q(x_0, \xi, x_1) \right) - \inf_{x_0 \in \mathbb{X}_0} \mathcal{R}_{\tilde{P}} \left(\inf_{x_1 \in \mathbb{X}_1} Q(x_0, \xi, x_1) \right) \\ & \leq \mathcal{R}_P \left(\inf_{x_0 \in \mathbb{X}_0} Q(\tilde{x}_0, \xi, x_1) \right) - \mathcal{R}_{\tilde{P}} \left(\inf_{x_1 \in \mathbb{X}_1} Q(\tilde{x}_0, \xi, x_1) \right) + \varepsilon \\ & \leq H_\beta \cdot d_r(P, \tilde{P})^\beta \cdot \sup_{\sigma \in S} \|\sigma\|_{r_\beta^*} + \varepsilon. \end{aligned}$$

Again, as $\varepsilon > 0$ was chosen arbitrarily, the assertion follows by interchanging the roles of P and \tilde{P} . \square

6.2 Approximation in the Nested Distance Sense

The multistage distance is a suitable distance for multistage stochastic optimization problems. To elaborate the relation consider the value function $v(\mathbb{P})$ of the stochastic optimization problem

$$v(\mathbb{P}) = \min \{ \mathbb{E}_{\mathbb{P}} Q(x, \xi) : x \in \mathbb{X}, x \triangleleft \mathfrak{F}; (\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P} \} \quad (6.1)$$

$$= \min \left\{ \int Q(x, \xi) P(d\xi) : x \in \mathbb{X}, x \triangleleft \mathfrak{F} \right\}$$

of expectation-minimization type.

The following theorem is the main theorem to bound stochastic optimization problems by the nested distance. This multistage distance links smoothness properties of the loss function Q with smoothness of the value function v .

Theorem 6.4 (Hölder Property of the Value Function). *Let $\mathbb{P}, \tilde{\mathbb{P}}$ be two nested distributions. Assume that \mathbb{X} is convex, and the cost function Q is convex in x for any fixed ξ , i.e.,*

$$Q((1-\lambda)x + \lambda\tilde{x}, \xi) \leq (1-\lambda)Q(x, \xi) + \lambda Q(\tilde{x}, \xi), \quad 0 < \lambda < 1.$$

Moreover let Q be uniformly Hölder continuous ($\beta \leq 1$) with constant H_β , that is

$$|Q(x, \xi) - Q(x, \tilde{\xi})| \leq H_\beta \cdot d(\xi, \tilde{\xi})^\beta$$

for all $x \in \mathbb{X}$ and vectors $\xi = (\xi_t)_{t \in \{0, \dots, T\}}$ and $\tilde{\xi} = (\tilde{\xi}_t)_{t \in \{0, \dots, T\}}$.

Then the value function v in (6.1) inherits the Hölder constant with respect to the nested distance, that is

$$\left| v(\mathbb{P}) - v(\tilde{\mathbb{P}}) \right| \leq H_\beta \cdot \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^\beta$$

for any $r \geq 1$.¹

Proof. Let $x \triangleleft \mathfrak{F}$ be a decision vector for problem (6.1) and the nested distribution \mathbb{P} , and let π be a bivariate probability measure on $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$, which satisfies the conditions (2.35), i.e., which is an optimal transportation measure for the nested distance. Note that π has marginal P , whence

$$\begin{aligned} \mathbb{E}_P [Q(x, \xi)] &= \int Q(x(\omega), \xi(\omega)) P(d\omega) \\ &= \int Q(x(\omega), \xi(\omega)) \pi(d\omega, d\tilde{\omega}) = \mathbb{E}_\pi [Q(x \circ \text{id}, \xi \circ \text{id})] \end{aligned} \tag{6.2}$$

(id is the projection $\text{id}(\omega, \tilde{\omega}) = \omega$, while $\tilde{\text{id}}(\omega, \tilde{\omega}) = \tilde{\omega}$).

Define next the new decision function

$$\tilde{x}_t(\tilde{\omega}) := \int x_t(\omega) \pi(d\omega | \tilde{\omega}),$$

which has the desired measurability, that is $\tilde{x} \triangleleft \tilde{\mathcal{F}}$ due its definition² and the conditions (2.35) imposed on π . By convexity of Q it follows from Jensen's inequality that

$$\begin{aligned} Q(\tilde{x}(\tilde{\omega}), \tilde{\xi}(\tilde{\omega})) &= Q\left(\int x_t(\omega) \pi(d\omega | \tilde{\omega}), \tilde{\xi}(\tilde{\omega})\right) \\ &\leq \int Q(x_t(\omega), \tilde{\xi}(\tilde{\omega})) \pi(d\omega | \tilde{\omega}) \end{aligned}$$

Integrating with respect to \tilde{P} one obtains

$$\begin{aligned} \mathbb{E}_{\tilde{P}} Q(\tilde{x}, \tilde{\xi}) &= \int Q(\tilde{x}(\tilde{\omega}), \tilde{\xi}(\tilde{\omega})) \tilde{P}(d\tilde{\omega}) \\ &\leq \iint Q(x_t(\omega), \tilde{\xi}(\tilde{\omega})) \pi(d\omega | \tilde{\omega}) \tilde{P}(d\tilde{\omega}) \\ &= \iint Q(x_t(\omega), \tilde{\xi}(\tilde{\omega})) \pi(d\omega, d\tilde{\omega}) = \mathbb{E}_\pi Q(x \circ \text{id}, \tilde{\xi} \circ \tilde{\text{id}}), \end{aligned}$$

¹For $\beta = 1$ Hölder continuity is just Lipschitz continuity.

² \tilde{x} may be given by $\tilde{x} := \mathbb{E}_\pi(x \circ \text{id} | \tilde{\text{id}})$.

and together with (6.2) it follows that

$$\begin{aligned}\mathbb{E}_{\tilde{P}} Q(\tilde{x}, \tilde{\xi}) - \mathbb{E}_P Q(x, \xi) &\leq \iint Q(x \circ \text{id}, \tilde{\xi} \circ \tilde{\text{id}}) - Q(x \circ \text{id}, \xi \circ \text{id}) d\pi \\ &\leq \iint H_\beta \mathbf{d}(\text{id}, \tilde{\text{id}})^\beta d\pi.\end{aligned}$$

Now let x be an ε -optimal decision for $v(\mathbb{P})$, that is

$$\int Q(x, \xi) dP < v(\mathbb{P}) + \varepsilon.$$

It follows that

$$\begin{aligned}\mathbb{E}_{\tilde{P}} Q(\tilde{x}, \tilde{\xi}) - v(\mathbb{P}) &\leq \mathbb{E}_{\tilde{P}} Q(\tilde{x}, \tilde{\xi}) - \mathbb{E}_P Q(x, \xi) + \varepsilon \\ &\leq \varepsilon + H_\beta \cdot \mathbb{E}_\pi \mathbf{d}^\beta\end{aligned}$$

and whence

$$v(\tilde{\mathbb{P}}) - v(\mathbb{P}) < \varepsilon + H_\beta \cdot \mathbb{E}_\pi \mathbf{d}^\beta.$$

Letting $\varepsilon \rightarrow 0$, taking the infimum over all π and interchanging the roles of \mathbb{P} and $\tilde{\mathbb{P}}$ gives that

$$\left| v(\mathbb{P}) - v(\tilde{\mathbb{P}}) \right| \leq H_\beta \cdot \mathbf{d}_\beta(\mathbb{P}, \tilde{\mathbb{P}})^\beta.$$

The statement for general index $r \geq 1 \geq \beta$ finally follows by applying Lemma 2.41. \square

Remark 6.5. In applications it is important to choose the distance in an adaptive way to the problem itself. As an example suppose the function Q takes the particular form

$$Q(x, \xi) = \sum_{t=0} e^{-rt} Q_t(x_t, \xi_t),$$

as is often the case in dynamic programming for a discount factor e^{-r} . If the functions Q_t have a common Lipschitz constant L , then

$$\left| Q(x, \xi) - Q(x, \tilde{\xi}) \right| \leq \sum_{t=0} e^{-rt} L \mathbf{d}_t(\xi_t, \tilde{\xi}_t) = L \cdot \mathbf{d}(\xi, \tilde{\xi})$$

for the distance

$$\mathbf{d}(\xi, \tilde{\xi}) := \sum_{t=0} e^{-rt} \cdot \mathbf{d}_t(\xi_t, \tilde{\xi}_t), \quad (6.3)$$

which is an adaptive choice.

If the trees are further bounded, i.e., $\mathbf{d}_t(\xi_t, \tilde{\xi}_t) \leq C$, then the distance (6.3) downweights nodes of later stages. This notably justifies rougher approximations towards the far future of the process. In extreme situations, a continuation by a single scenario (i.e., a tree of bushiness 1) from a certain stage of the tree can be justified for the particular distance (6.3).

Remark 6.6 (Best Possible Bound). Theorem 6.4 is adapted from Pflug and Pichler [94]. This reference contains a further statement ensuring that the bound presented in Theorem 6.4 cannot be improved.

As a corollary we get the following result for usual stochastic optimization.

Corollary 6.7. *Let \mathbb{X} be convex, suppose that the mapping $x \mapsto Q(x, \xi)$ is convex and uniformly Lipschitz continuous with constant L . Then the value function is continuous with respect to the nested distance,*

$$\left| \inf_{\tilde{x} \triangleleft \tilde{\xi}} \mathbb{E}_{\tilde{P}} Q(\tilde{x}, \tilde{\xi}) - \inf_{x \triangleleft \xi} \mathbb{E}_P Q(x, \xi) \right| \leq L \cdot \mathbf{d}_1(\mathbb{P}, \tilde{\mathbb{P}})$$

where $(\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}$ and $(\tilde{\Omega}, \tilde{\mathfrak{F}}, \tilde{P}, \tilde{\xi}) \sim \tilde{\mathbb{P}}$.

The presented results about error bounds imply in particular that making the nested distance between the problem based on \mathbb{P} and the approximate problem based on $\tilde{\mathbb{P}}$ smaller will lead to better bounds for the optimal value. Of course, the distance can be decreased by making the approximating tree bushier, see Example 4.31 and Fig. 4.5. To illustrate the approximation error as a discrete state model approximates a model with continuous state space consider the following example.

Example 6.8. We consider again the flowergirl problem of the introduction (Example 1.4), where negative profits are minimized, i.e., profits are maximized. For a tree of height 3 the problem was solved analytically (since the exact solution is available) and then the approximate tree based problem was solved with increasing bushiness. The result can be seen in Fig. 6.1a. The little squares indicate the exact optimal value and the solid line shows the optimal value based on an optimal approximation on the basic process by trees of increasing bushiness. The approximation error decreases with increasing bushiness.

As a comparison, we have also generated approximating trees by pure random sampling. The optimal values on these random trees are of course random variables themselves. Figure 6.1b shows boxplots of the distribution of these optimal values.

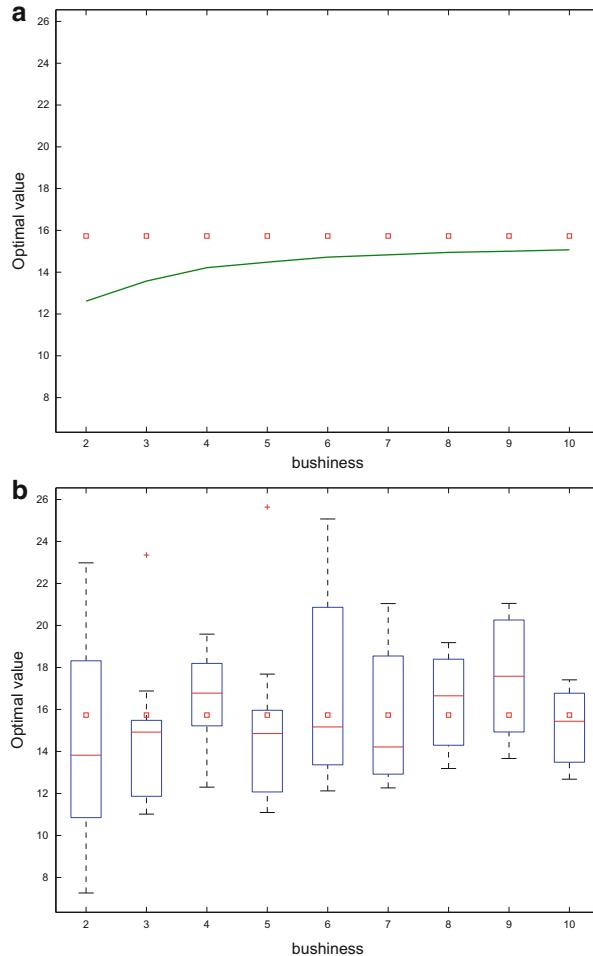


Fig. 6.1 Quality of the objective value for different approximating trees. (a) Convergence of the optimal value of the approximate tree based problems (*solid line*) to the true optimal value (*little squares*) by increasing the bushiness. (b) “Approximation” of the true objective value by the objective values of randomly generated trees. These values vary a lot as can be seen from the *boxplots*

They vary quite much and even their mean is often far away from the true value (again represented as little squares).³

The basic Theorem 6.4 may be extended to risk-sensitive optimization problems with some risk functionals as objective. If the objective is to minimize the Average Value-at-Risk, the following corollary gives the approximation result.

³The two pictures were prepared by Anna Timonina.

Corollary 6.9. Consider the optimization problem

$$v(\mathbb{P}) = \min \{ \text{AV@R}_{\alpha, \mathbb{P}}[Q(x, \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}; \mathbb{P} = (\Omega, \mathfrak{F}, P, \xi) \}. \quad (6.4)$$

If Q is convex in x and Hölder continuous ($\beta \leq 1$) in ξ with constant H_β , then

$$\left| v(\mathbb{P}) - v(\tilde{\mathbb{P}}) \right| \leq \frac{1}{1-\alpha} H_\beta \cdot \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}})^\beta.$$

Proof. Recall that $\text{AV@R}_\alpha = \min \{ z + \frac{1}{1-\alpha} \mathbb{E}[Y - z]_+ : z \in \mathbb{R} \}$. Extend the cost function Q to

$$\hat{Q}(z, x, \xi) = z + \frac{1}{1-\alpha} (Q(x, \xi) - z)_+$$

by augmenting the variable $z \in \mathbb{R}$. Problem (6.4) can be written as

$$v(\mathbb{P}) = \min \left\{ \mathbb{E}_{\mathbb{P}} \hat{Q}(z, x, \xi) : z \in \mathbb{R}, x \in \mathbb{X}, x \triangleleft \mathfrak{F}; \mathbb{P} = (\mathfrak{F}, P, \xi) \right\}.$$

Notice that $(z, x) \mapsto \hat{Q}(z, x, \xi)$ is convex, since $u \mapsto (u)_+$ is monotonic and convex. The conclusion follows from the previous theorem, since $(z, x) \mapsto \hat{Q}(z, x, \xi)$ is Hölder continuous with constant $\frac{1}{1-\alpha} H_\beta$:

$$\begin{aligned} \left| \hat{Q}(z, x, \xi) - \hat{Q}(z, x, \tilde{\xi}) \right| &\leq \frac{1}{1-\alpha} \left| Q(x, \xi) - Q(x, \tilde{\xi}) \right| \\ &\leq \frac{1}{1-\alpha} H_\beta \cdot \mathbf{d}(\xi, \tilde{\xi})^\beta. \end{aligned}$$

□

The continuity results can be extended to more general risk measures.

Theorem 6.10. Let \mathcal{R}_σ be a distortion risk functional with bounded distortion, $\sigma \in L^\infty$. Moreover suppose that \mathbb{X} is convex, and $x \mapsto Q(x, \xi)$ is convex and uniformly Lipschitz continuous with constant L , that is,

$$\left| Q(x, \xi) - Q(x, \tilde{\xi}) \right| \leq L \cdot \mathbf{d}(\xi, \tilde{\xi}).$$

Then the value function is continuous with respect to the nested distance,

$$\left| \inf_{\tilde{x} \triangleleft \tilde{\mathcal{F}}} \mathcal{R}_{\tilde{P}}(Q(\tilde{x}, \tilde{\xi})) - \inf_{x \triangleleft \mathcal{F}} \mathcal{R}_P(Q(x, \xi)) \right| \leq L \cdot \|\sigma\|_\infty \cdot \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}}).$$

Proof. Let i (\tilde{i} , resp.) denote the natural projection $i : \Xi \times \tilde{\Xi} \rightarrow \Xi$ ($\tilde{i} : \Xi \times \tilde{\Xi} \rightarrow \tilde{\Xi}$, resp.). Let π be a transport plan satisfying the conditions for the nested distribution,

(2.35). For a decision $x = (x_t)_{t=0}^T$ define $\tilde{x} := (\tilde{x}_t)_{t=0}^T$ with components $\tilde{x}_t := \mathbb{E}_\pi(x_t \circ i | \tilde{i})$ (note that \tilde{x}_t is a function on $\tilde{\Xi}$ with the same state space as x_t). As $x_t \triangleleft \mathcal{F}_t$, it follows by construction and (2.35) that $\tilde{x}_t \triangleleft \tilde{\mathcal{F}}_t$.

To apply Theorem 3.22 let h be a nondecreasing, convex \mathbb{R} -valued function with Lipschitz constant $\bar{L}(h)$. From convexity of Q ,

$$Q((1-\lambda)x_0 + x_1) \leq (1-\lambda)Q(x_0) + \lambda Q(x_1),$$

and monotonicity and convexity of h it follows that

$$\begin{aligned} h(Q((1-\lambda)x_0 + x_1)) &\leq h((1-\lambda)Q(x_0) + \lambda Q(x_1)) \\ &\leq (1-\lambda)h(Q(x_0)) + \lambda h(Q(x_1)), \end{aligned}$$

such that the composition $h \circ Q$ is convex as well. Hence, by the conditional Jensen inequality (cf. [142]) again,

$$h(Q(\tilde{x}, \tilde{\xi})) = h(Q(\mathbb{E}_\pi(x \circ i | \tilde{i}), \tilde{\xi})) \leq \mathbb{E}_\pi(h(Q(x \circ i, \tilde{\xi})) | \tilde{i}).$$

Integrating with respect to π gives that

$$\begin{aligned} \mathbb{E}_{\tilde{P}}h(Q(\tilde{x}, \tilde{\xi})) &= \mathbb{E}_\pi h(Q(\tilde{x}, \tilde{\xi})) \\ &\leq \mathbb{E}_\pi \mathbb{E}_\pi(h(Q(x \circ i, \tilde{\xi})) | \tilde{i}) = \mathbb{E}_\pi h(Q(x \circ i, \tilde{\xi} \circ \tilde{i})). \end{aligned}$$

Note that the composition $h \circ Q$ is Lipschitz continuous as well with constant

$$|h(Q(x, \xi)) - h(Q(x, \tilde{\xi}))| \leq \bar{L}(h) \cdot |Q(x, \xi) - Q(x, \tilde{\xi})| \leq \bar{L}(h) \cdot L \cdot d(\xi, \tilde{\xi}),$$

and hence

$$\begin{aligned} \mathbb{E}_{\tilde{P}}h(Q(\tilde{x}, \tilde{\xi})) - \mathbb{E}_P h(Q(x, \xi)) &\leq \mathbb{E}_\pi h(Q(x, \tilde{\xi})) - h(Q(x, \xi)) \\ &\leq \bar{L}(h) \cdot L \cdot \mathbb{E}_\pi d(\xi, \tilde{\xi}). \end{aligned}$$

Taking the infimum over all respective measures thus

$$\mathbb{E}_{\tilde{P}}h(Q(\tilde{x}, \tilde{\xi})) - \mathbb{E}_P h(Q(x, \xi)) \leq \bar{L}(h) \cdot L \cdot d_1(\mathbb{P}, \tilde{\mathbb{P}}).$$

Now choose $x^0 \triangleleft \mathfrak{F}$ and h with $\int_0^1 h^*(\sigma(u)) du \leq 0$ such that

$$\mathbb{E}_P h(Q(x^0, \xi)) \leq \inf_{x \triangleleft \mathcal{F}} \mathcal{R}_P(Q(x, \xi)) + \varepsilon$$

for some arbitrary $\varepsilon > 0$. Note, by Remark 3.24, that one may assume that $\bar{L}(h) \leq \|\sigma\|_\infty$. With $\tilde{x}_t^0 := \mathbb{E}_\pi(x_t^0 \circ i | \tilde{i})$ it follows thus that

$$\begin{aligned} \inf_{\tilde{x} \triangleleft \tilde{\mathcal{F}}} \mathbb{E}_{\tilde{P}} h(Q(\tilde{x}, \tilde{\xi})) - \inf_{x \triangleleft \mathcal{F}} \mathcal{R}_P(Q(x, \xi)) &\leq \mathbb{E}_{\tilde{P}} h(Q(\tilde{x}^0, \tilde{\xi})) \\ &\quad - \mathbb{E}_P h(Q(x^0, \xi)) + \varepsilon. \\ &\leq \|\sigma\|_\infty \cdot L \cdot \mathbf{d}_1(\mathbb{P}, \tilde{\mathbb{P}}) + \varepsilon. \end{aligned}$$

Now one may take the infimum over all functions $h : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $\int_0^1 h^*(\sigma(u)) du \leq 0$ and $\bar{L}(h) \leq \|\sigma\|_\infty$ to obtain

$$\begin{aligned} \inf_{\tilde{x} \triangleleft \tilde{\mathcal{F}}} \mathcal{R}_{\tilde{P}}(Q(\tilde{x}, \tilde{\xi})) - \inf_{x \triangleleft \mathcal{F}} \mathcal{R}_P(Q(x, \xi)) &\leq L \cdot \|\sigma\|_\infty \cdot \mathbf{d}_1(\mathbb{P}, \tilde{\mathbb{P}}) \\ &\leq L \cdot \|\sigma\|_\infty \cdot \mathbf{d}_r(\mathbb{P}, \tilde{\mathbb{P}}). \end{aligned}$$

Exchanging the roles of P and \tilde{P} reveals the desired assertion. \square

6.3 Bounds

For very large multistage stochastic programs, the solution to optimality may be quite tedious and requires often high performance or parallel software to be carried out. Bounding techniques give quick information about lower and upper bounds for the objective value. Given that the gap between these two bounds is acceptably small, one may even stop and never fully optimize. It is quite simple to get upper and lower bounds for minimization problems.

- *Lower bounds* are obtained if one or several constraints are relaxed:

$$\min \{F(x) : x \in \mathbb{X}'\} \leq \min \{F(x) : x \in \mathbb{X}\},$$

if $\mathbb{X}' \supseteq \mathbb{X}$.

- *Upper bounds* are obtained, if constraints are tightened. In particular—and this is the most important case—if a feasible solution is inserted:

$$F(x^+) \geq \min \{F(x) : x \in \mathbb{X}\},$$

if $x^+ \in \mathbb{X}$.

Convex analysis allows to derive lower and upper bounds for expectations.

- Jensen's inequality gives a *lower bound*: for a convex function Q

$$Q(\mathbb{E}[\xi]) \leq \mathbb{E}[Q(\xi)],$$

if all integrals are well defined.

- The Edmundson–Madansky inequality gives an *upper bound*: for a convex function Q and a random variable ξ , which takes its values in the interval $[a, b]$. It holds that

$$\mathbb{E}[Q(\xi)] \leq Q(a) \frac{b - \mathbb{E}(\xi)}{b - a} + Q(b) \frac{\mathbb{E}(\xi) - a}{b - a}.$$

In what follows we assume that the stochastic program $Opt(\mathbb{P})$ is of the form

$$Opt(\mathbb{P}) : v^*(\mathbb{P}) = \min \{ \mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}, \mathbb{P} \sim (\Omega, \mathfrak{F}, P, \xi) \}. \quad (6.5)$$

As usual, the nested distribution \mathbb{P} comprises the filtration \mathfrak{F} , the probability measure P and the scenario process ξ . In what follows we consider bounds for changing the probability measure and bounds for changing the filtration separately.

6.3.1 Lower Bounds by Changing the Probability Measure

Given the optimal value mapping $(\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P} \mapsto v^*(\mathbb{P}) = v^*(\Omega, \mathfrak{F}, P, \xi)$ we keep the filtration and the process ξ fixed and consider only the mapping

$$P \mapsto v^*(P).$$

One main structural property of some stochastic programs (including those which are of the expectation type) is that this mapping is *concave*.

Lemma 6.11. *Suppose that in the basic problem (6.5) the functional \mathcal{R} is compound concave (i.e., $P \mapsto \mathcal{R}_P(\cdot)$ is concave). Then the mapping $P \mapsto v^*(P)$ is also concave.*

Proof. Let $P = \sum_{i=1}^k p_i P_i$ with $\sum p_i = 1$, $p_i > 0$, and let the x^+ be an ε -solution of problem $Opt(P) = Opt(\mathfrak{F}, P, \xi)$.⁴ Since x^+ is feasible for all $Opt(P_i)$ we have that

$$v^*(P) + \varepsilon \geq \mathcal{R}_P[Q(x^+, \xi)] \geq \sum p_i \mathcal{R}_{P_i}[Q(x^+, \xi)] \geq \sum p_i v^*(P_i).$$

Since ε was arbitrary, the result follows. \square

⁴In case that the minimum is not attained in (6.5) only ε -solutions are available.

Numerous consequences result from this simple lemma. For instance, let $\mathcal{N}_T = \Omega = \{\omega_1, \dots, \omega_k\}$ be the leaf set of a finite tree such that the corresponding node process v_T at the final stage takes the k values $\omega_1, \dots, \omega_k$ with probabilities $P(\omega_1), \dots, P(\omega_k)$. Let δ_{ω_i} be the point mass at ω_i . Under δ_{ω_i} , the scenario ω_i has probability 1 and all other scenarios have probability zero (cf. this scenario splitting in Fig. 5.8). Thus solving the problem under δ_{ω_i} is a deterministic dynamic optimization problem, its optimal value is denoted by $v^*(\delta_{\omega_i})$. The value

$$v_1^* := \sum_i P(\omega_i) v^*(\delta_{\omega_i})$$

is the clairvoyant's solution, which may alternatively be written as

$$v_1^* = v^*(\mathfrak{F}^T, P),$$

where \mathfrak{F}^T is the clairvoyant's filtration $\mathfrak{F}^T = (\mathcal{F}_T, \dots, \mathcal{F}_T)$. Thus we get the (quite trivial) lower bound by the inequality

$$v_1^* = v^*(\mathfrak{F}^T, P) \leq v^*(\mathfrak{F}, P). \quad (6.6)$$

Some authors would call this inequality

$$\text{wait-and-see} \leq \text{here-and-now},$$

which is a bit misleading, since the essence of the “wait-and-see” solution is not to wait, but to decide immediately with complete knowledge of the future, an ability which we may attribute only to clairvoyants.

Refinement Chains. While the dissection of P into its atoms is the most extreme dissection, one may also dissect it into a convex combination of probabilities sitting on 2 or more leaves. For instance, let ω_1 be a specific leaf node (without loss of generality call it ω_1) and set, for $i \neq 1$,

$$P_i^{(2)} := P(\omega_1)\delta_{\omega_1} + (1 - P(\omega_1))\delta_{\omega_i}. \quad (6.7)$$

Defining $\pi_i^{(2)} := P(\omega_i)/(1 - P(\omega_1))$ one sees that $P = \sum_{i \neq 1} \pi_i^{(2)} P_i^{(2)}$ and hence

$$v_2^* := \sum_{i \neq 1} \pi_i^{(2)} v^*(P_i^{(2)}) \leq v^*(P). \quad (6.8)$$

Thus a lower bound for the basic problem can be found by solving $k - 1$ multistage problems with only 2 scenarios each. Since by (6.7) every subproblem with two scenarios is a convex combination of two (clairvoyant's) problems with just one scenario each, one gets

$$v^* \left(P_i^{(2)} \right) \geq P(\omega_1) v^*(\delta_{\omega_1}) + (1 - P(\omega_1)) v^*(\delta_{\omega_i})$$

for $i \neq 1$ and one arrives at a chain of inequalities

$$v_1^* = \sum_{i=1}^k P(\omega_i) v^*(\delta_{\omega_i}) \leq v_2^* \leq v^*(P).$$

Pairs of scenarios have been considered for the first time in Birge and Louveaux [11], they are further elaborated in Maggioni, Allevi and Bertocchi [76].

The method may be refined by considering longer *refinement chains*. Such a chain is of the structure

$$\begin{aligned} \Omega &= \Omega^{(\ell+1)}, \\ \left(\Omega_1^{(\ell)}, \Omega_2^{(\ell)}, \dots, \Omega_{m_\ell}^{(\ell)} \right), \\ &\vdots \\ \left(\Omega_1^{(2)}, \Omega_2^{(2)}, \dots, \Omega_{m_2}^{(2)} \right), \\ \left(\{\omega_1\}, \{\omega_2\}, \dots, \{\omega_k\} \right), \end{aligned}$$

where each row is a collection of subsets of the probability space Ω with the property that their union covers the whole space $\Omega = \bigcup_i \Omega_i^{(j)}$ for every j and that each set $\Omega_i^{(j+1)}$ is the union of sets from the next more refined collection

$$\Omega_i^{(j)} = \bigcup_{\Omega_s^{(j-1)} \subseteq \Omega_i^{(j)}} \Omega_s^{(j-1)}.$$

To a refinement chain of the probability space Ω there corresponds a dissection of the probability P into probability measures $P_i^{(j)}$,

$$\begin{aligned} P, \\ \left(P_1^{(\ell)}, \dots, P_{m_\ell}^{(\ell)} \right), \\ &\vdots \\ \left(P_1^{(2)}, \dots, P_{m_2}^{(2)} \right), \\ \left(P_1^{(1)} = \delta_{\omega_1}, \dots, P_k^{(1)} = \delta_{\omega_k} \right), \end{aligned}$$

such that

- $P_i^{(j)}$ has support $\Omega_i^{(j)}$,
- P can be written as $P = \sum_{i=1}^{m_j} \pi_i^{(j)} P_i^{(j)}$ and
- each $P_i^{(j)}$ can be written as a convex combination of probabilities from the refined collection $\{P_i^{(j-1)}\}$.

It is evident that such a refinement chain leads to a chain of lower bounds. Denoting

$$v_j^* = \sum_{i=1}^{m_j} \pi_i^{(j)} v^*(P_i^{(j)})$$

one arrives at a chain of inequalities

$$v_1^* \leq v_2^* \leq \cdots \leq v_\ell^* \leq v_{\ell+1}^*(P) = v^*(P).$$

Notice that the higher the index j , the fewer problems have to be solved but with an increasing number of scenarios. Notice also that the clairvoyant solution v_1^* is always the smallest lower bound and further bounds are improvements over the more or less trivial bound (6.6).

There are many ways to construct refinement chains. One may keep one or several scenarios fixed in all subsets $\Omega_i^{(j)}$ or choose them disjoint. Some examples will illustrate possible choices.

Example 6.12 (One fixed scenario). Suppose that scenario ω_1 is fixed and appears in all subsets. Then the structure of the refinement chain is

$$\begin{aligned} & (\{\omega_1, \omega_2, \dots, \omega_k\}), \\ & (\{\omega_1, \omega_2, \dots, \omega_s\}, \{\omega_1, \omega_{s+1}, \dots, \omega_k\}), \\ & \quad \vdots \\ & (\{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \dots, \{\omega_1, \omega_k\}). \end{aligned}$$

The situation is illustrated in Fig. 6.2. The scenarios are represented as points on a circle. The support sets $\Omega_i^{(j)}$ are shown as complete subgraphs (all its elements are linked by a straight line), the chain ranges from the full set $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ to the collection of singletons $\{\{\omega_1\}, \{\omega_2\}, \dots, \{\omega_k\}\}$.

As a generalization we consider the case that two or more scenarios may be fixed, i.e., may appear in all subsets. In order to calculate the correct weights $\pi_i^{(\cdot)}$ of the dissections, assume without loss of generality that the first f scenarios $\{\omega_1, \dots, \omega_f\}$ are fixed and appear in all subsets. Assume further that each subset contains exactly $r > f$ scenarios, meaning that $(k - f)/(r - f) = s$ must be an integer. Let $x = r - f > 0$. The corresponding s elements of the refinement chain are

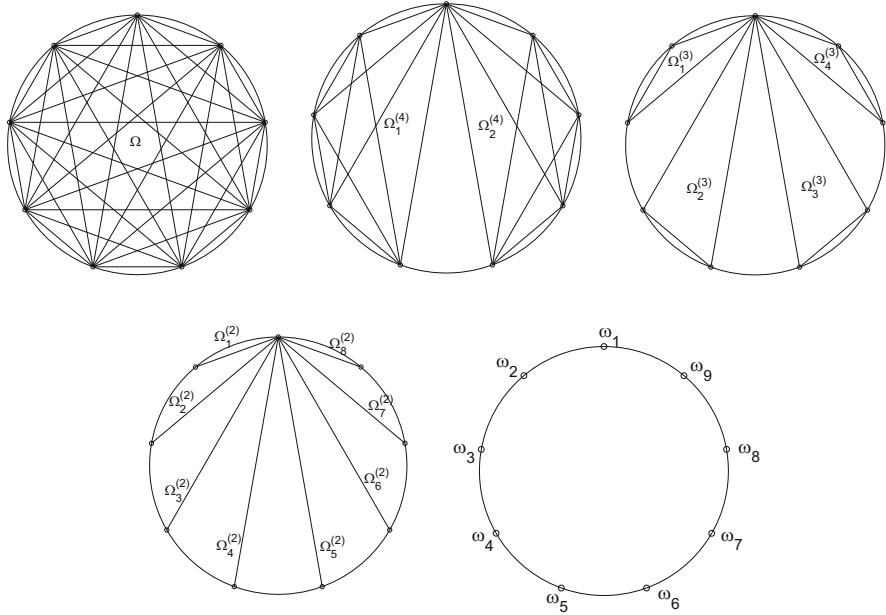


Fig. 6.2 A refinement chain with just one scenario fixed. The collection $(\Omega_1^{(2)}, \dots, \Omega_8^{(2)})$ is a “pairs of scenarios” collection

$$\Omega_1^{(\cdot)} = \{\omega_1, \dots, \omega_f, \omega_{f+1}, \dots, \omega_{f+x}\}, \Omega_2^{(\cdot)} = \{\omega_1, \dots, \omega_f, \omega_{f+x+1}, \dots, \omega_{f+2x}\},$$

or in general

$$\Omega_i^{(\cdot)} = \{\omega_1, \dots, \omega_f, \omega_{f+(i-1)x+1}, \dots, \omega_{f+i.x}\}, \quad i = 1, \dots, s.$$

The probabilities $P_i^{(\cdot)}$, as well as the corresponding weights $\pi_i^{(\cdot)}$ can be calculated as follows:

$$P_i^{(\cdot)} = \sum_{m=1}^f P(\omega_m) \cdot \delta_{\omega_m} + \frac{(1 - \sum_{m=1}^f P(\omega_m))}{\sum_{m=f+(i-1)x+1}^{f+i.x} P(\omega_m)} \sum_{m=f+(i-1)x+1}^{f+i.x} P(\omega_m) \cdot \delta_{\omega_m}$$

and

$$\pi_i^{(\cdot)} = \frac{\sum_{m=f+(i-1)x+1}^{f+i.x} P(\omega_m)}{(1 - \sum_{m=1}^f P(\omega_m))}; \quad i = 1, \dots, m.$$

With this choice one gets $\sum_{m=1}^s \pi_i^{(\cdot)} = 1$ and $P = \sum_{i=1}^s \pi_i^{(\cdot)} P_i^{(\cdot)}$, and therefore the corresponding lower bound is

$$\sum_{i=1}^s \pi_i^{(\cdot)} v^*(P_i^{(\cdot)}) \leq v^*(P).$$

Notice that (6.8) is the special case of $f = 1$ and $r = 2$. Figure 6.3 shows an example with two scenarios fixed.

Example 6.13. Alternatively one may also consider disjoint partitions

$$\Omega = \bigcup_i \Omega_i^{(j)} \quad \text{with} \quad \Omega_{i_1}^{(j)} \cap \Omega_{i_2}^{(j)} = \emptyset \quad \text{for} \quad i_1 \neq i_2.$$

In this case, the refinement chain corresponds to a filtration (which might be chosen different from the filtration \mathfrak{F}) and the probabilities $P_i^{(j)}$ are given by

$$P_i^{(j)} = \sum_{\omega_s \in \Omega_i^{(j)}} \pi_i^{(j)} \delta_{\omega_s},$$

where the weights $\pi_i^{(j)}$ are

$$\pi_i^{(j)} = \sum_{\omega_s \in \Omega_i^{(j)}} P(\omega_s).$$

Figure 6.4 illustrates such a chain.

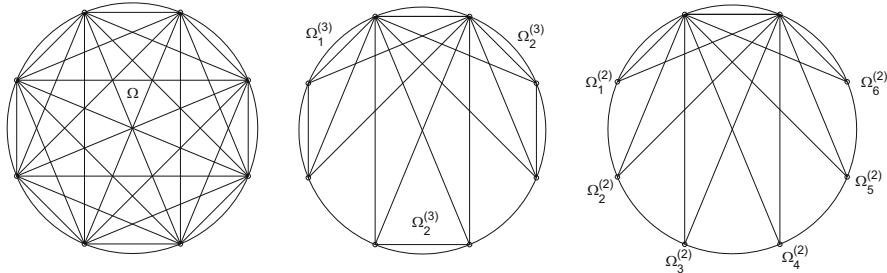


Fig. 6.3 A refinement chain with two scenarios fixed

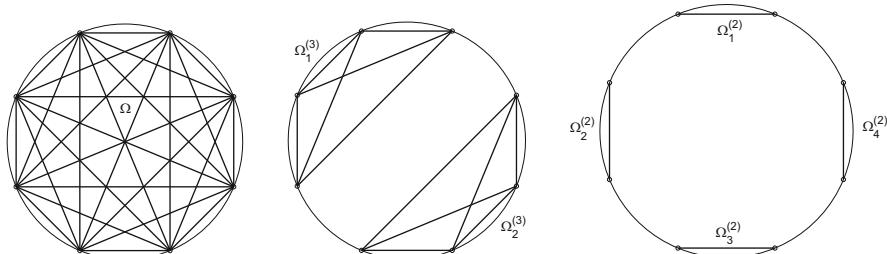


Fig. 6.4 A refinement chain with disjoint subsets

For more material on lower bounds see Maggioni and Pflug [77].

6.3.2 Lower Bounds for Replacing the Scenario Process by Its Expectation

A popular choice for replacing a stochastic dynamic problem by a deterministic one is to replace the scenario process $(\xi_t)_{t=1,\dots,T}$ by its expectation $(\mathbb{E}[\xi_t])_{t=1,\dots,T}$, if it exists. The basic problem (6.5) is replaced by

$$\text{Opt}(\mathbb{E}_P) : \min \{\mathcal{R}[Q(x, \mathbb{E}_P \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}, (\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}\}. \quad (6.9)$$

We call problem (6.9) the *expectation-reduced* problem. Notice that we have replaced the stochastic scenario process by a deterministic one, but kept the possibility for the decisions to depend on the filtration. One might argue that such *random decisions for nonrandom scenarios* are not appropriate. A closer look reveals in fact that the optimal decision for deterministic compound concave problems can be chosen as deterministic as well, i.e., it is not necessary to consider randomized decisions.

Theorem 6.14. *Suppose that the risk functional \mathcal{R} is compound concave (i.e., $P \mapsto \mathcal{R}_P$ is concave). Then the solution of the expectation-reduced problem (6.9) can be found for a deterministic decision sequence $(x_0^*, \dots, x_{T-1}^*)$.*

Proof. Notice that the expectation-reduced problem has a deterministic objective $\bar{Q}(x) = Q(x, \mathbb{E}\xi)$. Dissect the probability P into the atoms sitting on the leaves $P = \sum_i p_i P^{v_T=\omega_i}$. Then for any random x , $\mathcal{R}_P[\bar{Q}(x)] \geq \sum_i p_i \mathcal{R}_{P^{v_T=\omega_i}}[\bar{Q}(x)]$. Thus the optimal decision can be found by solving the clairvoyant problem. But since all clairvoyant problems are equal, the same solution may be inserted for all scenarios, i.e., the solution may be chosen deterministic as well. \square

Let $\mathbb{P}_{\mathbb{E}(\xi)}$ be the degenerate (nested) distribution, which assigns probability one to the sequence $\mathbb{E}\xi_1, \dots, \mathbb{E}\xi_T$. Under the assumption of Theorem 6.14 one may assume, without loss of generality, that the solution of the deterministic problem

$$\begin{aligned} v^*(\mathbb{P}_{\mathbb{E}(\xi)}) &= \min \{\mathcal{R}_{\mathbb{P}_{\mathbb{E}(\xi)}}[Q(x, \xi)] : x \triangleleft \mathfrak{F}, x \in \mathbb{X}\} \\ &= \min \{Q(x, \mathbb{E}_P \xi) : x \triangleleft \mathfrak{F}, x \in \mathbb{X}\} \end{aligned}$$

is deterministic itself. Here we have assumed that $\mathcal{R}(c) = c$ for all constants c .

Lemma 6.15. *Suppose that*

- (i) $\xi \mapsto Q(x, \xi)$ is convex for all x and that

(ii) \mathcal{R} is monotonic (property (M) in Definition 3.2) and antitonic with respect to negative second order stochastic dominance, see Definition 3.9.⁵

Then

$$v^*(\mathbb{P}_{\mathbb{E}(\xi)}) \leq v^*(\mathbb{P}).$$

Proof. Let V be a convex and monotonic utility function. Then $\xi \mapsto V(Q(x, \xi))$ is convex for all x and therefore $\mathbb{E}[V(Q(x, \xi))] \geq V(Q(\mathbb{E}\xi, x))$. Let $U(y) = -V(-y)$. Then $\mathbb{E}[U(-Q(x, \xi))] \leq U(-Q(\mathbb{E}\xi, x))$, i.e., $-Q(x, \xi) \prec_{SSD} -Q(x, \mathbb{E}\xi)$. By assumption $\mathcal{R}[Q(x, \xi)] \geq \mathcal{R}[Q(x, \mathbb{E}\xi)]$ and this implies the assertion. \square

Example 6.16. Recall that the linear multistage stochastic decision model is formulated as

$$\min \left\{ c_0 x_0 + \mathbb{E} \left[\min c_1(\omega) x_1 + \mathbb{E} \left[\min c_2(\omega) x_2 \right. \right. \right. \\ \left. \left. \left. + \mathbb{E} [\cdots + \mathbb{E} [\min c_T(\omega) x_T] \dots] \right] \right] : x \triangleleft \mathfrak{F}, x \in \mathbb{X} \right\},$$

where the feasible set \mathbb{X} is given by

$$\begin{aligned} Ax_0 &\geq h_0, \\ A_1(\omega) x_0 + W_1(\omega) x_1 &\geq h_1(\omega), \quad x_1 \triangleleft \mathcal{F}_1, \\ A_2(\omega) x_1 + W_2(\omega) x_2 &\geq h_2(\omega), \quad x_2 \triangleleft \mathcal{F}_2, \\ &\vdots \quad \vdots \\ A_T(\omega) x_{T-1} + W_T(\omega) x_T &\geq h_T(\omega), \quad x_T \triangleleft \mathcal{F}_T. \end{aligned}$$

Introducing the indicator function

$$\mathbb{I}_{\mathbb{X}}(x) = \begin{cases} 0 & \text{if } x \in \mathbb{X}, \\ \infty & \text{otherwise} \end{cases}$$

one may identify the cost function $Q(x, \xi)$ as

$$\begin{aligned} Q(x, \xi) &= c_0(\omega) x_0 + c_1(\omega) x_1 + \cdots + c_T(\omega) x_T + \mathbb{I}_{Ax_0 \geq h_0} \\ &\quad + \mathbb{I}_{A_1(\omega) x_0 + W_1(\omega) x_1 \geq h_1(\omega)} + \cdots + \mathbb{I}_{A_T(\omega) x_{T-1} + W_T(\omega) x_T \geq h_T(\omega)}. \end{aligned}$$

⁵It was shown in Lemma 3.10 that all distortion functionals are antitonic with respect to negative second order stochastic dominance.

Here the scenario process is composed of the four components $\xi_t(\omega) = [c_t(\omega), A_t(\omega), W_t(\omega), h_t(\omega)]$. The constraints are called to have only random right-hand side, if A and W are deterministic and only h depends on randomness.

Lemma 6.17. *If the model has only random right-hand side, then $\xi \mapsto Q(x, \xi)$ is convex.*

Proof. In this case $\xi_t = [c_t, h_t]$. The function $c \mapsto c_0 x_0 + \dots + c_T x_T$ is even linear and $h \mapsto \sum_t \mathbb{I}_{\eta_t \geq h_t}$ is convex for fixed η_t . \square

6.3.3 Bounds for Changing the Filtration

As we have seen in the previous section,

$$v^*(\mathfrak{F}^T, P) \leq v^*(\mathfrak{F}, P)$$

with $\mathfrak{F}^T = (\mathcal{F}_T, \dots, \mathcal{F}_T)$. This is just one example of the fundamental but quite trivial inequality

$$\mathfrak{F}_1 \subseteq \mathfrak{F}_2 \quad \text{implies that} \quad v^*(P, \mathfrak{F}_2) \leq v^*(P, \mathfrak{F}_1). \quad (6.10)$$

This inequality can produce lower and upper bounds: by refining the filtration, i.e., relaxing the nonanticipativity constraints, one gets upper bounds and by coarsening it, one gets lower bounds.

6.3.4 Upper Bounds by Inserting (Sub)Solutions

Denote by $v(\mathbb{P}, x)$ the value of the objective, when the decision x is inserted

$$v(\mathbb{P}, x) := \mathcal{R}_{\mathbb{P}}[Q(x, \cdot)],$$

then it always holds that

$$v(\mathbb{P}, x) \geq v^*(\mathbb{P}), \quad x \in \mathbb{X}.$$

Inserting a fixed decision $x_{0:T}^+$ for times $0, \dots, t - 1$ as fixed and optimizing only the decisions for times t, \dots, T is denoted by

$$v^*(\mathbb{P}, x_{0:t-1}^+) := \min \{ \mathcal{R}_{\mathbb{P}}[Q([x_{0:t-1}^+, x_{t:T}], \cdot)] : x_{t:T} \prec \mathfrak{F}_{t:T} \}.$$

Trivially,

$$v(\mathbb{P}, x_{0:T}^+) \geq v^*(\mathbb{P}, x_{0:t-1}^+) \geq v^*(\mathbb{P}).$$

6.4 Martingale Properties

Rolling horizon solutions (see Chap. 5) induce (super)martingale properties as it will be seen below: to each value of the tree process $v_t = u$, one associates a subproblem in the following way: the full problem is solved and its solution is $x^*(\mathbb{P})$. Then the solution up to time t is inserted into the conditional problem

$$v_t^*(u) = \min \left\{ \mathcal{R}_{\mathbb{P}^{v_t=u}} [Q([x_{0:t-1}^*, x_{t:T}], \cdot)] : x_{t:T} \triangleleft \mathfrak{F}_{t:T} \right\}. \quad (6.11)$$

Notice that the random variables $v_t^* = v_t^*(v_t)$ are \mathcal{F}_t -measurable. Notice also that the overall optimal value is $v_0^* = v^*(\mathbb{P})$. Recall that a functional \mathcal{R} is called compound concave if the mapping $P \mapsto \mathcal{R}_P(Y)$ is concave (property (CC) in Appendix A) for all random variables Y for which \mathcal{R} is defined. Alternatively one may define compound concavity by the validity of $\mathcal{R}(Y) \geq \mathbb{E}[\mathcal{R}(Y | \mathcal{F})]$ for all random variables Y and any σ -algebra \mathcal{F} (see Lemma 5.9).

Lemma 6.18. *If \mathcal{R} is compound concave, then the stochastic process $(v_0^*, v_1^*, \dots, v_{T-1}^*)$ is a supermartingale. If \mathcal{R} is the expectation, then this process is a martingale.*

Proof. Notice that $\mathbb{P}^{v_t=u}$ is the convex combination of $\mathbb{P}^{v_{t+1}=w}$, in symbols

$$\mathbb{P}^{v_t=u} = \int \mathbb{P}^{v_{t+1}=w} P_t^u(dw),$$

where P_t^u is the conditional distribution of v_{t+1} given that $v_t = u$. Consequently, by compound concavity for every x ,

$$\mathcal{R}_{\mathbb{P}^{v_t=u}} [Q(x, \xi)] \geq \mathbb{E}_{P_t^u} (\mathcal{R}_{\mathbb{P}^{v_{t+1}=w}} [Q(x, \xi)]). \quad (6.12)$$

Inserting $x^+ = [x_{0:t-1}^*(\mathbb{P}), x_{t:T}^+]$, an optimal solution of (6.11), one gets

$$\begin{aligned} v_t^* &= \mathcal{R}_{\mathbb{P}^{v_t=u}} [Q(x^+, \xi)] \geq \mathbb{E}_{P_t^u} (\mathcal{R}_{\mathbb{P}^{v_{t+1}=w}} [Q(x^+, \xi)]) \\ &\geq \mathbb{E}_{P_t^u} \min \left\{ \mathcal{R}_{\mathbb{P}^{v_{t+1}=w}} [Q([x_{0:t}^+, x_{t+1:T}], \cdot)] : x_{t+1:T} \triangleleft \mathfrak{F}_{t+1:T} \right\} \\ &= \mathbb{E}_{P_t^u} (v_{t+1}^*). \end{aligned} \quad (6.13)$$

If $\mathcal{R} = \mathbb{E}$, then (6.12) is an equality. Moreover, since for the expectation problem the optimal solutions of all subproblems coincide, equality holds in (6.13). \square

Chapter 7

The Problem of Ambiguity in Stochastic Optimization

In this chapter we consider again the basic multistage decision problem

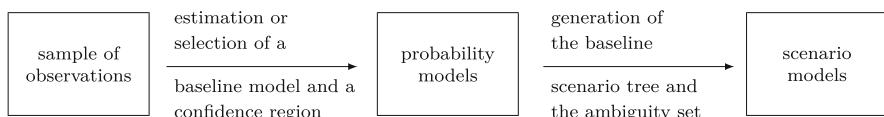
$$\min \{\mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}\}, \quad (7.1)$$

or in extensive form

$$\min \{\mathcal{R}_{\mathbb{P}}[Q(x_0, \dots, x_{T-1}, \xi_1, \dots, \xi_T)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}\},$$

where Q is the cost function, \mathbb{X} is the feasible set of decision functions, $\mathcal{R}_{\mathbb{P}}$ is a risk functional, and \mathbb{P} is the probability model (a nested distribution). As usual, $x \triangleleft \mathfrak{F}$ denotes the nonanticipativity conditions in the multistage case.

It is mainly the exception and not the rule that the model \mathbb{P} is fully known. Data driven model selection and parameter estimation are statistical techniques which are subject to errors. Even if statistical estimation or model selection results in a unique model (a point estimate) one feels that typically a set of models (i.e., confidence sets) is compatible with the data observations and this could and should be taken into account for the decision making process. Reconsidering the steps necessary to go from observed data to the approximate tree model (see Introduction, Sect. 1.2.1) one may identify a baseline model by statistical model selection, but in addition a confidence set of models, called the ambiguity set.



Following Ellsberg [42] we distinguish between the

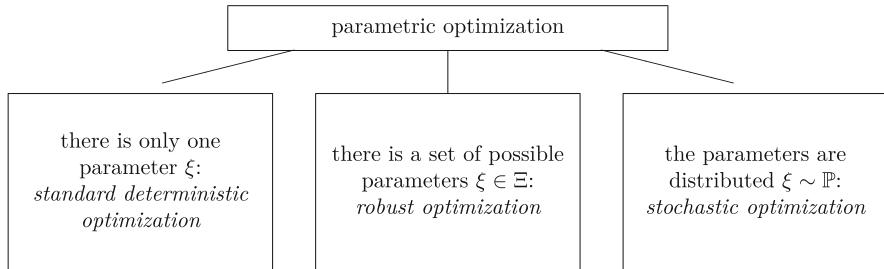
- *uncertainty problem*, if the model is fully known, but the realizations of the random variables are unknown; and the
- *ambiguity problem*, if the probability model itself is unknown. Another name for ambiguity is *Knightian uncertainty* (referring to F. Knight's 1921 book [68]).

In the ambiguity formulation there is not just one probability model, but a set of models \mathcal{P} (the ambiguity set), which are all possible descriptions of the reality. Typically \mathcal{P} is a certain neighborhood of $\hat{\mathbb{P}}$, the most appropriate model. To safeguard against the possible deviation from the baseline probability model we extend the baseline problem (7.1) to the ambiguity problem

$$\min \{ \max \{ \mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : \mathbb{P} \in \mathcal{P} \} : x \in \mathbb{X} \}, \quad (7.2)$$

which is of minimax type. An optimal solution of the ambiguity problem (7.2) is called a *distributionally robust solution*.

Robustness. The word *robustness* is used in different ways. A specification of this notion in our context is necessary. Minimizing $x \mapsto Q(x, \xi)$ for a fixed ξ is a standard optimization problem, minimizing $x \mapsto \max \{Q(x, \xi) : \xi \in \Xi\}$ is an ordinary robust optimization problem. If ξ is a random variable, then minimizing $\mathcal{R}_{\mathbb{P}}[Q(x, \xi)]$ in x is a standard stochastic optimization problem. This is illustrated below.



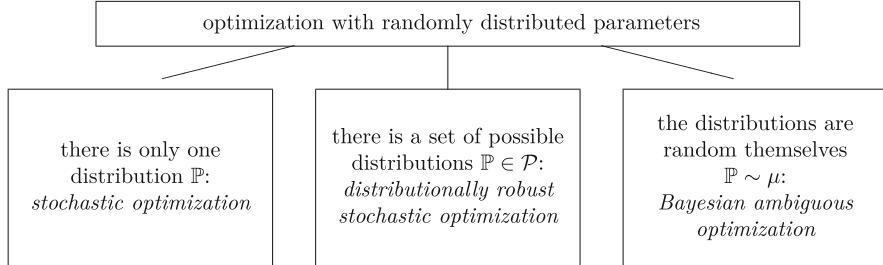
In the ambiguity extension, the new objective is

$$x \mapsto \max \{ \mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : \mathbb{P} \in \mathcal{P} \}.$$

The stochastic ambiguity problem contains both aspects: while assuming that the realizations of the parameter vectors come from some probability distribution, we allow, on the other hand, to vary this distribution within a certain set \mathcal{P} without further structuring it. Imposing a probability distribution on this set of probabilities (called a *prior distribution*) we would still solve an uncertainty problem, but of Bayesian type

$$x \mapsto \mathbb{E} \{ \mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : \mathbb{P} \sim \mu \},$$

where μ is the Bayesian prior distribution on the set of models. See the following illustration:



The ambiguity problem (7.2), however, does not specify a prior and therefore it has the structure of a *distributionally robust stochastic* problem. It is a combination of a robust and a stochastic problem (see Žáčková [138] for the earliest occurrence of problems of this form).

Convexity. Replacing the objective $\mathcal{R}_{\mathbb{P}}[Q(x, \xi)]$ by $\max_{\mathbb{P} \in \mathcal{P}} \mathcal{R}_{\mathbb{P}}[Q, x, \xi]$ does not destroy the convexity according to the following lemma.

Lemma 7.1. *Assume that*

- (i) $x \mapsto Q(x, \xi)$ is convex for every ξ ,
- (ii) $Y \mapsto \mathcal{R}_{\mathbb{P}}(Y)$ is monotonic (i.e., $Y_1 \leq Y_2$ implies that $\mathcal{R}_{\mathbb{P}}(Y_1) \leq \mathcal{R}_{\mathbb{P}}(Y_2)$, property (M) in Appendix A), and
- (iii) $Y \mapsto \mathcal{R}_{\mathbb{P}}(Y)$ is convex (property (C) in the Appendix).

Then both, $x \mapsto \mathcal{R}_{\mathbb{P}}[Q(x, \xi)]$ and $x \mapsto \max \{\mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : \mathbb{P} \in \mathcal{P}\}$, are convex functions.

Proof. The first assertion follows from

$$\begin{aligned} \mathcal{R}_{\mathbb{P}}[Q((1 - \lambda)x_0 + \lambda x_1, \cdot)] &\leq \mathcal{R}_{\mathbb{P}}[(1 - \lambda)Q(x_0, \cdot) + \lambda Q(x_1, \cdot)] \\ &\leq (1 - \lambda)\mathcal{R}_{\mathbb{P}}[Q(x_0, \cdot)] + \lambda\mathcal{R}_{\mathbb{P}}[Q(x_1, \cdot)]. \end{aligned}$$

Since the maximum of convex functions is convex, the second assertion follows as well. \square

Stress Testing. Let x^* be the optimal solution of the baseline problem (7.1). The robustness with respect to deviation from the assumed probability model of this solution is often checked by *stress testing*. A stress test consists in defining alternative probability models $\bar{\mathbb{P}}$ and calculating $\mathcal{R}_{\bar{\mathbb{P}}}[Q(x^*, \bar{\mathbb{P}})]$ in order to judge the change of the objective value under model variation, when the decision $x^*(\bar{\mathbb{P}})$, the optimal decision for the model $\hat{\mathbb{P}}$, is kept fixed. For a given ambiguity set \mathcal{P} , stress testing looks at the worst model, i.e., $\operatorname{argmax} \{\mathcal{R}_{\bar{\mathbb{P}}}[Q(x^*, \bar{\mathbb{P}})] : \bar{\mathbb{P}} \in \mathcal{P}\}$ and therefore helps in assessing the robustness of a given baseline decision under model error. Notice, however, that the distributionally robust case is quite different

from pure stress testing. While pure stress testing considers the worst case among some stress scenarios *while the decisions are kept fixed*, the distributionally robust decisions are *minimax decisions taken in view of the ambiguity set* and are therefore good decision compromises for all models in the ambiguity set.

Bibliographical Remarks. In his well-known 1961 paper on risk and ambiguity Ellsberg [42] writes:

We shall suppose that by compounding various probability judgements of varying degrees of reliability the decision maker can eliminate certain probability distributions over the states of nature as unreasonable . . . and arrive at an estimated distribution P that represents his available information. But let us assume that the situation is ambiguous for him. Out of the set \mathcal{P} of all possible distributions there remains a set \mathcal{P}_0 that still seem reasonable, reflecting judgements he might almost as well have made. . . It might now occur to him to ask: “What might happen to me if my best estimates (P) do not apply? What is the worst of the reasonable distributions of pay-off that I might associate with my decision?”

This more subjective notion of ambiguity was followed later by linking ambiguity sets to statistical model uncertainty. The idea of optimal decisions under several stochastic models (i.e., min–max solutions) appears for the first time in Scarf [121] in a linear inventory problem seeking the stockage policy which maximizes the minimum profit considering all demand distributions with given mean and given standard deviation.

More thorough studies of ambiguous decision problems as minimax problems were initiated by Žáčková (a.k.a. Dupačová) [32, 33, 138] for the class of stochastic linear problems with recourse under general assumptions for the ambiguity set. The formulation was in a game theoretic setup, where the first player chooses the decision and the second player chooses the probability model.

There are alternative names used in literature for the ambiguity problem, such as *minimax stochastic optimization*, *model uncertainty problem*, or *distributional robustness problem*.¹

Many proposals for ambiguity sets in the two-stage case have been made and analyzed. A list of popular classes of ambiguous models is presented in Dupačová [35]. There is a fast growing literature dealing with ambiguity either from theory or application point of view. Some authors consider only the fixed support case, where the possible values ξ_1, \dots, ξ_n are fixed and only their probabilities p_1, \dots, p_n are subject to ambiguity, others consider more general models for the probability measures P . We list different approaches here:

- The case when the ambiguity set consists of all probabilities with given first two moments was studied by Jagannathan [61] for the linear case.
- Shapiro and Kleywegt [130] define an ambiguity set as the convex hull of a finite collection of models P_i ,

¹A related problem is that of stability, which studies the dependency of the optimal solution on the probability model, i.e., the mapping $P \mapsto x^*(P)$ (see, e.g., [106, 109, 115]; the latter paper uses probability metrics for defining neighborhoods of models).

$$\mathcal{P} = \left\{ P : P = \sum_{i=1}^n \lambda_i P_i, \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0 \right\}.$$

- Ahmed and Shapiro [128] consider the following set of models

$$\mathcal{P} = \left\{ P \text{ is a probability s.t. } \mu_1 \prec P \prec \mu_2, \right. \\ \left. \int \phi_i \, dP = b_i, i = 1, \dots, k, \int \psi \, dP \leq c_i, i = 1, \dots, \ell \right\},$$

where $\mu_1 \prec \mu_2$ means that $\mu_1(A) \leq \mu_2(A)$ for all measurable sets A . In order to allow P to be probability measures, μ_1 must be a positive measure with total mass smaller than 1 and μ_2 must have mass larger than 1.

- Calafiore [15] uses the Kullback–Leibler divergence to define

$$\mathcal{P}_\epsilon = \left\{ (p_1, \dots, p_n) : \sum_{i=1}^n p_i \log \frac{p_i}{\hat{p}_i} \leq \epsilon \right\}.$$

- Thiele [133] considers the set

$$\mathcal{P} = \{(p_1, \dots, p_n) : |p_i - \hat{p}_i| \leq \epsilon_i\}.$$

- Delage and Ye [21] consider the ambiguity set

$$\mathcal{P} := \left\{ P \left| \begin{array}{l} (\int x \, P(dx) - c)^\top \Sigma_0^{-1} (\int x \, P(dx) - c) \leq \gamma_1 \text{ and} \\ \int (x - c)(x - c)^\top P(dx) \preceq \gamma_2 \Sigma_0 \end{array} \right. \right\}.$$

Here, $\Sigma_1 \preceq \Sigma_2$ means that $\Sigma_2 - \Sigma_1$ is a positive definite matrix and Σ_0 is a given covariance matrix. Therefore \mathcal{P} is defined by a conical constraint.

- Edirshinge [39] considers ambiguity sets, which are defined by a finite number of generalized moment equalities,

$$\mathcal{P} = \left\{ P : \int f_i \, dP = c_i, i = 1, \dots, n \right\}.$$

- Wozabal and Pflug [98] use for the first time ambiguity sets, which are balls with respect to the Wasserstein distance.

Up to now, only little work was done for the multistage case.

- Hansen and Sargent (Nobel prize 2011) consider in their 2007 book [50] dynamic stochastic optimization problems from the viewpoint of economists. They consider alternative models given by maximal deviation from the baseline model in Kullback–Leibler divergence.

- Goh and Sim [47] study ambiguity sets which are defined by a mean, which must lie in some conical set, a given covariance matrix and some upper bounds on the exponential moments and extend this to multistage.

In the next sections we deal with decision problems, where ambiguity sets are defined via the nested distance.

7.1 Single- or Two-Stage Models: Wasserstein Balls

We begin with single- or two-stage problems, since for these problems it suffices to consider ordinary (and not nested) distributions P . For single-stage problems there is no need to consider filtrations.

The scenario variable ξ in

$$\min \{\mathcal{R}_P[Q(x, \xi)] : x \in \mathbb{X}\}$$

takes values in \mathbb{R}^m . Since only the image measure of ξ on \mathbb{R}^m matters, we agree that the basic probability space is \mathbb{R}^m with the image measure P , together with the Borel sigma algebra or a smaller sigma algebra \mathcal{F} .

Let $(\mathbb{R}^m, \mathcal{F}, \hat{P})$ be the baseline scenario model, let $(\mathbb{R}^m, \mathcal{F}, P)$ be an alternative distribution, and let $d(\cdot, \cdot)$ be a distance function in \mathbb{R}^m , typically $d(u, v) = \|u - v\|$ for some norm $\|\cdot\|$. Recall that the Wasserstein distance $d_r(\hat{P}, P)$ is defined as

$$\begin{aligned} d_r(P, \hat{P}) \\ = \inf \left\{ \left[\iint d(u, v)^r \pi(du, dv) \right]^{1/r} : \pi(\hat{A} \times \mathbb{R}^m) = \hat{P}(\hat{A}), \pi(\mathbb{R}^m \times A) = P(A) \right\}, \end{aligned}$$

where the infimum is taken over all feasible transportation plans π , which are probability measures on $(\mathbb{R}^m \times \mathbb{R}^m, \hat{\mathcal{F}} \otimes \mathcal{F})$ with given marginals \hat{P} and P (cf. Definition 2.5).

Each transportation plan π defines uniquely the conditional probability $K(\cdot|u)$ by the relation

$$\int_{\hat{A}} K(A|u) \hat{P}(du) = \pi(\hat{A} \times A), \quad A \in \mathcal{F}, \hat{A} \in \hat{\mathcal{F}}.$$

We call K a *transportation kernel*. Notice that

$$\pi(\mathbb{R}^m \times A) = P(A) = \int_{\mathbb{R}^m} K(A|u) \hat{P}(du). \tag{7.3}$$

We use the short notation $P = \hat{P} \circ K$ for (7.3).² Using the transportation kernel the Wasserstein distance may also be written as

$$\mathbf{d}_r(\hat{P}, P)^r = \inf \left\{ \iint \mathbf{d}(u, v)^r K(dv|u) \hat{P}(du) : P = \hat{P} \circ K \right\}.$$

Here, the infimum is taken among all transportation kernels, i.e., kernels $K \geq 0$ which satisfy

- $A \mapsto K(A|u)$ is a probability measure on $(\mathbb{R}^m, \mathcal{F})$ for all $u \in \mathbb{R}^m$ and
- $u \mapsto K(A|u)$ is $\hat{\mathcal{F}}$ -measurable for all $A \in \mathcal{F}$.

The Wasserstein ball $\mathcal{B}(\hat{P}, \epsilon) := \{P : \mathbf{d}_r(\hat{P}, P) \leq \epsilon\}$ can also be represented as

$$\begin{aligned} & \mathcal{B}(\hat{P}, \epsilon) \\ &= \left\{ \hat{P} \circ K : K \text{ is a transportation kernel such that } \iint \mathbf{d}(u, v)^r K(dv|u) \hat{P}(du) \leq \epsilon^r \right\}. \end{aligned}$$

For the ambiguity set $\mathcal{B}(\hat{P}, \epsilon)$ being a Wasserstein ball, the ambiguity problem (7.2) reads

$$\min_x \max_K \left\{ \mathcal{R}_{\hat{P} \circ K}[Q(x, \xi)] : \iint \mathbf{d}(u, v)^r K(dv|u) \hat{P}(du) \leq \epsilon^r \right\}, \quad (7.4)$$

where the infimum is over all appropriate transportation kernels.

The minimax problem (7.4) is much too involved to allow a practical solution, since the maximum runs over all possible transportation kernels without any restriction on the support of the image measure P . Possible simplifications include the following cases:

- **The finite support case.** Assume that the probability measure \hat{P} is discrete and sits on n points $\hat{\xi}_1, \dots, \hat{\xi}_n$. In the finite support case, the Wasserstein ball is restricted to alternative probability measures, which sit on at most n points too,

$$\mathcal{B}(\hat{P}, \epsilon) = \left\{ P = \sum_{j=1}^n P(j) \delta_{\xi_j} : \mathbf{d}_r(\hat{P}, P) \leq \epsilon \right\}. \quad (7.5)$$

The alternative models in (7.5) are characterized by the choice of the new points (ξ_j) , $j = 1, \dots, n$ and the new probabilities $P(j)$, $j = 1, \dots, n$. The constraint $\mathbf{d}_r(\hat{P}, P) \leq \epsilon$ can be reformulated as: there exists a joint probability π such that

²The notation refers to the finite case, where P is a row vector and the transportation kernel K is a Markovian transition matrix but may of course be extended to the general case.

$$\begin{aligned} \sum_{i,j} \pi_{i,j} d(\hat{\xi}_i, \xi_j)^r &\leq \varepsilon^r, \\ \sum_j \pi_{i,j} &= \hat{P}(i) \text{ and} \\ \sum_i \pi_{i,j} &= P(j). \end{aligned} \tag{7.6}$$

Notice that the variables in the optimization problem (7.6) are ξ_j and $\pi_{i,j}$. Alternatively, using the notion of kernels K , the constraint may be written as: there exists a transportation kernel K , such that

$$\begin{aligned} \sum_{i,j} \hat{P}(i) K(j|i) d(\hat{\xi}_i, \xi_j)^r &\leq \varepsilon^r, \\ \sum_j K(j|i) &= 1, K \geq 0 \text{ and} \\ \sum_i \hat{P}(i) K(j|i) &= P(j). \end{aligned} \tag{7.7}$$

Here, the variables are ξ_j and $K(j|i)$. Because of the products between the kernel and the distance in (7.6) or (7.7), the constraint set is nonconvex. For smaller sizes of the finite probability space it is possible to solve the ambiguity problem by DC (difference of convex) functions, see Wozabal [143]. However, the computational effort is quite considerable.

- **The fixed support case.** To simplify the ambiguity problem further we assume that the baseline model is discrete as before,

$$\hat{P} = \sum_{i=1}^n \hat{P}(i) \delta_{\xi_i},$$

but assume in addition that the alternative models P sit on the same n points. Only the baseline probabilities $\hat{P}(i)$ are reweighted to a new probability measure

$$P = \sum_{i=1}^n P(i) \delta_{\xi_i},$$

such that $d_r(\hat{P}, P) \leq \epsilon$. Notice that since the distances $d(\xi_i, \xi_j)$ are fixed in (7.7), the objective function (7.7) is linear in K and the constraint set is polyhedral. For this reason the fixed support case is much easier to handle than the general case. It is also known under the name of *stress test case*, since stress tests are defined as probability models where “bad” scenarios have increased probability and the probability of “good” scenarios is reduced.

Example 7.2. Pflug and Wozabal consider a single-stage portfolio optimization problem under ambiguity in [98]. They identify distributionally robust portfolio decisions under ambiguity sets which are Wasserstein balls. Pflug, Pichler, and Wozabal [87] showed that by increasing the radius of the ambiguity sets, the optimal portfolio decisions tend to equal weights for all investment possibilities.

Remark 7.3. How large should ϵ be chosen? An answer can be found in Sect. 4.1.1, where confidence sets for the distance between the true probability and the empirical measure (which is a nonparametric estimate) are given. If the discrete measure \hat{P} is generated via an estimate $\hat{\theta}$ from a parametric model P_θ , then the situation is more complicated. The confidence set is indirectly determined through a confidence set for the parameter estimates $C = \{\theta : P_\theta(\|\theta - \hat{\theta}\| \leq \eta) \geq 1 - \alpha\}$ and the relation between the parameter and the Wasserstein distance

$$\epsilon = \sup \{d(P_{\hat{\theta}}, P_\theta) : \theta \in C\}.$$

7.2 Solution Methods for the Single- or Two-Stage Case

Consider the single-stage or two-stage problem with fixed values. That is, a finite number of possible scenario values ξ_1, \dots, ξ_n is kept fixed and only the probabilities vary. For two probability measures

$$P^{(1)} = \sum_{i=1}^n P_i^{(1)} \delta_{\xi_i} \quad \text{and} \quad P^{(2)} = \sum_{i=1}^n P_i^{(2)} \delta_{\xi_i},$$

compounding $P^{(1)}$ and $P^{(2)}$ with compound probability λ is equivalent to forming the convex combination

$$\lambda P^{(1)} + (1 - \lambda) P^{(2)} = \sum_{i=1}^n (\lambda P_i^{(1)} + (1 - \lambda) P_i^{(2)}) \delta_{\xi_i}.$$

Consequently, the Wasserstein ball

$$\mathcal{P} = B_r(\hat{P}, \varepsilon) = \left\{ P : d_r(P, \hat{P}) \leq \varepsilon \right\}$$

is convex in this case. Under the conditions of Lemma 7.1, the objective function $\mathcal{R}_P[Q(x, \xi)]$ is convex in x . If \mathcal{R} is convex-concave (see property (C-CC) in Appendix A), then $(x, P) \mapsto \mathcal{R}_P[Q(x, \xi)]$ is also convex-concave.

Consider first a general minimax problem in $\mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$. If $f(x, y)$ is convex-concave, i.e., convex in $x \in \mathcal{X} \subseteq \mathbb{R}^{m_1}$ and concave in $y \in \mathcal{Y} \subseteq \mathbb{R}^{m_2}$, where \mathcal{X} and \mathcal{Y} are convex sets, then the solution $x^* \in \mathcal{X}$ of the minimax problem

$\min_x \max_y f(x, y)$ is the first component of a saddle point. That is, there exists a $y^* \in \mathcal{Y}$ such that

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*)$$

for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Gradient / subgradient methods for finding a saddle point have been studied by extensively in literature since the 1970s [5, 22, 23]. Under convex-concavity and differentiability of f a necessary and sufficient condition for the solution is the validity of the simultaneous system of equations $\mathcal{E}(x, y) \equiv \begin{pmatrix} \nabla_x f(x, y) \\ \nabla_y f(x, y) \end{pmatrix} = 0$. Demyanov [23] and Danilin [20] proposed a gradient based saddle point algorithm based on direction d_k and step size strategy α_k such that sufficient progress at each iteration is ensured. Rustom and Howe [117] proposed to solve the problem

$$\min_{x,y} \left\{ \frac{1}{2} \|\mathcal{E}(x, y)\|_2^2 \right\}$$

rather than $\mathcal{E}(x, y) = 0$ using Quasi-Newton algorithms. Qi [104] introduces a quadratic approximation algorithm and Sasai [120] uses interior point saddle point algorithms.

In distributionally robust stochastic optimization, (sub-) gradient methods are less important, since the decision space \mathbb{X} and the model space \mathcal{P} are quite different in nature. We propose to use successive convex programming (SCP) (see Zillober et al. [145]), using a finitely generated inner approximation of the ambiguity set \mathcal{P} , see Algorithm 7.1.

Notice that $(u^{(k)})$ is an increasing sequence of numbers and $\mathcal{P}^{(k)}$ is an increasing set of models. The convergence of this algorithm is stated below. Since one cannot exclude that there are several saddle points (in this case the set of saddle points is closed and convex), only a weak limit result is available in general.

Proposition 7.4. *Assume that \mathcal{P} and \mathbb{X} are compact and convex and that $(x, P) \mapsto \mathcal{R}_P[Q(x, \xi)]$ is convex-concave and jointly continuous. Then every cluster point of $(x^{(k)})$ is a solution of (7.2). If the saddle point is unique, then the algorithm converges to the optimal solution.*

Proof. The proof follows from Proposition B.6 in Appendix B. □

7.3 The Multistage Case

The extension to multistage problems is more involved. The basic multistage problem is formulated as

Algorithm 7.1

The successive convex programming algorithm for the solution of the minimax problem

- **INITIALIZATION.** Set $k = 0$ and $\mathcal{P}^{(0)} = \{\hat{P}\}$ with $\hat{P} \in \mathcal{P}$.
- **OUTER OPTIMIZATION.** Solve the outer problem

$$\begin{array}{ll} \text{Minimize (in } x \text{ and } u) & u \\ \text{subject to} & \mathcal{R}_P[Q(x, \xi)] \leq u \text{ for all } P \in \mathcal{P}^{(k)}, \\ & x \in \mathbb{X} \end{array}$$

and call the solution $(x^{(k)}, u^{(k)})$. If the solution is not unique, choose any element of the solution set.

- **INNER OPTIMIZATION.** For the fixed $x^{(k)}$ solve the inner problem

$$\begin{array}{ll} \text{Maximize (in } P) & \mathcal{R}_P[Q(x^{(k)}, \xi)] \\ \text{subject to} & P \in \mathcal{P}, \end{array} \quad (7.8)$$

call the solution $P^{(k)}$ and let $\mathcal{P}^{(k+1)} := \mathcal{P}^{(k)} \cup \{P^{(k)}\}$. If the solution is not unique, choose any element of the solution set.

- **STOPPING CRITERION.** If

1. $\mathcal{P}^{(k+1)} = \mathcal{P}^{(k)}$ or
2. the optimal value of (7.8) equals $u^{(k)}$, then stop.

Otherwise set $k := k + 1$ and goto **OUTER OPTIMIZATION**.

In a practical implementation one may also stop if $u^{(k)} - u^{(k-1)} \leq \eta$ for some predefined η .

$$\min \{\mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}, (\Omega, \mathfrak{F}, P, \xi) \sim \mathbb{P}\},$$

where the probability model is given by the nested distribution \mathbb{P} .

As a natural generalization of the two-stage model consider a baseline model given by the nested scenario process distribution $\hat{\mathbb{P}} \sim (\Omega, \mathfrak{F}, \hat{P}, \hat{\xi})$ and alternative models $\mathbb{P} \sim (\Omega, \mathfrak{F}, P, \xi)$, which are nested distributions of the same depth and with the same filtration \mathfrak{F} (i.e., the same tree-structure). Notice that the decisions x for different models are then comparable, since they satisfy all $x \triangleleft \mathfrak{F}$. We define the set of ambiguity as the nested ball

$$\mathcal{B}_r(\hat{\mathbb{P}}, \varepsilon) := \left\{ \mathbb{P} : \mathbf{dL}_r(\hat{\mathbb{P}}, \mathbb{P}) \leq \varepsilon \right\}. \quad (7.9)$$

The multistage ambiguity problem is then

$$\min_x \max_{\mathbb{P}} \left\{ \mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F}, \mathbf{dL}_r(\hat{\mathbb{P}}, \mathbb{P}) \leq \varepsilon \right\}. \quad (7.10)$$

Problem (7.10) is quite complex, in particular nonconvex. In general, for the alternative models \mathbb{P} while the same number T of stages and equivalent filtrations,

the values of the scenario process and the probabilities may be different as long as the nested distance to the baseline model is smaller than ϵ .

While the given approach is in principle valid for arbitrary probability spaces, we restrict ourselves to finite trees. However, even for small trees problem (7.10) is quite difficult to solve because of size and nonconvexity.

For this reason we will consider only the fixed values case here.³ To this end, introduce the following notation: let \mathbb{T} denote a tree with given structure, which is valued by the scenario process ξ (meaning that the values of the process sit on the nodes of the tree). The leaf set (the scenarios) of \mathbb{T} is denoted by $\Omega = \mathcal{N}_T$. The scenario probabilities are $P = (P_i)_{i \in \mathcal{N}_T}$, where $\hat{P} = (\hat{P}_i)_{i \in \mathcal{N}_T}$ is the baseline model. The full tree, valued with the scenario values and the scenario probabilities, is denoted by $\mathbb{P}(\mathbb{T}, P)$. Notice that it would be inconsistent to define simply ambiguity sets as neighborhoods of P in the ℓ_r -distance, i.e.,

$$\left\{ P : \sum_{i \in \mathcal{N}_T} |P_i - \hat{P}_i|^r \leq \epsilon^r \right\} \quad (7.11)$$

as some authors propose it. The reason is that an ambiguity set of the form (7.11) does not respect the tree structure and is therefore not appropriate for the multistage model. In particular, a neighborhood as in (7.11) is not invariant with respect to topologically equivalent permutations of the tree.

Example 7.5. Consider the example presented in Fig. 7.1. In terms of the distance $d_1(P, \bar{P}) = \sum_{i \in \mathcal{N}_T} |P_i - \bar{P}_i|$, one finds that $d_1(P^{(1)}, \bar{P}) = 0.12 > 0.105 = d_1(P^{(2)}, \bar{P})$ but in terms of the nested distance $d_1(\mathbb{P}^{(1)}, \bar{\mathbb{P}}) = 0.12 < 1.2 = d_1(\mathbb{P}^{(2)}, \bar{\mathbb{P}})$. Therefore $\mathbb{P}^{(1)}$ is much closer to $\hat{\mathbb{P}}$ than $\mathbb{P}^{(2)}$ with respect to the nested distance. Since by Theorem 6.10 the nested distance can be viewed as the “correct” concept of nearness, the distance used in (7.11) is inappropriate to measure closeness.

As before, the ambiguity set is defined as a ball in the nested distance sense, i.e., we specify (7.9) to

$$B_\epsilon = \left\{ P : d_r(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \hat{P})) \leq \epsilon \right\}$$

and set the ambiguity set to

$$\mathcal{P}_\epsilon := \{ \mathbb{P}(\mathbb{T}, P) : P \in B_\epsilon \}. \quad (7.12)$$

Unfortunately, B_ϵ is typically nonconvex: for an example, where

³See also the paper [4] forthcoming in Computational Management Science 2014.

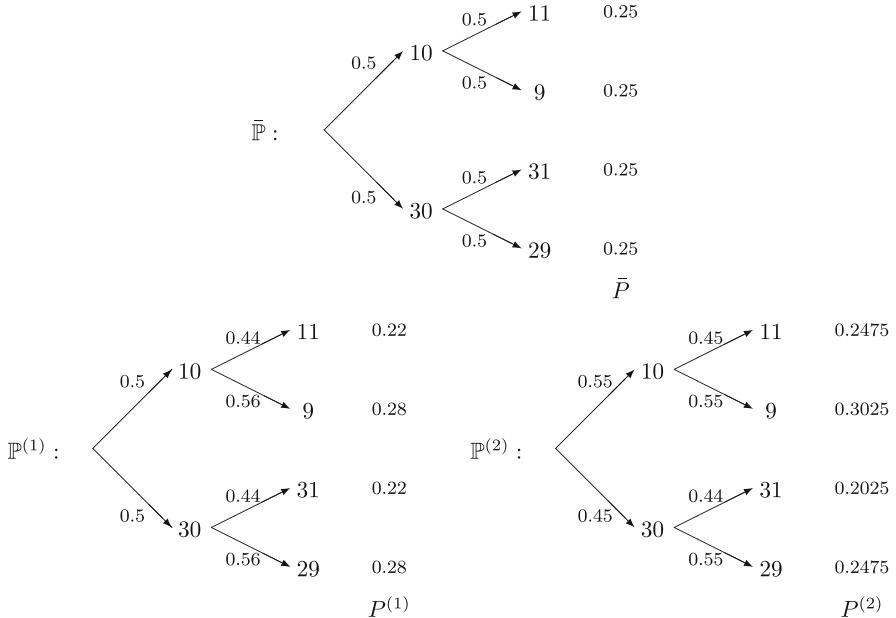


Fig. 7.1 An example for neighborhoods of trees. $\mathbb{P}^{(1)}$ is much closer to $\bar{\mathbb{P}}$ than $\mathbb{P}^{(2)}$ in nested distance, but in terms of the ℓ_1 -distances of the scenario probabilities $P^{(2)}$ is closer to \bar{P} than $P^{(1)}$

$$\begin{aligned} & \text{dil}_r \left(\mathbb{P}(\mathbb{T}, (1-\lambda)P^{(0)} + \lambda P^{(1)}), \mathbb{P}(\mathbb{T}, \hat{P}) \right) \\ & \geq (1-\lambda) \cdot \text{dil}_r (\mathbb{P}(\mathbb{T}, P^{(0)}), \mathbb{P}(\mathbb{T}, \hat{P})) + \lambda \cdot \text{dil}_r (\mathbb{P}(\mathbb{T}, P^{(1)}), \mathbb{P}(\mathbb{T}, \hat{P})) \end{aligned}$$

see Fig. 1.17 in the Introduction.

The final formulation of the ambiguity extension problem is now

$$\min_x \max_{\mathbb{P} \in \mathcal{P}_\epsilon} \{ \mathcal{R}_{\mathbb{P}}[Q(x, \xi)] : x \in \mathbb{X}, x \triangleleft \mathfrak{F} \}. \quad (7.13)$$

In the next section we demonstrate how one can get around the nonconvexity problem by amplifying the ambiguity set to its convex hull with respect to compounding. In addition, we show that the *worst case* model is always contained in the original, non-amplified ambiguity set.

7.3.1 A Minimax Theorem

The famous minimax theorems by von Neumann [82], Fan [43], Sion [132], and all the references therein assert that under some conditions the min and the max can

be interchanged in (7.13). The validity of such theorems is related to convexity / concavity properties of the criterion function and topological properties of feasible sets. While convexity for the feasible set \mathbb{X} is not problematic, one has to define the notion of convexity for scenario models. As was already said, it would be incorrect to just form convex combinations of the scenario probabilities. The correct notion of convex combinations, however, is compounding in the sense of (1.24) of the Introduction. We repeat the definition here.

Definition 7.6 (Compound Tree). If $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ are nested distributions, then their *compound* with probability λ is given by

$$\mathcal{C}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \lambda) = \begin{cases} \mathbb{P}^{(1)} & \text{with probability } \lambda, \\ \mathbb{P}^{(2)} & \text{with probability } 1 - \lambda. \end{cases}$$

If $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ are tree models, then $\mathcal{C}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}, \lambda)$ is also a tree model, where the subtree $\mathbb{P}^{(1)}$ can be reached from a new root with probability λ , and the subtree $\mathbb{P}^{(2)}$ can be reached with probability $1 - \lambda$.

It turns out that our ambiguity set \mathcal{P}_ϵ given by (7.12) is not convex with respect to compounding. Therefore we consider its closed convex hull $\bar{\mathcal{P}}_\epsilon$. To this end, the notion of compounding trees must be generalized to infinitely many elements: let \mathfrak{P} be the family of all probability measures on \mathcal{N}_T , which is—since \mathcal{N}_T is a finite set—a simplex. Let Λ be a probability measure from \mathfrak{P} . The compound $\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda)$ is defined as

$$\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda) = \mathbb{P}(\mathbb{T}, P), \quad \text{where } P \text{ is distributed according to } \Lambda,$$

meaning that the compound is obtained by first sampling a distribution P according to Λ and then taking the model $\mathbb{P}(\mathbb{T}, P)$. Figure 7.2 displays $\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda)$ for a probability measure Λ with finite support. If Λ sits on $P^{(1)}, P^{(2)}, \dots, P^{(k)}$ with probabilities λ_l for $1 \leq l \leq k$, then the compound model has height $T + 1$ and k nodes at stage 1, such that to the l th node of stage 1 the subtree $\mathbb{P}(\mathbb{T}, P^{(l)})$ is

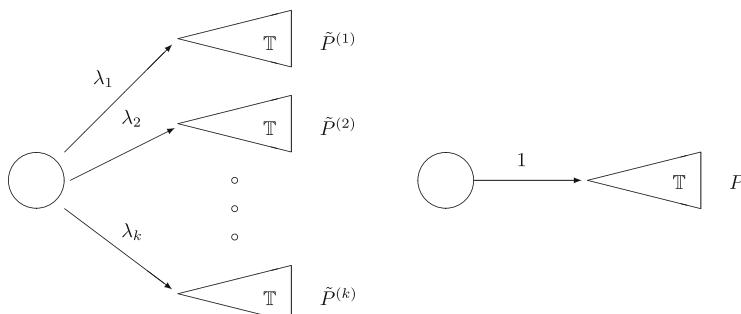


Fig. 7.2 The compound convex structure of the trees $\mathbb{P}(\mathbb{T}, \tilde{P}^{(l)})$ and the augmented tree $\mathbb{P}_+(\mathbb{T}, P)$

associated, i.e.,

$$\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda) = \mathbb{P}(\mathbb{T}, P^{(l)}) \quad \text{with probability } \lambda_l.$$

Denote by $\mathbb{P}_+(\mathbb{T}, \hat{P})$ the degenerated compound model, where the baseline model $\mathbb{P}(\mathbb{T}, \hat{P})$ is chosen with probability 1. It is equivalent to $\mathbb{P}(\mathbb{T}, \hat{P})$, but has an additional root, from which subtree $\mathbb{P}(\mathbb{T}, \hat{P})$ can be reached with probability 1. By construction, all random mixture trees $\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda)$ have height $T + 1$ and also $\mathbb{P}_+(\mathbb{T}, \hat{P})$ has height $T + 1$. Thus the distance between $\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda)$ and $\mathbb{P}(\mathbb{T}, \hat{P})$ is well defined.

The convex hull of the set

$$\mathcal{P}_\epsilon = \{\mathbb{P}(\mathbb{T}, P) : P \in B_\epsilon\}$$

with

$$B_\epsilon = \left\{ P : d(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \hat{P})) \leq \epsilon \right\}$$

is the set

$$\bar{\mathcal{P}}_\epsilon = \{\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda) : \Lambda \text{ is a probability measure on } B_\epsilon\}.$$

The convexified problem now is

$$\min_{x \in \mathbb{X}} \max_{\mathbb{P}(\mathbb{T}, P) \in \bar{\mathcal{P}}_\epsilon} \{\mathcal{R}_{\mathbb{P}(\mathbb{T}, P)}[Q(x, \xi)] : x \triangleleft \mathfrak{F}\}. \quad (7.14)$$

Notice that in the formulation (7.14) the decision variables x must coincide in all randomly sampled subproblems. By safeguarding ourselves against any random selection of elements of B_ϵ , we automatically safeguard ourselves against the worst case in B_ϵ .

The next step is to calculate the nested distance between two elements of $\bar{\mathcal{P}}_\epsilon$. For two leaves i (j , resp.) of the tree \mathbb{T} the inherited distance is defined as the distance of the corresponding paths leading to i (j , resp.), i.e.

$$d(i, j) = \sum_{t=1}^T \sum_{\ell=1}^m w_t^\ell d(\xi_\ell(\text{pred}_t(i)), \xi_\ell(\text{pred}_t(j))),$$

where w_t^ℓ are some positive weights. Assume that there exist constants c and C such that $0 < c \leq d(i, j) \leq C$ for all $i \neq j$. This condition is equivalent to the condition that the distance on $\mathcal{N}_T = \Omega$ inherited from the process ξ is a distance and not just a pseudodistance.

Theorem 7.7. *Let $Q(x, \xi)$ be convex in x with a convex and compact decision set \mathbb{X} and let the functional \mathcal{R} be compound concave (see Definition 3.25 or (CC) in the Appendix). Then*

$$\min_{x \in \mathbb{X}} \max_{\mathbb{P} \in \bar{\mathcal{P}}_\epsilon} \mathcal{R}_{\mathbb{P}}[Q(x, \xi)] = \max_{\mathbb{P} \in \bar{\mathcal{P}}_\epsilon} \min_{x \in \mathbb{X}} \mathcal{R}_{\mathbb{P}}[Q(x, \xi)]$$

and a saddle point (x^*, \mathbb{P}^*) exists, i.e., a pair (x^*, \mathbb{P}^*) with $x^* \in \mathbb{X}$ and $\mathbb{P}^* \in \bar{\mathcal{P}}_\epsilon$ such that

$$\mathcal{R}_{\mathbb{P}}[Q(x^*, \xi)] \leq \mathcal{R}_{\mathbb{P}^*}[Q(x^*, \xi)] \leq \mathcal{R}_{\mathbb{P}^*}[Q(x, \xi)]$$

for all $x \in \mathbb{X}$ and $\mathbb{P} \in \hat{\mathcal{P}}_\epsilon$. Moreover, $\mathbb{P}^* \in \mathcal{P}_\epsilon$ (and not just in $\bar{\mathcal{P}}_\epsilon$).⁴

Proof. Let

$$\|P - \hat{P}\| := \sum_{i \in \mathcal{N}_T} |P_i - \hat{P}_i| = 2 - 2 \sum_{i \in \mathcal{N}_T} \min(P_i, \hat{P}_i).$$

It follows that

$$\frac{c}{2} \cdot \|P - \hat{P}\| \leq \text{dL}_1(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \hat{P})) \leq \frac{C}{2} \cdot \|P - \hat{P}\|, \quad (7.15)$$

because an optimal transportation plan can transport a mass of $\min(P_i, \hat{P}_i)$ from i to i with distance 0, and thus only the masses $1 - \sum_{i \in \mathcal{N}_T} \min(P_i, \hat{P}_i)$ have to be transported over distances between c and C . Notice that the use of the distance $\|P - \tilde{P}\|$ is only to demonstrate compactness. While the topologies generated by the two metrics $\|P - \tilde{P}\|$ and $\text{dL}(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \tilde{P}))$ are the same (due to relation (7.15)), balls are quite different in the two metrics and only the latter metric is appropriate for nested distributions as we have argued in Example 7.5.

Next we see that $\bar{\mathcal{P}}_\epsilon$ is compact, since it is the continuous image of the set of all probability measures on B_ϵ , which is a compact set, since B_ϵ itself is compact. Thus all conditions for the validity of the minimax theorem are fulfilled and a saddle point (x^*, \mathbb{P}^*) must exist. \mathbb{P}^* is of the form $\mathbb{P}(\mathbb{T}, \Lambda^*)$ for some probability measure Λ^* on B_ϵ .

Introduce the shorter notation $\mathbb{P}(\mathbb{T}, \Lambda)$ for $\mathcal{C}(\mathbb{P}(\mathbb{T}, \cdot), \Lambda)$. The risk functional \mathcal{R} can be linearly extended, i.e.,

$$\mathcal{R}_{\mathbb{P}(\mathbb{T}, \Lambda)}[\cdot] = \int \mathcal{R}_{\mathbb{P}(\mathbb{T}, P)}[\cdot] \Lambda(dP).$$

Now we prove the equation

$$\text{dL}(\mathbb{P}(\mathbb{T}, \Lambda), \mathbb{P}_+(\mathbb{T}, \hat{P})) = \int \text{dL}(\mathbb{P}(\mathbb{T}, P), \mathbb{P}(\mathbb{T}, \hat{P})) \Lambda(dP). \quad (7.16)$$

In order to see this, assume first that Λ is finite, say $\mathbb{P}(\mathbb{T}, \Lambda) = \sum_{l=1}^k \lambda_l \mathbb{P}(\mathbb{T}, P^{(l)})$. Notice that by the recursive structure of the nested distance (see Algorithm 2.1),

⁴More precisely: the saddle point set contains a \mathbb{P}^* , which is in \mathcal{P}_ϵ .

the distance calculation at stage 1 is based on optimally transporting the single stage-1 node of $\mathbb{P}_+(\mathbb{T}, \hat{P})$ to the k stage-1 nodes of $\mathbb{P}(\mathbb{T}, \Lambda)$. These relations have a distance of $\text{dL}\left(P(\mathbb{T}, P^{(l)}), P(\mathbb{T}, \hat{P})\right)$, $l = 1, \dots, k$. There is no other choice than to transport quantities λ_l , $l = 1, \dots, k$, since this transportation is determined by the marginals. Hence

$$\begin{aligned}\text{dL}\left(P(\mathbb{T}, \Lambda), P_+(\mathbb{T}, \hat{P})\right) &= \text{dL}\left(\sum_{l=1}^k \lambda_l P(\mathbb{T}, P^{(l)}), P_+(\mathbb{T}, \hat{P})\right) \\ &= \sum_{l=1}^k \lambda_l \text{dL}\left(P(\mathbb{T}, P^{(l)}), P(\mathbb{T}, \hat{P})\right).\end{aligned}$$

If Λ is not finite, then it can be approximated by finite measures and therefore the relation (7.16) holds in general.

Finally we show that a saddle point model \mathbb{P}^* can be chosen as a single tree and not as a mixture of trees. Let x^* be the minimax decision, i.e.,

$$\mathcal{R}_{\mathbb{P}}[Q(x^*, \xi)] \leq \mathcal{R}_{\mathbb{P}^*}[Q(x^*, \xi)] \leq \mathcal{R}_{\mathbb{P}^*}[Q(x, \xi)]$$

for all $x \in \mathbb{X}$ and all $\mathbb{P} \in \bar{\mathcal{P}}_\epsilon$. Let the saddle point model be $\mathbb{P}^* = \mathbb{P}(\mathbb{T}, \Lambda^*)$. The support of Λ^* is closed (hence compact) and the continuous function $P \mapsto \mathcal{R}_{\mathbb{P}(\mathbb{T}, P)}[Q(x^*, \xi)]$ takes its maximum at some distribution P^* . Since $\text{dL}\left(\mathbb{P}(\mathbb{T}, P^*), \mathbb{P}(\mathbb{T}, \hat{P})\right) \leq \epsilon$ by construction, $\mathbb{P}(\mathbb{T}, P^*) \in \mathcal{P}_\epsilon$ and therefore, by the saddle point property, $\mathcal{R}_{\mathbb{P}(\mathbb{T}, P^*)}[Q(x^*, \xi)] \leq \mathcal{R}_{\mathbb{P}^*}[Q(x^*, \xi)]$. On the other hand,

$$\mathcal{R}_{\mathbb{P}^*}[Q(x^*, \xi)] = \int \mathcal{R}_{\mathbb{P}(\mathbb{T}, P)}[Q(x^*, \xi)] \Lambda^*(dP) \leq \mathcal{R}_{\mathbb{P}(\mathbb{T}, P^*)}[Q(x^*, \xi)].$$

Consequently, $\mathcal{R}_{\mathbb{P}(\mathbb{T}, P^*)}[Q(x^*, \xi)] = \mathcal{R}_{\mathbb{P}^*}[Q(x^*, \xi)]$, which shows that the saddle point model can be chosen from \mathcal{P}_ϵ . This concludes the proof. \square

7.3.2 Ambiguity Sets Defined by Nested Transportation Kernels

We have seen that problem (7.13) has a complex structure in its general form. In construction of models $\mathbb{P}(\mathbb{T}, P)$ only scenario probabilities differ from the baseline model $\mathbb{P}(\mathbb{T}, \hat{P})$ as long as the respective nested distance remains small.

In order to describe the nested distance in a recursive form we introduce the notion of transportation subplans as it was introduced in Sect. 2.10.3. A transportation subplan $\pi(\cdot, \cdot | k, l)$ indexed with a pair of nodes (k, l) transports the elements of $k+$, the set of direct successors of k , into $l+$, the set of direct successors of l and must satisfy the marginal constraints

$$\begin{aligned} \sum_{j' \in l^+} \pi(i', j'|k, l) &= \hat{Q}(i') \text{ and} \\ \sum_{i' \in k^+} \pi(i', j'|k, l) &= Q(j'), \end{aligned} \tag{7.17}$$

where $\hat{Q}(i') = \hat{P}(i'|i'-)$ and $Q(j') = P(j'|j'-)$ are the conditional node probabilities. Denoting by i_t the predecessor of terminal node i at stage t , i.e., $i_t = \text{pred}_t(i)$ and similarly for j , one can recover the unconditional probabilities from the conditional ones by

$$\hat{P}(i) = \hat{Q}(i) \cdot \hat{Q}(i_{T-1}) \cdots \hat{Q}(i_1) \quad \text{and} \quad P(j) = Q(j) \cdot Q(j_{T-1}) \cdots Q(j_1).$$

Because of this relation and by the validity of the constraints (7.17), the subplans can be *concatenated* to the full transportation plan as in (2.42)

$$\pi_{i,j} = \pi(i_1, j_1 | i_0, j_0) \cdots \pi(i_{T-1}, j_{T-1} | i_{T-2}, j_{T-2}) \cdot \pi(i, j | i_{T-1}, j_{T-1}).$$

Conversely, the transportation subplans for $i', j' \in \mathcal{N}_{t+1}$ and $k = i'-, l = j'-$ can be found from the total transportation plan $\pi_{i,j}$ for $i, j \in \mathcal{N}_T$ by

$$\pi(i', j'|k, l) = \frac{\sum_{\substack{i > i', j > j' \\ i > k, j > l}} \pi_{i,j}}{\sum_{i > k, j > l} \pi_{i,j}}.$$

We emphasize here again that we consider *only* the case where the alternative models are based on the same tree \mathbb{T} and this means that only the probabilities vary between the ambiguous models. For algorithmic purposes it is better to use the notion of transportation *subkernels* instead of transportation subplans.

For arbitrary nodes $k, l \in \mathcal{N}_t$ and $i' \in k^+, j \in l^+$ the subkernel $K_t(j'|i' : k, l)$ is defined as

$$K_t(j'|i'; k, l) = \frac{\pi(i', j'|k, l)}{\sum_{j'} \pi(i', j'|k, l)}.$$

Then, for fixed i' , k and l , $K_t(\cdot | i'; k, l)$ is a probability distribution on the set l^+ and therefore

$$K_t(j'|i'; k, l) \geq 0, \quad \sum_{j' \in l^+} K_t(j'|i'; k, l) = 1, ((i', j') \in \mathcal{N}_{t+1}, i' \in k^+, l).$$

The relation between transportation subkernels and transportation subplans is given by

$$\pi_{i,j} = K_1(j_1|i_1; i_0, j_0) \cdots K_{T-2}(j_{T-1}|i_{T-1}; i_{T-2}, j_{T-2}) \cdot K_{T-1}(j|i; i_{T-1}, j_{T-1}) \\ \cdot Q(i) \cdot Q(i_{T-1}) \cdots Q(i_1).$$

Therefore the transportation kernel $K(i, j)$ is the *composition* of subkernels K_t , $t = 1 \dots T - 1$

$$K(i, j) = [K_1 \circ \cdots \circ K_{T-1}](i, j) = K_1(j_1|i_1; i_0, j_0) \cdots K(j|i; i_{T-1}, j_{T-1}).$$

For a given baseline probability distribution $\hat{P} = (\hat{P}_i)_{i \in \mathcal{N}_T}$ we may define the new probability distribution P by $P(j) := \sum_{i,j \in \mathcal{N}_T} K(i, j) \cdot \hat{P}(i)$, in symbolic notation

$P = \hat{P} \circ K = \hat{P} \circ [K_1 \circ \cdots \circ K_{T-1}]$. Then problem (7.13) can be rewritten in the form

$$\min_{x \in \mathbb{X}} \max_K \{\mathcal{R}_{\hat{P} \circ K}[Q(x, \xi)] \text{ s.t. } K = K_1 \circ \cdots \circ K_{T-1}, \sum_{i,j \in \mathcal{N}_T} d_{i,j}^r \cdot K(i, j) \cdot \hat{P}(i) \leq \epsilon^r\}. \quad (7.18)$$

It is noticeable that the expression $\sum_{i,j \in \mathcal{N}_T} d_{i,j}^r \cdot K(i, j) \cdot \hat{P}(i) \leq \epsilon^r$ in (7.18) is multilinear in the transportation subkernels K_1, \dots, K_{T-1} . In the next section, algorithms are presented which solve problem (7.18) at least approximately.

Remark 7.8. How large should ϵ be chosen? A direct answer cannot be given, since the way how to get from an estimated parametric model to the finite scenario model is quite complicated. A partial answer can be given looking at the estimated conditional distributions $\hat{G}_t(\cdot|\xi_0, \dots, \xi_{t-1})$ and their Wasserstein distances. Typically the conditional probabilities are estimated via a parametric model and Remark 7.3 is relevant here.

7.3.3 Algorithmic Solution

As in the case of single- or two-stage models, the minimax problem is solved by an iterative procedure as shown in Algorithm 7.2. At each iteration a new model $\mathbb{P}(\mathbb{T}, P)$ (in short: \mathbb{P}), which is in ϵ nested distance of baseline model $\mathbb{P}(\mathbb{T}, \hat{P})$ (in short: $\hat{\mathbb{P}}$), is included in the model and therefore the size of the problem increases at each iteration.

Proposition 7.9. Let \mathbb{X} and $\bar{\mathcal{P}}_\epsilon$ be compact sets and $(x, \mathbb{P}) \mapsto \mathbb{E}_{\mathbb{P}}[Q(x, \xi)]$ be jointly continuous, then every cluster point of the iteration given by Algorithm 7.2 is a minimax solution.

Proof. The proof of this proposition can be found in the Appendix B, Proposition B.6. \square

Algorithm 7.2**Successive Programming**

- **INITIALIZATION.** Let $k = 1$ and determine the value of ϵ . Start with the “base line” model, i.e., $\mathcal{P}_\epsilon^{(k)} = \{\hat{\mathbb{P}}\}$.
- **OUTER OPTIMIZATION.** Solve the outer problem:
$$\begin{array}{ll} \min u \\ \text{s.t. } \mathbb{E}_{\mathbb{P}}[Q(x, \xi)] \leq u & \text{for all } \mathbb{P} \in \mathcal{P}_\epsilon^{(k)} \\ x \in \mathbb{X}, \\ x \triangleleft \mathfrak{F} \end{array}$$
 resulting in the solution $(x^{(k)}, u^{(k)})$. If the solution is not unique, choose any element of the solution set.
- **INNER OPTIMIZATION.** Fix $x^{(k)}$ and solve the inner problem:
$$\begin{array}{ll} \max \mathbb{E}_{\mathbb{P}}[Q(x^{(k)}, \xi)] \\ \text{s.t. } \mathbb{P} \in \mathcal{P}_\epsilon \end{array}$$
. Call the solution $\mathbb{P}^{(k)}$ and set $\mathcal{P}_\epsilon^{(k+1)} = \mathcal{P}_\epsilon^{(k)} \cup \{\mathbb{P}^{(k)}\}$. If the solution is not unique, choose any element of the solution set. Typically (but not necessarily) the inner problem is solved in a successive manner, see Algorithm 7.3.
- **STOPPING CRITERION.** We stop if there is no improvement in $u^{(k)}$, otherwise set $k := k + 1$ and goto **OUTER OPTIMIZATION**.

In a practical implementation we might

- choose a stopping criterion θ s.t. $u^{(k+1)} - u^{(k)} \geq \theta$, or
- specify in advance the number of iterations k (this means that the number of models included in set \mathcal{P}_ϵ to be determined at the beginning).

Algorithm 7.3

Stepwise linearization of the INNER OPTIMIZATION in Algorithm 7.2

At iteration step k the problem
$$\begin{array}{ll} \max \mathbb{E}_{\mathbb{P}}[Q(x^{(k)}, \xi)] \\ \text{s.t. } \mathbb{P} \in \mathcal{P}_\epsilon \end{array}$$
 has to be solved. Let $\epsilon_1 + \epsilon_2 + \dots + \epsilon_T = \epsilon$.

Suppose that $P^{(k-1)}$, the worse case of previous step, is of the form $P^{(k-1)} = \hat{P} \circ K_1^{(old)} \circ \dots \circ K_T^{(old)}$ (for $k = 1$ let $K_t^{(old)}$ be the identity matrix).

- For $t = 1$ to $t = T$ solve
$$\begin{array}{ll} \max_{K_t} \mathbb{E}_{\hat{P} \circ K_1^{(new)} \circ \dots \circ K_{t-1}^{(new)} \circ K_t \circ K_{t+1}^{(old)} \circ \dots \circ K_T^{(old)}} [Q(x^{(k)}, \xi)] \\ \text{s.t. } \mathbf{d}_r \left(\hat{P} \circ K_1^{(new)} \circ \dots \circ K_{t-1}^{(new)} \circ K_t \circ K_{t+1}^{(old)} \circ \dots \circ K_T^{(old)}, \hat{P} \right) \leq \epsilon_t \end{array}$$

and call the solution $K_t^{(new)}$.

- The next worse case is $P^{(k)} = \hat{P} \circ K_1^{(new)} \circ \dots \circ K_T^{(new)}$.

While Algorithm 7.2 is generally applicable, the INNER OPTIMIZATION step is typically quite complex, even in the case that the ambiguity problem is of the form (7.18), i.e., if alternative models are defined by kernels $K = K_1 \circ \dots \circ K_T$. The nonconvex problem may be approximated by a sequence of problems, where at each step only one transition kernel K_t is changed and the others are kept fixed. By fixing all but one of the kernels the problem becomes linear (see Algorithm 7.3).

The condition $\epsilon_1 + \epsilon_2 + \dots + \epsilon_T = \epsilon$ guarantees that all identified worst case models lie in the predetermined ambiguity set, however, one may not guarantee that the overall worst case model is attained by this procedure.

7.3.3.1 The Price of Ambiguity and the Gain for Distributional Robustness

Let $\hat{\mathbb{P}}$ be the baseline model and let $x^*(\hat{\mathbb{P}})$ be the optimal solution of the baseline problem (7.1). Likewise, let \mathcal{P} be the ambiguity set and let $x^*(\mathcal{P})$ be the solution of the minimax problem (7.2). Under convex-concavity, the solution of the minimax problem $x^*(\mathcal{P})$ together with the worst case model $\mathbb{P}^* \in \mathcal{P}$ forms a saddle point (see Proposition B.4 in the Appendix), meaning that the following inequality is valid for all feasible x and all $\mathbb{P} \in \mathcal{P}$,

$$\mathbb{E}_{\mathbb{P}}[Q(x^*(\mathcal{P}), \xi)] \leq \mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathcal{P}), \xi)] \leq \mathbb{E}_{\mathbb{P}^*}[Q(x, \xi)].$$

Let us call $\mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathcal{P}), \xi)]$ the *minimax value*. Notice that by construction

$$\mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathbb{P}), \xi)] - \mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathcal{P}), \xi)] \geq 0, \quad (7.19)$$

and this nonnegative quantity can be seen as the *gain for distributional robustness*. This value indicates the gain in optimal value, if the worse case model is true and the minimax solution is implemented instead of the optimal solution of the baseline model. On the other hand

$$\mathbb{E}_{\hat{\mathbb{P}}}[Q(x^*(\mathcal{P}), \xi)] - \mathbb{E}_{\hat{\mathbb{P}}}[Q(x^*(\hat{\mathbb{P}}), \xi)] \geq 0, \quad (7.20)$$

and this can be interpreted as the *price of ambiguity*. It measures the loss in optimal value, if the minimax solution is implemented, but the baseline model is true and one could have also implemented an optimal solution for the baseline model. Both values increase with the size of the ambiguity set.

The algorithm proposed in this section has been implemented. The following section presents and discusses computational results for a classical multiperiod production / inventory control problem.

7.4 Example: A Multiperiod Production / Inventory Control Problem

To picture the implications of the algorithms proposed, a simplified multistage stochastic optimization problem—a multiperiod production / inventory control problem—is implemented and numerical results are shown in this section. This

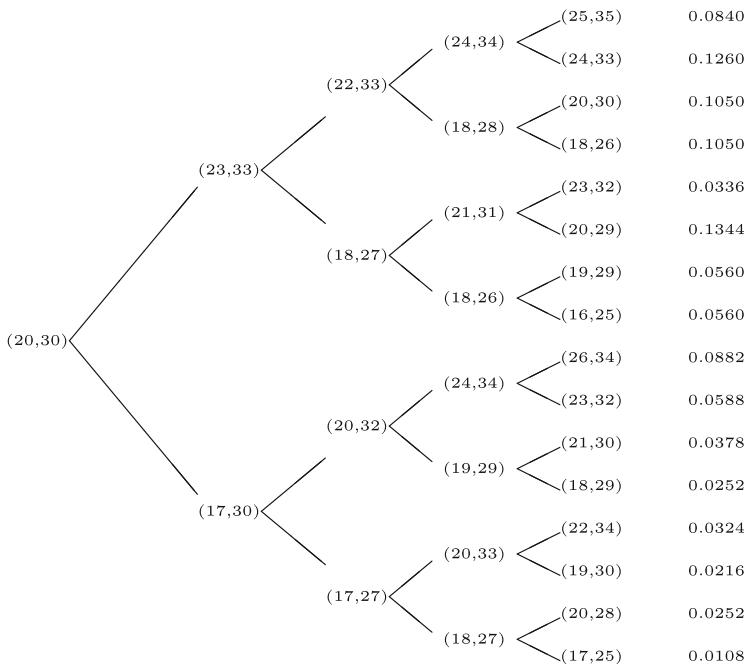


Fig. 7.3 The binary tree carrying the demands ($\text{product}_1, \text{product}_2$) and the scenario probabilities

example⁵ is used to illustrate the multistage approach to stochastic modeling and its ambiguity extension.

In this problem the production volume of two products is decided while maximizing the expected net profit derived from selling the products under stochastic demands of the subsequent weeks with fixed selling prices, production, inventory and external supply costs. Deciding on how much of each product types to produce during a particular week forms the decision variables. The production machine is designed to produce both types and there is an overall production capacity. The stochastic demand is characterized in terms of scenarios and a tree terminology is used to describe event probabilities and multistage scenarios. The demand scenarios are represented on a binary tree with not necessarily equal event probabilities. Figure 7.3 displays the tree structure, the demand requirements for both products as well as the scenario probabilities.

⁵The numerical example is taken from AIMMS optimization modeling [12, Chapter 17]. However, all computational procedures, solution algorithms, and resulting analysis are implemented in MATLAB R2012a. The implementations are due to Bita Analui.

Table 7.1 Nomenclature

Deterministic parameters	
π_b^s	Selling price for each product $b = 1, 2$
π_b^c	Production cost of each product $b = 1, 2$
π_b^i	Inventory cost of each product $b = 1, 2$
π_b^e	External supply cost of each product $b = 1, 2$
c	Maximum overall production capacity
\bar{x}^i	Maximum inventory capacity
$x_b(1)$	Initial stock level of product $b = 1, 2$
Stochastic parameters (for each node)	
$\xi_b(\cdot)$	Demand for product $b = 1, 2$
Decision variables (for each node)	
$x_b^f(\cdot)$	Production volume of product b for $b = 1, 2$
Decision dependent variables	
$x_b^i(\cdot)$	Inventory level of each product $b = 1, 2$
$x_b^e(\cdot)$	External supply of each product $b = 1, 2$
$v(\cdot)$	Profit

7.4.1 Mathematical Modeling Summary

In Table 7.1 the symbols defining the parameters, decisions, and decision dependent variables of the model are introduced.

The full mathematical model in nodal representation is formulated below. Note that decisions are only defined for nonterminal nodes. As usual, $n-$ denotes the direct predecessor of node n .

$$\text{maximize } \sum_n P(n) \cdot v(n) \quad (n \in \mathcal{N}) \quad (7.21a)$$

$$\text{subject to } \sum_b x_b^f(n-) \leq c, \quad (n \in \mathcal{N} \setminus \mathcal{N}_0) \quad (7.21b)$$

$$x_b^i(n-) + x_b^f(n-) + x_b^e(n) - \xi_b(n) = x_b^i(n), \quad (n \in \mathcal{N} \setminus \mathcal{N}_0) \quad (7.21c)$$

$$\sum_b x_b^i(n) \leq \bar{x}^i, \quad (n \in \mathcal{N}) \quad (7.21d)$$

$$x_b^i(n-) + x_b^e(n) \geq \xi_b(n), \quad (n \in \mathcal{N} \setminus \mathcal{N}_0) \quad (7.21e)$$

$$\sum_b \pi_b^s \cdot \xi_b(n) - \sum_b [\pi_b^c \cdot x_b^f(n-) + \pi_b^i \cdot x_b^i(n) + \pi_b^e \cdot x_b^e(n)] = v(n), \quad (n \in \mathcal{N} \setminus \mathcal{N}_0) \quad (7.21f)$$

Table 7.2 Parameter settings. Additionally, $\bar{x}^i = 52$ and $c = 46$

Product	π_b^s (€ / unit)	π_b^c (€ / unit)	π_b^i (€ / unit)	π_b^e (€ / unit)	$x_b^i(0)$
product ₁ ($b = 1$)	300	12	5	195	17
product ₂ ($b = 2$)	400	10	5	200	35

$$x_b^f \geq 0, x_b^i \geq 0, x_b^e \geq 0; b = 1, 2.$$

The objective of this inventory control model is to maximize the total expected net profit ($P(n)$ is the unconditional probability of reaching node $n \in \mathcal{N}$) under the following constraints: constraint (7.21b) ensures that the total production volume is bounded above with the overall capacity. Equation (7.21c) states that the inventory determined at each reachable node by the inventory at the predecessor node plus the production volume at the predecessor node plus the external supply at that node minus the demand pertaining to that node, while (7.21d) illustrates the maximum inventory capacity constraints. Constraint (7.21e) is added since for technical reasons it is not possible to sell the production of one period in the same period. Equation (7.21f) is an accounting equation for the net profit position at each node, which is derived from the sales revenue minus the total costs consisting of production, inventory, and external supply. The revenues and the cost parameters are presented in Table 7.2. In the following first the optimal solution of the original multistage problem (7.21a) is shown and further the maximin solution of distributionally robust extension of (7.21a) is presented and discussed.

7.4.2 Computational Results

Based on the multistage stochastic optimization problem developed and the input data provided in Table 7.2, the optimal value of expected net profit is 7,688(€). In Fig. 7.4, the trees of optimal productions x_1^f, x_2^f and the profits v are shown. Solution scenarios for both products follow a rather simple pattern. One direct effect of optimal decisions on profit scenarios is observed in the sudden decrease of net profit levels at stage one, since satisfying the emanating demand at stage two requires a compensatory act by external purchases for both products.

7.4.2.1 Worst Tree Visualizations

Using Algorithm 7.2 with the linearization approximation of Algorithm 7.3 worst trees were calculated as the saddle points for some ambiguity radii ϵ . We constructed ambiguity sets for $\epsilon = 1, 6, 11$ and $\epsilon = 16$; the reasoning behind this choice for the range of ϵ is simply that it varies between $\min d(i, j)$ and $\max d(i, j)$, where $d(i, j)$ is the distance between demand scenarios i and j . It turns out that the worst tree

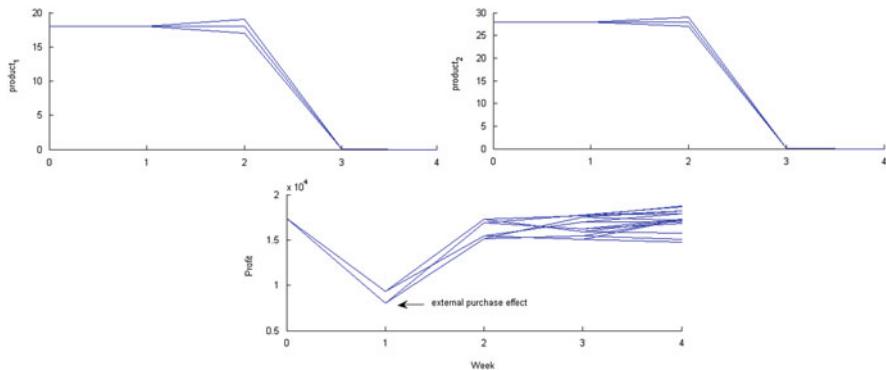


Fig. 7.4 Optimal solution scenarios for the baseline model: product₁ in top left, product₂ in top right and profit in bottom

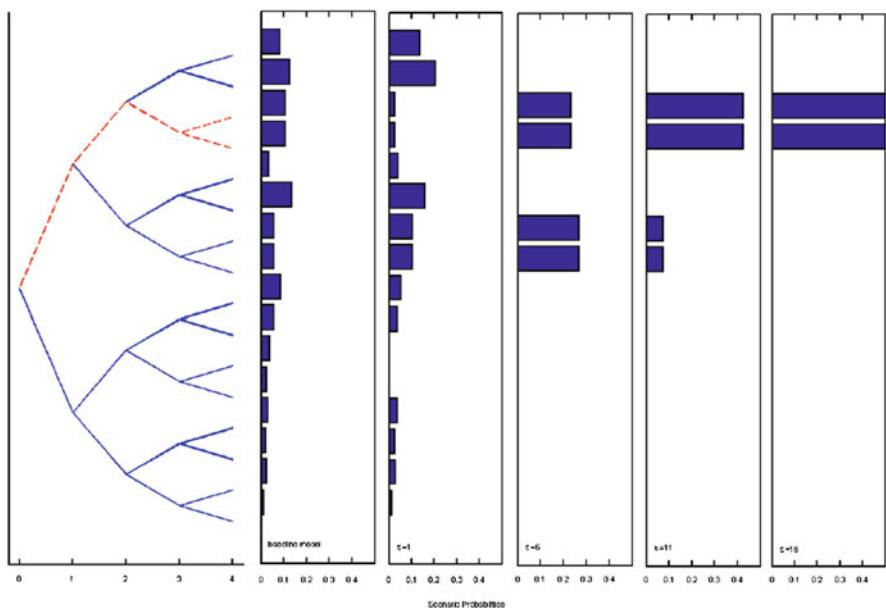


Fig. 7.5 The tree structure of problem (7.21a) and the worst trees (the saddle points) for increasing ambiguity radii ϵ . Quite typically, the worst case trees tend to be concentrated on few scenarios when ϵ increases

gets less and less complex as ϵ increases, since only bad scenarios “survive” and good scenarios successively “die out.”

Figure 7.5 shows the scenario probabilities of the worst tree (the saddle point tree) in dependence of the ambiguity radius. It is observed that at the largest radius ($\epsilon = 16$) only the bad scenarios 3 and 4 form the worst tree, all others have probability 0.

7.4.2.2 Maximin Solutions for Different Ambiguity Radii

In Figs. 7.6 and 7.7 the maximin solution of productions for product₁ and product₂ and their dependence on an increasing ambiguity radius ϵ is shown. It is noticeable that including more than only one “baseline” demand model leads to more diverse production scenarios as it can be seen for product₁ by comparing the top left graph in Figs. 7.4 with 7.6, and for product₂ by comparing the top right graph in Figs. 7.4 with 7.7. At first glance these results might seem quite unintuitive, since worst case scenario trees are getting simpler and simpler as ϵ increases, while the decision scenarios are becoming more complex. However, distributional robust decisions have to be good compromises for a variety of models and extreme decisions (i.e., produce maximum or nothing) are never optimal under model ambiguity.

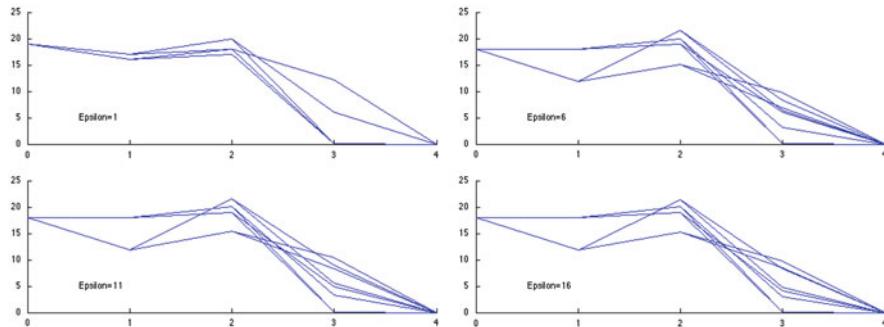


Fig. 7.6 The trees of optimal production decisions (the maximin decisions) for product₁ under the ambiguity radii of $\epsilon = 1, 6, 11, 16$

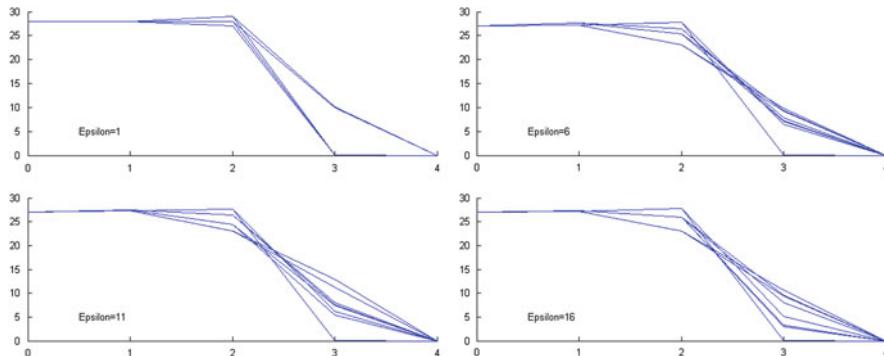


Fig. 7.7 The trees of optimal production decisions (the maximin decisions) for product₂ under the ambiguity radii of $\epsilon = 1, 6, 11$ and $\epsilon = 16$

7.4.2.3 The Price of Ambiguity

For the given inventory / production control problem, the gain for robustness (7.19) and the price of ambiguity (7.20) can be calculated. In order to be consistent with our general setup, we minimize here negative profits as costs. For the ambiguity radius of $\epsilon = 11$, e.g., the following values can be calculated:

$$\begin{aligned}\mathbb{E}_{\hat{\mathbb{P}}}[Q(x^*(\hat{\mathbb{P}}), \xi)] &= -7,688, \\ \mathbb{E}_{\hat{\mathbb{P}}}[Q(x^*(\mathcal{P}_\epsilon), \xi)] &= -7,514, \\ \max_{\mathbb{P} \in \mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{P}}[Q(x^*(\mathcal{P}_\epsilon), \xi)] &= -7,326 \text{ and} \\ \max_{\mathbb{P} \in \mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{P}}[Q(x^*(\hat{\mathbb{P}}), \xi)] &= -6,811,\end{aligned}$$

from which one gets

$$\begin{aligned}\text{price of ambiguity} &= -7,514 - (-7,688) = 174 \text{ and} \\ \text{gain of distributional robustness} &= -6,811 - (-7,326) = 515.\end{aligned}$$

Figure 7.8 shows the respective quantities in dependency of the ambiguity radius ϵ . For $\epsilon = 11$, the price of ambiguity is about 2.26 %, while the gain for distributional robustness is 7.56 %.

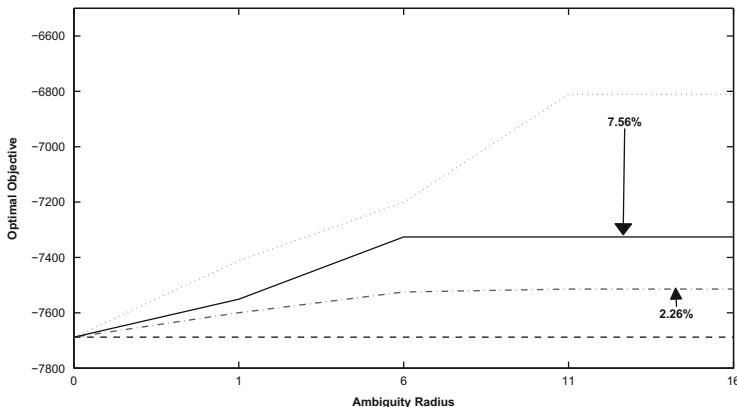


Fig. 7.8 The dashed lines show $\mathbb{E}[Q(x^*(\hat{\mathbb{P}}), \xi)]$, the baseline optimal value. The solid line is the minimax value $\max_{\mathbb{P} \in \mathcal{P}_\epsilon} \min_\xi \mathbb{E}_{\mathbb{P}}[Q(x, \xi)]$ as a function of the ambiguity radius ϵ . The dotted line shows $\epsilon \mapsto \max_{\mathbb{P} \in \mathcal{P}_\epsilon} \mathbb{E}_{\mathbb{P}}[Q(x^*(\hat{\mathbb{P}}), \xi)]$. Finally, the dash-dotted line shows $\epsilon \mapsto \mathbb{E}_{\hat{\mathbb{P}}}[Q(x^*(\mathcal{P}_\epsilon), \xi)]$ in dependency of the ambiguity radius ϵ

Chapter 8

Examples

In what follows we describe three examples of typical multistage stochastic optimization problems in an economic environment. While the first two examples relate to energy production, our third example discusses budget management for risk-prone countries.

8.1 Thermal Electricity Production

This example is due to Bita Analui:

Consider a power plant, which consists of several thermal units which may be fired with different fuels (typically oil and gas). The decisions to be made concern the energy production (separate for each fuel) and the purchase decisions of fuel. Fuel can be stored, but only up to a certain capacity. The only uncertain quantities are electricity spot prices. The objective is to maximize expected total net revenue over a period of seven production decisions (weeks or months). The random spot price process is represented as a scenario tree and the model is written in node-oriented form.

We outline the multistage optimization problem by listing the variables, the constraints, and the objective.

Decision Variables

- | | |
|------------|-----------------------------------------------------------------|
| $x_u^e(n)$ | energy to be produced in the next time period for each unit u |
| $x_f^b(n)$ | amount of fuel of type f bought at node n |

Scenario Process

- | | |
|----------|------------------------------------|
| $\xi(n)$ | stochastic electricity spot prices |
|----------|------------------------------------|

Deterministic Quantities

- $\pi_f(t)$ the fuel price at stage t
 s_f the storage costs per fuel of type f

Expressions

- $x_f^d(n)$ total amount of fuel to be burned (in the next period) at the node n
 $x_f^s(n)$ amount of stored fuel at the node n
 $x^c(n)$ cash at the node n

Energy Constraints

- The production of a power plant is limited by a maximum capacity θ_u and this amount represents the maximum power production of the corresponding unit u ,

$$x_u^e(n) \leq \theta_u \cdot \Delta\tau_{t(n)}. \quad (8.1a)$$

Here, $\Delta\tau_{t(n)}$ is the time span between decisions at the stage of the node n and the next stage (see (1.17)).

- The amount of fuel to be burned depends on the decisions about energy production $x_u^e(n)$ in the following way:

$$x_f^d(n) = \sum_u x_u^e(n) / [\text{eff}(u) \cdot \text{calval}(f)]. \quad (8.1b)$$

calval is the calorific value of one unit of fuel and $\text{eff}(u)$ is the efficiency of unit u . Their product decides how much energy can be produced out of one unit of fuel.

Fuel Constraints

- The stored fuel variable is formulated at the root node, and at intermediate and terminal nodes separately,

$$x_f^s(1) = x_f^b(1-) + x_f^b(1), \quad (8.1c)$$

$$x_f^s(n) = x_f^s(n-) - x_f^d(n-) + x_f^b(n), \quad n > 1, \quad (8.1d)$$

$$x_f^s(n) \leq \beta_f, \quad n \geq 1. \quad (8.1e)$$

$x_f^b(1-)$ is the initial amount of fuel and β_f is a capacity constraint.

- The amount of fuel to be purchased is bounded by γ_f ,

$$x_f^b(n) \leq \gamma_f, \quad n \geq 1. \quad (8.1f)$$

- The amount of fuel decided to be burned is bounded by the amount of fuel stored at the same node,

$$x_f^d(n) \leq x_f^s(n). \quad (8.1g)$$

Budget Constraints

- Cash at root node $n = 1$ (stage 0)

$$x^c(1) = x^c(1-) - \sum_f x_f^b(1) \cdot \pi_f(0), \quad (8.1h)$$

where $x^c(1-)$ is initial cash.

- Cash at the intermediate nodes $n \in \mathcal{N} \setminus \mathcal{N}_T$, with r being an interest rate

$$\begin{aligned} x^c(n) = & x^c(n-) \cdot (1 + r)^{\Delta \tau_{t(n-1)}} + \sum_u x_u^e(n-) \cdot \xi(n) \\ & - \sum_f x_f^b(n) \cdot \pi_f(t(n)) - \sum_f (x_f^s(n-) - \frac{1}{2} x_f^d(n-)) \cdot s_f \cdot \Delta \tau_{t(n-1)}. \end{aligned} \quad (8.1i)$$

- Cash at the terminal nodes $n \in \mathcal{N}_T$,

$$x^c(n) = x^c(n-) \cdot (1 + r)^{\Delta \tau_{t(n-)}} + \sum_u x_u^e(n-) \cdot \xi(n-), \quad n \in \mathcal{N}_T. \quad (8.1j)$$

The Objective and the Full Model Representation. It is the objective to find the amount of fuel to be purchased and the corresponding electricity production in order to maximize the expected terminal revenue (risk neutral objective),

$$\begin{aligned} & \text{maximize} \sum_{n \in \mathcal{N}_T} P(n) x^c(n) \\ & \text{subject to (8.1a)} - \text{(8.1j)}. \end{aligned}$$

The Scenario Tree. A tree model $\hat{\mathbb{P}}$ for the weekly average spot prices was constructed by our tree generation algorithm for one year. The tree with characteristics

Number of stages	52
Number of scenarios (leaves)	152
Number of nodes	2928

is shown in Fig. 8.1.

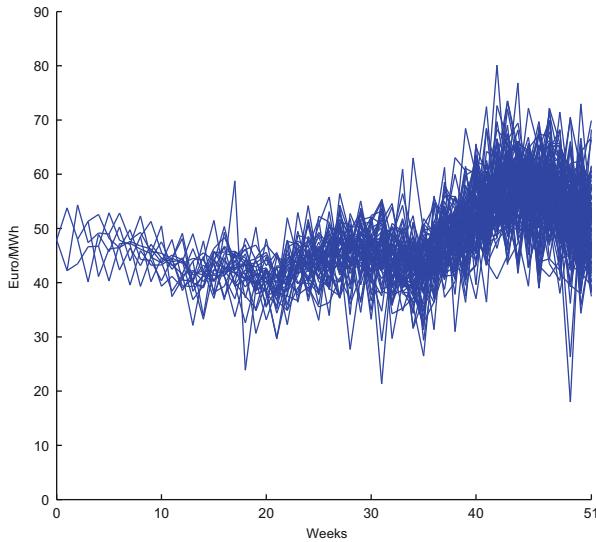


Fig. 8.1 The spotprice tree

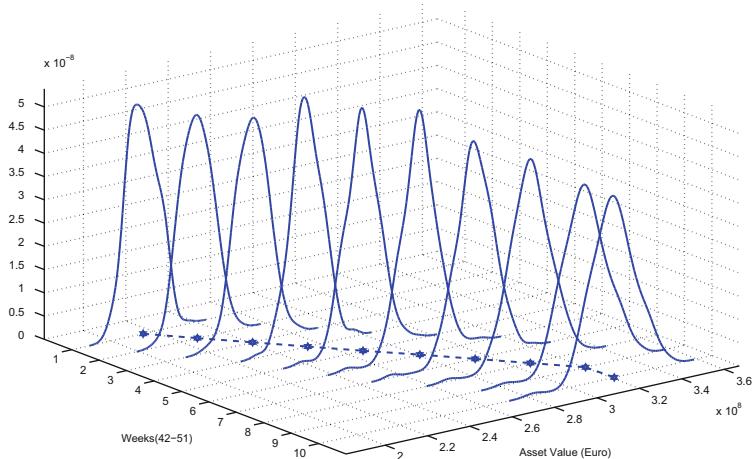


Fig. 8.2 The cash density over time

Numerical Results. The model was solved for the expected final cash as objective. Figure 8.2 shows density estimates of the accumulated cash variables for the last 10 weeks of the year. As can be seen, the uncertainty (e.g., measured by the variance) increases over time.

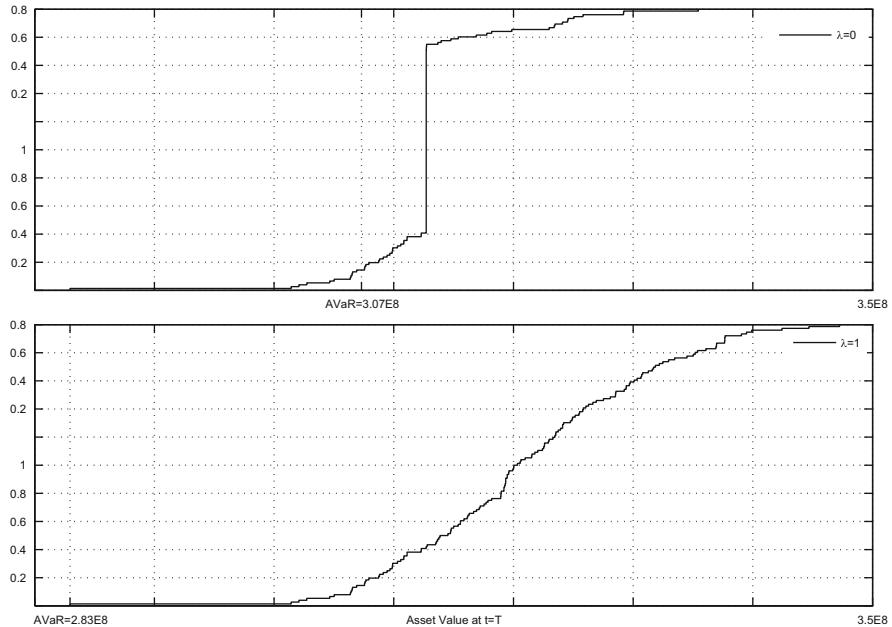


Fig. 8.3 Comparison of risk-neutral and risk-averse objective. The picture shows the cumulative distribution function of the final cash for the risk-averse situation (*top*) and the risk neutral case (*bottom*)

The same problem was also solved for a risk-averse objective, where the maximization of the expectation was replaced by the minimization of the average value-at-risk of the negative final cash at level $\alpha = 0.1$, see Fig. 8.3. The result is typical for risk-averse optimization: the distribution is more concentrated compared to the risk-neutral case; the lower tail is shifted to the right while the upper tail is shifted to the left. Safeguarding against losses on the left end of the distribution necessarily entails to give up some profit opportunities on the right end.

8.2 Hydro Electricity Production

Also this example is due to Bita Analui.

A multi-reservoir hydro generation system consisting of a cascade of interconnected reservoirs along with corresponding turbines, pumps, and spillways is considered. Such a system can be represented by a directed graph (see Fig. 8.4), where the nodes represent the reservoirs and the arcs represent possible water flows.

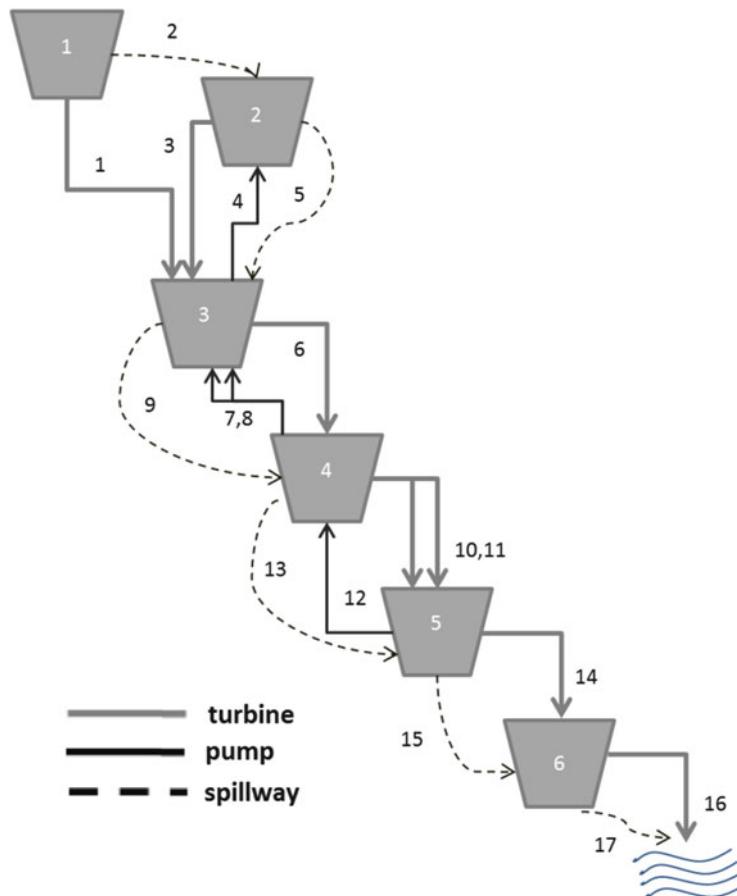


Fig. 8.4 The topology of the hydrosystem

Water flows can be either related to power generation by turbines, or to pumped water to higher reservoirs, or to spillage. Let J denote the set of reservoirs and I denote the set of arcs, then the arc-node incidence matrix, whose i, j -entry (denoted $A_{i,j}$) represents the interconnections among reservoirs and arcs, is as follows:

$$A_{i,j} = \begin{cases} 1 & \text{water flows into reservoir } j \text{ over arc } i, \\ -1 & \text{water flows out of reservoir } j \text{ over arc } i, \\ 0 & \text{arc } i \text{ is not connected to reservoir } j. \end{cases}$$

The potential decisions are given by the water flows over arcs—resulting in electric energy to be sold at the spot market—and the amount of water stored in reservoirs in order to maximize risk-adjusted expected terminal revenue. This is done by using a mixture of expectation and the average value at risk (AV@R_α).

The decision problem of a generator using a hydrosystem with given topology in a multistage stochastic optimization setting is formulated in time-oriented form as follows:

Decision Variables

$x_t^{i,f}$ flows through arc i in period t , where $i \in I_S$ denote the spillage arcs.

Expressions

$x_{t,i}^e$ energy (decided in period t) to be produced in the next time period $t + 1$ at arc i

$x_{t,j}^s$ reservoir j storage in period t

x_t^c cash in period t

Random Scenario Processes

ξ_t^e stochastic electricity spot prices (selling prices) in period t ¹

ξ_t^p pumping price, i.e., price for pumping the water into the next upper reservoir in period t ²

$\xi_{t,j}^f$ water inflows to the reservoirs in period t

Physical Constraints

- The following constraints put lower and upper bounds on the flow over arc i and storage volume of reservoir j . These bounds on storage must be assigned for flood control space and assuring minimum levels for dead storage and power plant operation

$$0 \leq x_{t,i}^f \leq \bar{x}_{t,i}^f, \quad i \in I, t = 1, \dots, T - 1,$$

$$\underline{x}_j^s \leq x_{t,j}^s \leq \bar{x}_j^s, \quad j \in J, t = 1, \dots, T - 1.$$

- A minimum content for each reservoir at the final stage needs to be fulfilled,

$$x_{end,j}^s \leq x_{T,j}^s, \quad j \in J.$$

- Water balance for all reservoirs: at each stage the water level at the end of the period depends on the water level at the beginning, the inflows and discharges during the period based on the system topology,

¹In a model with weekly decisions this will be an average over hourly spot prices.

²With weekly decisions we use an average over off-peak electricity prices.

$$\begin{aligned} x_{t,j}^s &= x_{t-1,j}^s + \xi_{t,j}^f + \sum_{\{i \notin I_S\}} A_{i,j} \cdot x_{t-1,i}^f \\ &\quad + \sum_{\{i \in I_S\}} A_{i,j} \cdot x_{t,i}^f \quad j \in J, t = 1, \dots, T-1. \end{aligned}$$

- The energy produced in period t is

$$x_{t,i}^e = x_{t-1,i}^f \cdot k^i \cdot \Delta \tau_{t-1},$$

where k^i represents the energy coefficient for arc i and $\Delta \tau_{t-1} = \tau_t - \tau_{t-1}$ the length of period t .

Budget Constraints

- The accounting equations for the cash position over time consider the interest rate r , and they depend on the gain from selling electricity and the cost of pumping water to the upstream reservoirs. The usage of $x_{t-1,i}^e$ versus (ξ_t^e, ξ_t^p) in this equation reflects the fact that decisions have to be made before knowing the actual spot and pumping prices, at which electricity is bought and sold:

$$x_t^c = x_{t-1}^c \cdot (1+r)^{\Delta t_{t-1}} + \sum_{\{i \in I \mid k^i > 0\}} x_{t-1,i}^e \cdot \xi_t^e + \sum_{\{i \in I \mid k^i < 0\}} x_{t-1,i}^e \cdot \xi_t^p.$$

The Objective and the Full Model Representation. The objective is to maximize the risk adjusted expected terminal revenue. The full model is represented in time-oriented form as follows, where all the constraints hold for $t = 0, \dots, T$ and all equations and inequalities are considered to hold almost-surely:

$$\begin{aligned} &\text{maximize } \lambda \mathbb{E}[x_T^c] - (1-\lambda) \text{AV@R}_{1-\alpha}[-x_T^c] \\ &\text{subject to } 0 \leq x_{t,i}^f \leq \bar{x}_i^f, \\ &\quad \underline{x}_j^s \leq x_{t,j}^s \leq \bar{x}_j^s, \\ &\quad x_{end,j}^s \leq x_{T,j}^s, \\ &\quad x_{t,j}^s = x_{t-1,j}^s + \xi_{t,j}^f + \sum_{\{i \notin I_S\}} A_{i,j} \cdot x_{t-1,i}^f \\ &\quad \quad + \sum_{\{i \in I_S\}} A_{i,j} \cdot x_{t,i}^f, \\ &\quad x_{t,i}^e = x_{t-1,i}^f \cdot k^i \cdot \Delta t_{(t-1)}, \\ &\quad x_t^c = x_{t-1}^c \cdot (1+r)^{\Delta t_{(t-1)}} + \sum_{\{i \in I \mid k^i > 0\}} x_{t-1,i}^e \cdot \xi_t^e \\ &\quad \quad + \sum_{\{i \in I \mid k^i < 0\}} x_{t-1,i}^e \cdot \xi_t^p. \end{aligned}$$

The scenario tree represents the information on weekly spot prices and pumping prices in addition to the weekly inflows to the reservoirs, where each path from the root to the terminal node of the tree corresponds to one scenario. Before setting up the stochastic optimization model it is necessary to identify the random input data ξ_1, \dots, ξ_T and to represent it by suitable statistical models and using the scenario

tree generation algorithm. The following sections present construction methods of stochastic inflow and prices based on history data.

Inflow data were available for the past 13 years. The observed data for the weeks 10–30 are shown in Fig. 8.5: Fig. 8.5a displays the inflows to the reservoirs, the hourly spot prices and forward prices observed for 1 year are shown in Fig. 8.5b.

A vector SARIMA model as in Papamichail and Georgiou [84] was fitted to represent the basic process for natural inflows into the reservoirs.

In a model selection stage the autocorrelation function and the partial autocorrelation function of the logarithmically transformed³ time series of inflows were used and by minimizing the AIC,⁴ a SARIMA(1, 0, 2)(2, 0, 2)₅₂ was found to be the appropriate candidate model.⁵

We used hourly spot prices from EEX (European Energy Exchange) as the basis for modeling weekly electricity and pumping prices, which were estimated as regime switching models. These models were used for the tree generation algorithm. We constructed a tree for 8 stages (weeks), namely weeks 16–24 in the year using the dynamic tree generation algorithm (Algorithm 4.9 of Chap. 4) with the specification of the vectors

$$b = (2, 2, 2, 1, 1, 1, 1, 1) \text{ and } d = (5, 5, 5, 7, 7, 7, 10, 10)$$

for the bushiness and the distances.

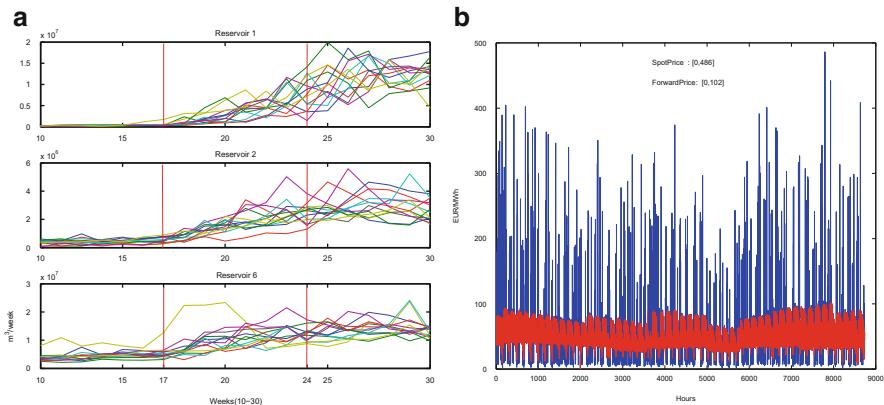


Fig. 8.5 Observed random data. **(a)** Historic inflows to reservoirs for weeks 10–30. **(b)** Historic electricity spot prices and future prices

³To stabilize the variance, the logarithmic transformation was applied to inflow time series.

⁴Akaike's Information Criterion (Akaike, 1974).

⁵The model selection was done using an algorithm from the forecast package in R statistical software.

The generated tree has the following characteristics:

Number of nodes	1,532
Number of scenarios (leaves)	392
Number of nodes per stage	1 4 15 51 97 196 387 389 392
Number of components per node	5 (3 inflows, spot- and pumping prices)

Figure 8.6 displays the first five components of the scenario tree. Since the reservoirs 3, 4, and 5 are downstream reservoirs, they do not have natural (or random) inflows.

We used AIMMS 3.12 for formulating and solving the linear multistage stochastic problem for the chain of 6 reservoirs with 17 arcs. In what follows we discuss the numerical results obtained for the specification $\lambda = 0.75$ and $\alpha = 0.05$ in the objective. Figure 8.7a shows the decision variables pumping and turbining (for reservoir 3) in tree form. The pertaining storage levels are depicted in Fig. 8.7b. The

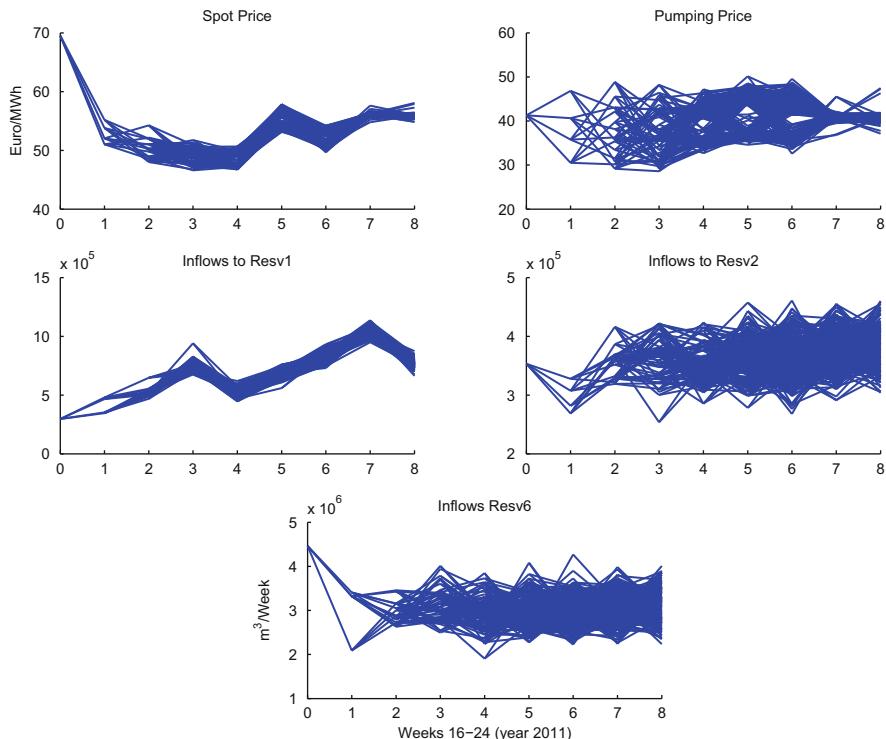


Fig. 8.6 The five tree components. Notice that only one tree was generated with 5 values sitting on each node

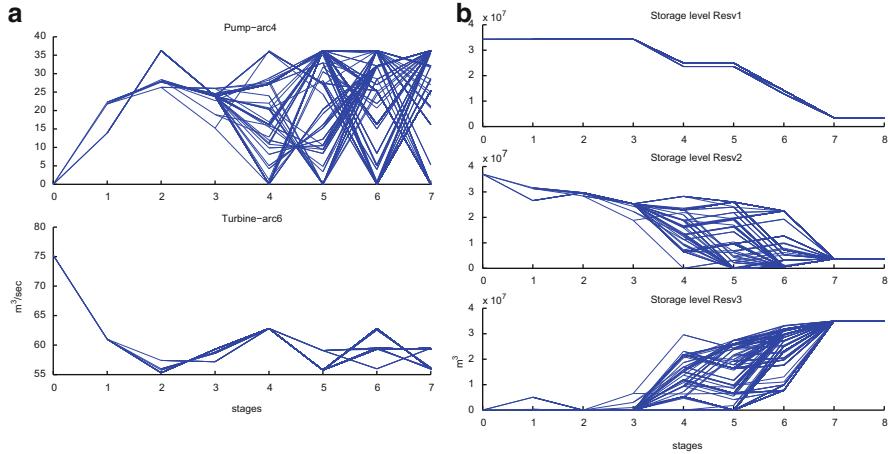


Fig. 8.7 Decision variables and storage levels. (a) The decision variables: pumping at arc 4 (top) and turbinating at arc 6 (bottom). (b) The storage levels for reservoirs 1,2, and 6

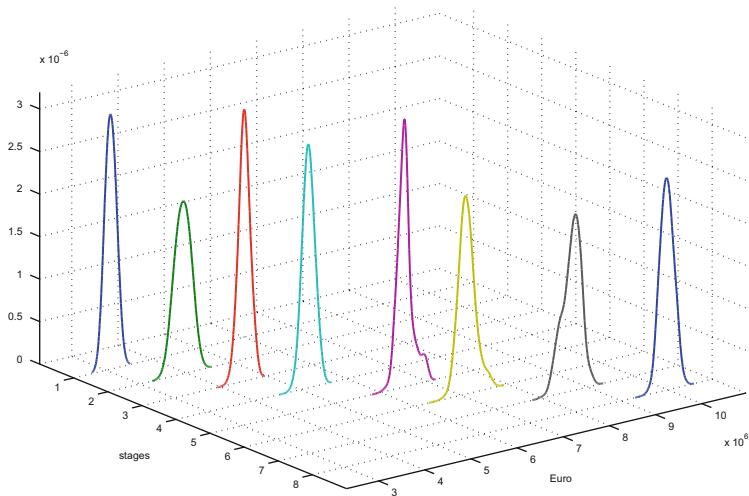


Fig. 8.8 The accumulated cash: density estimates

interconnection of reservoirs 2 and 3 can be observed from the storage level patterns, when, e.g., pumping up decisions show a decrease in storage level of lower reservoir, simultaneously the storage level in the upper reservoir is showing an increase.

The objective is to maximize risk-adjusted expected final cash. Density estimates may represent the random cash per period resulting from optimal storage management. Figure 8.8 depicts the density of accumulated cash over time. There is a clear increase in cash with some, but not too much variability.

In addition, a distributionally robust solution was also calculated. However, due to the complexity of the ambiguity problem, a smaller tree was estimated as the baseline model based on the same stochastic processes as before. The characteristics of the smaller tree are

Number of nodes	421
Number of scenarios (leaves)	119
Number of nodes per stage	1 3 11 35 51 82 119 119
Maximal distances d_t for the tree generation Algorithm 4.9	(5, 5, 5, 8, 8, 8, 12)

The minimax problem was solved as presented in Chap. 7. The worst case trees for different ambiguity radii are shown in Figs. 8.9 (the scenario probabilities) and 8.10 (the scenarios with positive probability). As already observed in Chap. 7, the worst case trees are more and more concentrated on fewer scenarios.

While the worst case scenario trees get simpler with increasing ambiguity radius, the minimax decisions get more complicated: see Fig. 8.11, where the reservoir levels are shown for the different ambiguity radii. Also this fact was already seen in Chap. 7 on ambiguity.

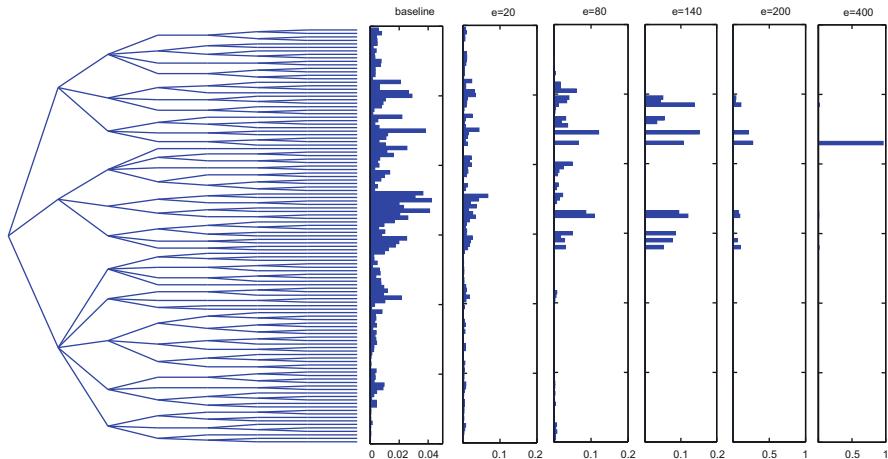


Fig. 8.9 As the ambiguity radius increases (0, 20, 80, 140, 200, 400), the worst case trees are more and more concentrated on few scenarios

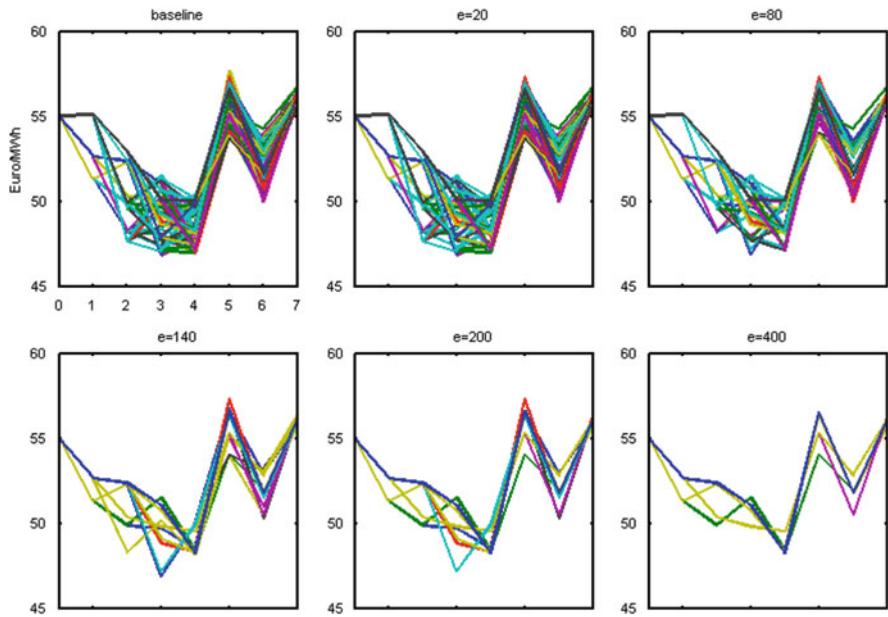


Fig. 8.10 The worst case spotprice models for the different ambiguity radii: 0, 20, 80, 140, 200, 400

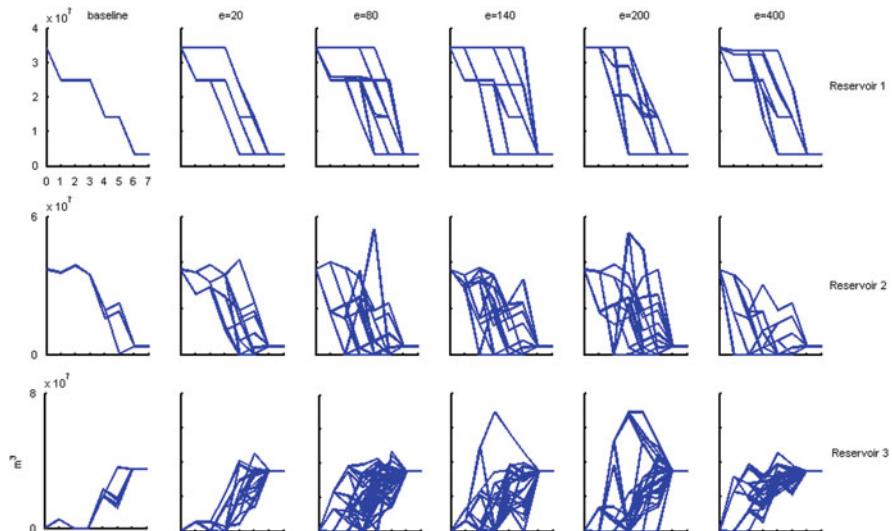


Fig. 8.11 The minimax decisions for the different ambiguity radii 0, 20, 80, 140, 200, 400. The reservoir levels are shown for 3 reservoirs

8.3 Budget Management for Risk-Prone Countries

This example is due to Anna Timonina. Many developing countries suffer from natural hazards such as hurricanes and other tropical storms, floods, or earthquakes (catastrophic events, called here CAT-events), which may destroy large parts of their investments and infrastructure. A big problem is the occurrence of several events within few years. Countries may be robust enough to absorb the first shock, but are typically not able to cope with a second one. For this reason, a multiperiod model is necessary, where one stage represents one budget year. We model the budget decisions of a government, which has a budget B_t to spend in year t , and must decide how to allocate funds for immediate consumption (i.e., the running costs), investment and insurance for infrastructure. The total government-owned infrastructure is called the total stock and is the basis of (tax) income through a proportional production function. Possibly credits can be taken, however, there is a debt limit.

We formulate the model first in time-oriented form and translate it later into a node-oriented one.

Decision Variables

d_t	new debt taken in period t
c_t	consumption in period t
z_t	proportion of insured value of the stock for period t
x_t	investment in period t

Scenario Process

ξ_t	the random relative loss of stock due to natural hazards
---------	----------------------------------------------------------

Expressions

S_t	the stock quantity at period t
-------	----------------------------------

Constants

r	proportional return on the stock
δ	depreciation rate
i_t	interest rate for loans
V	proportional load for insurance premium
\bar{d}	credit limit
\underline{c}	minimal consumption (non-discretionary budget)
ρ	discount factor

Problem Formulation

- (i) At the first stage the policy-maker decides on consumption, insurance, and investment for the next period and has the opportunity to take out a loan.
- (ii) At all following periods, except the last one, some random losses of the stock may occur. In case of existing insurance some compensation payments of the losses may be received. The decisions are again about consumption, insurance,

investments, and debts. Of course, interest payments for previously taken debts are due.

- (iii) At the last period all taken debts have to be repaid. The objective is to maximize the discounted expected utility for consumption plus expected utility for the final stock.

This can be written in the form of a multistage stochastic optimization program in time-oriented form:

$$\begin{aligned} & \text{maximize}_{\text{in } d_t, c_t, z_t, x_t} \quad (1 - \alpha) \sum_{t=1}^T \rho^{-t} \mathbb{E}[U_1(c_t)] + \alpha \mathbb{E}[U_2(S_T)] \\ & \text{subject to} \quad \begin{aligned} & x_t \triangleleft \mathcal{F}_t, \quad z_t \triangleleft \mathcal{F}_t, \quad c_t \geq \underline{c}, \quad d_t \leq \bar{d}, \quad t = 1, \dots, T, \\ & r S_t - \sum_{s=1}^{t-1} d_s i_s \geq c_t + x_t + \mathbb{E}(\xi_{t+1} | \xi^t)(1 + V) z_t S_t, \quad t = 1, \dots, T-1, \\ & r S_T - \sum_{t=1}^T d_t \geq c_T, \\ & S_{t+1} = [(1 - \delta)S_t + x_t](1 - \xi_t) + z_t \xi_t S_t + d_{t+1}, \quad t = 1, \dots, T-2, \\ & S_T = [(1 - \delta)S_{T-1} + x_{T-1}](1 - \xi_T) + z_T \xi_T S_T. \end{aligned} \end{aligned}$$

The utility functions U_1 and U_2 appearing here are chosen as power functions $U(c) = \frac{c^\gamma - 1}{\gamma}$, which have constant relative risk aversion $1 - \gamma$. Changing γ we can introduce risk-averse ($\gamma < 1$), risk-neutral ($\gamma = 1$) or risk-loving ($\gamma > 1$), policy-makers.

The tree approximation leads to the equivalent model in node-oriented way:

$$\begin{aligned} & \text{maximize in } d(\cdot), c(\cdot), z(\cdot), x(\cdot) \\ & \quad (1 - \alpha) \sum_{t=1}^T \rho^{-t} \sum_{n \in \mathcal{N}_t} P(n) U_1(c(n)) \\ & \quad + \alpha \sum_{n \in \mathcal{N}_T} P(n) [U_2(S(n))] \\ & \text{subject to} \quad \begin{aligned} & c(n) \geq \underline{c}, \quad d(n) \leq \bar{d}, \\ & r S(n) - \sum_{j \prec n} d(j) i_{t(j)} \geq c(n) + x(n) \\ & \quad + E(n)(1 + V) z(n) S(n), \quad n \in \mathcal{N} \setminus \mathcal{N}_T, \\ & r S(n) - \sum_{j \prec n} d(j) \geq c(n), \quad n \in \mathcal{N}_T, \\ & S(n) = [(1 - \delta)S(n-) + x(n-)](1 - \xi(n)) \\ & \quad + z(n-) \xi(n) S(n-) + d(n), \quad n \in \mathcal{N} \setminus (\mathcal{N}_T \cup \{1\}), \\ & S(n) = [(1 - \delta)S(n-) + x(n-)](1 - \xi(n)) \\ & \quad + z(n-) \xi(n) S(n-), \quad n \in \mathcal{N}_T, \end{aligned} \end{aligned}$$

with $E(n) = \sum_{m \in n+} \xi(m) \cdot Q(m)$ and $S(1)$ is the initial stock. The random variables ξ_t describing the losses after natural hazard events are distributed independently according to a lognormal distribution censored at 1 such that the values are between 0 and 1. Its expectation was set to 0.1. The time interval was 4 years with yearly decisions (3 decision stages, decisions are made at the beginning of each year). The scenario tree was constructed using optimal quantization of the loss variable with bushiness ranging from 2 (binary tree) to 9.

The optimization problem was solved. The first stage decisions are shown in Fig. 8.12.

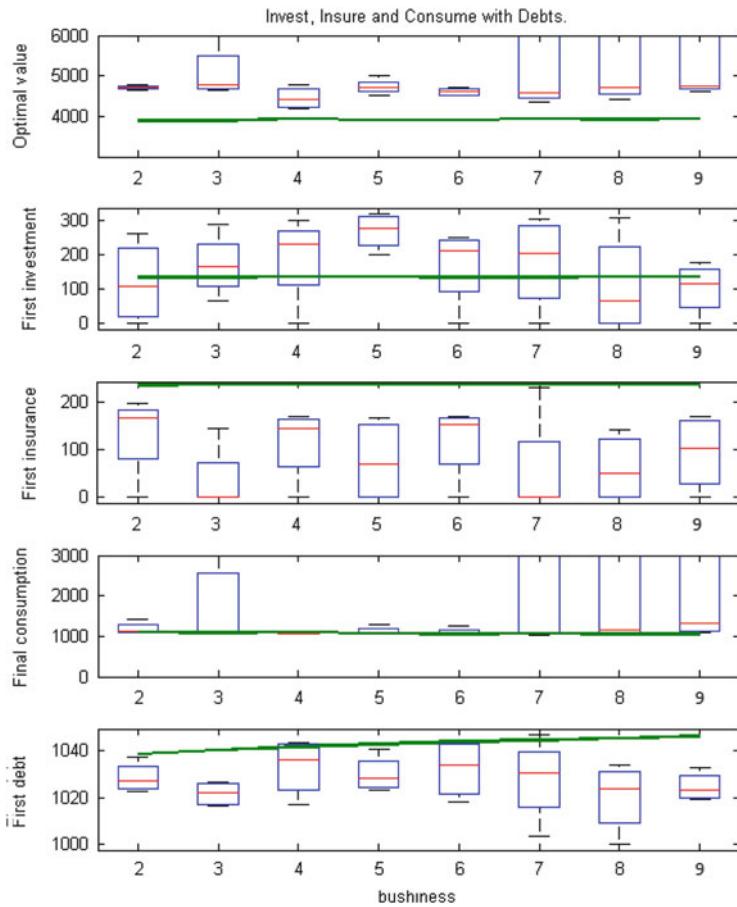


Fig. 8.12 The first stage decisions in dependency of the bushiness of the trees generated by optimal discretization. The *solid line(s)* show(s) the optimal value and the optimal first stage decisions. They are quite stable with respect to bushiness and even small bushiness gives good approximations. The *boxplots* show the range of solutions, when the tree is generated just by Monte Carlo simulation (10 replications). One sees that the values are quite dispersed over the different simulation runs. Monte Carlo sampling is not a reliable tree generation technique

Figure 8.13a illustrates the optimal decisions of a risk-neutral decision maker. Since the decisions are random variables, only their stagewise means are shown. From these two pictures one can derive that

- a risk-neutral policy-maker does not insure, but only invest;
- in case of lower possible deviations of losses the optimal debts tend to be maximal possible, while the consumption tends to be on the lower bound of the life standard in all the periods;

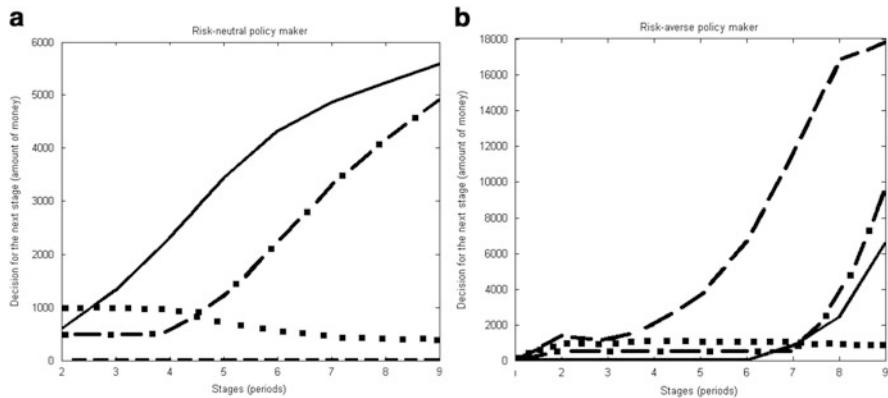


Fig. 8.13 The arithmetic means (over all values at the same stage) of the optimal decisions: investment (solid), consumption (dot-dashed), debt (dotted), insurance (dashed). **(a)** (Average) decisions for a risk neutral decision maker. **(b)** (Average) decisions by a risk averse decision maker

- in case of higher possible deviations of losses the optimal debts tend to decrease while the consumption tends to increase over periods.

In contrast, Fig. 8.13b illustrates the optimal decision for a risk-averse decision maker. From these pictures and other optimization runs with different parameters one can see that

- a risk-averse policy-maker should decide to finance both, investment and insurance;
- he/she should also avoid debts and takes only the amount that is necessary for the recovery after CAT-events;
- the higher the variances of the losses are, the more the policy-maker should insure.

Appendix A

Risk Functionals: Definitions and Notations

Single-Period Risk Functionals. The usual definition of risk functionals (risk measures) is based on a given probability space (Ω, \mathcal{F}, P) and a family of real-valued random variables \mathcal{Y} defined on Ω . \mathcal{Y} is the vector space of random variables described in Pichler [101], although the family of p -integrable functions L^p is a suitable choice in many instances.

A *risk functional* (risk measure) is a mapping $\mathcal{R} : \mathcal{Y} \rightarrow \mathbb{R} \cup \{\infty\}$ with the following properties (cf. Definition 3.2):

- (M) MONOTONICITY: $\mathcal{R}(Y_1) \leq \mathcal{R}(Y_2)$ whenever $Y_1 \leq Y_2$ almost surely;
- (C) CONVEXITY: $\mathcal{R}((1-\lambda)Y_0 + \lambda Y_1) \leq (1-\lambda)\mathcal{R}(Y_0) + \lambda\mathcal{R}(Y_1)$ for $0 \leq \lambda \leq 1$;
- (T) TRANSLATION EQUIVARIANCE:¹ $\mathcal{R}(Y + c) = \mathcal{R}(Y) + c$ if $c \in \mathbb{R}$;

If the following property holds in addition, the risk functional \mathcal{R} is called positively homogeneous, or a coherent risk functional.

- (H) POSITIVE HOMOGENEITY: $\mathcal{R}(\lambda Y) = \lambda\mathcal{R}(Y)$ whenever $\lambda > 0$.

Notice that in this definition the role of the probability space is explicit. However, in all our applications, the value of the risk functional will only depend on the distribution of Y or on the joint distribution of Y and some covariates Z . If it only depends on P^Y itself, the functional is called *version independent* or *law invariant*.

- (I) VERSION INDEPENDENCE (cf. Definition 3.12): $\mathcal{R}(Y) = \mathcal{R}(P^Y)$, where P^Y is the image measure of P under Y , i.e., the distribution of Y . Notice that the argument of \mathcal{R} can be seen as a probability measure on \mathbb{R} and not as a real

¹In an economic or monetary environment this is often called CASH INVARIANCE instead.

random variable on some Ω . $\mathcal{R}(Y) = \mathcal{R}(P^Y)$ indicates that the result is the same for all random variables which have coinciding image measure.

It is clear from the context whether a risk functional is written as a function of a random variable or a probability measure. For version independent risk functionals one may also go back to the original formulation and vary the probability measure P and Ω (which is forbidden for non-version independent risk functionals, since they are defined on $L^p(\Omega, \mathcal{F}, P)$ and not on any other $L^p(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ unless $L^p(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}) \subseteq L^p(\Omega, \mathcal{F}, P)$). Thus, if (I) is satisfied, we may consider the functional \mathcal{R} as having two arguments, Y and P , written

$$\mathcal{R}_P(Y),$$

since it is in any case $\mathcal{R}(P^Y)$. Therefore, for these functionals it makes sense to study the dependency on P in addition to the dependency on Y . Notice that if \mathcal{R} is a version independent risk functional and P is a probability distribution on \mathbb{R} , then the notations $\mathcal{R}(P)$ and $\mathcal{R}_P(\text{id})$, where id is the identity function, are equivalent.

A version-independent functional is called compound concave, if the mapping $P \mapsto \mathcal{R}_P(Y)$ is concave (cf. Definition 3.25).

(CC) COMPOUND CONCAVITY: $\mathcal{R}_{\lambda P + (1-\lambda)\tilde{P}}(Y) \geq \lambda \mathcal{R}_P(Y) + (1-\lambda)\mathcal{R}_{\tilde{P}}(Y)$, for $0 \leq \lambda \leq 1$ and all real random variables Y (we set $\mathcal{R}(Y) = \infty$ if Y is not in the domain of \mathcal{R}) and probability measures P and \tilde{P} .

The two properties (C) and (CC) can be put together to the notion of convex-concavity of risk functionals.

(C-CC) CONVEX-CONCAVITY: $(Y, P) \mapsto \mathcal{R}_P(Y) = \mathcal{R}(P^Y)$ is convex in Y (property (C)) and compound concave in P (property (CC)).

Example A.1. All distortion risk functionals (see Definition 3.6) are convex-concave (cf. Theorem 3.27).

Conditional Risk Functionals. Let Y be a real valued random variable on a probability space (Ω, \mathcal{F}, P) and let \mathcal{F}_1 be a sub sigma-algebra of \mathcal{F} . Then one may find the collections of (regular) conditional distributions $P^Y(\cdot | \mathcal{F}_1)$ with the following property:

For every measurable set A , $P^Y(A | \mathcal{F}_1)$ is a real valued, \mathcal{F}_1 -measurable random variable on Ω , such that $P(\cdot | \mathcal{F})(\omega)$ is a probability measure for all $\omega \in \Omega$.

In this case, the mapping $\omega \mapsto \mathcal{R}(P^Y(\cdot | \mathcal{F}_1))(\omega)$ is well defined (it may take the value $+\infty$) and this mapping is called the *conditional risk functional* and for short written as $\mathcal{R}(Y | \mathcal{F}_1)$. Notice that a conditional risk functional can only be defined for version invariant basic functionals \mathcal{R} but is itself not version invariant, because it depends on the particular Ω . However, the distribution of the random variable $\omega \mapsto \mathcal{R}(Y | \mathcal{F}_1)$ depends only on the distribution of the conditional distributions, i.e., the nested distribution \mathbb{P} of the structure $(\Omega, (\mathcal{F}_1, \mathcal{F}), P, Y)$, where $(\mathcal{F}_1, \mathcal{F})$ is a filtration. Now, the conditional risk functional $\mathcal{R}(Y | \mathcal{F}_1)$ may

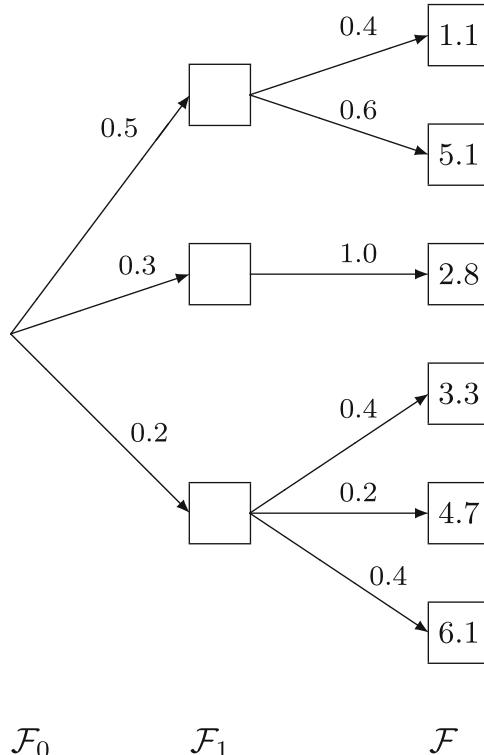


Fig. A.1 The nested structure in tree form

be concatenated with an arbitrary other version-independent risk functional, say \mathcal{R}_1 , giving $\mathcal{R}_1(\mathcal{R}(Y|\mathcal{F}_1))$. Since the nested distribution \mathbb{P} encodes the complete structure one may also take the nested distribution as the argument and write

$$\mathcal{R}_1(\mathcal{R}(Y|\mathcal{F}_1)) = \mathcal{R}_1 \circ \mathcal{R}(\mathbb{P}), \quad \text{where} \quad (\Omega, (\mathcal{F}_1, \mathcal{F}), P, Y) \sim \mathbb{P}.$$

The following example illustrates this.

Example. Consider the nested distribution in Fig. A.1. This nested distribution can also be written as

$$\mathbb{P} = \left[\begin{array}{c|c|c} 0.5 & 0.3 & 0.2 \\ \hline \left[\begin{array}{cc} 0.4 & 0.6 \\ \hline 1.1 & 5.1 \end{array} \right] & \left[\begin{array}{c} 1.0 \\ \hline 2.8 \end{array} \right] & \left[\begin{array}{ccc} 0.4 & 0.2 & 0.4 \\ \hline 3.3 & 4.7 & 6.1 \end{array} \right] \end{array} \right].$$

Notice that this nested distribution does not have intermediate values, but only values for the last “period.” The random variable Y defined through this structure has distribution

$$\begin{bmatrix} 0.20 & 0.30 & 0.08 & 0.04 & 0.30 & 0.08 \\ \hline 1.1 & 2.8 & 3.3 & 4.7 & 5.1 & 6.1 \end{bmatrix},$$

the conditional distributions $P^Y(\cdot | \mathcal{F}_1)$ take the values

$$\begin{cases} \begin{bmatrix} 0.4 & 0.6 \\ \hline 1.1 & 5.1 \end{bmatrix} & \text{with probability 0.5,} \\ \begin{bmatrix} 1.0 \\ \hline 2.8 \end{bmatrix} & \text{with probability 0.3,} \\ \begin{bmatrix} 0.4 & 0.2 & 0.4 \\ \hline 3.3 & 4.7 & 6.1 \end{bmatrix} & \text{with probability 0.2.} \end{cases}$$

Calculating the conditional upper $\text{AV@R}_{60\%}(Y | \mathcal{F}_1)$ one gets the distribution

$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ \hline 5.1 & 2.8 & 6.1 \end{bmatrix},$$

while the conditional expectation $\mathbb{E}(Y | \mathcal{F}_1)$ has the distribution

$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ \hline 3.5 & 2.8 & 4.7 \end{bmatrix}.$$

One may now calculate the concatenated risk functionals

$$\text{AV@R}_{0.6}(\text{AV@R}_{0.6}(Y | \mathcal{F}_1)) = \text{AV@R}_{0.6} \circ \text{AV@R}_{0.6}(\mathbb{P}) = 5.6,$$

$$\mathbb{E}(\text{AV@R}_{0.6}(Y | \mathcal{F}_1)) = \mathbb{E} \circ \text{AV@R}_{0.6}(\mathbb{P}) = 4.61,$$

$$\text{AV@R}_{0.6}(\mathbb{E}(Y | \mathcal{F}_1)) = \text{AV@R}_{0.6} \circ \mathbb{E}(\mathbb{P}) = 4.1,$$

$$\mathbb{E}(\mathbb{E}(Y | \mathcal{F}_1)) = \mathbb{E}(Y) = \mathbb{E} \circ \mathbb{E}(\mathbb{P}) = 3.53.$$

Lemma A.1. *Let \mathcal{R} be version-independent, compound concave and continuous with respect to weak convergence. Then for all σ -algebras \mathcal{F}_1*

$$\mathcal{R}(Y) \geq \mathbb{E}[\mathcal{R}(Y | \mathcal{F}_1)]. \tag{A.1}$$

Proof. If the σ -algebra is finite, then (A.1) is a direct consequence of the definition. The general case follows from approximating \mathcal{F}_1 by a finite σ -algebra. \square

A.1 Multiperiod Risk Functionals

Let now $Y = (Y_1, \dots, Y_T)$ be a real valued stochastic process defined on a filtered probability space $(\Omega, \mathfrak{F}, P)$ such that $Y \triangleleft \mathfrak{F}$, i.e., Y is \mathfrak{F} -adapted. A multiperiod risk functional is a mapping from a family \mathcal{Y} of \mathfrak{F} -adapted processes to the real line: $Y \mapsto \mathcal{R}(Y; \mathfrak{F})$. For stressing the fact that the process Y must be \mathfrak{F} -adapted we write the filtration as an argument of the risk functional. Multiperiod risk functionals may exhibit the similar properties as single-period ones. In particular, the properties (C) and (H) extend naturally. Monotonicity is meant in the almost surely (a.s.) sense for all components Y_t . Translation equivariance is typically formulated as

$$\mathcal{R}(Y_1, \dots, Y_t + c, \dots, Y_T; \mathfrak{F}) = \mathcal{R}(Y_1, \dots, Y_t, \dots, Y_T; \mathfrak{F}) + c$$

for all c and all t . The property of version independence (law invariance) can now be formulated in the following sense:

- (I) MULTIPERIOD VERSION INDEPENDENCE: A multiperiod risk functional is version independent, if its value depends only on the nested distribution, i.e.,

$$(\Omega, \mathfrak{F}, P, (Y_1, \dots, Y_T)) \sim \mathbb{P} \quad \text{and} \quad (\Omega', \mathfrak{F}', P', (Y'_1, \dots, Y'_T)) \sim \mathbb{P}$$

implies that

$$\mathcal{R}(Y; \mathfrak{F}) = \mathcal{R}(Y'; \mathfrak{F}') =: \mathcal{R}(\mathbb{P}), \quad (\text{A.2})$$

where \mathbb{P} is the nested distribution of the structure $(\Omega, \mathfrak{F}, P, Y = (Y_1, \dots, Y_t))$.

Examples for multiperiod risk functionals are given below. They are all multiperiod version independent and continuous with respect to the nested distance. For polyhedral functionals, a proof of this assertion is given.

- Sums of single-period functionals

$$\mathcal{R}_1(Y_1) + \dots + \mathcal{R}_T(Y_T);$$

- Compositions of conditional functionals (Ruszcyński and Shapiro [118])

$$\mathcal{R}_1(Y_1 + \mathcal{R}_2(Y_2 + \dots \mathcal{R}_T(Y_T | \mathcal{F}_{T-1}) \dots | \mathcal{F}_1) \dots);$$

- SEC (separable expected conditional) functionals (Pflug and Römisch [97])

$$\sum_{t=1}^{T-1} \mathbb{E}[\mathcal{R}_t(Y_{t+1} | \mathcal{F}_t)]; \quad (\text{A.3})$$

- Polyhedral functionals (Eichhorn and Römisch [40]). Polyhedral risk functionals are defined as solutions of stochastic optimization problems:

Definition A.2. A multiperiod probability functional \mathcal{R} is called polyhedral if there are dimensions $k_t, c_t \in \mathbb{R}^{k_t}$, nonempty polyhedral sets $V_t \subseteq \mathbb{R}^{k_t}, t = 0, \dots, T$, $w_{t,\tau} \in \mathbb{R}^{k_t - k_\tau}, \tau = 0, \dots, t, t = 1, \dots, T$, such that

$$\mathcal{R}(Y; \mathfrak{F}) = \inf \left\{ \sum_{t=0}^T \mathbb{E}(Y_t | Z_t) \middle| \begin{array}{l} c_0 - \sum_{\tau=1}^T w_{\tau,t} \mathbb{E}(Z_\tau) \in V_0 \\ c_t - \sum_{\tau=t}^T w_{\tau,t} \mathbb{E}(Z_\tau | \mathcal{F}_t) \in V_t \quad t = 1, \dots, T \end{array} \right\}. \quad (\text{A.4})$$

This is its dual representation (see Pflug and Römisch [97]).

Example A.3. Here are some examples of polyhedral risk functionals, see Eichhorn and Römisch [41]. The first three do not depend on the filtration \mathfrak{F} .

- $\mathcal{R}(Y_1, \dots, Y_T) = \sum_{t=1}^T \text{AV@R}_{\alpha_t}(Y_t)$
- $\mathcal{R}(Y_1, \dots, Y_T) = \text{AV@R}_\alpha \left(\sum_{t=1}^T Y_t \right)$
- $\mathcal{R}(Y_1, \dots, Y_T) = \text{AV@R}_\alpha(\min(Y_1, \dots, Y_T))$
- $\mathcal{R}(Y_1, \dots, Y_T; \mathfrak{F}) = \sum_{t=1}^T \mathbb{E}[\text{AV@R}_{\alpha_t}(Y_t | \mathcal{F}_{t-1})]$
- $\mathcal{R}(Y_1, \dots, Y_T; \mathfrak{F}) = \sum_{t=1}^T \mathbb{E}[\text{AV@R}_{\alpha_t}(Y_t - Y_{t-1} | \mathcal{F}_{t-1})]$

Obviously, all polyhedral risk functionals are version independent according to Definition (I), that is, they can be defined for nested distributions \mathbb{P} and one may write $\mathcal{R}(\mathbb{P})$. Moreover, the following proposition is valid.

Proposition A.4. *For polyhedral multiperiod risk functionals \mathcal{R} , the mapping $\mathbb{P} \mapsto \mathcal{R}(\mathbb{P})$ is Lipschitz continuous with respect to the nested distance \mathbf{d}_1 .*

Proof. The functional (A.4) is the minimal value of a multistage stochastic program. By renaming Y as ξ , Z as x and the constraint set as \mathbb{X} one sees that the assumptions of Theorem 6.4 are fulfilled, from which the Lipschitz continuity follows. \square

A.2 Information Monotonicity

Version independent risk functionals (see (I), (A.2)) depend only on the nested distribution $\mathbb{P} \sim (P, \mathfrak{F}, P, Y)$. While properties of the mapping $Y \mapsto \mathcal{R}(Y)$ are often considered (such as monotonicity or convexity), one may also look at the mapping $\mathfrak{F} \mapsto \mathcal{R}(Y; \mathfrak{F})$, where we have made the dependence on the filtration \mathfrak{F} explicit.

A quite fundamental property is information monotonicity. It states that better information reduces risk.

Definition A.5. A multiperiod risk functional $\mathcal{R}(Y; \mathfrak{F})$ is called *information monotone*, if

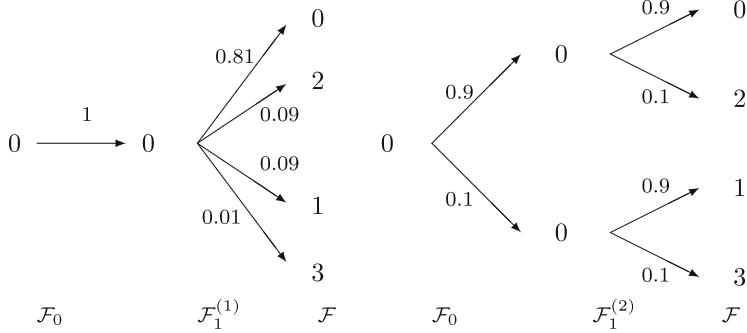


Fig. A.2 The concatenation of upper AV@R's is not information monotone

$$\mathfrak{F}' \subseteq \mathfrak{F} \quad \text{implies that } \mathcal{R}(Y; \mathfrak{F}') \geq \mathcal{R}(Y; \mathfrak{F}).$$

Example A.6 (Time Consistency Contradicts Information Monotonicity). Concatenations of conditional AV@R's are time consistent as is argued in Example 5.15. However, they are typically not information monotone. Cf. Fig. A.2. In both examples, the final costs Y are the same, but in the right tree, the filtration is finer. One calculates

$$\text{AV@R}_{90\%} \left[\text{AV@R}_{90\%} \left(Y | \mathcal{F}_1^{(1)} \right) \right] = 2.1 < 3 = \text{AV@R}_{90\%} \left[\text{AV@R}_{90\%} \left(Y | \mathcal{F}_1^{(2)} \right) \right]$$

showing that the finer filtration leads to a higher risk. Notice, however, that

$$\mathbb{E} \left[\text{AV@R}_{90\%} \left(Y | \mathcal{F}_1^{(1)} \right) \right] = \mathbb{E} \left[\text{AV@R}_{90\%} \left(Y | \mathcal{F}_1^{(2)} \right) \right] = 2.1.$$

Looking at (A.4) in Definition A.2 one sees that multiperiod polyhedral risk functionals are information monotone. Information monotonicity may also be defined for conditional risk functionals defined in Sect. 3.6.

Definition A.7. A conditional risk mapping $\mathcal{R}(\cdot)$ is called *information monotone*, if $\mathcal{F}' \subseteq \mathcal{F}$ implies that

$$\mathbb{E} [\mathcal{R}(Y | \mathcal{F}') \mathbf{1}_B] \geq \mathbb{E} [\mathcal{R}(Y | \mathcal{F}) \mathbf{1}_B]$$

for all $B \in \mathcal{F}'$ and $Y \in \mathcal{Y}$.

Example A.8 (See Kovacevic and Pflug [70]). A positively homogeneous conditional risk functional of the form

$$\mathcal{R}_t (Y | \mathcal{F}) = \text{ess sup} \{ \mathbb{E} (Y \cdot Z | \mathcal{F}) - \mathcal{R}_t^+ (Z | \mathcal{F}) : Z \in \mathcal{S}_t (\mathcal{F}) \}$$

$(\mathcal{S}_t(\mathcal{F})$ is the supergradient hull) is information monotone iff the following condition is fulfilled

$$\mathcal{F}' \subseteq \mathcal{F} \Rightarrow \mathcal{S}(\mathcal{F}) \subseteq \mathcal{S}(\mathcal{F}'). \quad (\text{A.5})$$

The family of positively homogeneous, information monotone conditional acceptability mappings forms a convex cone, which is closed under the operation of sup-convolution.

Proposition A.9. *Let $\bar{\mathcal{R}}(Y|\mathfrak{F}) = \mathcal{R}_1(Y_1 + \mathcal{R}_2(Y_2 + \dots \mathcal{R}_T(Y_T|\mathcal{F}_{T-1}) \dots |\mathcal{F}_1) \dots)$ be a composition of information monotone, positively homogeneous conditional risk mappings \mathcal{R}_t with supergradient hulls $\mathcal{S}_t(\cdot)$. Then $\bar{\mathcal{R}}(Y|\mathfrak{F})$ is information monotone if and only if*

$$\mathcal{S}_{t-1}(\mathcal{F}') \cdot \mathcal{S}_t(\mathcal{F}) \supseteq \mathcal{S}_t(\mathcal{F}') \quad (\text{A.6})$$

for every t , $1 \leq t \leq T - 1$ and every pair of σ -algebras $\mathcal{F}', \mathcal{F}$ such that $\mathcal{F}' \subseteq \mathcal{F}$.

Strict information monotonicity is achieved if and only if \supset holds instead of \supseteq for some t in (A.6).

Proof. See Kovacevic and Pflug [70]. □

Corollary A.10. *Consider a filtration \mathfrak{F} , and a sequence of positively homogeneous conditional risk functionals \mathcal{R}_t . A composition $\mathcal{R}_1(Y_1 + \mathcal{R}_2(Y_2 + \dots \mathcal{R}_T(Y_T|\mathcal{F}_{T-1}) \dots |\mathcal{F}_1) \dots)$ based on those mappings is information monotone only if the following conditions are fulfilled:*

- (i) Any occurrence of a conditional essential supremum can be preceded by any mapping \mathcal{R}_{t-1} that fulfills (A.5).
- (ii) Any other mapping \mathcal{R}_t than the essential supremum must be preceded by a conditional expectation.

In particular, all SEC functionals (A.3) are information monotone.

Proof. See Kovacevic and Pflug [70]. □

Example A.11. In rare cases, also non-positively homogeneous risk functionals may be information monotone. For instance, the nested entropic risk functionals, which are defined as concatenations $\mathcal{R}_0 \circ \mathcal{R}_1 \circ \dots \circ \mathcal{R}_{T-1}$, where $\mathcal{R}_t(Y|\mathcal{F}_t) = \frac{1}{\gamma_t} \log \mathbb{E}[\exp(\gamma_t Y)|\mathcal{F}_t]$, and $\gamma_t \geq 0$ for $Y \in \mathcal{Y}$. Nested entropic risk functionals are information monotone, if $\gamma_0 \leq \gamma_1 \leq \dots \leq \gamma_{T-1}$ as it is shown in [70].

Appendix B

Minimax Theorems

Let \mathbb{X} and \mathbb{Y} be two topological spaces and let f be a real function on $\mathbb{X} \times \mathbb{Y}$. We consider the minimax problem

$$\min_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} f(x, y). \quad (\text{B.1})$$

A value $x^* \in \mathbb{X}$ is a solution of (B.1), if

$$\max_{y \in \mathbb{Y}} f(x^*, y) \leq \max_{y \in \mathbb{Y}} f(x, y)$$

for all $x \in \mathbb{X}$. A basic requirement is that $x \mapsto f(x, y)$ is lower semicontinuous for all $y \in \mathbb{Y}$ and $y \mapsto f(x, y)$ is upper semicontinuous for all $x \in \mathbb{X}$. This requirement makes sure that the minima (maxima, resp.) are attained in (B.1). John von Neumann introduced in 1928 the concept of *minimax theorems*, which state the equality

$$\min_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} f(x, y) = \max_{y \in \mathbb{Y}} \min_{x \in \mathbb{X}} f(x, y) \quad (\text{B.2})$$

under various assumptions on \mathbb{X} , \mathbb{Y} , and f . Notice first that the following relation holds always

$$\min_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} f(x, y) \geq \max_{y \in \mathbb{Y}} \min_{x \in \mathbb{X}} f(x, y). \quad (\text{B.3})$$

The opposite inequality is related to the existence of a saddle point.

Definition B.1. A pair (x^*, y^*) is called a *saddle point* of f , if

$$f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*)$$

for all $x \in \mathbb{X}$ and $y \in \mathbb{Y}$.

Lemma B.2. *If there exists a saddle point, then a minimax theorem (B.2) holds.*

Proof. It suffices to show the reverse inequality of (B.3). It follows from

$$\min_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} f(x, y) \leq \max_{y \in \mathbb{Y}} f(x^*, y) \leq f(x^*, y^*) \leq \min_{x \in \mathbb{X}} f(x, y^*) \leq \max_{y \in \mathbb{Y}} \min_{x \in \mathbb{X}} f(x, y).$$

□

Here is an extension of von Neumann's original minimax theorem.

Theorem B.3 (Sion [132]). *Let \mathbb{X} and \mathbb{Y} be a compact convex subsets of a linear topological spaces. Let further f be a real valued function on $\mathbb{X} \times \mathbb{Y}$ such that*

- (i) $x \mapsto f(x, y)$ is quasiconvex and lower semicontinuous for all $y \in \mathbb{Y}$
- (ii) $y \mapsto f(x, y)$ is quasiconcave and upper semicontinuous for all $x \in \mathbb{X}$.

Then

$$\min_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} f(x, y) = \max_{y \in \mathbb{Y}} \min_{x \in \mathbb{X}} f(x, y). \quad (\text{B.4})$$

If the assumption about compactness of \mathbb{X} is dropped, then (B.4) has to be replaced by

$$\inf_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} f(x, y) = \max_{y \in \mathbb{Y}} \inf_{x \in \mathbb{X}} f(x, y).$$

Proof. see Sion [132].

□

Proposition B.4. *Let the assumptions of Theorem B.3 be fulfilled. Then $x^* \in \mathbb{X}$ is a solution of (B.1) if and only if there exists a $y^* \in \mathbb{Y}$ such that the pair (x^*, y^*) is a saddle point.*

Proof. The existence of a saddle point is sufficient for a minimax theorem to hold by Lemma B.2. We repeat the argument here: if (x^*, y^*) is a saddle point, then $\max_{y \in \mathbb{Y}} f(x^*, y) = f(x^*, y^*) \leq \max_{y \in \mathbb{Y}} f(x, y)$ and hence x^* is a solution of (B.1). Conversely, if x^* is a solution of the minimax problem (B.1) and if $y^* \in \operatorname{argmax}_y f(x^*, y)$, then

$$f(x^*, y) \leq f(x^*, y^*)$$

for all $y \in \mathbb{Y}$. Moreover, since $f(x^*, y^*) = \min_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} f(x, y) = \max_{y \in \mathbb{Y}} \min_{x \in \mathbb{X}} f(x, y)$,

$$f(x^*, y^*) = \min_{x \in \mathbb{X}} f(x, y^*) \leq f(x, y^*)$$

for all $x \in \mathbb{X}$.

□

Algorithms for finding saddle points are well studied in literature. Under the assumption of convex-concavity and smoothness, a necessary and sufficient condition for a saddlepoint $\zeta^* = (x^*, y^*)$ is that it solves the simultaneous system of equations: $\mathcal{E}(\zeta^*) \equiv \begin{pmatrix} \nabla_x f(x, y) \\ \nabla_y f(x, y) \end{pmatrix} = 0$. Sometimes it is even more convenient to solve the problem

$$\min_{\zeta} \left\{ \frac{1}{2} \|\mathcal{E}(\zeta)\|_2^2 \right\} \quad (\text{B.5})$$

rather than $\mathcal{E}(\zeta) = 0$ (cf. Rustom [117]).

For solving (B.5) iterative algorithms follow

$$\zeta_{k+1} = \zeta_k + \alpha_k d_k,$$

where d_k is the direction of the progress (typically the negative gradient with respect to x and the gradient with respect to y) and α_k is the step size.

Some authors [20, 23] proposed a gradient based algorithm for unconstrained problem based on direction d_k and step size strategy α_k , such that sufficient progress at each iteration is ensured. In [117], more saddle point computation algorithms are presented and discussed: quadratic approximation algorithm for constrained problems based on [104], interior point saddle point algorithm for constrained problems as elaborated in [120], and finally a Quasi-Newton algorithm for nonlinear systems.

One may think that a “coordinatewise” iteration

$$x^{(k+1)} = \operatorname{argmin}_{x \in \mathbb{X}} f(x, y^{(k)}) \quad (\text{B.6})$$

$$y^{(k+1)} = \operatorname{argmax}_{y \in \mathbb{Y}} f(x^{(k+1)}, y) \quad (\text{B.7})$$

would work. But even under the strict convexity-concavity of f and compactness of \mathbb{X} and \mathbb{Y} , the convergence of this algorithm cannot be guaranteed, as the following example shows.

Example B.5. Let $f(x, y) = x^2 - y^2 + 2xy$ and $\mathbb{X} = [-2, 2]$, $\mathbb{Y} = [-2, 2]$, and let $x^{(1)} = 1$, $y^{(1)} = 1$. Then the iteration (B.6)–(B.7) leads to

$$\begin{aligned} x^{(2)} &= -1, \\ y^{(2)} &= -1, \\ x^{(3)} &= 1, \\ y^{(3)} &= 1, \end{aligned}$$

which means that the algorithm oscillates and does not approach the saddle point $(0, 0)$.

The following modification of (B.6)–(B.7), however, works:

$$x^{(k+1)} \in \operatorname{argmin}_{x \in \mathbb{X}} \max_{1 \leq i \leq k} f(x, y^{(i)}) \quad (\text{B.8})$$

$$y^{(k+1)} \in \operatorname{argmax}_{y \in \mathbb{Y}} f(x^{(k+1)}, y). \quad (\text{B.9})$$

The procedure stops, if $x^{(k+1)} = x^{(k)}$ and $y^{(k+1)} = y^{(k)}$.

Proposition B.6. *Let \mathbb{X} and \mathbb{Y} be compact and let $(x, y) \mapsto f(x, y)$ be jointly continuous. Then every cluster point of the iteration given by ((B.8)–(B.9)) is a minimax solution.*

Proof. Denote by $f^k := \max_{1 \leq l \leq k} f(x^{(k+1)}, y^{(l)})$. Notice that

$$f^k = \min_{x \in \mathbb{X}} \max_{1 \leq i \leq k} f(x, y^{(i)}) \leq \min_{x \in \mathbb{X}} \max_{1 \leq i \leq k+1} f(x, y^{(i)}) = f^{k+1}.$$

Since the function f is bounded, f^k converges to $f^* := \sup f^k < \infty$. Moreover, by compactness, the sequence $x^{(k)}$ has one or several cluster points. Let x^* be such a cluster point. We show that $f^* = \max_{y \in \mathbb{Y}} f(x^*, y)$. Since trivially $f^* \leq \max_{y \in \mathbb{Y}} f(x^*, y)$, suppose that $f^* < \max_{y \in \mathbb{Y}} f(x^*, y)$. Then there must exist a y^+ such that $f(x^*, y^+) > f^*$. By continuity this inequality must then hold in a neighborhood of x^* and therefore there must exist a $x^{(k)}$ with $f(x^{(k)}, y^+) > f^*$, which contradicts the construction of the iteration.

Finally, we show that $x^* \in \operatorname{argmin}_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} f(x, y)$. If not, then there must exist x^+ such that

$$\max_{y \in \mathbb{Y}} f(x^+, y) < \max_{y \in \mathbb{Y}} f(x^*, y).$$

But by construction $\max_{1 \leq i \leq k} f(x^+, y^{(i)}) \geq \max_{1 \leq i \leq k} f(x^{(k+1)}, y^{(i)}) = f^k$ and letting k tend to infinity, one sees that $\max_{y \in \mathbb{Y}} f(x^+, y) \geq f^* = \max_{y \in \mathbb{Y}} f(x^*, y)$ and this contradiction shows that x^* — and thus every cluster point — is a solution of the minimax problem. \square

Appendix C

Comparison of Weighted Norms

It is well known that all norms are equivalent on \mathbb{R}^m (a finite dimensional vector space), because the closed unit ball is bounded and compact. Here we provide the constants for weighted norms.

Lemma C.1 (Comparison of ℓ^p -Norms). *Let $\|\cdot\|_p$ denote the weighted ℓ^p -norm, $\|x\|_p := \left(\sum_{t=0}^T w_t |x_t|^p\right)^{1/p}$. Then*

$$\|x\|_p \leq \left(\sum_{t=0}^T w_t\right)^{\frac{1}{p}-\frac{1}{r}} \cdot \|x\|_r \quad \text{whenever } p \leq r,$$

and

$$\|x\|_p \leq \left(\min_{i \in \{0, \dots, T\}} w_i\right)^{\frac{1}{p}-\frac{1}{r}} \cdot \|x\|_r \quad \text{whenever } p \geq r.$$

The bounds represent the best possible bounds.

Proof. Consider first the case $p \geq r$. Without loss of generality one may assume that $\|x\|_r^r = \min_{t \in \{0, \dots, T\}} w_t$. It follows that $|x_t| \leq 1$ for all $t \in \{0, \dots, T\}$, and consequently $|x_t|^p \leq |x_t|^r$, such that $\sum_{t=0}^T w_t |x_t|^p \leq \sum_{t=0}^T w_t |x_t|^r = \min_{t \in \{0, \dots, T\}} w_t$, or

$$\frac{1}{\min_{t \in \{0, \dots, T\}} w_t} \sum_{t=0}^T w_t |x_t|^p \leq \frac{1}{\min_{t \in \{0, \dots, T\}} w_t} \sum_{t=0}^T w_t |x_t|^r = 1.$$

Extracting the root it holds further that

$$\left(\frac{1}{\min_{t \in \{0, \dots, T\}} w_t} \sum_{t=0}^T w_t |x_t|^p \right)^{\frac{1}{p}} \leq 1 = \left(\frac{1}{\min_{t \in \{0, \dots, T\}} w_t} \sum_{t=0}^T w_t |x_t|^r \right)^{\frac{1}{r}},$$

that is

$$\|x\|_p \leq \left(\min_{t \in \{0, \dots, T\}} w_t \right)^{\frac{1}{p} - \frac{1}{r}} \cdot \|x\|_r,$$

the result. Equality holds for the vector $x_t = \delta_{t^*}(t)$, where t^* is the index for which $w_{t^*} = \min_{t \in \{0, \dots, T\}} w_t$.

As for $p \leq r$ the assertion follows from Hölder's inequality for the conjugate exponents $\alpha = \frac{r}{r-p}$ and $\alpha' = \frac{r}{p}$, as

$$\begin{aligned} \|x\|_p^p &= \sum_{t=0}^T w_t |x_t|^p = \sum_{t=1}^n w_t^{1-\frac{p}{r}} \cdot w_t^{\frac{p}{r}} |x_t|^p \\ &\leq \left(\sum_{t=0}^T w_t^{(1-\frac{p}{r})\alpha} \right)^{1/\alpha} \cdot \left(\sum_{t=0}^T \left(w_0^{\frac{p}{r}} |x_t|^p \right)^{\alpha'} \right)^{1/\alpha'} \\ &= \left(\sum_{t=0}^T w_t \right)^{\frac{r-p}{r}} \cdot \left(\sum_{t=0}^T w_t |x_t|^r \right)^{\frac{p}{r}}, \end{aligned}$$

and hence $\|x\|_p \leq \left(\sum_{t=0}^T w_t \right)^{\frac{1}{p} - \frac{1}{r}} \cdot \|x\|_r$. Equality is obtained for the constant vector $x = (1, 1, \dots, 1)$. \square

Appendix D

The Canonical Construction for Nested Distributions

Suppose that \mathbb{P} is a nested distributions of depth T . Recall that the nested distributions are a purely distributional concept. For every nested distribution one may construct in a canonical way a probability space and a valued tree process on it, such that its nested distribution coincides with the given one.

Compare this with the fact that a probability distribution P on \mathbb{R}^m is defined without a reference to a probability space. If one wants to introduce a probability space Ω on which a random variable X with distribution P is defined, one may take the probability space $(\mathbb{R}^m, \mathcal{B}^m, P)$, where \mathcal{B}^m is the Borel sigma-algebra and X is the identity. This is called the canonical construction ensuring that one may always construct a random variable X with a given distribution.

One may now ask the parallel question: given a nested distribution \mathbb{P} , does there exist a filtered probability space $(\Omega, \mathfrak{F}, P)$ and a random process $\xi \triangleleft \mathfrak{F}$ such that the nested distribution of the value-and-information structure $(\Omega, \mathfrak{F}, P, \xi)$ has nested distribution \mathbb{P} ? This question is answered by the *canonical construction*, where the name of each node is the nested distribution pertaining to the subtree, for which this node is the root. Two valued trees representing the nested distributions are equivalent, iff the respective canonical constructions are identical.

Example D.1. Consider the following nested distribution

$$\mathbb{P} = \begin{bmatrix} & 0.5 & 0.5 \\ & \hline & 2 & 2 \\ & \begin{bmatrix} 1.0 \\ \hline 3 \end{bmatrix} & \begin{bmatrix} 1.0 \\ \hline 1 \end{bmatrix} \end{bmatrix}. \quad (\text{D.1})$$

We construct a valued tree such that the names of the nodes are the corresponding nested distributions of the subtrees, see Fig. D.1.

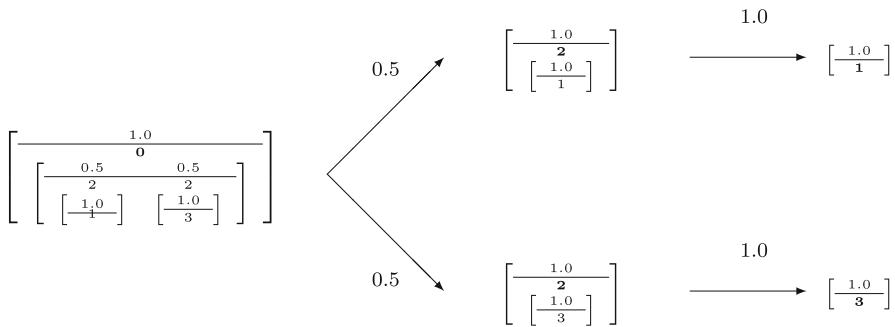


Fig. D.1 The canonical construction of the nested distribution (D.1)

It is evident that this valuated tree represents the nested distribution. The values of ξ on each node are the boldfaced numbers.

Bibliography

1. Acerbi, C.: Spectral measures of risk: A coherent representation of subjective risk aversion. *J. Bank. Finance* **26**, 1505–1518 (2002)
2. Acerbi, C., Simonetti, P.: Portfolio optimization with spectral measures of risk. *EconPapers* (2002)
3. Ambrosio, L., Gigli, N., Savaré, G.: *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd edn. Birkhäuser Verlag AG, Basel (2005)
4. Analui, B., Pflug, G.: On distributionnally robust multistage stochastic optimization. *Comput. Manag. Sci.* **11**(3), 197–220 (2014). DOI 10.1007/s10287-014-0213-y
5. Arrow, K., Gould, F., Howe, S.: General saddle point results for constrained optimization. *Math. Programm.* **5**, 225–234 (1973)
6. Artzner, P., Delbaen, F., Eber, J.M., Heath, D.: Coherent measures of risk. *Math. Finance* **9**, 203–228 (1999)
7. Artzner, P., Delbaen, F., Eber, J.M., Heath, D., Ku, H.: Coherent multiperiod risk adjusted values and Bellman’s principle. *Ann. Oper. Res.* **152**, 5–22 (2007). DOI 10.1007/s10479-006-0132-6
8. Bally, V., Pagés, G., Printems, J.: A quantization tree method for pricing and hedging multidimensional american options. *Math. Finance* **15**(1), 119–168 (2005)
9. Bellman, R.E.: *Dynamic Programming*. Princeton University Press, Princeton (1957)
10. Ben-Tal, A., Nemirovski, A.: Robust solution of uncertain linear programs. *Oper. Res. Lett.* **25**, 1–13 (1999)
11. Birge, J.R., Louveaux, F.: *Introduction to Stochastic Programming*. Springer, New York (1997)
12. Bisschop, J.: *AIMMS Optimization Modeling*. Lulu Enterprises Incorporated (2006)
13. Bolley, F.: Separability and completeness for the Wasserstein distance. In: Donati-Martin, C., Émery, M., Rouault, A., Stricker, C. (eds.) *Séminaire de Probabilités XLI*, Lecture Notes in Mathematics, vol. 1934, pp. 371–377. Springer, Berlin/Heidelberg (2008)
14. Bolley, F., Guillin, A., Villani, C.: Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probab. Theory Relat. Fields* **137**(3–4), 541–593 (2007). DOI 10.1007/s00440-006-0004-7. URL <http://dx.doi.org/10.1007/s00440-006-0004-7>
15. Calafiore, G.: Ambiguous risk measures and optimal robust portfolios. *SIAM J. Control Optim.* **18** (3), 853–877 (2007)
16. Carpentier, P., Chancelier, J.P., Cohen, G., de Lara, M., Girardeau, P.: Dynamic consistency for stochastic optimal control problems. *Ann. Oper. Res.* **200**(1), 247–263 (2012). DOI 10.1007/s10479-011-1027-8

17. Cheridito, P., Kupper, M.: Recursiveness of indifference prices and translation-invariant preferences. *Math. Financ. Econ.* **2**(3), 173–188 (2009). DOI [10.1007/s11579-009-0020-3](https://doi.org/10.1007/s11579-009-0020-3)
18. Cheridito, P., Kupper, M.: Composition of time-consistent dynamic monetary risk measures in discrete time. *Int. J. Theor. Appl. Finance (IJTAF)* **14**(1), 137–162 (2011). DOI <http://dx.doi.org/10.1142/S0219024911006292>
19. Collado, R.A., Papp, D., Ruszczyński, A.P.: Scenario decomposition of risk-averse multistage stochastic programming problems. *Ann. Oper. Res.* **200**(1), 147–170 (2012). <http://dx.doi.org/10.1007/s10479-011-0935-y>
20. Danilin, Y.M., Panin, V.M.: Methods for searching saddle points. *Kibernetika* **3**, 119–124 (1974)
21. Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* **58**, 596–612 (2010)
22. Demyanov, V.F., Malozemov, V.N.: *Introduction to Minimax*. Wiley, New York (1974)
23. Demyanov, V.F., Pevnyi, A.B.: Numerical methods for finding saddle points. *USSR Comput. Math. Math. Phys.* **12**, 1099–1127 (1972)
24. Denneberg, D.: Distorted probabilities and insurance premiums. In: *Proceedings of the 14th SOR*, Ulm. Athenäum, Frankfurt (1989)
25. Denneberg, D.: Distorted probabilities and insurance premiums. *Meth. Oper. Res.* **63**, 21–42 (1990)
26. Depraz, O., Gerber, H.U.: On convex principles of premium calculation. *Insur. Math. Econ.* **4**(3), 179–189 (1985). DOI [http://dx.doi.org/10.1016/0167-6687\(85\)90014-9](http://dx.doi.org/10.1016/0167-6687(85)90014-9)
27. Dick, J., Pillichshammer, F.: *Digital Nets and Sequences*. Cambridge University Press, Cambridge (2010)
28. Dobrushin, R.L.: Central limit theorem for non-standard Markov chains. *Dokl. Akad. Nauk SSSR* **102**(1), 5–8 (1956)
29. Dobrushin, R.L.: Prescribing a system of random variables by conditional distributions. *Theor. Probab. Appl.* **15**, 458–486 (1970)
30. Dudley, R.M.: The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Stat.* **40**(1), 40–50 (1969)
31. Dunford, N., Schwartz, J.T.: *Linear Operators. Part I. General Theory*. Wiley-Interscience, New York (1957)
32. Dupačová, J.: On minimax decision rule in stochastic linear programming. *Stud. Math. Program.* 47–60 (1980)
33. Dupačová, J.: The minimax approach to stochastic programming and an illustrative application. *Stochastics* **20**, 73–88 (1987)
34. Dupačová, J.: Stability and sensitivity-analysis for stochastic programming. *Ann. Oper. Res.* **27**, 115–142 (1990)
35. Dupačová, J.: Uncertainties in minimax stochastic programs. *Optimization* **1**, 191–220 (2010)
36. Dupačová, J., Gröwe-Kuska, N., Römisch, W.: Scenario reduction in stochastic programming. *Math. Programm. A* **95**(3), 493–511 (2003). DOI [10.1007/s10107-002-0331-0](https://doi.org/10.1007/s10107-002-0331-0)
37. Dupačová, J., Hurt, J., Štěpán, J.: *Stochastic Modeling in Economics and Finance. Applied Optimization*. Kluwer Academic, Dordrecht (2003)
38. Durrett, R.: *Probability: Theory and Examples*. Duxbury Advanced Series. Thompson, Belmont (2005)
39. Edirisinghe, N.C.P.: Stochastic Programming: The state of the Art in Honor of George B. Dantzig, chap. *Stochastic Programming Approximations using Limited Moment Information with Application to Asset Allocation*, pp. 97–138. International Series in Operations Research and Management Science. Springer (2011)
40. Eichhorn, A., Römisch, W.: Polyhedral risk measures in stochastic programming. *SIAM J. Optim.* **16**(1), 69–95 (2005)
41. Eichhorn, A., Römisch, W.: Dynamic risk management in electricity portfolio optimization via polyhedral risk functionals. In: *Proc. of the IEEE Power Engineering Society (PES) General Meeting*, Pittsburgh, PA, USA (2008)

42. Ellsberg, D.: Risk, ambiguity and Savage axioms. *Q. J. Econ.* **75**(4), 643–669 (1961)
43. Fan, K.: Minimax theorems. *Proc. N.A.S.* **39**, 42–47 (1953)
44. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Springer, New York (2006)
45. Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. *Int. Stat. Rev.* **70**(3), 419–435 (2002)
46. Glasserman, P.: Monte Carlo Methods in Financial Engineering, vol. 53. Springer, New York (2004)
47. Goh, J., Sim, M.: Distributionally robust optimization and its tractable approximations. *Oper. Res.* **58**, 902–917 (2010)
48. Graf, S., Luschgy, H.: Foundations of Quantization for Probability Distributions, Lecture Notes in Mathematics, vol. 1730. Springer, Berlin/Heidelberg (2000)
49. Gutjahr, W.J., Pichler, A.: Stochastic multi-objective optimization: a survey on non-scalarizing method. *Ann. Oper. Res.* 1–25 (2013). DOI 10.1007/s10479-013-1369-5
50. Hansen, L.P., Sargent, T.J.: Robustness. Princeton University Press, Princeton (2007)
51. Hartigan, J.A.: Clustering Algorithms. Wiley, New York (1975)
52. Heitsch, H., Römisch, W.: Scenario reduction algorithms in stochastic programming. *Comput. Optim. Appl. Stoch. Programm.* **24**(2–3), 187–206 (2003)
53. Heitsch, H., Römisch, W.: Scenario tree modeling for multistage stochastic programs. *Math. Program. A* **118**, 371–406 (2009)
54. Heitsch, H., Römisch, W.: Scenario tree reduction for multistage stochastic programs. *Comput. Manag. Sci.* **2**, 117–133 (2009)
55. Heitsch, H., Römisch, W., Strugarek, C.: Stability of multistage stochastic programs. *SIAM J. Optim.* **17**(2), 511–525 (2006)
56. Henrion, R., Strugarek, C.: Convexity of chance constraints with independent random variables. *Comput. Optim. Appl.* **41**, 263–276 (2008)
57. Henrion, R., Strugarek, C.: Convexity of chance constraints with dependent random variables: the use of copulae. In: Bertocchi, M., Consigli, G., Dempster, M. (eds.), *Stochastic Optimization Methods in Finance and Energy*, International Series in Operations Research and Management Science, Vol. 163, pp. 427–439. Springer, New York (2011)
58. Heyde, C.C.: On a property of the lognormal distribution. *J. Roy. Stat. Soc. B* **25**(2), 392–393 (1963)
59. Hochreiter, R., Pflug, G.Ch.: Financial scenario generation for stochastic multi-stage decision processes as facility location problems. *Ann. Oper. Res.* **152**(1), 257–272 (2007)
60. Hoffman, A.J., Kruskal, J.B.: Integral Boundary Points of Convex Polyhedra, chap. 3, pp. 49–76. Springer, Berlin/Heidelberg (2010)
61. Jagannathan, R.: Minimax procedure for a class of linear programs under uncertainty. *Oper. Res.* **25**, 173–177 (1977)
62. Jobert, A., Rogers, L.C.G.: Valuations and dynamic convex risk measures. *Math. Finance* **18**(1), 1–22 (2008). DOI <http://dx.doi.org/10.1111/j.1467-9965.2007.00320.x>
63. Jouini, E., Schachermayer, W., Touzi, N.: Law invariant risk measures have the Fatou property. In: Kusuoka, S., Yamazaki, A. (eds.) *Advances in Mathematical Economics*, vol. 9, pp. 49–71. Springer, Tokyo (2006). DOI 10.1007/4-431-34342-3_4
64. Kantorovich, L.: On the translocation of masses. *C.R. Acad. Sci. URSS* **37**, 199–201 (1942)
65. Karatzas, I., Shreve, S.E.: Methods of Mathematical Finance. Stochastic Modelling and Applied Probability. Springer, New York (1998)
66. Kersting, G.: Die Geschwindigkeit der Glivenko-Cantelli-Konvergenz gemessen in der Prohorov Metrik. *Math. Z.* **163**, 65–102 (1978). In German
67. King, A.J., Wallace, S.W.: Modeling with Stochastic Programming, Springer Series in Operations Research and Financial Engineering, vol. XVI. Springer, New York (2013)
68. Knight, F.: Risk, Uncertainty and Profit. Houghton Mifflin, Boston (1921)
69. Komiya, H.: Elementary proof for Sion's minimax theorem. *Kodai Math. J.* **11**, 5–7 (1988). Sion

70. Kovacevic, R., Pflug, G.Ch.: Are time consistent valuations information-monotone? *Int. J. Theor. Appl. Finan.* **17**(1) (2014). DOI 10.1142/S0219024914500034
71. Kovacevic, R., Pflug, G.Ch.: Time consistency and information monotonicity of multiperiod acceptability functionals. No. 8 in Radon Series on Computational and Applied Mathematics, pp. 347–369. de Gruyter, Berlin (2009)
72. Kudô, H.: A note on the strong convergence of σ -algebras. *Ann. Probab.* **2**(1), 76–83 (1974). URL <http://dx.doi.org/10.1214/aop/1176996752>
73. Kusuoka, S.: On law invariant coherent risk measures. *Adv. Math. Econ.* **3**, 83–95 (2001)
74. Lemieux, C.: Monte Carlo and Quasi Monte Carlo Sampling. Springer Series in Statistics. Springer, New York (2009)
75. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp. 281–297. University of California Press, Berkeley, California (1967). URL <http://projecteuclid.org/euclid.bsmsp/1200512992>
76. Maggioni, F., Allevi, E., Bertocchi, M.: Measures of information in multistage stochastic programming. *SPEPS Ser.* **2** (2012)
77. Maggioni, F., Pflug, G.: Bounds and approximations for multistage stochastic optimization. Manuscript, U Bergamo
78. Mirkov, R., Pflug, G.Ch.: Tree approximations of dynamic stochastic programs. *SIAM J. Optim.* **18**(3), 1082–1105 (2007)
79. Monge, G.: Mémoire sur la théorie des déblais et de remblais. *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pp. 666–704 (1781)
80. Müller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. Wiley Series in Probability and Statistics. Wiley, Chichester (2002)
81. Na, S., Neuhoff, D.L.: Bennett's integral for vector quantizers. *IEEE Trans. Inform. Theory* **41**(4), 886–900 (1995)
82. von Neumann, J.: Zur Theorie der Gesellschaftsspiele. *Math. Ann.* (100), 295–320 (1928). In German
83. Niederreiter, H.: Random Number Generation and Quasi Monte Carlo Methods. SIAM, Philadelphia (1992)
84. Papamichail, D.M., Georgiou, P.E.: Seasonal ARIMA inflow models for reservoir sizing. *J. Am. Water Res. Assoc.* **37**(4), 877–885 (2001)
85. Parthasarathy, K., Kalyanapuram, R.: Probability Measures on Metric Spaces. Academic press, New York (1972)
86. Penner, I.: Dynamic convex risk measures: Time consistency, prudence, and sustainability. Ph.D. thesis, Humboldt University of Berlin (2009)
87. Pflug, G.Ch., Pichler, A., Wozabal, D.: The 1/N investment strategy is optimal under high model ambiguity. *J. Bank. Finance* **36**, 410–417 (2012)
88. Pflug, G.Ch.: Optimization of Stochastic Models, The Kluwer International Series in Engineering and Computer Science, vol. 373. Kluwer Academic, Dordrecht (1996). URL <http://link.springer.com/book/10.1007/F978-1-4613-1449-3>
89. Pflug, G.Ch.: Some remarks on the value-at-risk and the conditional value-at-risk. In: Uryasev, S. (ed.) *Probabilistic Constrained Optimization: Methodology and Applications*, pp. 272–281. Kluwer Academic, Dordrecht (2000)
90. Pflug, G.Ch.: Scenario tree generation for multiperiod financial optimization by optimal discretization. *Math. Programm.* **89**, 251–271 (2001). DOI 10.1007/s101070000202
91. Pflug, G.Ch.: On distortion functionals. *Stat. Risk Model.* **24**, 45–60 (2006). DOI dx.doi.org/10.1524/stnd.2006.24.1.45
92. Pflug, G.Ch.: Version-independence and nested distribution in multistage stochastic optimization. *SIAM J. Optim.* **20**, 1406–1420 (2009). DOI <http://dx.doi.org/10.1137/080718401>
93. Pflug, G.Ch., Pichler, A.: Approximations for Probability Distributions and Stochastic Optimization Problems, International Series in Operations Research & Management

- Science, vol. 163, chap. 15, pp. 343–387. Springer, New York (2011). DOI 10.1007/978-1-4419-9586-5_15
94. Pflug, G.Ch., Pichler, A.: A distance for multistage stochastic optimization models. *SIAM J. Optim.* **22**(1), 1–23 (2012). DOI <http://dx.doi.org/10.1137/110825054>
 95. Pflug, G.Ch., Pichler, A.: Time consistent decisions and temporal decomposition of coherent risk functionals. *Optimization online* (2012)
 96. Pflug, G.Ch., Pichler, A.: Time-inconsistent multistage stochastic programs: martingale bounds. No. 3 in Stochastic Programming E-Print Series. Humboldt Universität, Institut für Mathematik (2012)
 97. Pflug, G.Ch., Römisch, W.: Modeling, Measuring and Managing Risk. World Scientific, River Edge (2007)
 98. Pflug, G.Ch., Wozabal, D.: Ambiguity in portfolio selection. *Quant. Finance* **7**(4), 435–442 (2007). DOI 10.1080/14697680701455410
 99. Pichler, A.: Distance of probability measures and respective continuity properties of acceptability functionals. Ph.D. thesis, University of Vienna, Vienna, Austria (2010)
 100. Pichler, A.: Evaluations of risk measures for different probability measures. *SIAM J. Optim.* **23**(1), 530–551 (2013). DOI <http://dx.doi.org/10.1137/110857088>
 101. Pichler, A.: The natural Banach space for version independent risk measures. *Insur. Math. Econ.* **53**(2), 405–415 (2013). DOI <http://dx.doi.org/10.1016/j.insmatheco.2013.07.005>. URL <http://www.sciencedirect.com/science/article/pii/S0167668713001054>
 102. Pichler, A.: Premiums and reserves, adjusted by distortions. *Scand. Actuarial J.* (2013). DOI 10.1080/03461238.2013.830228
 103. Prekopa, A.: On logarithmic concave measures and functions. *Acta Sci. Math.* **34**, 335–343 (1973)
 104. Qi, L., Sun, W.: An iterative method for the minimax problem. *Minimax and Applications*. Kluwer Academic, Boston (1995)
 105. Rachev, S.T.: Probability Metrics and the Stability of Stochastic Models. Wiley, West Sussex (1991)
 106. Rachev, S.T., Römisch, W.: Quantitative stability in stochastic programming: The method of probability method. *Math. Oper. Res.* **27**(4), 792–818 (2002)
 107. Rachev, S.T., Rüschendorf, L.: Mass Transportation Problems Vol. I: Theory, Vol. II: Applications, Probability and its applications, vol. XXV. Springer, New York (1998)
 108. Rachev, S.T., Stoyanov, S.V., Fabozzi, F.J.: A Probability Metrics Approach to Financial Risk Measures. Wiley, London (2011).
 109. Robinson, W., Wets, R.J.B.: Stability in two stage stochastic programming. *SIAM J. Control Optim.* **25**, 1409–1416 (1987)
 110. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
 111. Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. *J. Risk* **2**, 21–41 (2000)
 112. Rockafellar, R.T., Wets, R.J.B.: Nonanticipativity and L^1 -martingales in stochastic optimization problems. *Math. Programm. Study* **6**, 170–187 (1976)
 113. Rockafellar, R.T., Wets, R.J.B.: The optimal recourse problem in discrete time: L^1 -multipliers for inequality constraints. *SIAM J. Control Optim.* **16**(1), 16–36 (1978)
 114. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer, New York (1997)
 115. Römisch, W., Schultz, R.: Stability analysis for stochastic programs. *Ann. Oper. Res.* **30**, 241–266 (1991)
 116. Rüschendorf, L.: The Wasserstein distance and approximtion theorems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **70**, 117–129 (1985)
 117. Rustem, B., Howe, M.: Algorithms for Worst-Case Design and Applications to Risk Management. Princeton University Press, Princeton (2002)
 118. Ruszczyński, A., Shapiro, A.: Conditional risk mappings. *Math. Oper. Res.* **31**, 544–561 (2006)
 119. Ruszczyński, A.: Nonlinear Optimization. Princeton University Press, Princeton (2006)

120. Sasaki, H.: An interior penalty method for minimax for problems with constraints. *SIAM J. Control Optim.* **12**, 643–649 (1974)
121. Scarf, H.E.: A min-max solution of an inventory problem. In: Arrow, K.J., Karlin, S., Scarf, H. (eds.). *Studies in the Mathematical Theory of Inventory and Production*. Stanford University Press, Stanford (1958)
122. Schachermayer, W., Kupper, M.: Representation results for law invariant time consistent functions. *Math. Financ. Econ.* **2**, 189–210 (2009). DOI 10.1007/s11579-009-0019-9
123. Shaked, M., Shanthikumar, J.G.: *Stochastic Order*. Springer Series in Statistics. Springer, New York (2007)
124. Shapiro, A.: On complexity of multistage stochastic programs. *Oper. Res. Lett.* **34**, 1–8 (2006)
125. Shapiro, A.: On a time consistency concept in risk averse multistage stochastic programming. *Oper. Res. Lett.* **37**(37), 143–147 (2009)
126. Shapiro, A.: On Kusuoka representations of law invariant risk measures. *Math. Oper. Res.* **38**(1), 142–152 (2013). <http://dx.doi.org/10.1287/moor.1120.0563>
127. Shapiro, A.: Time consistency of dynamic risk measures. *Oper. Res. Lett.* **40**(6), 436–439 (2012). DOI 10.1016/j.orl.2012.08.007. URL <http://www.sciencedirect.com/science/article/pii/S0167637712001010>
128. Shapiro, A., Ahmed, Sh.: On a class of minimax stochastic programs. *SIAM J. Optim.* **14**, 1237–1249 (2004)
129. Shapiro, A., Dentcheva, D., Ruszczyński, A.: *Lectures on Stochastic Programming*. MOS-SIAM Series on Optimization. MQS-SIAM Series on Optimization 9 (2009). URL <http://pubs.siam.org/doi/book/10.1137/1.9780898718751>
130. Shapiro, A., Kleywegt, A.J.: Minimax analysis of stochastic problems. *Optim. Meth. Software* **17**, 523–542 (2002)
131. Shapiro, A., Nemirovski, A.: On complexity of stochastic programming problems. In: Jeyakumar, V., Rubinov, A. (eds.) *Continuous Optimization: Current Trends and Applications*, pp. 111–144. Springer, New York (2005)
132. Sion, M.: On general minimax theorems. *Pac. J. Math.* **8**(1), 171–176 (1958)
133. Thiele, A.: Robust stochastic programming with uncertain probabilities. *IMA J. Manag. Math.* **19**, 289–321 (2008)
134. van der Vaart, A.W.: *Asymptotic Statistics*. Cambridge University Press, Cambridge (1998)
135. Vallander, S.S.: Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.* **18**, 784–786 (1973)
136. Vershik, A.M.: Kantorovich metric: Initial history and little-known applications. *J. Math. Sci.* **133**(4), 1410–1417 (2006). DOI 10.1007/s10958-006-0056-3. URL <http://dx.doi.org/10.1007/s10958-006-0056-3>
137. Villani, C.: *Topics in Optimal Transportation*, Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (2003)
138. Žáčková, J.: On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování Matematiky* **91**, 423–430 (1966)
139. Wang, S.S., Young, V.R., Panjer, H.H.: Axiomatic characterization of insurance prices. *Insur. Math. Econ.* **21**, 173–183 (1997). DOI [http://dx.doi.org/10.1016/S0167-6687\(97\)00031-0](http://dx.doi.org/10.1016/S0167-6687(97)00031-0)
140. Weber, S.: Distribution-invariant risk measures, information, and dynamic consistency. *Math. Finance* **16**(2), 419–441 (2006)
141. Werner, A.S., Pichler, A., Midthun, K.T., Hellemo, L., Tomsgard, A.: Risk measures in multi-horizon scenario trees. In: Kovacevic, R., Pflug, G.Ch., Vespucci, M.T. (eds.) *Handbook of Risk Management in Energy Production and Trading*, International Series in Operations Research & Management Science, vol. 199, chap. 8, pp. 183–208. Springer, New York (2013). URL <http://www.springer.com/business+26+management/operations+research/book/978-1-4614-9034-0>
142. Williams, D.: *Probability with Martingales*. Cambridge University Press, Cambridge (1991)
143. Wozabal., D.: A framework for optimization under ambiguity. *Ann. Oper. Res.* **193**(1), 21–47 (2012)

144. Zador, P.L.: Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. Inform. Theory* **28**, 139–149 (1982)
145. Zillober, Ch., Schittkowski, K., Moritzen, K.: Very large scale optimization by sequential convex programming. *Optim. Math. Software* **19** (1), 103–120 (2004)

Index

- Acceptance consistent, 181
- Ambiguity
 - price of, 255
 - set, 230
- Ball
 - nested, 239
 - Wasserstein, 235
- Barycenter, 53
- Canonical construction, 289
- Centers of order r , 138
- Chance constrained problem, 4
- Clairvoyant
 - filtration, 220
 - solution, 220
- Co-monotone, 108
- Compound
 - concave, 181
 - convex, 181
- Conditional tail expectation, 98
- Conjugate function, 111
- Constraints
 - chance constraints, 9
 - implicit, 10
- Cost function, 46
- Decision
 - here-and-now, 9
 - wait-and-see, 9
- Decision problem
 - risk-averse, 4
 - risk-neutral, 4
- Decomposition theorem, 189
- Dirac measure, 50
- Discrepancy, 131
 - star, 130
- Discrete metric, 55
- Distance
 - bounded Lipschitz, 45
 - Fortet–Mourier, 45
 - inherited, 47
 - Kantorovich, 50
 - nested, 74
 - semi-distance, 42
 - uniform, 45
 - variational, 44
 - Wasserstein, 48
- Distributionally robust solution, 230
- Distributional robustness, 232
- Dual function, 106
- Dynamically decomposable, 178
- Dynamic programming principle, 196
- Empirical
 - distribution function, 127
 - measure, 127
- Ergodic coefficient, 152
- Expectation, 7
- Facility location, 136
- Fenchel–Moreau representation, 109

- Filtration, 12
 - natural filtration, 73
- Gain of distributional robustness, 249
- Here-and-now, 9
- History process, 73
- Inequality
 - Edmundson–Madansky, 219
 - Fenchel–Young, 112
 - Jensen, 219
- Information
 - available, 12
 - monotone, 281
- Information monotonicity, 280
- Kusuoka representation, 103
 - conditional, 195
- Legendre transformation, 105
- Lorentz curve, 98
- Marginal, 47
- Matrix
 - recourse matrix, 10
 - technology matrix, 10
- Measure
 - image, 7
 - pushforward, 7
- Minimax theorems, 283
- Model uncertainty, 232
- Moment matching, 58
 - caveat, 43
- Monte Carlo method, 127
- Nested distribution, 30
- Node-oriented modeling, 28
- Nonanticipativity constraint, 12
- Optimal quantizers, 138
- Points
 - principal, 136
 - representative, 136
- Price of ambiguity, 249
- Problem
 - flowergirl, 8
 - newsboy, 2
- Process
 - law of a stochastic process, 72
 - stochastic, 71
- Process distance, 74
- Product measure, 49
- Program
 - multistage stochastic, 13
 - two-stage, 9
- Quantile, 7
- Quantization error, 145
- Quantizer, 133
 - principal points, 136
 - representative points, 136
- Quasi-Monte Carlo, 130
- Recourse
 - complete, 9
 - function, 9
 - relative complete, 9
- Recourse problem, 9
- Recursive risk functional, 182
- Rejection consistent, 181
- Risk functional, 96
 - coherent risk functional, 275
 - conditional risk functional, 120
 - distortion risk functionals, 99
 - max-risk functional, 97
 - polyhedral, 280
 - spectral risk functional, 99
- Robust optimization, 4
- Saddle point, 283
- Scenario splitting, 206
- Scenario trees, 24
- Second order stochastic dominance, 102
- Set
 - argmin , 3
- Stochastic approximation, 142
- Stochastic basis, 71
- Time consistency, 177
- Time-oriented modeling, 28
- Transport
 - map, 133
 - nested transport plan, 74
 - transport plan, 48

- Transportation kernel, 247
Tree, 23
 - bushiness, 158
 - compound, 242
 - convex combination, 37
 - fully valued, 24
 - scenario, 23Tree process, 25
- Average Value-at-Risk at random risk level, 121
conditional Value-at-Risk, 98
Value of stochastic solution, 16
Version independent, 103
 - multiperiod, 279Voronoi tessellation, 135
- Wait-and-see, 9
- Value-at-Risk, 97
(upper) Average Value-at-Risk, 97