SCENARIO DECOMPOSITION OF RISK-AVERSE STOCHASTIC OPTIMIZATION PROBLEMS

BY RICARDO A. COLLADO SOTO

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Operations Research

Written under the direction of

Dr. Andrzej Ruszczyński

and approved by

_____

_____

_____

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2010

ABSTRACT OF THE DISSERTATION

Scenario Decomposition of Risk-Averse Stochastic Optimization Problems

by Ricardo A. Collado Soto

Dissertation Director: Dr. Andrzej Ruszczyński

In the last decade the theory of coherent risk measures established itself as an alternative to expected utility models of risk averse preferences in stochastic optimization. Recently, increased attention is paid to dynamic measures of risk, which allow for risk-averse evaluation of streams of future costs or rewards. When used in stochastic optimization models, dynamic risk measures lead to a new class of problems, which are significantly more complex than their risk-neutral counterparts. Decomposition, an established and efficient approach to risk-neutral multistage stochastic optimization problems, cannot be directly applied to risk-averse models. With dynamic risk measures, the main feature facilitating decomposition, the integral form of the objective function, is absent. Our main objective is to overcome this difficulty by exploiting specific structure of dynamic risk measures, and to develop new decomposition methods that extend the ideas of earlier approaches to risk-neutral problems.

In this work we develop generalizations of scenario decomposition methods, in the spirit of J.M. Mulvey and A. Ruszczyński, *A new scenario decomposition method for large-scale stochastic optimization*, Operations Research 43, 1995. The key to success

is the use of dual properties of dynamic measures of risk to construct a family of risk-neutral approximations of the problem. First, we define and analyze a two-stage risk-averse stochastic optimization problem. Next, we develop methods to solve efficiently this problem. Later, we formally define a multistage risk-averse stochastic optimization problem and we discuss its properties. We also develop efficient methods to solve the multistage problem and apply these to an inventory planning and assembly problem. Finally, we analyze and compare the results of our computational experiments.

Acknowledgements

This research project would not have been possible without the support of many people. I wish to express my gratitude to my advisor, Dr. Andrzej Ruszczyński who has been abundantly helpful and offered invaluable assistance, support and guidance. My deepest gratitude are also due to our director, Dr. Endre Boros and to all the faculty and staff of the Rutgers Center for Operations Research without whose knowledge, patience and assistance this study would not have been successful. I would also like to acknowledge Dr. Alexander Kelmans from the University of Puerto Rico and his Discrete Mathematics Seminar for getting me started in the long journey towards my degree.

Special thanks go to my family for their support and in particular to my lovely wife, Melba, without whose love, encouragement, help and dedication I would not have finished this thesis. I would also like to thank my friends: Dávid Papp, Anupama Reddy, Silas Clark, Ileana Perez, Liliana Perez and Anibal Soto for helping make my time in the Ph.D. program more fun and interesting.

Finally, I would like to acknowledge the United Negro College Fund and the NASA-Harriett G. Jenkins Pre-Doctoral Fellowship Program for the economic assistance that allowed me to focus on my research and continue with my studies uninterrupted.

Table of Contents

Chapter 1

Introduction

The traditional approach of optimizing the expectation operator in stochastic programs successfully introduces uncertainty of events in stochastic models but might fail to convey the element of risk that certain modeling problems face. A clear example is given by the problem of maximizing the return rate of a portfolio. The optimal solution obtained by optimizing the portfolio's return expectation will suggest to concentrate the investments in the assets with the highest expected return rate. This is not a reasonable solution because it fails to recognize the possibility of loosing money due to changes in the market that might not benefit that sole asset with highest expectation. Of course, a way to avoid this would be to diversify the investment taking somehow into account the risk involved in holding a position versus the random changes of the market.

During the last decade researchers have developed the coherent risk measures as an alternative to the expectation operator in the traditional stochastic programs. These operators are consistent with a systematized theory of risk as presented in [4, 5] and by substituting the expectation operator give rise to *risk-averse programs*. Coherent risk measures have a rich axiomatic theory including duality and differentiability, thus allowing the development of efficient methods for the solution of risk-averse programs, see for example [3, 11, 23]. In [38, 39, 37, 41] we can find a comprehensive treatment of the coherent risk measures and risk-averse optimization including the development of multi-stage risk-averse programs (see, e.g., [36] for a general development of multi staged stochastic problems).

Simply speaking, a *measure of risk* is a function $\rho : \mathcal{X} \to \mathbb{R}$ which assigns to each random variable X, from a set $\mathcal{X}$ of possible positions, a value that corresponds to

the assessment of the risk involved by holding said position. These scalar measures of risk allow the risk manager to order and compare different positions according to their corresponding risk value. The usefulness of a particular risk measure is dependent on its properties and how closely it models the perception of risk at the different positions for the particular application at hand.

Until 20 years ago most of the risk-related research was performed on a case-by-case basis where for particular applications different notions of risk measures were developed with varying range of success. In the pre-Markowitz era, financial risk was considered as a correcting factor of expected return and risk-adjusted returns were defined on an ad-hoc basis. Later, Markowitz proposed to measure the risk associated to the return of each investment by means of the deviation from the mean of the return distribution, the variance, and covariance between all pairs of investments, see [43].

More recently the Value at Risk (VaR) measure has been increasingly used to measure and manage the market risk. The importance of VaR cannot be overestimated. Reports and surveys from 1995 state that the measure was widely used by both financial and non-financial corporations in the U.S., see [19]. The VaR measure has such wide acceptance that regulators in both U.S and the European Union allow VaR models to be used provided they comply with restrictions, see [19]. Despite the fact that the VaR has proven to be a popular measure of risk it has its detractors who point out its lack of subadditivity and convexity as serious mathematical deficiencies that lead to undesirable properties of the models based on the measure. See [4, 20, 21] to get a in-depth view of the deficiencies of the VaR measure.

At the end of the 90's Artzner, Delbaen, Eber, and Heath studied the "capital requirements to regulate the risk assumed by market participants, traders, and insurance underwriters, as well as to allocate existing capital", see [4], and identified a handful of desirable properties that a risk measure for such applications should have. Needless to say that subadditivity and convexity are among these desirable properties. From this list of properties P. Artzner et al. developed the concept of a coherent measure of risk. In the last decade the theory of coherent risk measures established itself as an alternative to expected utility models of risk averse preferences in stochastic optimization. The

theory was further developed in numerous publications (see, e.g., [12, 13, 44, 34, 38] and the references therein). This development has established the coherent risk measures as an unified framework providing the grounds for a solid theoretical foundation on the notion of risk aversion on which optimization models can be based on.

The notion of risk-averseness conveyed by the theory of risk measures is general enough to model every risk-averse decision maker. The idea is to allow the decision maker to pick a risk measure that suits his priorities the most and optimize his actions by minimizing the risk assessed by the measure. One of the most attractive aspects of the theory is its potential for applicability. By developing models based on the theory of coherent risk measures one is in fact considering not one but a large family of models each one based on a different coherent risk measure which could potentially be used in different applications. The choice of which coherent risk measure will be used for a particular application should be based on additional considerations on the requirements of the application in question.

Recently, increased attention is paid to dynamic measures of risk, which allow for risk-averse evaluation of streams of future costs or rewards (see, e.g., [2, 10, 14, 25, 27, 34, 37, 40]). When used in stochastic optimization models, dynamic risk measures lead to a new class of problems, which are significantly more complex than their risk-neutral counterparts (see [38, 37, 39, 41]). The progression from risk-neutral models to risk-averse dynamic models follows a natural path. The coherent risk measures allow the evaluation and comparison of risk of uncertain outcomes and adding these measures to static risk-neutral stochastic programs gives rise to risk-averse optimization problems. From the coherent risk measures we develop the conditional risk-measures which in turn evolve into dynamic risk-measures. These measures paired with dynamic stochastic optimization models give rise to risk-averse dynamic models. Although natural, this progression is by no means trivial, requiring a full decade of research and development to reach to the state where risk-averse models are postulated and applied to relevant applications, see [23, 22, 11, 41, 3, 1]. Still there are major hurdles to overcome in the quest of developing a general theory of risk-averse optimization.

Decomposition, an established and efficient approach to risk-neutral multistage

stochastic optimization problems, see [8, 16, 26, 33] and the references therein, cannot be directly applied to risk-averse models. With dynamic risk measures, the main feature facilitating decomposition, the integral form of the objective function, is absent. Our main objective is to overcome this difficulty by exploiting specific structure of dynamic risk measures, and to develop new decomposition methods that extend the ideas of earlier approaches to risk-neutral problems. This research was initiated by N. Miller and A. Ruszczyński in [22], where they developed risk-averse counterparts of the primal (Benders-type) decomposition methods.

In this work we develop generalizations of scenario decomposition methods, in the spirit of [24]. The key to success is the use of dual properties of dynamic measures of risk to construct a family of risk-neutral approximations of the problem. In Chapter 2 we introduce the coherent and conditional measures of risk and its main properties; see the appendix for a brief overview of the necessary mathematical background. In Chapter 3 we introduce a general two-stage risk-averse stochastic optimization problem and in Chapter 4 we show a more particular version with mean upper semideviation. Chapter 5 develops the dual cutting plane method, which is our base decomposition method. In the development of efficient master algorithms we modify the bundle method, to better exploit the specifics of the problem at hand. The resulting algorithm, which we call the partial bundle method, is discussed in Chapter 6.

Chapter 7 develops the general multistage risk-averse stochastic optimization problem and methods. Since this is the main chapter of our work, it warrants an overview of its contents. In sections 7.1 and 7.2 we formally define a multistage risk-averse stochastic optimization problem and we discuss its properties. Section 7.3 discusses nonanticipativity constraints. In section 7.4 we advance the duality theory of dynamic measures of risk, by identifying the properties that are essential for our decomposition approach. In section 7.5 we present the main idea of our new decomposition methods. In section 7.6 we analyze properties of the master (coordination) problem of the method.

Finally, Chapter 8 is devoted to the application of two versions of our methods, with several coordination algorithms, to an inventory planning and assembly problem. In this chapter we also show numerical results from our implementations and compare the results obtained.

Chapter 2

Coherent Measures of Risk

In this chapter we introduce and develop the basic theory of coherent measures of risk. We follow closely the development from [34, 38, 39, 41].

2.1 Coherent Measures of Risk

Let $(\Omega, \mathcal{F}, P)$ be a probability space with a sigma algebra $\mathcal{F}$ and probability measure P. In this work we will restrict ourselves to the spaces of uncertain outcomes $\mathcal{Z} := \mathcal{L}_p(\Omega, \mathcal{F}, P)$, where $p \in [1, +\infty)$. Each element $Z := Z(\omega)$ of $\mathcal{Z}$ is viewed as an uncertain outcome on $(\Omega, \mathcal{F})$ and is by definition a random variable with finite p-th order moment with respect to the reference probability measure P.

A *risk measure* is a proper class function $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$. By this we mean that ρ is constant on the classes of functions which differ only on sets of P-measure zero, i.e. $\rho(Z) = \rho(Z')$ if $P\{\omega : Z(\omega) \neq Z'(\omega)\} = 0$. The function ρ is proper in the sense that $\rho(Z) > -\infty$ for all $Z \in \mathcal{Z}$ and its domain

$$\mathrm{dom}(\rho) := \{Z \in \mathcal{Z} : \rho(Z) < +\infty\}$$

is nonempty. We say that the probability space $(\Omega, \mathcal{F}, P)$ is the probability space *associated* with the risk measure ρ.

For $Z, Z' \in \mathcal{Z}$ we denote by $Z \preceq Z'$ the pointwise partial order, meaning $Z(\omega) \leq Z'(\omega)$ for a.e. $\omega \in \Omega$. We also assume that the smaller the realizations of Z, the better; for example, Z may represent a random cost.

A *coherent risk measure* is a risk measure $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ satisfying the following axioms:

(A1) *Convexity*: $\rho\left(tZ + (1-t)\,Z'\right) \leq t\rho(Z) + (1-t)\rho(Z')$, for all $Z, Z' \in \mathcal{Z}$ and all

$t \in [0, 1]$;

(A2) *Monotonicity*: If $Z, Z' \in \mathcal{Z}$ and $Z \preceq Z'$, then $\rho(Z) \leq \rho(Z')$;

(A3) *Translation equivariance*: If $a \in \mathbb{R}$ and $Z \in \mathcal{Z}$, then $\rho(Z + a) = \rho(Z) + a$;

(A4) *Positive homogeneity*: If $t > 0$ and $Z \in \mathcal{Z}$, then $\rho(tZ) = t\rho(Z)$.

These axioms were conceived by Artzner et al. [4] in the context of risky capital markets and it is under these circumstances that we can better understand the above definition. Suppose that our *perception of risk* is an amount of a *security capital* that we would set aside from our investments to protect us from the fluctuations of the market. Examining axioms (A1) - (A4) under this context makes the definition of coherent risk measures clear.

Axioms (A1) and (A4) imply subadditivity, which can be stated as "a merger does not create extra risk" [4], a natural requirement. Axiom (A2) states that investing more capital under any possible market scenario increases the investor's perception of risk. Similarly, Axiom (A3) says that investing an extra a units under any possible market scenario increases the investor's perception of risk by that same amount a. Finally, axiom (A4) states that a change of currency in the capital investments does not change the investor's perception of risk under the new currency.

All the measures of risk considered in this work will have a finite associated probability space $(\Omega, \mathcal{F}, P)$. That is, Ω has a finite number of elements and $\mathcal{F} = 2^{\Omega}$, i.e. the power set sigma algebra. In this case we will let $\mathcal{Z} := \mathcal{L}_1(\Omega, \mathcal{F}, P)$, which is just the set of all functions $f : \Omega \to \mathbb{R}$. For this reason we will identify the set $\mathcal{Z}$ with $\mathbb{R}^N$, where N is the number of elements of Ω.

Leveraging on the conjugate duality theory reviewed in the appendix and the definition of coherent risk measures we arrive to the *representation theorem of coherent risk measures*:

Theorem 1. *Suppose that $(\Omega, \mathcal{F}, P)$ is a finite probability space as described above with $|\Omega| = N$, $P = (p_1, \ldots, p_N)$ and $\mathcal{Z} := \mathcal{L}_1(\Omega, \mathcal{F}, P) = \mathbb{R}^N$. If $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ is a lower semicontinuous coherent risk measure then ρ is subdifferentiable and for every random*

variable $Z = (z_1, \ldots, z_N) \in \mathcal{Z}$*:*

$$\rho(Z) = \max_{\mu \in \partial\rho(0)} \sum_{i=1}^{N} \mu_i p_i z_i, \tag{2.1}$$

where

$$\partial\rho(0) \subseteq \left\{ \zeta \in \mathbb{R}^N \,\middle|\, \sum_{i=1}^{N} \zeta_i p_i = 1, \zeta \geq 0 \right\}.$$

Proof. First we should point out that the spaces in question are the spaces $\mathcal{Z} = \mathcal{L}_1(\omega, \mathcal{F}, P) = \mathbb{R}^N$ and $Z^* = \mathcal{L}_\infty(\Omega, \mathcal{F}, P) = \mathbb{R}^N$. For $Z \in \mathcal{Z}$ and $\zeta \in \mathcal{Z}^*$ their scalar product is defined as

$$\langle \zeta, Z \rangle := \sum_{i=1} \zeta_i p_i z_i. \tag{2.2}$$

Clearly, all the results presented in the appendix hold even if we replace the conventional scalar product by the product defined in (2.2).

Since ρ is convex, lower semicontinuous, and proper, Theorem 19 implies that $\rho = \rho^{**}$. More specifically, for all $Z \in \mathcal{Z}$,

$$\rho(Z) = \sup_{\mu \in \mathbb{R}^N} \left\{ \langle \mu, Z \rangle - \rho^*(\mu) \right\}. \tag{2.3}$$

We first show that $\mathrm{dom}(\rho^*) \subset \mathbb{R}^N_+$. Suppose there is $\overline{\mu} \in \mathrm{dom}(\rho^*)$ which is not an element of the cone $\mathbb{R}^N_+$. Then it can be strictly separated from $\mathbb{R}^N_+$. There exists $W \in \mathbb{R}^N$ and $\epsilon > 0$ such that:

$$\langle W, \mu \rangle \leq \langle W, \overline{\mu} \rangle - \epsilon \quad \text{for all} \quad \mu \in \mathbb{R}^N_+.$$

It follows that W is in the polar of $\mathbb{R}^N_+$ and so, $W \leq 0$. Setting $\mu = 0$ we get that $\langle W, \overline{\mu} \rangle \geq \epsilon > 0$. Consider an arbitrary point $Z \in \mathbb{R}^N$ and points $Y = Z + tW$, for $t \geq 0$. As $Y \leq Z$, the monotonicity axiom $(A2)$ implies that $\rho(Y) \leq \rho(Z)$. From the

definition (8.18) of the conjugate dual we deduce that

$$\begin{aligned}\rho^*(\overline{\mu}) &= \sup_{Z\in\mathbb{R}^N}\{\langle\overline{\mu},Z\rangle-\rho(Z)\} = \sup_{Z\in\mathbb{R}^N}\{\langle\overline{\mu},Z+tW\rangle-\rho(Z+tW)\}\\ &\geq \sup_{Z\in\mathbb{R}^N}\{\langle\overline{\mu},Z\rangle-\rho(Z)\}+t\langle\overline{\mu},W\rangle \geq \rho^*(\overline{\mu})+t\epsilon.\end{aligned}$$

Therefore, we conclude that $\rho^*(\overline{\mu})=\infty$, a contradiction. Consequently, $\mathrm{dom}(\rho^*)\subset\mathbb{R}^N_+$.

The translation equivariance axiom $(A3)$ implies that for all $a\in\mathbb{R}$:

$$\begin{aligned}\rho^*(\zeta) &= \sup_{Z\in\mathbb{R}^N}\{\langle\zeta,Z\rangle-\rho(Z)\} = \sup_{Z\in\mathbb{R}^N}\{\langle\zeta,Z+a\mathbb{1}\rangle-\rho(Z+a\mathbb{1})\}\\ &= \sup_{Z\in\mathbb{R}^N}\{\langle\zeta,Z\rangle-\rho(Z)\}+a\left(\langle\zeta,\mathbb{1}\rangle-1\right)=\rho^*(\zeta)+a\left(\langle\zeta,\mathbb{1}\rangle-1\right),\end{aligned}$$

where $\mathbb{1}$ is the vector with all entries equal to one. This means that $\rho^*(\zeta)<\infty$ only if $\langle\zeta,\mathbb{1}\rangle=1$. In other words, the domain of the conjugate function is included in the set of probability vectors

$$\mathbb{P}:=\left\{\zeta\in\mathbb{R}^N\,\middle|\,\sum_{i=1}^N\zeta_i p_i=1,\,\zeta\geq 0\right\}.$$

The positive homogeneity axiom $(A4)$ implies that for every $t>0$ and every $\mu\in\mathrm{dom}(\rho^*)$ we have

$$\begin{aligned}\rho^*(\mu) &= \sup_{Z\in\mathbb{R}^N}\{\langle\mu,Z\rangle-\rho(Z)\} = \sup_{Z\in\mathbb{R}^N}\{\langle\mu,tZ\rangle-\rho(tZ)\}\\ &= \sup_{Z\in\mathbb{R}^N}\{t\langle\mu,Z\rangle-t\rho(Z)\}=t\rho^*(\mu).\end{aligned}$$

This implies that $\rho^*(\mu)=0$. Consequently, ρ^* is the indicator function of certain set $\mathcal{A}\subset\mathbb{P}$. Formula (2.3) simplifies to

$$\rho(Z)=\sup_{\mu\in\mathcal{A}}\langle\mu,Z\rangle, \tag{2.4}$$

where $\langle\cdot,\cdot\rangle$ is as defined in (2.2). It is not difficult to see that the set $\mathcal{A}$ is closed and convex. Therefore Lemma 21 implies that $\mathcal{A}=\partial\rho(0)$. By combining this with (2.4) we

obtain representation (2.1), as we wanted. □

We see from theorem 1 that if ρ is a coherent risk measure then $\partial\rho(0)$ is a set of probability density functions. Consequently, for any $\mu \in \partial\rho(0)$ we can view $\sum_{i=1}^{N} \mu_i p_i z_i$ as the expectation $\mathbb{E}_\mu[Z]$ taken with respect to the probability measure μdP, defined by the density μ. For this reason representation (2.1) can be written as

$$\rho(Z) = \max_{\mu \in \partial\rho(0)} \mathbb{E}_\mu[Z].$$

Thus, evaluating a coherent risk measure amounts to select the "worst" expected value out of a set of given possible density functions, namely $\partial\rho(0)$. This is one of the main results in the theory of coherent risk measures and it will allow us to bridge the gap between the classical scenario decomposition methods and the coherent risk measures.

The following is the more general version of Theorem (1) for a general probability space Ω, see [41, ch. 6] for details and proof.

Theorem 2. *Let $(\Omega, \mathcal{F}, P)$ be a sample space with sigma algebra $\mathcal{F}$ and probability measure P. Let $p \in [1, +\infty)$, $q \in (1, +\infty]$ be such that $1/p + 1/q = 1$ and let $\mathcal{Z} := \mathcal{L}_p(\Omega, \mathcal{F}, P), \mathcal{Z}^* := \mathcal{L}_q(\Omega, \mathcal{F}, P)$ be a conjugate pair of spaces. Let $\rho : \mathcal{Z} \to \overline{\mathbb{R}}$ be a lower semicontinuous coherent risk measure and $\rho^* : \mathcal{Z}^* \to \overline{\mathbb{R}}$ be its conjugate dual. Then ρ is subdifferentiable and for every random variable $Z \in \mathcal{Z}$:*

$$\rho(Z) = \max_{\mu \in \partial\rho(0)} \int_\Omega \mu(\omega) Z(\omega) dP(\omega), \quad \forall Z \in \mathcal{Z} \tag{2.5}$$

where

$$\partial\rho(0) \subseteq \left\{ \zeta \in \mathbb{R}^N \,\middle|\, \int_\Omega \zeta(\omega) dP(\omega) = 1, \zeta \geq 0 \right\}.$$

2.2 Examples

In this section we will introduce several important examples of coherent risk measures which are relevant to this work. We closely follow the examples provided in [41, Ch. 6].

2.2.1 Conditional Value-at-Risk

Value-at-Risk

Let $H_Z(z) = \Pr(Z \leq z)$ be the cdf of the random variable Z and $\alpha \in (0,1)$. Recall that the left-side α-quantile of H_Z is defined as

$$H_Z^{-1}(\alpha) := \inf\{t : H_Z(t) \geq \alpha\},$$

and the right-side α-quantile as $\sup\{t : H_Z(t) \leq \alpha\}$. If Z represents losses, the (left-side) quantile $H_Z^{-1}(1-\alpha)$ is also called *Value-at-Risk* and denoted $\mathrm{VaR}_\alpha(Z)$. Its meaning is the following: *losses larger than* $\mathrm{VaR}_\alpha(Z)$ *occur with probability not exceeding* α. Note that

$$\mathrm{VaR}_\alpha(Z+\tau) = \mathrm{VaR}_\alpha(Z) + \tau, \quad \forall \tau \in \mathbb{R}.$$

The weighted mean deviation from a quantile is defined as follows:

$$q_\alpha[Z] := \mathbb{E}\left[\max\left\{(1-\alpha)(H_Z^{-1}(\alpha) - Z),\, \alpha(Z - H_Z^{-1}(\alpha))\right\}\right]. \tag{2.6}$$

The functional $q_\alpha[Z]$ is well-defined and finite-valued for all $Z \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$. It can be easily shown that

$$q_\alpha[Z] := \min_{t \in \mathbb{R}} \{\varphi(t) := \mathbb{E}\left[\max\{(1-\alpha)(t-Z),\, \alpha(Z-t)\}\right]\}. \tag{2.7}$$

Indeed, the right and left side derivatives of the function $\varphi(\cdot)$ are

$$\varphi'_+(t) = (1-\alpha)\Pr[Z \leq t] - \alpha\Pr[Z > t],$$
$$\varphi'_-(t) = (1-\alpha)\Pr[Z < t] - \alpha\Pr[Z \geq t].$$

At the optimal t the right derivative is nonnegative and the left derivative nonpositive, and thus

$$\Pr[Z < t] \leq \alpha \leq \Pr[Z \leq t].$$

This means that every α-quantile is a minimizer in (2.7).

The risk functional $q_\alpha[\cdot]$ can be used in *mean-risk models*, both in the case of minimization

$$\min_{x \in X} \mathbb{E}[Z_x] + cq_{1-\alpha}[Z_x], \tag{2.8}$$

and in the case of maximization

$$\max_{x \in X} \mathbb{E}[Z_x] - cq_\alpha[Z_x]. \tag{2.9}$$

We use $1-\alpha$ in the minimization problem and α in the maximization problem, because in practical applications we are interested in these quantities for small α.

Conditional Value-at-Risk

The mean-deviation from quantile model is closely related to the concept of Conditional Value at Risk. Suppose that Z represents losses and we want to satisfy the chance constraint:

$$\mathrm{VaR}_\alpha[Z_x] \leq 0. \tag{2.10}$$

We have that[1] $\Pr(Z_x > 0) = \mathbb{E}\left[\mathbf{1}_{(0,\infty)}(Z_x)\right]$, and hence constraint (2.10) can also be written as the expected value constraint:

$$\mathbb{E}\left[\mathbf{1}_{(0,\infty)}(Z_x)\right] \leq \alpha. \tag{2.11}$$

The source of difficulties with chance constraints is that the step function $\mathbf{1}_{(0,\infty)}(\cdot)$ is not convex and, even worse, it is discontinuous at zero. As a result, chance constraints are often nonconvex, even if the function $x \to Z_x$ is convex almost surely. One possibility is to approach such problems by constructing a convex approximation of the expected value on the left of (2.11).

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a nonnegative valued, nondecreasing, convex function such that $\psi(z) \geq \mathbf{1}_{(0,\infty)}(z)$ for all $z \in \mathbb{R}$. By noting that $\mathbf{1}_{(0,\infty)}(tz) = \mathbf{1}_{(0,\infty)}(z)$ for any $t > 0$ and

[1]Recall that $\mathbf{1}_{(0,\infty)}(z) = 0$ if $z \leq 0$, and $\mathbf{1}_{(0,\infty)}(z) = 1$ if $z > 0$.

$z \in \mathbb{R}$, we have that $\psi(tz) \geq \mathbf{1}_{(0,\infty)}(z)$ and hence the following inequality holds

$$\inf_{t>0} \mathbb{E}\left[\psi(tZ_x)\right] \geq \mathbb{E}\left[\mathbf{1}_{(0,\infty)}(Z_x)\right].$$

Consequently, the constraint

$$\inf_{t>0} \mathbb{E}\left[\psi(tZ_x)\right] \leq \alpha \tag{2.12}$$

is a *conservative* approximation of the chance constraint (2.10) in the sense that the feasible set defined by (2.12) is contained in the feasible set defined by (2.10).

Of course, the smaller the function $\psi(\cdot)$ is the better this approximation will be. From this point of view the best choice of $\psi(\cdot)$ is to take piecewise linear function $\psi(z) := [1 + \gamma z]_+$ for some $\gamma > 0$. Since constraint (2.12) is invariant with respect to scale change of $\psi(\gamma z)$ to $\psi(z)$, we have that $\psi(z) := [1 + z]_+$ gives the best choice of such a function. For this choice of function $\psi(\cdot)$, we have that constraint (2.12) is equivalent to

$$\inf_{t>0} \left\{t\mathbb{E}[t^{-1} + Z]_+ - \alpha\right\} \leq 0,$$

or equivalently

$$\inf_{t>0} \left\{\alpha^{-1}\mathbb{E}[Z + t^{-1}]_+ - t^{-1}\right\} \leq 0.$$

Now replacing t with $-t^{-1}$ we get the form:

$$\inf_{t\geq 0} \left\{t + \alpha^{-1}\mathbb{E}[Z - t]_+\right\} \leq 0. \tag{2.13}$$

The quantity

$$\mathrm{CVaR}_\alpha(Z) := \inf_{t\in\mathbb{R}} \left\{t + \alpha^{-1}\mathbb{E}[Z - t]_+\right\} \tag{2.14}$$

is called the *Conditional Value-at-Risk* of Z (at level α). Note that $\mathrm{CVaR}_\alpha(Z)$ is well defined and finite valued for every $Z \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$.

The function $\varphi(t) := t + \alpha^{-1}\mathbb{E}[Z - t]_+$ is convex. Its derivative at t is equal to $1 + \alpha^{-1}[1 - H_Z(t)]$ provided that the cdf $H_Z(\cdot)$ is continuous at t. If $H_Z(\cdot)$ is discontinuous at t, then the respective right and left side derivatives of $\varphi(\cdot)$ are given by the same formula with $H_Z(t)$ understood as the corresponding right and left side limits. Therefore

the minimum of $\varphi(t)$, over $t \in \mathbb{R}$, is attained on the interval $[t^*, t^{**}]$, where $t^* := \inf\{z : H_Z(z) \geq 1-\alpha\}$ and $t^{**} := \sup\{z : H_Z(z) \leq 1-\alpha\}$ are the respective left and right side quantiles. Recall that the left-side quantile $t^* = \mathrm{VaR}_\alpha(Z)$.

Since the minimum of $\varphi(t)$ is attained at $t^* = \mathrm{VaR}_\alpha(Z)$, we have that $\mathrm{CVaR}_\alpha(Z)$ is bigger than $\mathrm{VaR}_\alpha(Z)$ by the amount of $\alpha^{-1}\mathbb{E}[Z-t^*]_+$ (of course, $\mathbb{E}[Z-t^*]_+$ is positive, unless $P[Z=t^*] \geq \alpha$, in which case it is zero). Therefore

$$\inf_{t\in\mathbb{R}} \left\{t + \alpha^{-1}\mathbb{E}[Z-t]_+\right\} \leq 0 \text{ implies that } t^* \leq 0,$$

and hence constraint (2.13) is equivalent to $\mathrm{CVaR}_\alpha(Z) \leq 0$. It is easy to see that for any $a \in \mathbb{R}$,

$$\mathrm{CVaR}_\alpha(Z+a) = \mathrm{CVaR}_\alpha(Z) + a.$$

Therefore, the constraint

$$\mathrm{CVaR}_\alpha[Z_x] \leq 0 \tag{2.15}$$

is equivalent to the constraint (2.13) and gives a conservative approximation of the chance constraint (2.10).

The function $\rho(Z) := \mathrm{CVaR}_\alpha(Z)$, defined on a space of random variables, is convex, i.e., if Z and Z' are two random variables and $t \in [0,1]$, then

$$\rho(tZ + (1-t)Z') \leq t\rho(Z) + (1-t)\rho(Z').$$

This follows from that the function $t + \alpha^{-1}\mathbb{E}[Z-t]_+$ is convex jointly in t and Z. Also $\rho(\cdot)$ is monotone, i.e., if Z and Z' are two random variables such that with probability one $Z \geq Z'$, then $\rho(Z) \geq \rho(Z')$. It follows that if $G(\cdot,\xi)$ is convex for a.e. $\xi \in \Xi$, then the function $\rho[G(\cdot,\xi)]$ is also convex. Indeed, by convexity of $G(\cdot,\xi)$ and monotonicity of $\rho(\cdot)$, we have for any $t \in [0,1]$ that

$$\rho[G(tZ + (1-t)Z'),\xi)] \leq \rho[tG(Z,\xi) + (1-t)G(Z',\xi)],$$

and hence by convexity of $\rho(\cdot)$ that

$$\rho[G(tZ + (1-t)Z'), \xi)] \leq t\rho[G(Z, \xi)] + (1-t)\rho[G(Z', \xi)],$$

Consequently, (2.15) is a *convex conservative* approximation of the chance constraint (2.10). Moreover, from the considered point of view, (2.15) is the best convex conservative approximation of the chance constraint (2.10).

We can now relate the concept of Conditional Value at Risk to mean deviations from quantiles. See equation (2.6) for the definition of $q_\alpha[Z]$ and [41, page 259] for the proof of the next theorem.

Theorem 3. *Let $Z \in \mathcal{L}_1(\Omega, \mathcal{F}, P)$ and $H(z)$ be its cdf. Then the following identities hold true*

$$CVaR_\alpha(Z) = \frac{1}{\alpha}\int_{1-\alpha}^{1} VaR_{1-\beta}(Z)d\beta = \mathbb{E}[Z] + \frac{1}{\alpha}q_{1-\alpha}[Z]. \tag{2.16}$$

Moreover, if $H(z)$ is continuous at $z = VaR_\alpha(Z)$, then

$$CVaR_\alpha(Z) = \frac{1}{\alpha}\int_{VaR_\alpha(Z)}^{+\infty} zdH(z) = \mathbb{E}\left[Z | Z \geq VaR_\alpha(Z)\right]. \tag{2.17}$$

The last equation in (2.17) explains the origin of the term "Conditional Value at Risk." Also motivated by the first equation in (2.16), in some publications CVaR is called the *Average Value-at-Risk*. It follows from the development in this section that the function $\rho(Z) := \mathrm{CVaR}_\alpha(Z)$ is a coherent measure of risk for every $\alpha \in (0, 1)$.

2.2.2 Mean-Upper-Semideviation of Order p

Let $Z := \mathcal{L}_p(\Omega, \mathcal{F}, P)$ and for $c \geq 0$ consider

$$\rho(Z) := \mathbb{E}[Z] + c\left(\mathbb{E}\left[[Z - \mathbb{E}[Z]]_+^p\right]\right),$$

where $[a]_+^p := (\max\{0, a\})^p$. For any $c \geq 0$ this function satisfies conditions $(A1)$, $(A3)$ and $(A4)$, and it can be shown (see [41, p. 277]) that the representation from Theorem

2 holds with the set $\partial\rho(0)$ given by

$$\partial\rho(0) = \left\{\zeta' \in \mathcal{Z}^* \,\middle|\, \zeta' = 1 + \zeta - \mathbb{E}[\zeta],\ \|\zeta\|_q \leq c,\ \zeta \succeq 0\right\}, \tag{2.18}$$

where $\frac{1}{p}+\frac{1}{q} = 1$ and $Z^* := \mathcal{L}_q(\Omega, \mathcal{F}, P)$ is the dual space associated to Z. See [29, 30, 41] for more details. Since $|\mathbb{E}[\zeta]| \leq \mathbb{E}|\zeta| \leq \|\zeta\|_q$ for any $\zeta \in \mathcal{L}_q(\Omega, \mathcal{F}, P)$, we have that every element of $\partial\rho(0)$ is nonnegative and has its expected value equal to 1. It can be shown that the monotonicity condition $(A2)$ holds, if and only if $c \in [0, 1]$. That is, ρ is a coherent risk measure if $c \in [0, 1]$.

Since ρ is convex continuous, it is subdifferentiable. It can be shown that its subdifferential, $\partial\rho(Z)$, is formed by vectors $\zeta' = 1 + \zeta - \mathbb{E}[\zeta]$ such that

$$\zeta \in \arg\max\left\{\langle\zeta, Y\rangle \,|\, \|\zeta\|_q \leq c,\ \zeta \succeq 0\right\}, \tag{2.19}$$

where $Y := Z - \mathbb{E}[Z]$. Suppose that $p \in (1, +\infty)$. Then the set of maximizers on the right hand side of (2.19) is not changed if Y is replaced by Y_+, where $Y_+(\cdot) := [Y(\cdot)]_+$. Consequently, if $Z(\omega)$ is not constant for a.e. $\omega \in \Omega$, and hence $Y_+ \neq 0$, then $\partial\rho(Z)$ is a singleton and

$$\nabla\rho(Z) = 1 + c\,\zeta^*_{Y_+} - c\,\mathbb{E}[\zeta^*_{Y_+}],$$

where $\zeta^*_{Y_+}$ is the contact point of Y_+ (note that the contact point of Y_+ is nonnegative since $Y_+ \succeq 0$).

Suppose now that $p = 1$ and hence $q = +\infty$. Then the set on the right hand side of (2.19) is formed by $\zeta(\cdot)$ such that: $\zeta(\omega) = c$ if $Y(\omega) > 0$, $\zeta(\omega) = 0$, if $Y(\omega) < 0$, and $\zeta(\omega) \in [0, c]$ if $Y(\omega) = 0$. It follows that $\partial\rho(Z)$ is a singleton iff $Z(\omega) \neq \mathbb{E}[Z]$ for a.e. $\omega \in \Omega$, in which case

$$\nabla\rho(Z) = \left\{\zeta :\ \begin{array}{ll} \zeta(\omega) = 1 + c\,(1 - \Pr(Z > \mathbb{E}[Z])), & \text{if } \ Z(\omega) > \mathbb{E}[Z], \\ \zeta(\omega) = 1 - c\Pr(Z > \mathbb{E}[Z]), & \text{if } \ Z(\omega) < \mathbb{E}[Z]. \end{array}\right. \tag{2.20}$$

2.3 Coherent Conditional Risk Measures

This section is an overview of the theory of conditional risk measures developed by Ruszczyński and Shapiro in [37]. The intention is to develop the risk-measures support for models suitable for risk-averse sequential decision making. For an in-depth view at these topics see [37, 41].

The main issue here is our knowledge at the time when risk is evaluated. In the classical setting of multistage stochastic optimization, the main tool used to formulate the corresponding dynamic programming equations is the concept of conditional expectation. Given two sigma algebras $\mathcal{F}_1 \subset \mathcal{F}_2$ of subsets of Ω, with $\mathcal{F}_1$ representing our knowledge when the expectation is evaluated, and $\mathcal{F}_2$ representing all events under consideration, the conditional expectation can be defined as a mapping from a space of $\mathcal{F}_2$-measurable functions into a space of $\mathcal{F}_1$-measurable functions. Of course, the conditional expectation mapping is linear. The basic idea of this approach is to extend the concept of conditional expectation to an appropriate class of convex mappings.

In order to construct dynamic models of risk we need to extend the concept of risk functions. Similarly to the representation of coherent risk measures we use the framework of $\mathcal{L}_p$ spaces, $p \in [1, +\infty)$. Let $(\Omega, \mathcal{F}, P)$ be a probability space with sigma algebra $\mathcal{F}$ and probability measure P. Consider a filtration $\mathcal{F}_1 \subset \mathcal{F}_2 = \mathcal{F}$. As before, we consider spaces $\mathcal{Z}_t := \mathcal{L}_p = (\Omega, \mathcal{F}_t, P)$ of $\mathcal{F}_t$-measurable random outcomes, $t = 1, 2$.

A *coherent conditional risk measure* is a function $\rho : \mathcal{Z}_2 \to \mathcal{Z}_1$ satisfying the following axioms:

($A1'$) *Convexity*: $\rho\left(\alpha Z + (1-\alpha) Z'\right) \preceq \alpha\rho(Z) + (1-\alpha)\rho(Z')$, for all $Z, Z' \in \mathcal{Z}_2$ and all $\alpha \in [0, 1]$;

($A2'$) *Monotonicity*: If $Z, Z' \in \mathcal{Z}_2$ and $Z \preceq Z'$, then $\rho(Z) \preceq \rho(Z')$;

($A3'$) *Predictable Translation Equivariance*: If $V \in \mathcal{Z}_1$ and $Z \in \mathcal{Z}_2$, then $\rho(V + Z) = V + \rho(Z)$;

($A4'$) *Positive Homogeneity*: If $\gamma \geq 0$ and $Z \in \mathcal{Z}_2$, then $\rho(\gamma Z) = \gamma\rho(Z)$.

As before, we assume that the smaller the realizations of Z, the better and that Ω and all sigma-algebras are finite.

An example of coherent conditional risk measure is the *conditional mean–upper semideviation* model defined by

$$\rho_t(Z) = \mathbb{E}[Z|\mathcal{F}_t] + \kappa_t \, \mathbb{E}\left[\left(Z - \mathbb{E}[Z|\mathcal{F}_t]\right)_+ \middle| \mathcal{F}_t\right], \tag{2.21}$$

with an $\mathcal{F}_t$-measurable $\kappa_t \in [0,1]$. See [41, p. 277] for the details showing that the mean upper semideviation is a coherent conditional risk measure, and for other examples of conditional risk measures.

Consider a conditional risk measure $\rho : \mathcal{Z}_2 \to \mathcal{Z}_1$. With a set $A \in \mathcal{F}_1$, such that $P(A) \neq 0$, we associate the function

$$\rho_A(Z) := \mathbb{E}\left[\rho(Z) \mid A\right], \quad Z \in \mathcal{Z}_2, \tag{2.22}$$

where $\mathbb{E}\left[Y \mid A\right] := \frac{1}{P(A)} \int_A Y dP$ denotes the conditional expectation of random variable $Y \in \mathcal{Z}_1$ given event $A \in \mathcal{F}_1$. Conditions $(A1')$-$(A4')$ imply that the corresponding conditions (A1)-(A4) hold for ρ_A, and hence ρ_A is a coherent risk measure defined on the space $\mathcal{Z}_2 = \mathcal{L}_p(\Omega, \mathcal{F}_2, P)$. Moreover, for any $B \in \mathcal{F}_1$ we have by $(A3')$ and $(A4')$ that

$$\rho_A(Z + \alpha \mathbb{1}_B) := \mathbb{E}\left[\rho(Z) + \alpha \mathbb{1}_B \mid A\right] = \rho_A(Z) + \alpha P(B|A), \quad \forall \alpha \in \mathbb{R}. \tag{2.23}$$

Since ρ_A is a coherent risk measure, by Theorem (2) it can be represented in the form

$$\rho_A(Z) = \sup_{\zeta \in \mathcal{B}(A)} \int_\Omega \zeta(\omega) Z(\omega) dP(\omega), \tag{2.24}$$

for some set $\mathcal{B}(A)$ of probability density functions. Let us make the following observation:

Lemma 4. *Each density $\zeta \in \mathcal{B}(A)$ is supported on the set A.*

Proof. For any $B \in \mathcal{F}_1$ such that $P(B \cap A) = 0$ and any $\alpha \in \mathbb{R}$, we have by (2.23)

that $\rho_A(Z + \alpha \mathbb{1}_B) = \rho_A(Z)$. On the other hand, if there exists $\zeta \in \mathcal{B}(A)$ such that $\int_B \zeta dP > 0$ then it follows from (2.24) that $\rho_A(Z+\alpha \mathbb{1}_B)$ tends to $+\infty$ as $\alpha \to +\infty$. $\square$

Just as we obtained the representation theorem of coherent risk measures, we show that a conditional coherent risk measure can be represented as a maximum of a family of conditional expectations. We restrict ourselves to the situation where the subalgebra $\mathcal{F}_1$ has a countable number of elementary events. That is, there is a countable partition $\{A_i\}_{i\in\mathbb{N}}$ of the sample space Ω which generates $\mathcal{F}_1$, i.e, $\cup_{i\in\mathbb{N}} A_i = \Omega$, the sets A_i, $i \in \mathbb{N}$ are disjoint and form the family of elementary events of sigma algebra $\mathcal{F}_1$. Since $\mathcal{F}_1$ is a subalgebra of $\mathcal{F}_2$, we have that $A_i \in \mathcal{F}_2, i \in \mathbb{N}$. We also have that a function $Z : \Omega \to \mathbb{R}$ is $\mathcal{F}_1$-measurable iff it is constant on every set $A_i, i \in \mathbb{N}$.

Consider a conditional coherent risk measure $\rho : \mathcal{Z}_2 \to \mathcal{Z}_1$. Let

$$\mathfrak{N} := \{i \in \mathbb{N} \mid P(A_i) \neq 0\}$$

and $\rho_{A_i}, i \in \mathfrak{N}$, be the corresponding coherent risk measure defined in (2.22). By (2.24) with every $\rho_{A_i}, i \in \mathfrak{N}$, is associated set $\mathcal{B}(A_i)$ of probability density functions, supported on the set A_i such that

$$\rho_{A_i}(Z) = \sup_{\zeta\in\mathcal{B}(A_i)} \int_\Omega \zeta(\omega) Z(\omega) dP(\omega). \tag{2.25}$$

Now let $v = (v_i)_{i\in\mathbb{N}}$ be a probability distribution on $(\Omega, \mathcal{F}_1)$, assigning probability v_i to the event $A_i, i \in \mathbb{N}$. Assume that v is such that $v(A_i) = 0$ iff $P(A_i) = 0$ (i.e. P is absolutely continuous with respect to v on $(\Omega, \mathcal{F}_1)$); otherwise the probability measure v is arbitrary. Define the following family of probability measures on $(\Omega, \mathcal{F}_2)$:

$$\mathfrak{C} = \left\{ \hat{\mu} = \sum_{i\in\mathfrak{N}} v_i \mu_i \;\middle|\; d\mu_i = \zeta_i dP,\ \zeta_i \in \mathcal{B}(A_i),\ i \in \mathfrak{N} \right\}. \tag{2.26}$$

Note that since $\sum_{i\in\mathfrak{N}} v_i = 1$, every $\hat{\mu} \in \mathfrak{C}$ is a probability measure. For $\hat{\mu} \in \mathfrak{C}$, with

respective densities $\zeta_i \in \mathcal{B}(A_i)$ and $d\mu_i = \zeta_i dP$, and $Z \in \mathcal{Z}_2$ we have that

$$\mathbb{E}_{\mu_i}[Z \mid \mathcal{F}_1] = \sum_{i \in \mathfrak{N}} \mathbb{E}_{\mu_i}[Z \mid \mathcal{F}_1]. \tag{2.27}$$

Moreover, since ζ_i is supported on A_i,

$$\mathbb{E}_{\mu_i}[Z \mid \mathcal{F}_1](\omega) = \begin{cases} \int_{A_i} Z\zeta_i dP & \text{if } \omega \in A_i \\ 0 & \text{otherwise.} \end{cases} \tag{2.28}$$

By the max-representations (2.25) it follows that for $Z \in \mathcal{Z}_2$ and $\omega \in A_i$,

$$\sup_{\hat{\mu} \in \mathfrak{C}} \mathbb{E}_{\hat{\mu}}[Z \mid \mathcal{F}_1](\omega) = \sup_{\zeta_i \in \mathcal{B}(A_i)} \int_{A_i} Z\zeta_i dP = \rho_{A_i}(Z). \tag{2.29}$$

Also since $[\rho(Z)](\cdot)$ is $\mathcal{F}_1$-measurable, and hence is constant on every set A_i, we have that $[\rho(Z)](\omega) = \rho_{A_i}(Z)$ for every $\omega \in A_i, i \in \mathfrak{N}$. We obtain the following result.

Theorem 5. *Let $\mathcal{Z}_i := \mathcal{L}_p(\Omega, \mathcal{F}_i, P), i = 1, 2$, with $\mathcal{F}_1 \subset \mathcal{F}_2$, and let $\rho : \mathcal{Z}_2 \to \mathcal{Z}_1$ be a conditional coherent risk measure. Suppose that $\mathcal{F}_1$ has an countable number of elementary events and let $v = (v_i)_{i \in \mathbb{N}}$ be a probability distribution on $(\Omega, \mathcal{F}_1)$, assigning probability v_i to the event $A_i, i \in \mathbb{N}$. Assume that v is such that $v(A_i) = 0$ iff $P(A_i) = 0$ (i.e. P is absolutely continuous with respect to v on $(\Omega, \mathcal{F}_1)$); otherwise the probability measure v is arbitrary. Then*

$$\rho(Z) = \sup_{\hat{\mu} \in \mathfrak{C}} \mathbb{E}_{\hat{\mu}}[Z \mid \mathcal{F}_1], \quad \forall Z \in \mathcal{F}_2, \tag{2.30}$$

where $\mathfrak{C}$ is a family of probability measures on $(\Omega, \mathcal{F}_2)$, specified in (2.26), corresponding to a probability distribution v on $(\Omega, \mathcal{F}_1)$.

Chapter 3

Scenario Decomposition of Two-Stage Risk-Averse Problems

3.1 A Two-Stage Risk-Averse Problem

This chapter will be devoted to the study of a *two-stage risk-averse problem* of the form

$$\min_{x \in X} \ \rho_1 \left(c^\top x + \mathcal{Q}(x, \xi) \right), \tag{3.1}$$

where ρ_1 is a coherent risk measure, $X \subseteq \mathbb{R}^n$ is compact and polyhedral, and $\mathcal{Q}(x, \xi)$ is the optimal value of the *second stage problem*

$$\begin{aligned} \min_{y \in \mathbb{R}^m} \ & \rho_2[q^\top y] \\ \text{s.t. } & Tx + Wy = h, \ y \geq 0, \end{aligned} \tag{3.2}$$

where ρ_2 is a conditional coherent risk measure. Here $\xi := (q, h, T, W)$ is the data of the second stage problem. We view some or all elements of the vector ξ as well as the vector c as random. The ρ_1 operator at the first stage problem (3.1) is taken with respect to the probability distribution of $c^\top x + \mathcal{Q}(x, \xi)$.

We assume that the set

$$X^{\text{ind}} := \{x \in X \mid \mathcal{Q}(x, \xi) < \infty \ \text{ w.p. } 1\}$$

is nonempty. For each $x \in X^{\text{ind}}$, $\rho_1(x)$ is well defined. If for some x and ξ the second stage problem (3.2) is infeasible, then by definition $\mathcal{Q}(x, \xi) = +\infty$. It could also happen that the second stage problem is unbounded from below and hence $\mathcal{Q}(x, \xi) = -\infty$. This is somewhat pathological situation, meaning that for some value of the first stage

decision vector and a realization of the random data, the value of the second stage problem can be improved indefinitely. Models exhibiting such properties should be avoided.

We will assume that the distribution of ξ has finite support. That is, ξ has a finite number of realizations (called *scenarios*) $\xi_k = (q_k, h_k, T_k, W_k)$ with respective probabilities p_k. From now on we will assume that ξ has exactly N scenarios. In this case we will let $\mathcal{Z} := \mathcal{L}_1(\Omega, \mathcal{F}, P)$ which we will just identify with the space $\mathbb{R}^N$.

The representation theorem of coherent risk measures, Theorem (1), allows us to rewrite problem (3.1)-(3.2) in the form

$$\begin{aligned} \min_{x\in X} \max_{\mu\in\partial\rho_1(0)} \sum_{i=1}^{N} \mu_i p_i \left[c^\top x + \min_{y_i\in\mathbb{R}^m} \rho_2[q_i^\top y_i] \right] \\ \text{s.t. } T_i x + W_i y_i = h_i,\ i = 1, \dots, N \\ y_i \geq 0,\ i = 1, \dots, N. \end{aligned} \tag{3.3}$$

For $i = 1, \dots, N$, let $c_i \in \mathbb{R}^n$ and ρ_{2i} be coherent risk measures. From now on, we will focus on the following slight generalization of problem (3.3):

$$\begin{aligned} \min_{x\in X} \max_{\mu\in\partial\rho_1(0)} \sum_{i=1}^{N} \mu_i p_i \left[c_i^\top x + \min_{y_i\in\mathbb{R}^m} \rho_{2i}[q_i^\top y_i] \right] \\ \text{s.t. } T_i x + W_i y_i = h_i \\ y_i \geq 0\ ,\ \ i = 1, \dots, N. \end{aligned} \tag{3.4}$$

3.2 The Dual

In this section we will obtain a dual formulation of problem (3.4).

Consider the following problem:

$$\begin{aligned} \min_{(x_1,\dots,x_N)\in X^N} \max_{\mu\in\partial\rho_1(0)} \sum_{i=1}^{N} \mu_i p_i \left[c_i^\top x_i + \min_{y_i\in\mathbb{R}^m} \rho_{2i}[q_i^\top y_i] \right] \\ \text{s.t. } \ T_i x_i + W_i y_i = h_i \\ y_i \geq 0,\ x_i \in X,\ \ i = 1, \dots, N \end{aligned} \tag{3.5}$$

with the extra *nonanticipativity* constraints

$$x_i = \sum_{k=1}^{N} p_k x_k,\ i = 1, \ldots, N. \tag{3.6}$$

The nonanticipativity constraints (3.6) play a double role in the development of our theory and methods. First, they ensure that the first-stage decision variables of (3.5) do not depend on the second-stage realization of the random data and thus making problem (3.5)–(3.6) equivalent to our main problem (3.4). Second, these extra constraints will help us obtain a dual formulation of (3.4) through their appearance in the Lagrangian of (3.5)–(3.6). We aim to do this next.

Let $\mathcal{X} = \mathbb{R}^{n \cdot N}$ and $\mathcal{L} = \{x = (x_1, \ldots, x_N) \,|\, x_1 = \ldots = x_N\}$. Equip the space $\mathcal{X}$ with the scalar product

$$\langle x, y \rangle = \sum_{i=1}^{N} p_i x_i^\top y_i \tag{3.7}$$

and define the linear operator $P : \mathcal{X} \to \mathcal{X}$ as

$$Px = \left(\sum_{i=1}^{N} p_i x_i, \ldots, \sum_{i=1}^{N} p_i x_i \right).$$

The nonanticipativity constraints (3.6) can be compactly written as

$$x = Px.$$

It can be verified that P is the orthogonal projection operator of $\mathcal{X}$, equipped with the scalar product (3.7), onto its subspace $\mathcal{L}$.

For every $x \in \mathbb{R}^N$ and $i \in \{1, \ldots, N\}$ define

$$\begin{aligned} F(x, \omega_i) \triangleq\ & c_i^\top x + \inf_{y \in \mathbb{R}^m} \rho_{2i}[q_i^\top y] \\ & \text{s.t.}\ \ T_i x + W_i y = h_i \\ & y \geq 0. \end{aligned} \tag{3.8}$$

The convexity property (A1) of the coherent risk measure ρ_{2i} implies that $F(\cdot, \omega_i)$ is a

convex function. Also the assumptions made on the second stage problem $\mathcal{Q}(x,\xi)$ imply that $F(X,\omega_i) \subset \mathbb{R}$. Therefore the compactness of X implies that $F(\cdot,\omega_i)$ is continuous relative to X (see [28, Thm. 10.4]).

By assigning Lagrange multipliers $\lambda_k \in \mathbb{R}^n, k = 1,\ldots,N$, to the nonanticipativity constraints (3.6), we obtain that the Lagrangian of problem (3.5)–(3.6) is given by:

$$L(x,\lambda) \triangleq \max_{\mu\in\partial\rho_1(0)} \sum_{i=1}^{N} \mu_i p_i F(x_i,\omega_i) + \sum_{j=1}^{N} p_j \lambda_j^\top \left(x_j - \sum_{t=1}^{N} p_t x_t \right),$$

where $x = (x_1,\ldots,x_N) \in X^N$ and $\lambda^\top = (\lambda_1^\top,\ldots,\lambda_N^\top)$. Note that since P is an orthogonal projection, $I-P$ is also an orthogonal projection (onto the space orthogonal to $\mathcal{L}$), and hence

$$\sum_{j=1}^{N} p_j \lambda_j^\top \left(x_j - \sum_{t=1}^{N} p_t x_t \right) = \langle \lambda, (I-P)x \rangle = \langle (I-P)\lambda, x \rangle .$$

Therefore the above Lagrangian can be written in the following equivalent form

$$L(x,\lambda) = \max_{\mu\in\partial\rho_1(0)} \sum_{i=1}^{N} \mu_i p_i F(x_i,\omega_i) + \sum_{j=1}^{N} p_j \left(\lambda_j - \sum_{t=1}^{N} p_t \lambda_t \right)^\top x_j.$$

Observe that shifting the multipliers $\lambda_j, j = 1,\ldots,N$, by a constant vector does not change the value of the Lagrangian, because the expression $\lambda_j - \sum_{t=1}^{N} p_t \lambda_t$ is invariant to such shifts. Therefore, with no loss of generality we can assume that

$$\sum_{j=1}^{N} p_j \lambda_j = 0$$

or equivalently, that $P\lambda = 0$. Under the condition $P\lambda = 0$, the Lagrangian can be written simply as

$$L(x,\lambda) = \max_{\mu\in\partial\rho_1(0)} \sum_{i=1}^{N} p_i \left[\mu_i F(x_i,\omega_i) + \lambda_i^\top x_i \right].$$

Putting everything together we obtain the following dual formulation of problem

(3.5)–(3.6):

$$\max_{\lambda\in\mathbb{R}^{n\cdot N}} \min_{x\in X^N} \max_{\mu\in\partial\rho_1(0)} \sum_{i=1}^{N} p_i \left[\mu_i F(x_i,\omega_i) + \lambda_i^\top x_i\right] \tag{3.9}$$
$$\text{s.t.} \quad \sum_{j=1}^{N} p_j\lambda_j = 0.$$

Note that X is polyhedral, $\partial\rho_1(0)$ is convex and compact, and $\mu_i F(x_i,\omega_i) + \lambda_i^\top x_i$ is convex in x and linear in μ. Therefore we can interchange the innermost max by the min in (3.9) (see [42, Thm. 3.1]) and obtain the following equivalent formulation for the dual of (3.5)–(3.6):

$$\max_{\substack{\lambda\in\mathbb{R}^{n\cdot N}\\ \mu\in\partial\rho_1(0)}} \sum_{i=1}^{N} p_i \min_{x_i\in X} \left[\mu_i F(x_i,\omega_i) + \lambda_i^\top x_i\right] \tag{3.10}$$
$$\text{s.t.} \quad \sum_{j=1}^{N} p_j\lambda_j = 0.$$

We will require that our problem satisfies the Slater's constraint qualification condition and in this way guaranteeing that the duality gap will be zero. We should first point out that the convexity of the functions $F(\cdot,\omega_i)$ imply that the primal objective function

$$\max_{\mu\in\partial\rho_1(0)} \sum_{i=1}^{N} \mu_i p_i\, F(x,\omega_i)$$

is convex too. Due to this fact and the continuity of $F(\cdot,\omega_i)$ relative to X, problem (3.5)–(3.6) satisfies the Slater's constraint qualification condition if there is $x = (x_1,\ldots,x_N) \in \operatorname{int} X^N$ such that

$$x_i = \sum_{k=1}^{N} p_k x_k,\ i = 1,\ldots,N.$$

In this case the duality gap is zero (see [35, Thm. 4.7]) and so, the optimal value of the primal problem (3.5)–(3.6) is the same as the optimal value of the dual problem (3.10).

Suppose that once we are in scenario i, there are exactly N_i possible sub-scenarios that could occur each with probability $\pi_{ij}, j = 1,\ldots,N_i$ and its own vector q_{ij}. Clearly

$\pi_{ij} > 0$ and $\sum_{j=1}^{N_i} \pi_{ij} = 1$. The representation theorem of coherent risk measures shows that

$$\rho_{2i}[q_i^\top y] = \max_{\delta \in \partial\rho_{2i}(0)} \sum_{j=1}^{N_i} \delta_j \pi_{ij} q_{ij}^\top y, \tag{3.11}$$

where

$$\partial\rho_{2i}(0) \subseteq \left\{ \delta \in \mathbb{R}^{N_i} \,\middle|\, \sum_{j=1}^{N_i} \delta_j \pi_{ij} = 1, \delta \geq 0 \right\}.$$

From (3.8), (3.10), and (3.11) we obtain that the dual of the main problem is equivalent to

$$\begin{aligned} \max_{\substack{\lambda \in \mathbb{R}^{n\cdot N} \\ \mu \in \partial\rho_1(0)}} \min_{x \in X^N} & \sum_{i=1}^{N} p_i \left[\mu_i c_i^\top x_i + \mu_i \min_{y_i \in \mathbb{R}^m} \left[\max_{\delta_i \in \partial\rho_{2i}(0)} \sum_{j=1}^{N_i} \delta_{ij} \pi_{ij} q_{ij}^\top y_i \right] + \lambda_i^\top x_i \right] \\ \text{s.t.} \quad & T_i x_i + W_i y_i = h_i, \ i = 1, \ldots, N \\ & y_i \geq 0, \ i = 1, \ldots, N \\ & \sum_{j=1}^{N} p_j \lambda_j = 0. \end{aligned} \tag{3.12}$$

This, in turn, is equivalent to

$$\begin{aligned} \max_{\substack{\lambda \in \mathbb{R}^{n\cdot N} \\ \mu \in \partial\rho_1(0)}} \min_{\substack{x \in X^N \\ y \in \mathbb{R}^{m\cdot N}}} \max_{\delta \in \partial\rho_2(0)} & \sum_{i=1}^{N} p_i \left[\mu_i c_i^\top x_i + \mu_i \sum_{j=1}^{N_i} \delta_{ij} \pi_{ij} q_{ij}^\top y_i + \lambda_i^\top x_i \right] \\ \text{s.t.} \quad & T_i x_i + W_i y_i = h_i, \ i = 1, \ldots, N \\ & y_i \geq 0, \ i = 1, \ldots, N \\ & \sum_{j=1}^{N} p_j \lambda_j = 0, \end{aligned} \tag{3.13}$$

where $\partial\rho_2(0) := \partial\rho_{21}(0) \times \cdots \times \partial\rho_{2N}(0)$, $y = (y_1, \ldots, y_N)$, and $\delta = (\delta_1, \ldots, \delta_N)$. As before (see [42, Thm. 3.1]), we can interchange the innermost max and min and obtain

the following equivalent formulation

$$\max_{\substack{\lambda\in\mathbb{R}^{n\cdot N}\\ \mu\in\partial\rho_1(0)\\ \delta\in\partial\rho_2(0)}} \min_{\substack{x\in X^N\\ y\in\mathbb{R}^{m\cdot N}}} \sum_{i=1}^{N} p_i \left[\mu_i c_i^\top x_i + \mu_i \sum_{j=1}^{N_i} \delta_{ij}\pi_{ij} q_{ij}^\top y_i + \lambda_i^\top x_i\right] \tag{3.14}$$

$$\begin{aligned} \text{s.t.}\quad & T_i x_i + W_i y_i = h_i,\ i = 1,\dots,N \\ & y_i \geq 0,\ i = 1,\dots,N \\ & \sum_{j=1}^{N} p_j \lambda_j = 0. \end{aligned}$$

Let

$$\mathcal{S} := \left\{ \begin{pmatrix} \pi_{11}\mu_1\delta_{11} \\ \vdots \\ \pi_{ij}\mu_i\delta_{ij} \\ \vdots \\ \pi_{NN_N}\mu_N\delta_{NN_N} \end{pmatrix} \middle| \ \mu_i \in [\partial\rho_1(0)]^i, \begin{pmatrix} \delta_{i1} \\ \vdots \\ \delta_{iN_i} \end{pmatrix} \in \partial\rho_{2i}(0),\ i = 1,\dots,N \right\},$$

where $[\partial\rho_1(0)]^i$ is the projection of $\partial\rho_1(0)$ on the ith axis. Note that $\mathcal{S}$ is a convex and compact set. Then (3.14) is equivalent to:

$$\max_{\substack{\lambda\in\mathbb{R}^{n\cdot N}\\ \mu\in\partial\rho_1(0)\\ \sigma\in\mathcal{S}}} \min_{\substack{x\in X^N\\ y\in\mathbb{R}^{m\cdot N}}} \sum_{i=1}^{N} p_i \left[\mu_i c_i^\top x_i + \sum_{j=1}^{N_i} \sigma_{ij} q_{ij}^\top y_i + \lambda_i^\top x_i\right] \tag{3.15}$$

$$\begin{aligned} \text{s.t.}\quad & T_i x_i + W_i y_i = h_i,\ i = 1,\dots,N \\ & y_i \geq 0,\ i = 1,\dots,N \\ & \sum_{j=1}^{N} p_j \lambda_j = 0. \\ & \sum_{j=1}^{N_i} \sigma_{ij} = \mu_i\,,\ i = 1,\dots,N \end{aligned}$$

Let $\mathbb{R}^{\widetilde{N}} := \mathbb{R}^{N_1} \times \cdots \times \mathbb{R}^{N_N}$. For every scenario $i = 1,\dots,N$ define the function

$\chi_i : \mathbb{R}^{n \cdot N} \times \mathbb{R}^N \times \mathbb{R}^{\widetilde{N}} \to \overline{\mathbb{R}}$ such that

$$\chi_i(\lambda, \mu, \sigma) \triangleq \min_{x,y} \ \mu_i c_i^\top x + \sum_{j=1}^{N_i} \sigma_{ij} q_{ij}^\top y + \lambda_i^\top x \tag{3.16}$$

$$\text{s.t.} \quad T_i x + W_i y = h_i \tag{3.17}$$

$$x \in X, \ y \in \mathbb{R}^m, \ y \geq 0.$$

Then the dual of our main problem is given by:

$$\max_{\lambda,\mu,\sigma} \ \sum_{i=1}^{N} p_i \, \chi_i(\lambda, \mu, \sigma) \tag{3.18}$$

$$\text{s.t.} \quad \sum_{j=1}^{N} p_j \lambda_j = 0$$

$$\sum_{j=1}^{N_i} \sigma_{ij} = \mu_i \,, \ i = 1, \ldots, N$$

$$\lambda \in \mathbb{R}^{n \cdot N}, \ \mu \in \partial\rho_1(0), \ \sigma \in \mathcal{S}.$$

Let $\mathcal{A}_i$ be the set of pairs (x, y) satisfying the system of constraints (3.17), by assumption $\mathcal{A}_i \neq \emptyset$. The compactness of X implies that A_i is compact. From this we conclude that $\chi_i(\lambda, \mu, \sigma) \in \mathbb{R}$, for every $(\lambda, \mu, \sigma) \in \mathbb{R}^{n \cdot N} \times \mathbb{R}^N \times \mathbb{R}^{\widetilde{N}}$ and χ_i is proper, concave, and Lipschitz continuous (see [28, Thm. 10.4]).

3.3 The Subgradient of the Dual Function

Define the sets

$$\Lambda \triangleq \left\{ \lambda \in \mathbb{R}^{n \cdot N} \;\middle|\; \sum_{i=1}^{N} p_i \lambda_i = 0 \right\}, \tag{3.19}$$

$$\Delta \triangleq \left\{ (\mu, \sigma) \in \partial\rho_1(0) \times \mathcal{S} \;\middle|\; \sum_{j=1}^{N_i} \sigma_{ij} = \mu_i \,, \ i = 1, \ldots, N \right\},$$

and the function $\vartheta_i(\cdot,\cdot) : \left(\mathbb{R}^{n\cdot N} \times \mathbb{R}^N \times \mathbb{R}^{\widetilde{N}}\right) \times \mathcal{A}_i \to \mathbb{R}$ by

$$\vartheta_i\left[(\lambda,\mu,\sigma),(x,y)\right] \triangleq \mu_i c_i^\top x + \sum_{j=1}^{N_i} \sigma_{ij} q_{ij}^\top y + \lambda_i^\top x. \tag{3.20}$$

The above definition implies that for every $(\lambda,\mu,\sigma) \in \mathbb{R}^{n\cdot N} \times \mathbb{R}^N \times \mathbb{R}^{\widetilde{N}}$

$$\chi_i(\lambda,\mu,\sigma) = \min_{(x,y)\in\mathcal{A}_i} \vartheta_i\left[(\lambda,\mu,\sigma),(x,y)\right]. \tag{3.21}$$

Let

$$\chi(\lambda,\mu,\sigma) \triangleq \sum_{i=1}^{N} p_i\, \chi_i(\lambda,\mu,\sigma). \tag{3.22}$$

Then, we can rewrite our main dual problem (3.18) as

$$\max_{(\lambda,\mu,\sigma)\in\Lambda\times\Delta} \chi(\lambda,\mu,\sigma), \tag{3.23}$$

where $\lambda \in \lambda$ and $(\mu,\sigma) \in \Delta$.

The main purpose of this section is to calculate the subdifferential of χ, which is key in the development of methods to solve efficiently problem (3.23). Remember that for each $i \in 1,\ldots,N$, the function $\chi_i(\cdot,\cdot)$ is concave, continuous, and takes only real values. Therefore we can apply the Moreau-Rockafellar theorem and obtain

$$\partial\chi(\lambda,\mu,\sigma) = \sum_{i=1}^{N} p_i \partial\chi_i(\lambda,\mu,\sigma). \tag{3.24}$$

Because of this we will focus on the subdifferentials of the χ_i's. Definition (3.20) allows us to see that the following proposition holds.

Proposition 6. *The function $\vartheta_i\left[\,\cdot\,,(x,y)\right]$ is concave for every $(x,y) \in \mathcal{A}_i$. Also, the function $\vartheta_i\left[(\lambda,\mu,\sigma),\,\cdot\,\right]$ is lower semicontinuous for every $(\lambda,\mu,\sigma) \in \mathbb{R}^{n\cdot N} \times \mathbb{R}^N \times \mathbb{R}^{\widetilde{N}}$.*

Let $\partial\vartheta_i\left[(\lambda_0,\mu_0,\sigma_0),(x,y)\right]$ be the subdifferential of $\partial\vartheta_i\left[\,\cdot\,,(x,y)\right]$ at the point $(\lambda_0,\mu_0,\sigma_0)$ and let Q_i be the $N_i \times m$ matrix with rows $q_{ij}^\top$. Then

$$\partial\vartheta_i\left[(\lambda_0,\mu_0,\sigma_0),(x,y)\right]^\top = \left(\hat{x}_i^\top, e_i^\top \cdot c_i^\top x, \overline{Q_i y}^\top\right), \tag{3.25}$$

where $\hat{x}_i^\top := (0,\dots,0,\, x^\top, 0,\dots,0) \in \mathbb{R}^{n\cdot N}$ such that the x is in the ith position and each 0 is a vector of $\mathbb{R}^n$, and $\overline{Q_i y}^\top := \big(0,\dots,0,\, (Q_i y)^\top, 0,\dots,0\big) \in \mathbb{R}^{\widetilde{N}}$ such that the $(Q_i y)^\top$ is in the ith position and a 0 in position t is a vector of $\mathbb{R}^{N_t}$. For every $(\lambda,\mu,\sigma) \in \mathbb{R}^{n\cdot N} \times \mathbb{R}^N \times \mathbb{R}^{\widetilde{N}}$ define

$$\mathcal{A}_i(\lambda,\mu,\sigma) \triangleq \arg\min_{(x,y)\in\mathcal{A}_i} \left\{ \mu_i c_i^\top x + \sum_{j=1}^{N_i} \sigma_{ij} q_{ij}^\top y + \lambda_i^\top x \right\}.$$

The set $\mathcal{A}_i(\lambda,\mu,\sigma)$ is the set of optimal solutions to the right hand side of (3.21). Then, since the function χ_i is a minimum function (see (3.21) and [35, Thm. 2.87]), we get

$$\begin{aligned} \partial\chi_i(\lambda_0,\mu_0,\sigma_0)^\top &= \operatorname{conv}\left[\bigcup_{(x,y)\in\mathcal{A}_i(\lambda_0,\mu_0,\sigma_0)} \partial\vartheta_i\left[(\lambda_0,\mu_0,\sigma_0),(x,y)\right]^\top \right] \\ &= \left\{ \left(\hat{x}_i^\top, e_i^\top \cdot c_i^\top x, \overline{Q_i y}^\top \right) \,\middle|\, (x,y) \in \mathcal{A}_i(\lambda_0,\mu_0,\sigma_0) \right\}. \end{aligned} \tag{3.26}$$

Using (3.24) and (3.26), we obtain

$$\partial\chi(\lambda_0,\mu_0,\sigma_0)^\top = \sum_{i=1}^{N} p_i \left\{ \left(\hat{x}_i^\top, e_i^\top \cdot c_i^\top x, \overline{Q_i y}^\top \right) \,\middle|\, (x,y) \in \mathcal{A}_i(\lambda_0,\mu_0,\sigma_0) \right\}. \tag{3.27}$$

Usually we need to find a subgradient of χ at a given point $(\lambda_0,\mu_0,\sigma_0)$. From (3.27) we can derive a simple procedure to accomplish this:

Step 1. For every $i = 1,\dots,N$, solve the linear program

$$\min_{(x,y)\in\mathcal{A}_i} (\mu_0)_i\, c_i^\top x + \sum_{j=1}^{N_i} (\sigma_0)_{ij}\, q_{ij}^\top y + (\lambda_0)_i^\top x$$

and call the obtained optimal solution (x_i, y_i). Note that $(x_i, y_i) \in \mathcal{A}_i(\lambda_0,\mu_0,\sigma_0)$, for every $i = 1,\dots,N$.

Step 2. Compute $\alpha := \sum_{i=1}^{N} p_i \left((\hat{x}_i)_i^\top, e_i^\top \cdot c_i^\top(x_i), \overline{Q_i(y_i)_i}^\top \right)$.

Then (3.27) implies that $\alpha \in \partial\chi(\lambda_0,\mu_0,\sigma_0)$, as we wanted. Notice that the obtained subgradient α has the following structure:

$$\alpha = \begin{pmatrix} p_1 x_1 \\ \vdots \\ p_N x_N \\ p_1 c_1^\top x_1 \\ \vdots \\ p_N c_N^\top x_N \\ p_1 Q_1 y_1 \\ \vdots \\ p_N Q_n y_n \end{pmatrix}. \tag{3.28}$$

Chapter 4

The Two-Stage Risk-Averse Problem With Mean Upper Semideviation

4.1 Specifying the Coherent Risk Measures

It is possible to obtain different formulations of the dual problem when we restrict to consider specific risk measures in formulation (3.4). In this section we will consider problem (3.4) where ρ_1 and ρ_{2i} are all mean upper semideviations of the first order. Let $\rho_1(Z) := \mathbb{E}[Z] + a_1 \mathbb{E}\left[Z - \mathbb{E}[Z]\right]_+$ and $\rho_{2i}(Z) := \mathbb{E}[Z] + b_i \mathbb{E}\left[Z - \mathbb{E}[Z]\right]_+$, where $i = 1, \ldots, N$ and $a_1, b_i \in [0, 1]$.

Equation (2.18) shows that the subdifferential at 0 of ρ_1 is

$$\partial\rho_1(0) = \left\{ \mathbb{1} - \mathbb{1}\sum_{i=1}^{N} p_i\xi_i + \xi \;\middle|\; \xi = (\xi_i)_{i=1}^{N} \text{ and } 0 \leq \xi_i \leq a_1 \right\}, \tag{4.1}$$

where $\mathbb{1}$ is the vector with all entries equal to 1. Applying the representation theorem of the coherent risk measures (Theorem 1), we obtain that

$$\rho_{2i}(q_i^\top y) = \max_{\delta \in \partial\rho_{2i}(0)} \sum_{j=1}^{N_i} \delta_j \pi_{ij} q_{ij}^\top y, \tag{4.2}$$

where, by (2.18),

$$\partial\rho_{2i}(0) = \left\{ \mathbb{1} - \mathbb{1}\sum_{j=1}^{N_i} \pi_{ij}\xi_j + \xi \;\middle|\; \xi = (\xi_j)_{j=1}^{N_i} \text{ and } 0 \leq \xi_j \leq b_i \right\}. \tag{4.3}$$

Substituting (4.3) into (4.2) gives

$$\rho_{2i}(q_i^\top y) = \max_{\xi \in [0,b_i]^{N_i}} \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y + \sum_{k=1}^{N_i} \xi_k \pi_{ik} \left[q_{ik}^\top y - \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y \right]. \tag{4.4}$$

Since $\pi_{ik} > 0$ and $\xi \in [0, b_i]^{N_i}$, the maximum on the right hand side of (4.4) is given by ξ^* such that

$$\xi_k^* = \begin{cases} b_i & \text{if } \; q_{ik}^\top y - \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

$k = 1, \ldots, N_i$. Therefore, we can obtain $\rho_{2i}(q_i^\top y)$ by solving the following linear program

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y + \sum_{k=1}^{N_i} d_k \\ \text{subject to:} \quad & d_k \geq 0, \\ & d_k \geq b_i \pi_{ik} \left[q_{ik}^\top y - \mathbb{E}[q_i^\top y] \right], \\ & \text{for all} \;\; k = 1, \ldots, N_i. \end{aligned} \tag{4.5}$$

where $\mathbb{E}[q_i^\top y] := \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y$.

For every scenario $i = 1, \ldots, N$, define the function $\varphi_i : \mathbb{R}^{n \cdot N} \times \mathbb{R}^N \to \overline{\mathbb{R}}$ such that

$$\varphi_i(\lambda, \mu) \triangleq \min_{x \in X} \left[\mu_i F(x, \omega_i) + \lambda_i^\top x \right],$$

where $\lambda = (\lambda_1, \ldots, \lambda_N)$. Then by (3.10), the dual of our problem is given by:

$$\begin{aligned} \max_{\substack{\lambda \in \mathbb{R}^{n \cdot N} \\ \mu \in \partial \rho_1(0)}} \quad & \sum_{i=1}^{N} p_i \, \varphi_i(\lambda, \mu) \\ \text{s.t.} \quad & \sum_{j=1}^{N} p_j \lambda_j = 0. \end{aligned} \tag{4.6}$$

Using the definition of $F(x, \omega_i)$ and the characterization of $\rho_{2i}(q_i^\top y)$ obtained in (4.5),

we get

$$\varphi_i(\lambda,\mu) = \min_{x,y,d} \mu_i \left[c_i^\top x + \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y + \sum_{k=1}^{N_i} d_k \right] + \lambda_i^\top x \tag{4.7}$$

$$\begin{aligned} &\text{subject to:} \qquad (4.8)\\ &d_k \geq c_k \pi_{ik} \left[q_{ik}^\top y - \mathbb{E}[q_i^\top y] \right], \;\; k = 1, \ldots, N_i \\ &T_i x + W_i y = h_i \\ &x \in X, \;\; y \geq 0, \;\; d \geq 0. \end{aligned}$$

Note that the compactness of X implies that $\varphi_i : \mathbb{R}^{n \cdot N} \times \mathbb{R}^N \to \mathbb{R}$ is a proper concave function and therefore φ_i is Lipschitz continuous (see [28, Thm. 10.4]). Formulation (4.6)–(4.8) is practical for the application of cutting plane methods.

For every $i = 1, \ldots, N$, let $\mathcal{B}_i$ be the set of triples (x, y, d) satisfying the system of inequalities (4.8). The set $\mathcal{B}_i$ is closed but not bounded. Nevertheless, our assumption of the second stage problem $\mathcal{Q}(x, \xi)$ having always a real optimal solution guarantees that we could assume (by adding extra constraints) that $\mathcal{B}_i$ is bounded. So, from now on we will assume that $\mathcal{B}_i$ is closed and bounded.

Define the function $\psi_i(\cdot, \cdot) : \left(\mathbb{R}^{n \cdot N} \times \mathbb{R}^N\right) \times \mathcal{B}_i \to \mathbb{R}$ by

$$\psi_i \left[(\lambda, \mu), (x, y, d) \right] \triangleq \mu_i \left[c_i^\top x + \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y + \sum_{k=1}^{N_i} d_k \right] + \lambda_i^\top x. \tag{4.9}$$

The above definition implies that for every $(\lambda, \mu) \in \mathbb{R}^{n \cdot N} \times \mathbb{R}^N$

$$\varphi_i(\lambda, \mu) = \min_{(x,y,d) \in \mathcal{B}_i} \psi_i \left[(\lambda, \mu), (x, y, d) \right]. \tag{4.10}$$

Let

$$\varphi(\lambda, \mu) \triangleq \sum_{i=1}^{N} p_i \, \varphi_i(\lambda, \mu). \tag{4.11}$$

Then, we can rewrite the dual problem (4.6) as

$$\max_{(\lambda,\mu)\in\Lambda\times\partial\rho_1(0)} \varphi(\lambda,\mu). \tag{4.12}$$

By definition $\varphi(\cdot,\cdot)$ is proper, concave, and continuous, so applying the Moreau-Rockafellar theorem we obtain:

$$\partial\varphi(\lambda,\mu) = \sum_{i=1}^{N} p_i \partial\varphi_i(\lambda,\mu). \tag{4.13}$$

By definition (4.9) we can easily see that the following proposition holds.

Proposition 7. *The function $\psi_i[\,\cdot\,,(x,y,d)]$ is concave for every $(x,y,d)\in\mathcal{B}_i$. Also, the function $\psi_i[(\lambda,\mu),\,\cdot\,]$ is upper semicontinuous for every $(\lambda,\mu)\in\mathbb{R}^{n\cdot N}\times\mathbb{R}^N$.*

For every (x,y,d) let

$$\phi_i(x,y,d) = c_i^\top x + \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y + \sum_{k=1}^{N_i} d_k$$

and let $\partial\psi_i[(\lambda_0,\mu_0),(x,y,d)]$ be the subdifferential of $\partial\psi_i[\,\cdot\,,(x,y,d)]$ at the point (λ_0,μ_0). Then

$$\partial\psi_i[(\lambda_0,\mu_0),(x,y,d)]^\top = \left(\hat{x}_i^\top, e_i^\top\cdot\phi_i(x,y,d)\right), \tag{4.14}$$

recall that $\hat{x}_i^\top = (0,\ldots,0,\,x^\top,0,\ldots,0)\in\mathbb{R}^{n\cdot N}$ has the x is in the ith position and each 0 is a vector of $\mathbb{R}^n$. Also, for every $(\lambda,\mu)\in\mathbb{R}^{n\cdot N}\times\mathbb{R}^N$ define

$$\mathcal{B}_i(\lambda,\mu) \triangleq \arg\min_{(x,y,d)\in\mathcal{B}_i} \left\{\mu_i\phi_i(x,y,d) + \lambda_i^\top x\right\}.$$

Clearly the set $\mathcal{B}_i(\lambda,\mu)$ is the set of optimal solutions to the right hand side of (4.10).

Then, as before,

$$\partial\varphi_i(\lambda_0,\mu_0)^\top = \left\{\left(\hat{x}_i^\top, e_i^\top\cdot\phi_i(x,y,d)\right)\,\middle|\,(x,y,d)\in\mathcal{B}_i(\lambda_0,\mu_0)\right\}, \tag{4.15}$$

and

$$\partial\varphi(\lambda_0,\mu_0)^\top = \sum_{i=1}^{N} p_i \left\{ \left(\hat{x}_i^\top, e_i^\top \cdot \phi_i(x,y,d) \right) \,\middle|\, (x,y,d) \in \mathcal{B}_i(\lambda_0,\mu_0) \right\}. \tag{4.16}$$

Just as before, (4.16) give us a simple procedure to obtain a subgradient of φ at (λ_0,μ_0):

Step 1. For every $i = 1,\ldots,N$, solve the linear program

$$\min_{(x,y,d)\in\mathcal{B}_i} (\mu_0)_i\, \phi_i(x,y,d) + (\lambda_0)_i^\top x$$

and call the obtained optimal solution (x_i,y_i,d_i). Note that $(x_i,y_i,d_i) \in \mathcal{B}_i(\lambda_0,\mu_0)$, for every $i = 1,\ldots,N$.

Step 2. Compute $\alpha := \sum_{i=1}^{N} p_i \left((\widehat{x_i})_i^\top, e_i^\top \cdot \phi_i(x_i,y_i,d_i) \right)$.

Then (4.16) implies that $\alpha \in \partial\varphi(\lambda_0,\mu_0)$, as we wanted. Notice that the obtained subdifferential α is very simple:

$$\alpha = \begin{pmatrix} p_1\, x_1 \\ \vdots \\ p_N\, x_N \\ p_1\, \phi(x_1,y_1,d_1) \\ \vdots \\ p_N\, \phi(x_N,y_N,d_N) \end{pmatrix}. \tag{4.17}$$

Chapter 5

The Dual Cutting Plane Method

We will apply the cutting plane method to the main dual problem:

$$\max_{(\lambda,\mu)\in\Lambda\times\partial\rho_1(0)} \varphi(\lambda,\mu), \tag{5.1}$$

where

$$\varphi(\lambda,\mu) = \sum_{i=1}^{N} p_i\,\varphi_i(\lambda,\mu), \tag{5.2}$$

and for every $i = 1,\ldots,N$ and $(\lambda,\mu)\in\Lambda\times\partial\rho_1(0)$,

$$\varphi_i(\lambda,\mu) = \min_{x,y,d} \mu_i \left[c_i^\top x + \sum_{j=1}^{N_i} \pi_{ij} q_{ij}^\top y + \sum_{k=1}^{N_i} d_k \right] + \lambda_i^\top x \tag{5.3}$$

subject to:

$$d_k \geq c_k \pi_{ik} \left[q_{ik}^\top y - \mathbb{E}[q_i^\top y] \right],\ \ k = 1,\ldots,N_i$$

$$T_i x + W_i y = h_i$$

$$x \in X,\ \ y \geq 0,\ \ d \geq 0.$$

As we stated before the functions $\varphi_i : \mathbb{R}^{n\cdot N}\times\mathbb{R}^N \to \overline{\mathbb{R}}$ are proper, concave and Lipschitz continuous. The set $\Lambda\times\partial\rho_1(0)$ is closed but might not be bounded. In order to be assured of finding a solution with the cutting plane method we should add, if necessary, linear constraints to Λ in such a way that $\Lambda\times\partial\rho_1(0)$ is compact. The set $\Lambda\times\partial\rho_1(0)$ should be large enough so that it contains an optimal solution to problem (5.1). From now on we will assume that $\Lambda\times\partial\rho_1(0)$ is a compact set.

All the properties mentioned in the previous paragraph are the theoretical requirements that guarantee the convergence to an optimal solution of the cutting plane method when applied to problem (5.1) (see [35] p. 357). The idea behind the *cutting plane method* is to use the subgradient inequality,

$$\varphi(\lambda,\mu) \leq \varphi(\lambda_0,\mu_0) + \langle g, (\lambda,\mu) - (\lambda_0,\mu_0)\rangle\,,$$

which holds true for every $(\lambda,\mu) \in \mathbb{R}^{n\cdot N} \times \mathbb{R}^N$ and each subgradient $g \in \partial\varphi(\lambda_0,\mu_0)$, for constructing upper approximations of $\varphi(\cdot)$ (remember that $\varphi(\cdot)$ is concave). At each step the method refines the approximation to $\varphi(\cdot)$ and selects point which is the "best so far" approximation to an optimal solution of (5.1).

The method starts at a given point $(\lambda^1,\mu^1) \in \Lambda \times \partial\rho_1(0)$, calculates $g^1 \in \partial\varphi(\lambda^1,\mu^1)$, and constructs a linear approximation of $\varphi(\cdot)$:

$$\varphi^1(\lambda,\mu) \triangleq \varphi(\lambda^1,\mu^1) + \left\langle g^1, (\lambda,\mu) - (\lambda^1,\mu^1)\right\rangle.$$

In a general iteration k, having already generated points $(\lambda^1,\mu^1),\ldots,(\lambda^k,\mu^k)$, values of the function $\varphi(\lambda^1,\mu^1),\ldots,\varphi(\lambda^k,\mu^k)$, and corresponding subgradients $g^1,\ldots,g^k$, the method construct an upper approximation of the function $\varphi(\cdot)$

$$\varphi^k(\lambda,\mu) \triangleq \min_{1\leq j\leq k} \left[\varphi(\lambda^j,\mu^j) + \left\langle g^j, (\lambda,\mu) - (\lambda^j,\mu^j)\right\rangle\right]. \tag{5.4}$$

Then it solves the *master problem*:

$$\underset{(\lambda,\mu)\in\Lambda\times\partial\rho_1(0)}{\text{maximize}}\ \varphi^k(\lambda,\mu), \tag{5.5}$$

and add its solution $(\lambda^{k+1},\mu^{k+1})$ to the set of points. After evaluating $\varphi(\lambda^{k+1},\mu^{k+1})$ and $g^{k+1} \in \partial\varphi(\lambda^{k+1},\mu^{k+1})$, it increases k by one, and continue the calculations. If

$$\varphi(\lambda^{k+1},\mu^{k+1}) = \varphi^k(\lambda^{k+1},\mu^{k+1}),$$

then the method stops; at this moment the point $\varphi(\lambda^{k+1},\mu^{k+1})$ is optimal ([35, sec.

7.2]).

The master problem (5.5) is equivalent to the following constrained optimization problem:

$$\begin{aligned} &\text{maximize} && z \\ &\text{subject to} && z \le \varphi(\lambda^j,\mu^j) + \left\langle g^j, (\lambda,\mu) - (\lambda^j,\mu^j)\right\rangle, \quad j = 1,\dots,k, \\ &&& (\lambda,\mu) \in \Lambda \times \partial\rho_1(0), \end{aligned} \tag{5.6}$$

whose solution $\left[(\lambda^{k+1},\mu^{k+1}), z^{k+1}\right]$ is the next approximation to the solution of (5.1) and an upper bound for $\varphi(\cdot)$ on $\Lambda\times\partial\rho_1(0)$. This new formulation of the master problem has the advantage that, after passing to iteration $k+1$, just one constraint is added to this problem, and re-optimization by a dual method is an attractive option. This is particularly useful since the set $\Lambda \times \partial\rho_1(0)$ is polyhedral and problem (5.6) is a linear program, for which efficient linear programming techniques can be employed.

Now we will show an explicit reformulation of (5.6). At each iteration k, solve for every $i = 1,\dots,N$ the problem (4.7)-(4.8) with $(\lambda,\mu) := (\lambda^k,\mu^k)$. Denote correspondingly by β_i^k and (x_i^k, y_i^k, d_i^k) the obtained optimal value and optimal solution. Then $\phi(\lambda^k,\mu^k) = \sum_{i=1}^N p_i\,\beta_i^k$ and by equation (4.17) we obtain that

$$\begin{pmatrix} p_1\,x_1^k \\ \vdots \\ p_N\,x_N^k \\ p_1\,\phi(x_1^k, y_1^k, d_1^k) \\ \vdots \\ p_N\,\phi(x_N^k, y_N^k, d_N^k) \end{pmatrix} \in \partial\varphi(\lambda^k,\mu^k).$$

This, the definition of Λ (see 3.19), and the characterization of $\partial\rho_1(0)$ (see (4.1)) gives the following reformulation of the master problem:

maximize z

subject to:

$$
\begin{aligned}
z \leq \varphi(\lambda^j, \mu^j) + \left\langle \begin{pmatrix} p_1\, x_1^j \\ \vdots \\ p_N\, x_N^j \end{pmatrix}, \lambda - \lambda^j \right\rangle + \left\langle \begin{pmatrix} p_1\, \phi(x_1^j, y_1^j, d_1^j) \\ \vdots \\ p_N\, \phi(x_N^j, y_N^j, d_N^j) \end{pmatrix}, \mu - \mu^j \right\rangle, \\
\text{for all } j = 1, \ldots, k, \\
\sum_{i=i}^{N} p_i \lambda_i = 0, \\
\mu_i = 1 - \sum_{i=1}^{N} p_i \xi_i + \xi_i,\ i = 1, \ldots, N, \\
0 \leq \xi_i \leq a_1,\ i = 1, \ldots, N,
\end{aligned}
\tag{5.7}
$$

where $\mu = (\mu_1, \ldots, \mu_N)$. This formulation is concise and practical for implementations of the method.

Chapter 6

The Partial Bundle Method

The dual cutting plane method took advantage of all the particular calculations that we developed for problem (3.4), its dual, and their restatements. There are, however, some more possible routes of optimization that we have not explored yet. First, we could apply a more sophisticated "cutting plane type" methods such at the bundle method to the dual problem (5.1). Second, we could exploit the geometrical properties of the feasible region of the problem to simplify the required calculations on the selected method. This is exactly what we set to do in the following sections. The new method exploits the features in the domain of the objective function to reduce the number of variables that will be involved in the quadratic master problem. We call this new method the *partial bundle* method.

The methods developed in this section follow the literature and act upon convex functions. Despite this, our main goal is to apply these methods on problem (5.1) which is concave. This should not present any problem since most of the "convex" results could be easily translated into "concave" and later we will do so without explicitly stating it.

6.1 Partial Proximal Point Method

The methods and ideas appearing in this section are a slight variation of the ones introduced by D. Bertsekas and P. Tseng in [7]. The reader is referred to [7] for an in-depth exposition of proximal-point methods.

6.1.1 Partial Moreau-Yoshida Regularization

Consider a convex function $f : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ that is proper, and lower semicontinuous.

For a fixed number $\varrho > 0$, we define the function $f_\varrho : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ by

$$f_\varrho(v, w) \triangleq \min_{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m} \left\{ \frac{\varrho}{2}\|y - w\|^2 + f(x, y) \right\}. \tag{6.1}$$

The function f_ϱ is called the *partial Moreau-Yosida regularization* of f. Since $f(x, y)$ is convex and lower semicontinuous the function

$$F(y) \triangleq \inf_{x\in\mathbb{R}^n} f(x, y)$$

is also convex and lower semicontinuous.

Unfortunately the properties of f do not imply that F is a proper function. For example the function $f(x, y) = x$ satisfies all the properties stated above but $F(y) = \inf_{x\in\mathbb{R}^m} x$ is not proper. A proper function $f : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ is *x-bounded* if for every bounded $Y \subset \mathbb{R}^m$ the set $X := \{(x, y) \in \mathbb{R}^n \times Y \mid f(x, y) \in \mathbb{R}\}$ is bounded. Notice that if f is x-bounded then the corresponding function F is proper. From now on we will assume that f is x-bounded.

Let $\overline{f}_\varrho : \mathbb{R}^m \to \overline{\mathbb{R}}$ be defined by

$$\overline{f}_\varrho(w) \triangleq \min_{y\in\mathbb{R}^m} \left\{ \frac{\varrho}{2}\|y - w\|^2 + F(y) \right\}. \tag{6.2}$$

Then $\overline{f}_\varrho$ is the Moreau-Yoshida regularization of F and it is well known (see [35, Lemma 7.10]) that $\overline{f}_\varrho$ is a real-valued, convex and continuously differentiable function with $\nabla\overline{f}_\varrho(w) = \varrho\,(w - y_\varrho(w))$, where $y_\varrho(w)$ is the solution of (6.2). It is not difficult to see that from the properties of f it follows that $f_\varrho(v, w) = \overline{f}_\varrho(w)$ for all $(v, w) \in \mathbb{R}^n \times \mathbb{R}^m$. Therefore we can conclude the following about the partial Moreau-Yoshida regularization of x-bounded functions:

Theorem 8. *For every $\varrho > 0$, the function f_ϱ is real-valued, convex and continuously differentiable with $\nabla f_\varrho(v, w) = [0, \varrho\,(w - y_\varrho(w))]$, where $(x_\varrho(v), y_\varrho(w))$ is any solution of* (6.1).

6.1.2 The Partial Proximal Point Method

Let us consider the convex optimization problem

$$\underset{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m}{\text{minimize}} \quad f(x,y), \tag{6.3}$$

in which $f : \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ is convex, proper, lower semicontinuous and x-bounded. Using the partial Moreau-Yoshida regularization of f, we construct the following iterative process. At iteration k, given the current approximation (v^k, w^k) to the solution of (6.3), we find a point $(x^k, y^k) = (x_\varrho(v^k), y_\varrho(w^k))$, which is a solution of the problem

$$\underset{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m}{\text{minimize}} \quad \frac{\varrho}{2}\|y - w^k\|^2 + f(x,y). \tag{6.4}$$

The next approximation is defined according to the formula:

$$\left(v^{k+1}, w^{k+1}\right) = \left(x_\varrho(v^k), y_\varrho(w^k)\right), \ k = 1, 2, \ldots \tag{6.5}$$

The iterative method (6.5) is called the *partial proximal point method*, see [7]. Although we will not directly apply this method, it is of theoretical importance and a natural progression in the development of the partial bundle method in the following section.

Let us recall that it follows from Theorem 8 that if f is x-bounded then problem (6.4) has a solution. Thus the partial proximal point method is well defined. Since $f_\varrho(v^k, w^k) \le f(v^k, w^k)$ by construction, we have $f(v^{k+1}, w^{k+1}) \le f(v^k, w^k)$, $k = 1, 2 \ldots$ Actually, the progress made at each iteration can be estimated:

Lemma 9. *Assume that there exists* $(\tilde{x}, \tilde{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ *such that* $f(\tilde{x}, \tilde{y}) < f(v, w)$. *Then if* $\tilde{y} = w$,

$$f_\varrho(v, w) \le f(v, w) - \left(f(v, w) - f(\tilde{x}, \tilde{y})\right),$$

else

$$f_\varrho(v, w) \le f(v, w) - \varrho\|\tilde{y} - w\|^2 \varphi\left(\frac{f(v, w) - f(\tilde{x}, \tilde{y})}{\varrho\|\tilde{y} - w\|^2}\right),$$

where

$$\varphi(\tau) = \begin{cases} 0 & \text{if } \tau < 0, \\ \tau^2 & \text{if } 0 \le \tau \le 1, \\ -1 + 2\tau & \text{if } \tau > 1. \end{cases}$$

We conclude that in any case $f_\varrho(v, w) < f(v, w)$.

Proof. Consider the segment containing points

$$(x, y) = (v, w) + t\left((\tilde{x}, \tilde{y}) - (v, w)\right),$$

where $0 \le t \le 1$. Restricting the minimization in (6.4) to these (x, y)'s provides an upper bound for the optimal value:

$$\begin{aligned} f_\varrho(v, w) &\le \min_{0 \le t \le 1} \left[f\left((1-t)(v, w) + t(\tilde{x}, \tilde{y})\right) + \frac{\varrho t^2}{2} \|\tilde{y} - w\|^2 \right] \\ &\le f(v, w) + \min_{0 \le t \le 1} \left[t\left(f(\tilde{x}, \tilde{y}) - f(v, w)\right) + \frac{\varrho t^2}{2} \|\tilde{y} - w\|^2 \right]. \end{aligned} \tag{6.6}$$

In the last estimate we also used the convexity of f.

If $\tilde{y} = w$ then (6.6) implies that $f_\varrho(v, w) \le f(v, w) - (f(v, w) - f(\tilde{x}, \tilde{y}))$. Else, $\tilde{y} \ne w$ and the value of t that minimizes (6.6) is equal to

$$\hat{t} = \min\left(1, \frac{f(v, w) - f(\tilde{x}, \tilde{y})}{\varrho \|\tilde{y} - w\|^2}\right).$$

Our assertion follows now from a straightforward calculation. □

At the solution $(x_\varrho(w), y_\varrho(w))$ of problem (6.4), Lemma 9 shows that $f\left(x_\varrho(w), y_\varrho(w)\right) \le f_\varrho(v, w) < f(v, w)$. Therefore problem (6.4) will always find a better point if exists. Consequently, $(x, y) = (v, w)$ is the minimizer in (6.4) if and only if (v, w) is a minimizer of f.

We say that a sequence $\left\{(x^k, y^k)\right\} \subset \mathbb{R}^n \times \mathbb{R}^m$ *approximates* an optimal solution (x^*, y^*) of (6.3) if $\lim_{k \to \infty} f(x^k, y^k) = f(x^*, y^*)$. In fact, the partial proximal point method must approximate an optimal solution, if an optimal solution exists.

Theorem 10. *Assume that problem (6.3) has an optimal solution. Then the following holds.*

1. *The sequence* $\left\{(v^k, w^k)\right\}$ *generated by the partial proximal point method approximates and optimal solution of (6.3).*

2. *The sequence* $\left\{w^k\right\}$ *converges to a point* $\tilde{y}$ *such that there is an optimal solution of (6.3) of the form* $(\tilde{x}, \tilde{y})$.

Proof. Let $(x^*, y^*) \in \mathbb{R}^n \times \mathbb{R}^m$ be an optimal solution. We have the identity

$$\|w^{k+1} - y^*\|^2 = \|w^k - y^*\|^2 + 2\left\langle w^{k+1} - w^k, w^{k+1} - y^* \right\rangle - \|w^{k+1} - w^k\|^2. \tag{6.7}$$

Theorem (8) implies that:

$$\left[0, \varrho\left(w^k - w^{k+1}\right)\right] \in \partial f(v^{k+1}, w^{k+1}). \tag{6.8}$$

By the definition of the subgradient,

$$f(x^*, y^*) \geq f(v^{k+1}, w^{k+1}) + \varrho\left\langle w^{k+1} - w^k, w^{k+1} - y^* \right\rangle. \tag{6.9}$$

Using this inequality in (6.7) (and skipping the last term) we obtain

$$\|w^{k+1} - y^*\|^2 \leq \|w^k - y^*\|^2 - \frac{2}{\varrho}\left[f(v^{k+1}, w^{k+1}) - f(x^*, y^*)\right] \tag{6.10}$$

Several conclusions follow from this estimate. First, summing up (6.10) from $k = 1$ to ∞, we get

$$\sum_{k=2}^{\infty}\left(f(v^k, w^k) - f(x^*, y^*)\right) \leq \frac{\varrho}{2}\|w^1 - y^*\|^2,$$

so $f(v^k, w^k) \to f(x^*, y^*)$ as $k \to \infty$. Therefore the sequence $\left\{(v^k, w^k)\right\}$ approximates and optimal solution of (6.3).

Second, the sequence $\left\{w^k\right\}$ is bounded and so it has accumulation points. Similarly, the x-boundedness of f implies that $\left\{v^k\right\}$ also has accumulation points. Consequently the lower semicontinuity of f implies that for every accumulation point $(\tilde{x}, \tilde{y})$

of $\{(v^k, w^k)\}$ we have $f(\tilde{x}, \tilde{y}) = f(x^*, y^*)$. We choose one such $(\tilde{x}, \tilde{y})$, substitute it for (x^*, y^*) in (6.10), and conclude that the sequence $\{w^k\}$ is convergent to $\tilde{y}$. □

6.2 The Partial Bundle Method

6.2.1 The Method

We consider the problem

$$\underset{(x,y)\in A}{\text{minimize}}\, f(x, y), \tag{6.11}$$

in which the set $A \subseteq \mathbb{R}^n \times \mathbb{R}^m$ is closed convex and the function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is convex, proper, and lower semicontinuous. We assume that the set A is x-bounded in the sense that for every bounded subset $Y \subset \mathbb{R}^m$ the intersection $A \cap (\mathbb{R}^n \times Y)$ is bounded.

We define the following *regularized master problem*:

$$\underset{(x,y)\in A}{\text{minimize}}\, \frac{\varrho}{2}\|y - w^k\|^2 + f^k(x, y), \tag{6.12}$$

where the model f^k is defined by:

$$f^k(x, y) \triangleq \max_{j\in J_k} \left[f(x^j, y^j) + \left\langle g^j, (x, y) - (x^j, y^j)\right\rangle\right], \tag{6.13}$$

with $g^j \in \partial f(x^j, y^j)$, $j \in J_k$. The set J_k is a subset of $\{1, \ldots, k\}$ determined by a procedure for selecting cuts. At this moment we may think of J_k as being equal to $\{1, \ldots, k\}$.

In the proximal term $(\varrho/2)\|y - w^k\|^2$, where $\varrho > 0$, the center (v^k, w^k) is updated depending on the relations between the value of $f(x^{k+1}, y^{k+1})$ at the master's solution, (x^{k+1}, y^{k+1}), and its prediction provided be the current model, $f^k(x^{k+1}, y^{k+1})$. If these values are equal or close, we set $(v^{k+1}, w^{k+1}) := (x^{k+1}, y^{k+1})$ (*descent step*); otherwise $(v^{k+1}, w^{k+1}) := (v^k, w^k)$ (*null step*). In any case, the collection of cuts is updated, and the iteration continues.

The regularized master problem can be equivalently written as a problem with a

quadratic objective function and linear constraints:

$$\begin{aligned} \text{minimize} \quad & z + \frac{\varrho}{2}\|y - w^k\|^2 \\ \text{subject to} \quad & z \geq f(x^j, y^j) + \left\langle g^j, (x,y) - (x^j, y^j) \right\rangle, \ j \in J_k, \\ & (x,y) \in A. \end{aligned} \tag{6.14}$$

If the set A is a convex polyhedron, the master problem can be readily solved by specialized techniques, enjoying the finite termination property.

Let us observe that problem (6.14) satisfies Slater's constraint qualification condition. Indeed, for every $(x_S, y_S) \in A$ we can choose z_s so large that all constraints are satisfied as strict inequalities. Therefore the optimal solution of the master problem satisfies the necessary and sufficient conditions of optimality with Lagrange multipliers (see [35, Thm. 3.34]). We denote by $\lambda_j^k \in J_k$, the Lagrange multipliers associated with the constraints of problem (6.14).

The detailed *partial bundle method* is stated below. The parameter $\gamma \in (0,1)$ is a fixed constant used to compare the observed improvement in the objective value to the predicted improvement.

Step 0 Set $k := 1$, $J_0 := \emptyset$, and $z^1 := -\infty$.

Step 1 Calculate $f(x^k, y^k)$ and $g^k \in \partial f(x^k, y^k)$. If $f(x^k, y^k) > z^k$ then set $J_k := J_{k-1} \cup \{k\}$; otherwise set $J_k := J_{k-1}$.

Step 2 If $k = 1$ or if

$$f(x^k, y^k) \leq (1-\gamma) f(v^{k-1}, w^{k-1}) + \gamma f^{k-1}(x^k, y^k),$$

then set $(v^k, w^k) := (x^k, y^k)$; otherwise Step 2 is a null step and we set $(v^k, w^k) := (v^{k-1}, w^{k-1})$.

Step 3 Solve the master problem (6.14). Denote by (x^{k+1}, y^{k+1}) and z^{k+1} its solution and set $f^k(x^{k+1}, y^{k+1}) := z^{k+1}$.

Step 4 If $f^k(x^{k+1}, y^{k+1}) = f(v^k, w^k)$ then stop (the point (v^k, w^k) is an optimal solution); otherwise continue.

Step 5 If Step 2 was a null step then go to Step 6. Else (Step 2 was a descent step) remove from the set of cuts J_{k-1} some (or all) cuts whose Lagrange multipliers λ_j^k at the solution of (6.14) are 0.

Step 6 Increase k by one, and go to Step 1.

6.2.2 Convergence

First we prove that if the algorithm gets stuck at a (v, w)-center then it will approximate an optimal solution.

Lemma 11. *Let f^* be an optimal solution to* (6.11) *and suppose that the sequence, $\{(x^k, y^k)\}$, obtained by the partial bundle method consists of only null steps from iteration t on. Then*

$$\lim_{k\to\infty} f^{k-1}(x^k, y^k) = f^* = \lim_{k\to\infty} f(x^k, y^k).$$

Proof. For any $\epsilon > 0$, let

$$\mathcal{K}_\epsilon := \left\{k : k > t \text{ and } f^{k-1}(x^k, y^k) + \epsilon < f(x^k, y^k)\right\}$$

and let $k_1, k_2 \in \mathcal{K}_\epsilon$ with $t < k_1 < k_2$.

Since we only have null steps we get that for every $k > t$, $(v^k, w^k) = (x^t, y^t)$ and the cutting plane generated at k will remain on the master problem from k on. This implies that the sequence $\{f^{k-1}(x^k, y^k)\}$ is non-decreasing from $t + 1$ on. Also, since the cutting plane generated at (x^{k_1}, y^{k_1}) will remain in the master problem at iteration $k_2 - 1$, we get:

$$f(x^{k_1}, y^{k_1}) + \left\langle g^{k_1}, (x^{k_2}, y^{k_2}) - (x^{k_1}, y^{k_1})\right\rangle \leq f^{k_2-1}(x^{k_2}, y^{k_2}).$$

On the other hand, $\epsilon < f(x^{k_2}, y^{k_2}) - f^{k_2-1}(x^{k_2}, y^{k_2})$ which combined with the last

inequality yields

$$\epsilon < f(x^{k_2}, y^{k_2}) - f(x^{k_1}, y^{k_1}) + \left\langle g^{k_1}, (x^{k_1}, y^{k_1}) - (x^{k_2}, y^{k_2}) \right\rangle.$$

Since all the steps made are null, the points y^k, with $k > t$, are contained in a bounded neighborhood of $w^k = y^t$. This and the x-boundedness of f guarantee us that $B := \operatorname{Conv}\left\{(x^j, y^j) \;\middle|\; j \in \mathcal{K}_\epsilon\right\}$ is bounded. The function f is subdifferentiable in $\overline{B}$, so there exists a constant C such that $f(x_1, y_1) - f(x_2, y_2) \leq C\|(x_1, y_1) - (x_2, y_2)\|$, for all $x_1, x_2 \in \overline{B}$. Subgradients on bounded sets are bounded, and thus we can choose C large enough so that $\|g^j\| \leq C$, for all $j \in \mathcal{K}_\epsilon$. It follows from the last displayed inequality that

$$\epsilon < 2C\|(x^{k_1}, y^{k_1}) - (x^{k_2}, y^{k_2})\| \;\text{ for all }\; k_1, k_2 \in \mathcal{K}_\epsilon,\; k_1 \neq k_2.$$

As the set $\overline{B}$ is compact, there can exist only finitely many points in $\mathcal{K}_\epsilon \subset \overline{B}$ having distance at least $\epsilon/(2C)$ from each other. Thus the last inequality implies that the set $\mathcal{K}_\epsilon$ is finite for each $\epsilon > 0$. This means that

$$\lim_{k\to\infty} f(x^k) - f^{k-1}(x^k) = 0. \tag{6.15}$$

By construction the sequences $\left\{f^{k-1}(x^k)\right\}$ and $\left\{f(x^k)\right\}$ satisfy the relation

$$f^{k-1}(x^k) \leq f^* \leq f(x^k), \text{ for every } k \in \mathbb{N}.$$

Therefore the eventual monotonicity of $\left\{f^{k-1}(x^k)\right\}$ and (6.15) imply that

$$\lim_{k\to\infty} f^{k-1}(x^k, y^k) = f^* = \lim_{k\to\infty} f(x^k, y^k).$$

□

Next we prove another intermediate step towards convergence.

Lemma 12. *Assume that problem* (6.11) *has an optimal solution and suppose that the*

sequence $\{(x^k, y^k)\}$ *obtained by the partial bundle method has infinitely many descent steps. Then the following holds.*

1. *The sequence* $\{(v^k, w^k)\}$ *approximates an optimal solution of* (6.11).
2. *The sequence* $\{w^k\}$ *converges to a point* $\tilde{y}$ *such that there is an optimal solution of* (6.11) *of the form* $(\tilde{x}, \tilde{y})$.

Proof. Let us denote by $\mathcal{K}$ the set of iterations at which descent steps occur. If $(v^{k+1}, w^{k+1}) = (x^{k+1}, y^{k+1})$ is the optimal solution of the master problem (6.12), we have the necessary condition of optimality

$$0 \in \partial \left[\frac{\varrho}{2} \|y - w^k\|^2 + f^k(x, y) \right] + N_A(x, y) \quad \text{at} \quad (x, y) = (v^{k+1}, w^{k+1}).$$

Hence

$$-\left[0, \varrho(w^{k+1} - w^k)\right] \in \partial f^k(v^{k+1}, w^{k+1}) + N_A(v^{k+1}, w^{k+1}).$$

Let (x^*, y^*) be an optimal solution of (6.11). By the subgradient inequality for f^k we get (for some $h \in N_A(v^{k+1}, w^{k+1})$) the estimate

$$\begin{aligned} f^k(x^*, y^*) &\geq f^k(v^{k+1}, w^{k+1}) - \left\langle \left[0, \varrho\left(w^{k+1} - w^k\right)\right], (x^*, y^*) - (v^{k+1}, w^{k+1}) \right\rangle \\ &\quad - \left\langle h, (x^*, y^*) - (v^{k+1}, w^{k+1}) \right\rangle \\ &\geq f^k(v^{k+1}, w^{k+1}) - \varrho \left\langle w^{k+1} - w^k, y^* - w^{k+1} \right\rangle. \end{aligned} \tag{6.16}$$

Suppose a descent step from (v^k, w^k) to (v^{k+1}, w^{k+1}) occurs, so the test of Step 2 is satisfied (for $k+1$):

$$f(v^{k+1}, w^{k+1}) \leq (1 - \gamma) f(v^k, w^k) + \gamma f^k(v^{k+1}, w^{k+1}).$$

After elementary manipulations we can rewrite it as

$$f^k(v^{k+1}, w^{k+1}) \geq f(v^{k+1}, w^{k+1}) - \frac{1 - \gamma}{\gamma} \left[f(v^k, w^k) - f(v^{k+1}, w^{k+1}) \right]. \tag{6.17}$$

Combining the last inequality with (6.16) and using the relation $f(x^*, y^*) \geq f^k(x^*, y^*)$

we obtain

$$f(x^*, y^*) \;\geq\; f(v^{k+1}, w^{k+1}) \;+\; \frac{1-\gamma}{\gamma}\left[f(v^{k+1}, w^{k+1}) - f(v^k, w^k)\right] \\ -\; \varrho\left\langle w^{k+1} - w^k,\, y^* - w^{k+1}\right\rangle.$$

This can be substituted into the identity:

$$\|w^{k+1} - y^*\|^2 = \|w^k - y^*\|^2 + 2\left\langle w^{k+1} - w^k, w^{k+1} - y^*\right\rangle - \|w^{k+1} - w^k\|^2.$$

After skipping the last term we get

$$\begin{aligned}\|w^{k+1} - y^*\|^2 \leq\; & \|w^k - y^*\|^2 - \frac{\varrho}{2}\left[f(v^{k+1}, w^{k+1}) - f(x^*, y^*)\right] \\ & + \frac{2(1-\gamma)}{\gamma\varrho}\left[f(v^k, w^k) - f(v^{k+1}, w^{k+1})\right] \quad \text{for all} \quad k \in \mathcal{K}.\end{aligned} \tag{6.18}$$

The series $\sum_{k=1}^{\infty}[f(v^k, w^k) - f(v^{k+1}, w^{k+1})]$ is convergent, because $\{f(v^k, w^k)\}$ is non-increasing and bounded from below by $f(x^*, y^*)$. Therefore we obtain from (6.18) that the distance $\|w^{k+1} - y^*\|$ is uniformly bounded, and so $\{w^k\}$ must have accumulation points. This and the x-boundedness of f imply that the sequence $\{v^k, w^k\}$ has accumulation points.

Summing up (6.18) for $k \in \mathcal{K}$ we get

$$\sum_{k\in\mathcal{K}}\Big(f(v^{k+1}, w^{k+1}) - f(x^*, y^*)\Big) \leq \frac{\varrho}{2}\|w^1 - y^*\|^2 + \frac{1-\gamma}{\gamma}\left[f(v^1, w^1) - \lim_{k\to\infty} f(v^k, w^k)\right],$$

so $f(v^{k+1}, w^{k+1}) \to f(x^*, y^*)$, $k \in \mathcal{K}$. Consequently, at every accumulation point $(\tilde{x}, \tilde{y})$ of $\{(v^k, w^k)\}$ one has $f(\tilde{x}, \tilde{y}) = f(x^*, y^*)$. Since $(\tilde{x}, \tilde{y})$ is optimal, we can substitute it for (x^*, y^*) in (6.18). Skipping the negative term we get

$$\|w^{k+1} - \tilde{y}\|^2 \leq \|w^k - \tilde{y}\|^2 + \frac{2(1-\gamma)}{\gamma\varrho}\left[f(v^k, w^k) - f(v^{k+1}, w^{k+1})\right].$$

It is true not only for $k \in \mathcal{K}$ but for all k, because at $k \notin \mathcal{K}$ we have a trivial equality

here. Summing these inequalities from $k = l$ to $k = q > l$ we get

$$\|w^{q+1} - \tilde{y}\|^2 \leq \|w^l - \tilde{y}\|^2 + \frac{2(1-\gamma)}{\gamma\varrho}\left[f(v^l, w^l) - f(v^{q+1}, w^{q+1})\right].$$

Since $\tilde{y}$ is an accumulation point, for $\epsilon > 0$ we can find l such that $\|w^l - \tilde{y}\| \leq \epsilon$. Also, if l is large enough, $f(v^l, w^l) - f(v^{q+1}, w^{q+1}) \leq \epsilon$ for all $q > l$, because $\{f(v^k, w^k)\}$ is convergent. Then $\|w^{q+1} - \tilde{y}\|^2 \leq \epsilon^2 + 2\epsilon(1-\gamma)/(\gamma\varrho)$ for all $q > l$, so the sequence $\{w^k\}$ is convergent to $\tilde{y}$. □

Now we are ready to prove convergence of the partial bundle method.

Theorem 13. *Assume that problem* (6.11) *has an optimal solution, f^*, and let $\{(x^k, y^k)\}$ be the sequence obtained by the partial bundle method. Then*

$$\liminf_{k\to\infty} f(x^k, y^k) = f^*.$$

Proof. If there are only finitely many descent steps then Lemma 11 gives the desired result. Thus we assume that the number of descent steps is infinite and by Lemma 12, $\lim_{k\to\infty} f(v^k, w^k) = f^*$. Clearly, the sequence $\{f(v^k, w^k)\}$ is an infinite subsequence of $\{f(x^k, y^k)\}$. Then, since $f(x^k, y^k) \geq f^*$ for every k, we obtain that $\liminf_{k\to\infty} f(x^k, y^k) = f^*$. □

Chapter 7

Scenario Decomposition of Multistage Risk-Averse Problems and Methods

7.1 A Multistage Risk-Averse Problem

Let $(\Omega, \mathcal{F}, P)$ be a probability space with a sigma algebra $\mathcal{F}$ and probability measure P. Consider a filtration $\{\emptyset, \Omega\} = \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_T = \mathcal{F}$. A random vector $x = (x_1, \dots, x_T)$, where each x_t has values in $\mathbb{R}^{n_t}$, $t = 1, \dots, T$, is called a *policy*. If each x_t is $\mathcal{F}_t$-measurable, $t = 1, \dots, T$, a policy x is called *implementable*. A policy x is called *feasible*, if it satisfies the following conditions:

$$
\begin{array}{lllllll}
A_1 x_1 & & & & & = & b_1, \\
B_2 x_1 & + & A_2 x_2 & & & = & b_2, \\
& & B_3 x_2 & + & A_3 x_3 & = & b_3, \\
\multicolumn{7}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\
& & & & B_T x_{T-1} \;+\; A_T x_T & = & b_T, \\
x_1 \in X_1, & & x_2 \in X_2, & & x_3 \in X_3, \quad \dots \quad x_T \in X_T. & &
\end{array}
\tag{7.1}
$$

In these equations, for every $t = 1, \dots, T$, the matrices A_t of dimensions $m_t \times n_t$, the matrices B_t of dimensions $m_t \times n_{t-1}$, and the vectors b_t of dimensions m_t are $\mathcal{F}_t$-measurable data. Each set X_t is a random convex and closed polyhedron which is measurable with respect to $\mathcal{F}_t$ (in the sense of measurability of multifunctions; see [6]).

Suppose c_t, $t = 1, \dots, T$, is an adapted sequence of random cost vectors, that is, each c_t is $\mathcal{F}_t$-measurable. A policy x results in a cost sequence

$$
Z_t = \langle c_t, x_t \rangle, \quad t = 1, \dots, T. \tag{7.2}
$$

Our intention is to formulate and analyze a risk-averse multistage stochastic programming problem, to minimize a dynamic measure of risk, $\varrho(Z_1, \ldots, Z_T)$, over all implementable and feasible policies x. In order to define the functional $\varrho(\cdot)$, we recall some basic concepts of the theory of dynamic measures of risk. We follow the development given in [38, 37, 39, 41].

Consider vector spaces $\mathcal{Z}_t$ of $\mathcal{F}_t$-measurable random outcomes. As $\mathcal{F}_1$ is trivial, $\mathcal{Z}_1 = \mathbb{R}$. Suppose we observe a random sequence Z_t, $t = 1, \ldots, T$, adapted to the filtration $\{\mathcal{F}_t\}$. Its risk can be evaluated by using the following *dynamic coherent measure of risk*

$$\varrho_{1,T}(Z_1, Z_2, \ldots, Z_T) = Z_1 + \rho_1\Big(Z_2 + \rho_2\big(Z_3 + \cdots + \rho_{T-1}(Z_T)\ldots\big)\Big), \tag{7.3}$$

where each $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ is a coherent conditional measure of risk. The structure (7.3) was postulated in [37] and derived in [31] from abstract principles of monotonicity and time consistency of dynamic risk measures.

Our problem is to minimize (7.3) with each Z_t given by (7.2), over all implementable and feasible policies x. In order to complete the problem formulation, we need to be more specific about the vector spaces $\mathcal{Z}_t$, the vector spaces of random vectors in which the components x_t of the policy live, as well as integrability conditions on the problem data A_t, B_t, b_t and c_t, so that $Z_t \in \mathcal{Z}_t$ for all $t = 1, \ldots, T$. In this work, we assume that all sigma-algebras are finite and all vector spaces are finite-dimensional. We discuss it in the next section.

7.2 Scenario Trees and Recursive Risk Evaluation

In the finite distribution case, possible realizations of data form a *scenario tree*. It has nodes organized in levels which correspond to stages $1, \ldots, T$. At level $t = 1$ we have only one *root node* $\nu = 1$. Nodes at levels $t = 2, \ldots, T$ correspond to elementary events in $\mathcal{F}_t$. Each node ν at level $t = 2, \ldots, T$ is connected to a unique node $a(\nu)$ at level $t - 1$, called the *ancestor node*, which corresponds to the elementary event in $\mathcal{F}_{t-1}$ that contains the event associated with ν. Thus, every node ν at levels $t = 1, \ldots, T - 1$ is

connected to a set $C(\nu)$ of nodes at level $t+1$, called *children nodes*, which correspond to elementary events in $\mathcal{F}_{t+1}$ included in the event corresponding to ν. We denote by Ω_t the set of all nodes at stage $t = 1, \ldots, T$. We have the relations $\Omega_{t+1} = \cup_{\nu \in \Omega_t} C(\nu)$ and $C(\nu) = \{\eta \in \Omega_{t+1} : \nu = a(\eta)\}$. The sets $C(\nu)$ are disjoint, i.e., $C(\nu) \cap C(\nu') = \emptyset$ if $\nu \neq \nu'$. A *scenario* is a path s from the root to a node at the last stage T. By construction, there is one-to-one correspondence between the scenarios and the set $\Omega_T = \Omega$. Let $\mathcal{S}(\nu)$ be the set of scenarios passing through node ν. These sets satisfy the recursive relation:

$$\mathcal{S}(\nu) = \{\nu\}, \quad \nu \in \Omega_T,$$
$$\mathcal{S}(\nu) = \bigcup_{\eta \in C(\nu)} \mathcal{S}(\eta), \quad \nu \in \Omega_t, \quad t = T-1, \ldots, 1.$$

As the nodes of the tree correspond to events defining nested partitions of Ω, the measure P can be specified by conditional probabilities:

$$p_{\nu\eta} = P[\eta|\nu], \quad \nu \in \Omega_t, \quad \eta \in C(\nu), \quad t = 1, \ldots, T-1.$$

Every node ν at level t has a *history*: the path $(\nu_1, \ldots, \nu_{t-1}, \nu)$ from the root to ν. The probability of the node ν is thus the product of the corresponding conditional probabilities

$$p_\nu = p_{\nu_1\nu_2} p_{\nu_2\nu_3} \cdots p_{\nu_{t-1}\nu}. \tag{7.4}$$

In particular, when $t = T$, formula (7.4) describes the probability of a scenario $\nu \in \Omega_T$.

For every node $\nu \in \Omega_t$, an $\mathcal{F}_t$-measurable random variable Z has identical values on all scenarios $s \in \mathcal{S}(\nu)$. It can, therefore, be equivalently represented as a function of a node at level Ω_t, which we write Z^{Ω_t}.

Consider a conditional measure of risk $\rho_t(\cdot)$. Its value is $\mathcal{F}_t$-measurable, and thus we can consider its representation as a function of a node at level t. It follows from [37, Thm. 3.2] that for every $\mathcal{F}_t$-measurable nonnegative function Γ a stronger version of (A4) holds:

$$\Gamma \rho_t(Z_{t+1}) = \rho_t(\Gamma Z_{t+1}).$$

Let $\nu \in \Omega_t$, and let $\mathbb{1}_\nu$ be the characteristic function of the event ν. Setting $\Gamma = \mathbb{1}_\nu$ in the last equation, for all $Z_{t+1}, W_{t+1} \in \mathcal{Z}_{t+1}$ we obtain

$$\mathbb{1}_\nu \rho_t(\mathbb{1}_\nu Z_{t+1} + (1 - \mathbb{1}_\nu)W_{t+1}) = \rho_t(\mathbb{1}_\nu Z_{t+1}) = \mathbb{1}_\nu \rho_t(\mathbb{1}_\nu Z_{t+1}).$$

In the last equation we multiplied both sides by $\mathbb{1}_\nu$. We see that W_{t+1} plays no role here. The value of $\rho_t(Z_{t+1})$ at elementary events associated with node ν depends only on the values of $Z_{t+1}^{\Omega_{t+1}}$ at nodes $\eta \in C(\nu)$. We denote the vector of these values by $Z_{t+1}^{C(\nu)}$, and we write the conditional risk measure equivalently as $\rho_t^\nu\big(Z_{t+1}^{C(\nu)}\big)$.

Let us define the random variables

$$V_t = \rho_t\Big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{T-1}(Z_T)\dots\big)\Big), \quad t = 1, \dots, T. \tag{7.5}$$

They are $\mathcal{F}_t$-measurable, and thus we only need to consider their values V_t^ν associated with scenarios $s \in \mathcal{S}(\nu)$. It follows that the value of the measure of risk (7.3) can be written on the scenario tree in a recursive manner:

$$\varrho_{1,T}(Z_1, Z_2, \dots, Z_T) = Z_1 + V_1^1, \tag{7.6}$$

$$V_t^\nu = \rho_t^\nu\big(Z_{t+1}^{C(\nu)} + V_{t+1}^{C(\nu)}\big), \quad \nu \in \Omega_t, \quad t = 1, \dots, T. \tag{7.7}$$

7.3 Nonanticipativity Constraints

A standard approach to multistage stochastic programming is based on *scenario decomposition*. With every scenario s in the tree, we associate a sequence of decision vectors

$$x^s = (x_1^s, \dots, x_T^s), \quad s \in \Omega.$$

Such a collection of sequences forms a policy which is not necessarily implementable, unless it satisfies a certain linear equation, called the *nonanticipativity constraint*. It requires that the process x be adapted to the filtration $\{\mathcal{F}_t\}$. Abstractly, we can write

$$x_t = \mathbb{E}\big[x_t | \mathcal{F}_t\big], \quad t = 1, \dots, T - 1. \tag{7.8}$$

For the scenario model, the nonanticipativity constraint can be written as a system of linear equations at the nodes of the tree. For every node ν at level $t = 1, \ldots, T-1$ the values x_t^s should be identical for all $s \in \mathcal{S}(\nu)$. Direct specification of (7.8) yields

$$x_t^s = \mathbb{E}\big[x_t | \mathcal{S}(\nu)\big] = \frac{\sum_{\omega \in \mathcal{S}(\nu)} p_\omega x_t^\omega}{\sum_{\omega \in \mathcal{S}(\nu)} p_\omega}, \quad s \in \mathcal{S}(\nu), \quad \nu \in \Omega_t, \quad t = 1, \ldots, T-1. \tag{7.9}$$

Other constraints of problem (1)-(2)-(3) decompose by scenario:

$$x \in \mathcal{X} = \mathcal{X}^1 \times \cdots \times \mathcal{X}^{|\Omega|}, \tag{7.10}$$

where for each $s \in \Omega$ we have

$$\mathcal{X}^s = \big\{ x \in X_1^s \times \cdots \times X_T^s : B_t^s x_{t-1}^s + A_t^s x_t^s = b_t^s,\ t = 1, \ldots, T \big\}. \tag{7.11}$$

In (7.11) the symbols A_t^s, B_t^s, b_t^s, and X_t^s denote realizations of problem data at stage t in scenario s, and the term $B_t^s x_{t-1}^s$ is omitted for $t = 1$.

In risk-neutral multistage stochastic programming, we can write the corresponding optimization problem:

$$\begin{gathered} \min \sum_{s \in \Omega} p^s \sum_{t=1}^{T} \langle c_t^s, x_t^s \rangle \\ \text{s.t. (7.9) and (7.11).} \end{gathered} \tag{7.12}$$

Then, Lagrange multipliers λ_t^s are associated with the nonanticipativity constraints (7.9), and the following Lagrangian function is constructed:

$$L(x, \lambda) = \sum_{s \in \Omega} p^s \sum_{t=1}^{T} \langle c_t^s, x_t^s \rangle + \sum_{t=1}^{T-1} \sum_{\nu \in \Omega_t} \sum_{s \in \mathcal{S}(\nu)} p^s \big\langle \lambda_t^s, x_t^s - \mathbb{E}\big[x_t | \mathcal{S}(\nu)\big] \big\rangle. \tag{7.13}$$

The problem

$$\min_{x \in \mathcal{X}} L(x, \lambda)$$

decomposes into scenario subproblems, one for each $s \in \Omega$. We shall not go into these details here; the reader can find them in [41, Sec. 3.2.4]. The dual problem is to find the optimal values of Lagrange multipliers associated with (7.9). It can be solved

by nonsmooth optimization methods or by augmented Lagrangian methods. As the constraints (7.9) are redundant, we can restrict the multipliers to the subspace defined by the equations

$$\mathbb{E}[\lambda_t|\mathcal{F}_t] = 0, \quad t = 1, \ldots, T-1. \tag{7.14}$$

In the scenario tree case, these conditions translate into

$$\sum_{s \in \mathcal{S}(\nu)} p^s \lambda_t^s = 0, \quad \nu \in \Omega_t, \quad t = 1, \ldots, T-1. \tag{7.15}$$

Again, the reader is referred to [41, Ch. 3] for the details.

The difficulty with the scenario decomposition in the risk-averse setting is the definition and nonlinear character of the dynamic risk measure (7.3). If a policy x is not implementable, the sequence $\{Z_t\}$ is not adapted to the filtration $\{\mathcal{F}_t\}$ and formula (7.3) makes no sense, because of the definition of ρ_t as a function acting on $\mathcal{F}_{t+1}$-measurable random variables. We cannot just substitute the dynamic risk measure for the objective function in (7.12).

7.4 Transition Multikernels and Their Compositions

We first recall the dual representation of conditional measures of risk. Let $\mathcal{P}(C)$ denote the set of probability distributions on a set of nodes $C \subset \Omega_t$. By theorems (2) and (5), for every $t = 1, \ldots, T-1$ and every node $\nu \in \Omega_t$ there exists a convex closed set $\mathcal{A}_t(\nu) \subset \mathcal{P}(C(\nu))$ such that

$$\rho_t^\nu\big(Z_{t+1}^{C(\nu)}\big) = \max_{\mu \in \mathcal{A}_t(\nu)} \big\langle \mu, Z_{t+1}^{C(\nu)} \big\rangle. \tag{7.16}$$

In fact, $\mathcal{A}_t(\nu) = \partial \rho_t^\nu\big(0\big)$.

We shall call a set mapping $\mathcal{K} : \Omega_t \rightrightarrows \mathcal{P}(\Omega_{t+1})$ a *transition multikernel*. IWe call it *convex*, if for all $\nu \in \Omega_t$ the set $\mathcal{K}(\nu)$ is convex. We call it *closed*, if for all $\nu \in \Omega_t$ the set $\mathcal{K}(\nu)$ is closed. The transition multikernels $\mathcal{A}_t$ associated with the conditional risk measures $\rho_t(\cdot)$ are convex and closed, as subdifferentials of convex functions $\rho_t^\nu(\cdot)$ at 0,

$\nu \in \Omega_t$. They also satisfy the conditions

$$\mathcal{A}_t(\nu) \subset \mathcal{P}(C(\nu)), \quad \forall\, \nu \in \Omega_t. \tag{7.17}$$

For $t = 1$ there is only one node $\nu = 1 \in \Omega_1$, and thus $\mathcal{A}_1$ is simply a set probability distributions on Ω_2. If a kernel μ_t is a selection of $\mathcal{A}_t$, that is, $\mu_t(\nu) \in \mathcal{A}_t(\nu)$ for all $\nu \in \Omega_t$, we shall simply write $\mu_t \in \mathcal{A}_t$. The value of $\mu(\nu)$ at an node $\eta \in C(\nu)$ will be written as $\mu(\nu, \eta)$.

Compositions of transition multikernels are germane for our analysis. Let us start from a composition of a measure $q_t \in \mathcal{P}(\Omega_t)$ with a kernel $\mu_t \in \mathcal{A}_t$. It is a measure on Ω_{t+1} given by the following relations:

$$(\mu_t \circ q_t)(\eta) = q_t(a(\eta))\mu_t(a(\eta), \eta),\ \eta \in \Omega_{t+1}; \tag{7.18}$$

recall that $a(\eta)$ is the ancestor of η. If we have a set of probability distributions $\mathcal{Q}_t \subset \mathcal{P}(\Omega_t)$ and a transition multikernel $\mathcal{A}_t$ satisfying (7.17), we can define their composition $\mathcal{A}_t \circ Q_t$ as the following set of probability distributions on Ω_{t+1}:

$$\mathcal{A}_t \circ Q_t = \big\{\mu_t \circ q_t : q_t \in Q_t,\, \mu_t \in \mathcal{A}_t\big\}. \tag{7.19}$$

Lemma 14. *Suppose Q_t is a convex and compact set of probability measures on Ω_t and a transition multikernel $\mathcal{A}_t$ satisfies* (7.17) *and is convex and compact. Then the set $Q_{t+1} = \mathcal{A}_t \circ Q_t$ is convex and compact.*

Proof. To prove convexity, let $q^k_{t+1}(\eta) = q^k_t(a(\eta))\mu^k_t(\eta)$, with $\mu^k_t \in \mathcal{A}_t$, $q^k_t \in Q_t$, $k = 1, 2$, and consider their convex combination,

$$q_{t+1} = \alpha q^1_{t+1} + (1 - \alpha) q^2_{t+1}, \quad \alpha \in (0, 1).$$

Define $q_t = \alpha q^1_t + (1 - \alpha) q^2_t$. By the convexity of Q_t, we have $q_t \in Q_t$, and thus the set $\mathcal{A}_t \circ \{q_t\}$ is included in Q_{t+1}. To show that $q_{t+1} \in Q_{t+1}$, it is sufficient to prove that

$q_{t+1} \in \mathcal{A}_t \circ \{q_t\}$. This amounts to verifying for all $\eta \in \Omega_{t+1}$ the following relation:

$$\alpha q_t^1(a(\eta))\mu_t^1(\eta) + (1-\alpha)q_t^2(a(\eta))\mu_t^2(\eta) \in q_t(a(\eta))\mathcal{A}_t(a(\eta)). \tag{7.20}$$

Let $\eta \in \Omega_{t+1}$ and $\nu = a(\eta)$. Observe that $q_t^1(\nu) \geq 0$ and $q_t^2(\nu) \geq 0$. If $q_t(\nu) = 0$, we must have $q_t^1(\nu) = q_t^2(\nu) = 0$ and (7.20) is trivial. It remains to consider the case of $q_t(\nu) > 0$. Define

$$\beta(\nu) = \frac{\alpha q_t^1(\nu)}{q_t(\nu)}.$$

By the definition of q_t, $\beta(\nu) \in [0,1]$. The left hand side of (7.20) can be written as follows

$$\alpha q_t^1(\nu)\mu_t^1(\eta) + (1-\alpha)q_t^2(\nu)\mu_t^2(\eta) = q_t(\nu)\big(\beta(\nu)\mu_t^1(\eta) + (1-\beta(\nu))\mu_t^2(\eta)\big).$$

Due to the convexity of $\mathcal{A}_t$, the right hand side is an element of $\mathcal{A}_t \circ \{q_t\}$, which proves (7.20).

The compactness of Q_{t+1} follows from the compactness of Q_t and $\mathcal{A}_t$. □

We can now prove a useful dual representation of a dynamic measure of risk.

Theorem 15. *Suppose a dynamic risk measure $\varrho(\cdot)$ is given by* (7.3) *with conditional risk measures $\rho_t(\cdot)$ satisfying conditions* (A1)–(A4). *Then for every adapted sequence $Z_1, \ldots, Z_T$ we have the relation*

$$\varrho(Z_1, \ldots, Z_T) = \max_{q_T \in Q_T} \big\langle q_T, Z_1 + Z_2 + \cdots + Z_T \big\rangle, \tag{7.21}$$

where

$$Q_T = \mathcal{A}_{T-1} \circ \ldots \mathcal{A}_2 \circ \mathcal{A}_1 \tag{7.22}$$

is a convex and closed set of probability measures on Ω.

Proof. Recursive composition of transition multikernels μ_t yields a sequence of sets of measures:

$$Q_{t+1} = \mathcal{A}_t \circ Q_t, \quad t = 1, \ldots, T-1, \tag{7.23}$$

with $Q_1 = \{1\}$. Each Q_t is a set of probability measures on Ω_t. Lemma 14 implies that they are all convex and compact.

The multikernel representation (7.16) allows us to rewrite the definition of a dynamic risk measure (7.3) as follows:

$$\varrho(Z_1, \dots, Z_T) = Z_1 + \max_{\mu_1 \in \mathcal{A}_1} \Big(\big\langle \mu_1, Z_2^{\Omega_2} \big\rangle + \max_{\mu_2 \in \mathcal{A}_2} \big(\big\langle \mu_2 \circ \mu_1, Z_3^{\Omega_3} \big\rangle + \dots \\ \cdots + \max_{\mu_{T-1} \in \mathcal{A}_{T-1}} \big\langle \mu_{T-1} \circ \cdots \circ \mu_2 \circ \mu_1, Z_T \big\rangle \cdots \big) \Big). \quad (7.24)$$

All the maximum operations can be put at the beginning, and we obtain:

$$\varrho(Z_1, \dots, Z_T) = Z_1 + \max_{\substack{\mu_t \in \mathcal{A}_t \\ t=1,\dots,T-1}} \Big(\big\langle \mu_1, Z_2^{\Omega_2} \big\rangle + \big\langle \mu_2 \circ \mu_1, Z_3^{\Omega_3} \big\rangle + \dots \\ \cdots + \big\langle \mu_{T-1} \circ \cdots \circ \mu_2 \circ \mu_1, Z_T \big\rangle \Big). \quad (7.25)$$

Let $q_t = \mu_{t-1} \circ \cdots \circ \mu_2 \circ \mu_1$, $t = 2, \dots, T$. Each of them is an element of the corresponding set Q_t. Consider the product

$$\big\langle q_t, Z_t^{\Omega_t} \big\rangle = \sum_{\nu \in \Omega_t} q_t(\nu) Z_t^{\Omega_t}(\nu).$$

Suppose $\mu_t \in \mathcal{A}_t$ and $\nu \in \Omega_t$. Then $\mu_t(\nu)$ is a probability distribution on $C(\nu)$. Since Z_t is $\mathcal{F}_t$-measurable, $Z_t^{\Omega_{t+1}}$ has identical values on the nodes $\eta \in C(\nu)$. Therefore,

$$Z_t^{\Omega_t}(\nu) = \langle \mu_t(\nu), Z_t^{\Omega_{t+1}} \rangle.$$

Recalling the definition (7.18), we conclude that

$$\big\langle q_t, Z_t^{\Omega_t} \big\rangle = \big\langle \mu_t \circ q_t, Z_t^{\Omega_{t+1}} \big\rangle = \big\langle q_{t+1}, Z_t^{\Omega_{t+1}} \big\rangle.$$

Applying this relation recursively to all terms of (7.25), we obtain the identity

$$\begin{aligned} \varrho(Z_1, \ldots, Z_T) &= \max_{\substack{\mu_t \in \mathcal{A}_t \\ t=1,\ldots,T-1}} \big\langle \mu_{T-1} \circ \cdots \circ \mu_2 \circ \mu_1, Z_1 + Z_2 + \cdots + Z_T \big\rangle \\ &= \max_{q_T \in Q_T} \big\langle q_T, Z_1 + Z_2 + \cdots + Z_T \big\rangle, \end{aligned} \tag{7.26}$$

as postulated. □

7.5 Duality and Decomposition

An advantage of formula (7.21) is that its right hand side remains well-defined also for sequences $\{Z_t\}$, which are not adapted to the filtration $\{\mathcal{F}_t\}$. This allows for the development of the corresponding duality theory and decomposition.

Consider the extended problem formulation corresponding to to the risk-neutral formulation (7.12). The nonanticipativity constraints (7.9) can be compactly written as a system of linear equations $x = \Pi x$, where Π is the projection on the implementable subspace:

$$\Pi(x_1, \ldots, x_T) = \big(\mathbb{E}x_1, \mathbb{E}[x_2|\mathcal{F}_2], \ldots, \mathbb{E}[x_{T-1}|\mathcal{F}_{T-1}], x_T\big).$$

Employing the dual representation of the dynamic measure of risk $\varrho(\cdot)$, we obtain the following problem:

$$\min_x \max_{q \in Q_T} \sum_{s \in \Omega} q^s \langle c^s, x^s \rangle \tag{7.27}$$

$$\text{s.t. } x - \Pi x = 0, \tag{7.28}$$

$$x^s \in \mathcal{X}^s, \quad s \in \Omega. \tag{7.29}$$

We write $\langle c^s, x^s \rangle$ for the sum $\sum_{t=1}^T \langle c_t^s, x_t^s \rangle$. By Theorem 15, this problem is equivalent to the problem of minimizing (7.3), subject to (7.2) and (7.1).

We now develop duality relations for problem (7.27)–(7.29), extending to the risk-averse case the approach outlined in [41, Sec. 3.2.4]. After associating Lagrange multipliers λ with the nonanticipativity constraints (7.28), we obtain the following Lagrangian function:

$$L(x, \lambda) = \max_{q \in Q_T} \sum_{s \in \Omega} \left(q^s \langle c^s, x^s \rangle + p^s \langle \lambda^s, x^s - \Pi^s x \rangle \right).$$

It is sufficient to consider λ such that $\Pi\lambda = 0$, because any shift of λ by by a vector in the range of Π does not affect the last term. More specifically, we require that

$$\sum_{s \in \mathcal{S}(\nu)} p^s \lambda_t^s = 0, \quad \nu \in \Omega_t, \quad t = 1, \ldots, T-1. \tag{7.30}$$

Under this condition, the Lagrangian simplifies:

$$L(x, \lambda) = \max_{q \in Q_T} \sum_{s \in \Omega} \left(q^s \langle c^s, x^s \rangle + p^s \langle \lambda^s, x^s \rangle \right). \tag{7.31}$$

The dual function is defined as follows:

$$L_D(\lambda) = \inf_{x \in \mathcal{X}} L(x, \lambda),$$

and the dual problem is to find

$$\max_{\Pi\lambda = 0} \inf_{x \in \mathcal{X}} \max_{q \in Q_T} \sum_{s \in \Omega} \left(q^s \langle c^s, x^s \rangle + p^s \langle \lambda^s, x^s \rangle \right). \tag{7.32}$$

The function under the "$\max - \inf - \max$" operations is bilinear in x and q, the set Q_T is convex and compact, and the set $\mathcal{X}$ is convex. Therefore, we can interchange the inner "inf" and "max" operations, see. [42, Thm. 3.1], to write the dual problem as follows:

$$\max_{\Pi\lambda = 0} \max_{q \in Q_T} \left[\inf_{x \in \mathcal{X}} \sum_{s \in \Omega} \left(q^s \langle c^s, x^s \rangle + p^s \langle \lambda^s, x^s \rangle \right) \right]. \tag{7.33}$$

It is convenient to replace the measure q with its density δ with respect to p. Clearly,

δ lives in a convex compact set

$$\Delta = \Big\{\delta \in \mathbb{R}^{|\Omega|} : \big(p^s\delta^s\big)_{s\in\Omega} \in Q_T\Big\}. \tag{7.34}$$

The dual problem takes on the form:

$$\max_{\Pi\lambda=0}\max_{\delta\in\Delta}\left[\inf_{x\in\mathcal{X}}\sum_{s\in\Omega}p^s\left(\delta^s\langle c^s,x^s\rangle+\langle\lambda^s,x^s\rangle\right)\right]. \tag{7.35}$$

The problem in brackets has the same structure as in the risk-neutral case, but with scenario costs re-scaled by δ^s.

Theorem 16. *If Problem* (7.27)-(7.29) *has an optimal solution then the dual problem* (7.35) *has an optimal solution, and the optimal values of both problems coincide.*

The theorem follows from the duality theory in convex programming (see, e.g., [35, Thms. 4.7 and 4.8]). No constraint qualification is needed, because the constraints (7.28) are linear and the sets $\mathcal{X}^s$, $s \in \Omega$, are convex closed polyhedra.

Observe that the inner problem (in brackets) in (7.35) decomposes into individual scenario subproblems

$$\min_{x^s\in\mathcal{X}^s}\langle\delta^s c^s+\lambda^s,x^s\rangle, \quad s\in\Omega. \tag{7.36}$$

These subproblems can be readily solved by specialized techniques, exploiting the structure of the deterministic version of the dynamic problem in question.

Our approach can be interpreted as a construction of a family of risk-neutral approximations of the problem, one for each $\delta \in \Delta$.

7.6 Master Problem

Let us denote by $\Psi^s(\lambda^s,\delta^s)$ the optimal value of problem (7.36). The main difficulty is to solve the dual problem:

$$\max_{\Pi\lambda=0}\max_{\delta\in\Delta}\sum_{s\in\Omega}p^s\Psi^s(\lambda^s,\delta^s). \tag{7.37}$$

As each $\Psi^s(\cdot,\cdot)$ is concave and piecewise-linear, problem (7.37) is a convex programming problem.

The optimal value of the scenario subproblem (7.36) is a composition of the linear map $(\lambda^s, \delta^s) \mapsto \delta^s c^s + \lambda^s$ with the support function of the set $\mathcal{X}_s$. Using rules of subdifferential calculus we obtain

$$\partial\Psi^s(\lambda^s, \delta^s) = \left\{ \left(x^s, \langle c^s, x^s \rangle\right) : x^s \text{ is a solution of (7.36)} \right\}. \tag{7.38}$$

As the objective of (7.37),

$$D(\lambda, \delta) = \sum_{s\in\Omega} p^s \Psi^s(\lambda^s, \delta^s),$$

is a sum of terms that have no variables in common, we get

$$\partial D(\lambda, \delta) = \partial\Psi^1(\lambda^1, \delta^1) \times \cdots \times \partial\Psi^{|\Omega|}(\lambda^{|\Omega|}, \delta^{|\Omega|}). \tag{7.39}$$

Therefore, to calculate a subgradient at a point (λ, δ) we need to solve subproblems (7.36) and apply formula (7.39). In principle, problem (7.37) can be solved by any nonsmooth optimization method. One simple possibility would be the cutting plane method (see, e.g., [15, 35]); another choice is the bundle method (see [15, 18, 17, 35]).

The essence of the bundle method is the application of regularization with respect to the decision variables, which are in our case λ and δ, similarly to the proximal point method. This allows to localize the iterations and makes the bundle method more reliable for problems of higher dimension, where the cutting plane method becomes very slow.

Here, the specificity of problem (7.37) is that regularization is mainly needed for the nonanticipativity multipliers λ. The densities δ are restricted to live in a compact set Δ; in the extreme case of the risk-neutral problem we simply have $\Delta = \{(1, 1, \ldots, 1)\}$. We therefore propose a *partial bundle method*, which employs regularization with respect to the variables λ only. Exactly as the bundle method, it collects for every scenario s optimal solutions x^{sj} of the scenario subproblems and corresponding solutions $(\lambda^{sj}, \delta^{sj})$

of the master problem at iterations $j \in J_s$. The set J_s may be the set of all previous iterations, or its subset determined by the cut selection rules of the bundle method. The method also has the regularization center $\bar{\lambda}$, which is updated depending on the success of the current iteration, and uses a regularization coefficient $r > 0$.

The master problem of the partial bundle method has the following form

$$\begin{aligned}
&\max_{v_s,\lambda,\delta} \sum_{s\in\Omega} p^s \left(v_s - \frac{r}{2}\|\lambda^s - \bar{\lambda}^s\|^2 \right) \\
&\text{s.t. } v_s \le \left\langle \delta^{sj} c^s + \lambda^{sj}, x^{sj} \right\rangle + \left\langle \left(x^{sj}, \langle c^s, x^{sj} \rangle \right), (\lambda^s, \delta^s) - (\lambda^{sj}, \delta^{sj}) \right\rangle, \\
&\qquad\qquad s \in \Omega, \quad j \in J_s, \\
&\Pi\lambda = 0, \\
&\delta \in \Delta.
\end{aligned} \tag{7.40}$$

After its solution, the regularization center $\bar{\lambda}$, the regularization coefficient r, and the sets of cuts are updated in exactly the same way as in the bundle method (see [17, 35]). Convergence analysis of the partial bundle method were presented in Chapter 6 for the basic problem of minimizing a convex function of two decision vectors, without the complications of dealing with the sum of functions, over $s \in \Omega$. Our master problem (7.40) uses disaggregated subgradients, as in [9, 32]: each v_s is an upper bound on the corresponding function $\Psi^s(\lambda^s, \delta^s)$.

Chapter 8

Numerical Illustration

8.1 The Model

Our aim is to illustrate the scenario decomposition approach and the methods discussed in previous sections on the following inventory and assembly problem. A product line consists of several different models. Each model has its own list of parts, but different models may have some parts in common. At the first stage, we decide how many units of each part will be bought. After the purchase is complete, the actual demand for the different models is revealed. Then we decide how many units of each model will be produced, while keeping within the constraints defined by the numbers of parts available.

There is a penalty for each unit of unsatisfied demand and there is a "storage cost" associated to each unit that is produced over the demand. The storage cost involves product depreciation and is a random variable which will become known only after the second stage decisions have been made. It is assumed that all the products will eventually be sold and the storage cost is paid only once.

Let z_i be the number of parts of type i that will be purchased and let u_j be the number of units of model j that will be produced. Let M be the integer nonnegative matrix that describes the parts needed to assemble each different model, i.e. Mu is the vector of parts necessary to assemble the vector of models u. Random demand for product j is denoted by D_j and random unit storage cost is denoted by H_j. Other problem parameters, which are deterministic, are: r_j - selling price of product j, c_i - cost of part i, l_j - penalty for uncovered demand of product j.

Our goal is to minimize the negative of the profit, which is composed of three parts:

$$
\begin{aligned}
Z_1 &= \sum_i c_i z_i, \\
Z_2 &= -\sum_j r_j u_j, \text{ and} \\
Z_3 &= \sum_j \left[l_j (D_j - u_j)_+ + H_j (u_j - D_j)_+ \right].
\end{aligned}
$$

Since the components Z_2 and Z_3 are random, and our decisions u depend on the demand vector observed, we express the production problem as a three stage risk-averse optimization problem. In fact, there are no third stage decisions: only random cost evaluation. At stages 1 and 2 we use the conditional mean–semideviation risk measures of the first order of the form (2.21) with coefficients $\kappa_1 \in [0, 1]$ and $\kappa_2 \in [0, 1]$, respectively.

Assume that there are N possible demand realizations each occurring with corresponding probability p_s. Moreover, suppose that each demand realization s there are N_s possible storage cost realizations each occurring with probability $p_{s\eta}$, $\eta = 1, \ldots, N_s$. For given decisions u^s at node s, the cost equals:

$$
Z_2^s + Z_3 = -\langle r, u^s \rangle + \langle l, w^s \rangle + \langle H^{s\eta}, v^s \rangle,
$$

where w^s and v^s are the under and over production due to decision y^s at node s. In this

case a straightforward linear programming formulation of the problem is the following:

$$
\begin{aligned}
\min_{\substack{z,u,w,v\\ \rho,\sigma,\zeta,\gamma}} \ & \langle c, z\rangle + \sum_{s=1}^{N} p_s \rho^s + \kappa_1 \sum_{s=1}^{N} p_s \sigma^s \\
\text{s.t. } & \rho^s = \sum_{\eta=1}^{N_s} p_{s\eta} \zeta^{s\eta} + \kappa_2 \sum_{\eta=1}^{N_s} p_{s\eta} \gamma^{s\eta}, \\
& \sigma^s \geq \rho^s - \sum_{k\in\Omega_2} p_k \rho^k, \quad \sigma^s \geq 0, \\
& \zeta^{s\eta} = -\langle r, u^s\rangle + \langle l, w^s\rangle + \langle H^{s\eta}, v^s\rangle, \\
& \gamma^{s\eta} \geq \zeta^{s\eta} - \sum_{k=1}^{N_s} p_{sk} \zeta^{sk}, \quad \gamma^{s\eta} \geq 0, \\
& Mu^s - z \leq 0, \quad u^s \geq 0, \\
& w^s \geq D^s - u^s, \quad w^s \geq 0, \\
& v^s \geq u^s - D^s, \quad v^s \geq 0, \\
& \text{for all } s = 1, \ldots, N \text{ and } \eta = 1, \ldots, N_s.
\end{aligned}
\tag{8.1}
$$

In the problem above, $D^s := (D_1^s, \ldots, D_m^s)$ is the sth realization of product demands, and $H_i^{s\eta}$ is the storage cost of product i under demand realization s and storage realization η. The variable ρ^s represents the value of the conditional risk measure $\rho_2(Z_2 + Z_3)$ at node s, and the value of the risk measure $\rho_1(\cdot)$ is calculated directly in the objective function. The variables ζ represent cost realizations in the corresponding scenarios. The variables σ and γ represent the upper semideviations of the costs at stage 1 and 2, respectively.

The size of the linear programming representation of the production problem shows the importance of developing efficient methods to solve multi stage risk-averse problems. We applied to our problem the cutting plane, the classical bundle, and the partial bundle method. Whenever possible, we compared the results obtained by these methods with the result of solving the linear programming problem (8.1) directly by a simplex algorithm. For the scenario decomposition methods, we considered two versions. One was the full three-stage version, which is most general and applies also to problems

involving decisions at the last stage and general non-polyhedral measures of risk. Another version was a model with a truncated two-stage tree, in which the problems at the second stage are risk-averse problems themselves. This was possible due to the polyhedral structure of the mean–semideviation risk measure and to the absence of third stage decisions.

8.2 The Partial Bundle Method

To obtain explicitly the master problem of the partial bundle method for our application we need to calculate the set Δ appearing in (7.40). The structure of the subdifferential was already given in (4.1) and in this particular case it takes the following form:

$$\partial\rho_1(0) = \left\{ \mathbb{1} - \mathbb{1}\sum_{s=1}^{N} p_s\tau_s + \tau \;\middle|\; \tau = (\tau_s)_{s=1}^{N} \text{ and } 0 \le \tau_s \le \kappa_1 \right\} \tag{8.2}$$

and

$$\partial\rho_2^s(0) = \left\{ \mathbb{1} - \mathbb{1}\sum_{\eta=1}^{N_s} p_{s\eta}\iota_{s\eta} + \iota^s \;\middle|\; \iota^s = (\iota_{s\eta})_{\eta=1}^{N_s} \text{ and } 0 \le \iota_{s\eta} \le \kappa_2 \right\}, \tag{8.3}$$

where $\mathbb{1}$ is the vector with all entries equal to 1. Let $\partial\rho_2(0) := \partial\rho_2^1(0) \times \cdots \times \partial\rho_2^N(0)$ and $\pi = (p_{s\eta})_{s\in\Omega_1,\eta\in C(s)}$. Then $Q_2 = \mathcal{A}_2 \circ \mathcal{A}_1 = \partial\rho_2(0) \circ \partial\rho_1(0)$ and $\Delta = \left\{\delta : \left(p^s\delta^s\right)_{s=1}^{N} \in Q_2\right\}$. Thanks to the structure of the subdifferentials (8.2) and (8.3) the set Δ is polyhedral, and so, $\Delta = \left\{\left(\delta_\eta^s\right)_{s=1,\ldots,N,\eta=1,\ldots,N_s}\right\}$ such that

$$\begin{aligned}
\delta_\eta^s &= p_{s\eta}\left[1 - \sum_{k=1}^{N} p_k\tau_k + \tau_s - \sum_{k=1}^{N_s} p_{sk}\epsilon_{sk} + \epsilon_{s\eta}\right] \\
&0 \le \tau_i \le \kappa_1, \quad i = 1,\ldots,N, \\
&0 \le \epsilon_{ij} \le \kappa_2\left(1 - \sum_{k=1}^{N_s} p_k\tau_k + \tau_i\right), \quad i = 1,\ldots,N, \quad j = 1,\ldots,N_s.
\end{aligned}$$

The master problem of the partial bundle method for our application is:

$$
\begin{aligned}
\max_{v_s,\lambda,\delta} \; & \sum_{s=1}^{N} p^s \left(v_s - \frac{r}{2} \|\lambda^s - \bar{\lambda}^s\|^2 \right) \\
\text{s.t. } & v_s \leq \left\langle \delta^{sj} c^s + \lambda^{sj}, x^{sj} \right\rangle + \left\langle \left(x^{sj}, \langle c^s, x^{sj} \rangle \right), (\lambda^s, \delta^s) - (\lambda^{sj}, \delta^{sj}) \right\rangle, \\
& \Pi\lambda = 0, \\
& \delta^s = \left(p_{s\eta} \left[1 - \sum_{k=1}^{N} p_k \tau_k + \tau_s - \sum_{k=1}^{N_s} p_{sk} \epsilon_{sk} + \epsilon_{s\eta} \right] \right)_{\eta=1}^{N_s}, \qquad (8.4) \\
& 0 \leq \tau_s \leq \kappa_1, \\
& 0 \leq \epsilon_{s\eta} \leq \kappa_2 \left(1 - \sum_{k=1}^{N_s} p_k \tau_k + \tau_s \right), \quad \eta = 1, \ldots, N_s, \\
& \text{for all } j \in J_s, \quad s = 1, \ldots, N.
\end{aligned}
$$

At every iteration j of the partial bundle method the obtained subgradient have the following form

$$
\left[\left(p_1 z^{1j}\right)^\top, \ldots, \left(p_N z^{Nj}\right)^\top, p_1 \langle c^1, z^{1j} \rangle, \ldots, p_N \langle c^N, z^{Nj} \rangle, \left(p_1 G^1 y^{1j}\right)^\top, \ldots, \left(p_N G^N y^{Nj}\right)^\top \right],
$$

where $x^{sj} := (z^{sj}, y^{sj})$ is the optimal solution of subproblem (7.36) for scenario s at iteration j with z^{sj} corresponding to the first stage components of x^{sj}, and y^{sj} corresponding to the second and third stage components of x^{sj}. Also, c^s is the cost vector of the first stage scenario s, and G^s is the matrix of second stage scenario costs corresponding to the first stage scenario s. In our example $c^s = c$ and the rows of G^s are $(g_{s_\eta})^\top = \left(r^\top, l^\top, (H^{s\eta})^\top \right)$, for every $s = 1, \ldots, N$, $\eta = 1, \ldots, N_s$.

After a few algebraic simplifications we derive from (8.11) the individual scenario

subproblems for each scenario $s = 1, \ldots, N$,

$$\begin{aligned} \min_{z,u,w,v,\zeta} \ & \alpha^s \langle c^s, z\rangle + \langle \beta^s, \zeta\rangle + \langle \lambda^s, z\rangle \\ \text{s.t. } & \zeta^\eta = -\langle r, u\rangle + \langle l, w\rangle + \langle H^{s\eta}, v\rangle, \quad \eta = 1, \ldots, N_s, \\ & Mu - z \le 0, \quad u \ge 0, \\ & w \ge D^s - u, \quad w \ge 0, \\ & v \ge u - D^s, \quad v \ge 0, \end{aligned} \tag{8.5}$$

where each y^{sj} component of x^{sj} in (8.4) has been subdivided according to (8.1), i.e., $x := (z, y) := (z, u, v, w)$. Similarly, α^s, β^s are the corresponding z, y components of δ^s.

8.3 The Truncated Tree Method

In order to obtain the truncated two-stage tree method we need to find an efficient way of evaluating the second stage upper semideviation risk measure. Applying (7.16), we obtain for every $s = 1, \ldots, N$,

$$\rho_2^s(G^s y) = \max_{\delta \in \partial \rho_2^s(0)} \sum_{\eta=1}^{N_s} \delta_\eta p_{s\eta} g_{s\eta}^\top y, \tag{8.6}$$

where $\partial \rho_2^s(0)$ is obtained from (8.3). Substituting (8.3) into (8.6) gives

$$\rho_2^s(G^s y) = \max_{\iota \in [0,\kappa_2]^{N_s}} \sum_{\eta=1}^{N_s} p_{s\eta} g_{s\eta}^\top y + \sum_{\eta=1}^{N_s} \iota_\eta p_{s\eta} \left[g_{s\eta}^\top y - \sum_{\zeta=1}^{N_s} p_{s\zeta} g_{s\zeta}^\top y \right]. \tag{8.7}$$

Therefore $\rho_2^s(G^s y)$ can be obtained by solving the following linear program

$$\begin{aligned} \min \ & \sum_{\eta=1}^{N_s} p_{s\eta} g_{s\eta}^\top y + \sum_{\eta=1}^{N_s} d_\eta \\ \text{s.t. } & d_\eta \ge \kappa_2 p_{s\eta} \left[g_{s\eta}^\top y - \sum_{\zeta=1}^{N_s} p_{s\zeta} g_{s\zeta}^\top y \right], \quad \eta = 1, \ldots, N_s, \\ & d_\eta \ge 0, \quad \eta = 1, \ldots, N_s. \end{aligned} \tag{8.8}$$

The main idea of the *truncated tree* method is that instead of minimizing (7.3)

subject to (7.1) and (7.2), we minimize

$$\widetilde{\varrho_{1,3}} = Z_1 + \rho_1\left(\widetilde{Z_2}\right), \tag{8.9}$$

subject to (7.1) and (7.2), and

$$\widetilde{Z_2} = Z_2 + \rho_2\left(Z_3\right). \tag{8.10}$$

We consider the truncated problem as a two-stage problem and apply to it the same dual analysis that we did before. At the end we obtain formulation (7.37) with a few key differences. First, λ and δ refer to the random variables Z_1 and $\widetilde{Z_2}$ and have no components directly relating to either Z_2 or Z_3. More importantly, the individual scenario subproblems should take into consideration the cost of the new random variable $\widetilde{Z_2}$ and thus (7.36) is replaced by

$$\min_{(z^s,y^s)\in\mathcal{X}^s} \langle \delta^s c^s + \lambda^s, z^s\rangle + \delta^s \rho_2^s(g_s^\top y^s), \quad s = 1,\ldots,N, \tag{8.11}$$

where z^s and y^s are the decision variables corresponding to the first and second stage scenarios. By substituting (8.8) and (8.11) into (7.37) we obtain the following problem formulation for our application

$$\max_{\Pi\lambda=0} \max_{\delta\in\Delta} \sum_{s=1}^{N} p^s \Psi^s(\lambda^s,\delta^s), \tag{8.12}$$

where $\Psi^s(\lambda^s,\delta^s)$ is the optimal value of the following problem

$$\begin{aligned}
\min_{z,y,d}\ & \delta^s\left[(c^s)^\top z + \sum_{\eta=1}^{N_s} p_{s\eta} g_{s\eta}^\top y + \sum_{\eta=1}^{N_s} d_\eta\right] + (\lambda^s)^\top z \\
\text{s. t. }\ & d_\eta \geq \kappa_2 p_{s\eta}\left[g_{s\eta}^\top y - \sum_{\zeta=1}^{N_s} p_{s\zeta} g_{s\zeta}^\top y\right], \quad \eta = 1,\ldots,N_s, \\
& B_3^s z + A_3^s y = b_3^s, \\
& z\in X,\ y\geq 0,\ d\geq 0.
\end{aligned} \tag{8.13}$$

At every iteration j of the truncated tree partial bundle method, the subgradient has the following form

$$\left[\left(p_1\, z^{1j}\right)^\top, \ldots, \left(p_N\, z^{Nj}\right)^\top, p_1\, \phi_s\left(z^{1j}, y^{1j}, d^{1j}\right), \ldots, p_N\, \phi_s\left(z^{N,j}, y^{Nj}, d^{Nj}\right)\right], \tag{8.14}$$

where for every $s = 1, \ldots, N$,

$$\phi_s(z, y, d) = (c^s)^\top z + \sum_{\eta=1}^{N_s} p_{s\eta} g_{s\eta}^\top y + \sum_{\eta=1}^{N_s} d_\eta,$$

and $x^{sj} := (z^{sj}, y^{sj}, d^{sj})$ is the optimal solution of subproblem (8.13) for scenario s at iteration j.

By construction, we only consider the first scenarios for the decomposition in (8.12) and so $\Delta = \left\{\delta : \left(p^s\delta^s\right)_{s=1}^N \in \partial\rho_1(0)\right\}$, where $\partial\rho_1(0)$ was shown in (8.2). Therefore the master problem of partial bundle method for the truncated tree method has the following form:

$$\begin{aligned}
&\max_{v,\lambda,\delta} \sum_{s=1}^{N} p^s \left(v_s - \frac{r}{2}\|\lambda^s - \bar{\lambda}^s\|^2\right) \\
&\text{s.t. } v_s \le \left\langle \delta^{sj} c^s + \lambda^{sj}, x^{sj}\right\rangle + \left\langle \left(x^{sj}, \langle c^s, x^{sj}\rangle\right), (\lambda^s, \delta^s) - (\lambda^{sj}, \delta^{sj})\right\rangle, \\
&\quad \Pi\lambda = 0, \\
&\quad \delta^s = p_s\left[1 - \sum_{k=1}^{N} p_k\tau_k + \tau_s\right], \\
&\quad 0 \le \tau_s \le \kappa_1, \\
&\quad \text{for all } j \in J_s, \quad s = 1, \ldots, N.
\end{aligned} \tag{8.15}$$

After a few algebraic simplifications we derive from (8.11) the individual truncated tree

scenario subproblems for each scenario $s = 1, \ldots, N$,

$$
\begin{aligned}
\min_{z,u,w,v,\zeta} \ & \delta^s t + \langle \lambda^s, z \rangle \\
\text{s.t. } & \zeta^\eta = -\langle r, u \rangle + \langle l, w \rangle + \langle H^{s\eta}, v \rangle, \\
& t = \langle c^s, z \rangle + \left\langle (p_{sk})_{k=1}^{N_s}, \ \zeta + \kappa_2 S \right\rangle \\
& S_\eta \geq \zeta^\eta - \left\langle (p_{sk})_{k=1}^{N_s}, \ \zeta \right\rangle, \quad S_\eta \geq 0, \\
& Mu - z \leq 0, \quad u \geq 0, \\
& w \geq D^s - u, \quad w \geq 0, \\
& v \geq u - D^s, \quad v \geq 0, \\
& \text{for all } \eta = 1, \ldots, N_s,
\end{aligned}
\tag{8.16}
$$

where each decision variable y^{sj} from (8.11) has been subdivided according to (8.1), i.e. $x := (z, y) := (z, u, v, w)$.

8.4 Numerical Comparisons

Following the development in previous sections, we coded the methods in AMPL and compared the running time in seconds, number of iterations, and the average time in seconds per iteration of each method. For the numerical experiments we used Mosek linear and non-linear solvers with the AMPL interface. Our test machine has an Intel Core i7-920 processor with 8MB of L3 cache and 2.66GHz speed coupled with 8GB DDR3 SDRAM memory. The data sets for the experiments was randomly generated using a random number generator tailored for this appliction.

Table 8.1 shows the comparison of all the methods on a problem with 10 parts and 5 products, for different numbers of first-stage and second-stage scenarios. The classical cutting plane method was inefficient and failed to converge in a reasonable time on most instances, while being outperformed by all the other methods when it converged. For this reason we omitted it from Table 8.1. Clearly, small problems are best solved directly by linear programming in formulation (8.1). The usefulness of decomposition

Size	**LP**	**Truncated Tree**			**Partial Trunc. Tree**			**General Bundle**		
$N \times N_s$	Time	Time	Iter.	T/I	Time	Iter.	T/I	Time	Iter.	T/I
6×3	0	106	476	0.223	15	97	0.155	84	492	0.171
5×5	0	95	451	0.211	36	194	0.186	61	419	0.146
5×6	0	75	388	0.193	13	86	0.151	48	270	0.178
6×6	0	134	574	0.233	133	441	0.302	109	521	0.209
10×10	0	313	435	0.720	287	419	0.685	309	501	0.617
50×50	5	1381	510	2.708	1652	485	3.406	3283	414	7.930
100×100	98	5570	660	8.439	1547	300	5.157	28316	579	48.91
200×200	5767	5975	240	24.89	4722	200	23.61	54336	291	186.7
300×300	-	19910	255	78.08	20622	255	80.87	-	-	-

Table 8.1: **LP**: Linear Programming formulation. **Truncated Tree**: Bundle Method applied to the Truncated Tree formulation. **Partial Trunc. Tree**: Partial Bundle Method applied to the Truncated Tree formulation. **General Bundle**: Bundle method applied to the general multistage decomposition formulation. Tests were performed for N first-stage scenarios, and N_s second-stage scenarios following each first-stage scenario.

is shown when we consider large problems. For example on the instance with 200 first-stage scenarios, with each followed by 200 second-stage scenarios, the general bundle and the partial truncated tree methods outperformed the linear programming formulation. More important is the case with 300 first-stage scenarios, with 300 second-stage scenarios after each of them, where the linear programming approach failed, but the truncated tree and partial truncated tree methods were able to find a solution. In this case the meager memory requirements of these methods allowed us to obtain a solution even when the linear programming formulation was too large for our computer memory. In general, we saw the partial truncated tree method outperforming the truncated tree method but this improvement might be problem-specific.

Notice that the truncated tree method moves the calculation of the second stage risk measure from the master problem to the subproblems resulting in a smaller master problem but larger subproblems. This is the main difference between the truncated tree and general bundle methods. In larger instances, the dimension of the master problem affects the number of iterations necessary to find a solution, as well as time to solve the master problem at each iteration. For these reason, the truncated tree method with its simpler master problem outperforms the general bundle method on the largest instances.

8.5 Conclusions

We defined a multistage risk-averse stochastic problem with conditional risk measures. Considerable work was devoted to develop the theoretical foundation of the risk-averse problem and we used Lagrangian duality to formulate a dual representation of the problem. Along the way we developed the tools necessary for the development of dual cutting plane and bundle methods.

To test our techniques we considered a simple problem in manufacturing and transportation with upper semideviations as risk measures and three stages. For this problem we developed a specialized version of the bundle methods where the tree of decisions was truncated into only two stages. We also implemented the cutting plane, general bundle, truncated tree, and partial truncated tree methods for our application. Comparisons of the results allow us to reach conclusions of when these algorithms would be better applied.

Appendix: Mathematical Background

This appendix contains an overview of some mathematical facts needed for the development of our theory and methods. The reader is referred to [35] for an in-depth look at these important mathematical facts.

A.1 Conjugate Duality

Let $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ and $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a function. Let $\alpha \in \mathbb{R}$ and $s \in \mathbb{R}^n$. An *affine minorant* of f is an affine function

$$l_\alpha(x) = \langle s, x \rangle - \alpha, \tag{8.17}$$

such that $l_\alpha(x) \leq f(x)$ for all $x \in \mathbb{R}$. We call s the *slope* of the affine minorant l_α. Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a function. The function $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$ defined by

$$f^*(s) \triangleq \sup_{x \in \mathbb{R}} \{\langle s, x \rangle - f(x)\} \tag{8.18}$$

is called the *conjugate function* of f. Clearly, a proper function f has an affine minorant with slope s if and only if $f^*(s) < +\infty$.

Lemma 17. *Suppose that $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and has an affine minorant. Then the conjugate function f^* is proper, convex, and lower semicontinuous.*

Let f be a proper convex function. Then the domain of f, denoted by dom f, is a convex set and f is subdifferentiable in the interior of its domain. By restricting ourselves to the linear manifold of smallest dimension containing the domain of f we can assume that the interior of dom f is nonempty. It follows that after the restriction there is a point $x_0 \in$ dom f such that f is subdifferentiable at x_0. So, there is $s_0 \in \partial f(x_0)$

such that for every $x \in \mathbb{R}^n$

$$f(x) \geq f(x_0) + \langle s_0, x - x_0 \rangle ,$$

i.e. f has an affine minorant with slope s_0. An application of Lemma 17 then gives

Theorem 18. *If $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex and proper then its conjugate, $f^* : \mathbb{R}^n \to \overline{\mathbb{R}}$, is convex, proper, and lower semicontinuous.*

The *biconjugate* function is the conjugate function of the conjugate function f^*, that is

$$f^{**}(x) \triangleq \sup_{s \in \mathbb{R}} \{\langle s, x \rangle - f^*(s)\} . \tag{8.19}$$

The following result is known as the *Fenchel-Moreau Theorem*.

Theorem 19. *Suppose that $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ has at least one affine minorant. Then*

$$\text{epi } f^{**} = \overline{\text{conv}}(\text{epi} f).$$

In particular, if f is a proper, convex and lower semicontinuous function, then

$$f^{**} = f.$$

The relation between a proper convex function f and its dual f^* can be used to characterize the subgradients of both f and f^*.

Theorem 20. *Suppose that $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper convex function. Then the following two statements are equivalent:*

(i) $s \in \partial f(x)$;

(ii) $f(x) + f^*(s) = \langle s, x \rangle .$

If, in addition, f is lower semicontinuous then both statements are equivalent to

(iii) $x \in \partial f^*(s)$.

A.2 Indicator and Support Functions

Let Z be a set in $\mathbb{R}^n$. The *indicator* and *support* functions of Z are defined by

$$\delta_Z(x) = \begin{cases} 0 & \text{if } x \in Z \\ +\infty & \text{otherwise} \end{cases}, \ \forall x \in \mathbb{R}^n$$

and

$$\sigma_Z(s) = \sup_{x \in Z} \langle s, x \rangle , \ \forall s \in \mathbb{R}^n,$$

respectively. Notice that

$$\delta_Z^*(s) = \sup_{x \in \mathbb{R}^n} \{\langle s, x \rangle - \delta_Z(x)\} = \sup_{x \in Z} \langle s, x \rangle = \sigma_Z(s),$$

and so $\sigma_Z^* = \delta_Z^{**}$. Theorem 19 implies that

$$\operatorname{epi} \sigma_Z^* = \operatorname{epi} \delta_Z^{**} = \overline{\operatorname{conv}}(\operatorname{epi} \delta_Z) = \operatorname{epi} \delta_{\overline{\operatorname{conv}}(Z)}.$$

It follows that if Z is closed and convex then the functions σ_Z and δ_Z are mutually conjugate, i.e. $\sigma_Z^* = \delta_Z$ and $\delta_Z^* = \sigma_Z$.

Lemma 21. *Let $Z \subset \mathbb{R}^n$ be nonempty, closed and convex. Then σ_Z is subdifferentiable and $\partial \sigma_Z(0) = Z$.*

A.3 The Subdifferential of the Maximum Function

Consider the function

$$F(x) = \sup_{y \in Y} f(x, y),$$

where $f : \mathbb{R}^n \times Y \to \mathbb{R}$ satisfies the following conditions:

(i) $f(\cdot, y)$ is convex for all $y \in Y$;

(ii) $f(x, \cdot)$ is upper semicontinuous for all x in a certain neighborhood of a point x_0;

(iii) The set $Y \subset \mathbb{R}^m$ is compact.

The maximum function F is convex and by (ii) it is also proper and thus it is subdifferentiable at the point x_0.

Let $\hat{Y}(x)$ denote the set of $y \in Y$ at which $f(x,y) = F(x)$. Clearly conditions (i)-(iii) imply that the set $\hat{Y}(x)$ is nonempty and compact in a certain neighborhood of x_0. Let $\partial_x f(x_0, y)$ denote the subdifferential of the function $f(\cdot, y)$ at x_0.

The following result gives an useful reformulation of the subdifferential of F at x_0.

Theorem 22. *Assume conditions* (i)-(iii). *Then*

$$\partial F(x_0) \supset \operatorname{conv}\left(\bigcup_{y\in\hat{Y}(x_0)} \partial_x f(x_0, y)\right).$$

If, in addition, the function $f(\cdot, y)$ is continuous at x_0 for all $y \in Y$, then

$$\partial F(x_0) = \operatorname{conv}\left(\bigcup_{y\in\hat{Y}(x_0)} \partial_x f(x_0, y)\right). \tag{8.20}$$

References

[1] Shabbir Ahmed, Ulas akmak, and Alexander Shapiro, *Coherent risk measures in inventory problems*, European Journal of Operational Research **182** (2007), no. 1, 226 – 238.

[2] P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, , and H. Ku, *Coherent multiperiod risk adjusted values and bellman's principle*, Annals of Operations Research **152** (2007), 5–22.

[3] Philippe Artzner, *Application of coherent risk measures to capital requirements in insurance*, N. Am. Actuar. J. **3** (1999), no. 2, 11–25, SOA Seminar: Integrated Approaches to Risk Measurement in the Financial Services Industry (Atlanta, GA, 1997).

[4] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath, *Coherent measures of risk*, Math. Finance **9** (1999), no. 3, 203–228.

[5] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath, *Coherent measures of risk*, Risk management: value at risk and beyond (Cambridge, 1998), Cambridge Univ. Press, Cambridge, 2002, pp. 145–175.

[6] J.-P. Aubin and H. Frankowska, *Set-valued analysis*, Birkhäuser, Boston, 1990.

[7] D. Bertsekas, P. Tseng, *Partial proximal minimization algorithms for convex programming*, SIAM J. Optim. **4** (1994), no. 3, 551-572.

[8] Birge, J. R., and F. V. Louveaux, *Introduction to stochastic programming*, Springer, New York, 1997.

[9] J. R. Birge and F. V. Louveaux, *A multicut algorithm for two-stage stochastic linear programs*, European Journal of Operational Research **34** (1988), 384 – 392.

[10] P. Cheridito, F. Delbaen, and M. Kupper, *Dynamic monetary risk measures for bounded discrete-time processes*, Electronic Journal of Probability **11** (2006), 57–106.

[11] Sungyong Choi and Andrzej Ruszczyński, *A risk-averse newsvendor with law invariant coherent measures of risk*, Oper. Res. Lett. **36** (2008), no. 1, 77–82.

[12] H. Föllmer and A. Schied, *Stochastic finance. an introduction in discrete time*, Walter de Gruyter, Berlin (2004).

[13] M. Fritelli and E. Rosazza Gianin, *Putting order in risk measures*, Journal of Banking and Finance **26** (2002), 1473–1486.

[14] M. Fritelli and G. Scandolo, *Risk measures and capital requirements for processes*, Mathematical Finance **16** (2006), 589–612.

[15] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms*, Springer, Berlin, 1993.

[16] P. Kall and J. Mayer, *Stochastic linear programming*, Springer, New York, 2005.

[17] K. C. Kiwiel, *Methods of descent for nondifferentiable optimization*, Lecture Notes in Mathematics, vol. 1133, Springer-Verlag, Berlin, 1985.

[18] K. C. Kiwiel, *An aggregate subgradient method for nonsmooth convex minimization*, Mathematical Programming **27** (1983), 320 – 341.

[19] T.J. Linsmeier and Pearson N.D., *Risk measurement: An introduction to value-at-risk*, Technical report 96-04, OFOR.

[20] H. Mauser and D. Rosen, *Beyond var: From measuring risk to managing risk*, ALGO Research Quaterly **1** (1999), no. 2, 5 – 20.

[21] R. McKay and T.E. Keefer, *Var is a dangerous technique*, Corporate Finance Searching for Systems Integration Supplement, Sep., pp. 30 (1996).

[22] N. Miller and A. Ruszczyński, *Risk-averse two-stage stochastic linear programming: modeling and decomposition*, Operations Research, to appear in 2011.

[23] Naomi Miller and Andrzej Ruszczyński, *Risk-adjusted probability measures in portfolio optimization with coherent measures of risk*, European J. Oper. Res. **191** (2008), no. 1, 193–206.

[24] J.M. Mulvey and A. Ruszczyński, *A new scenario decomposition method for large-scale stochastic optimization*, Operations Research **43** (1995), 477–490.

[25] G. Ch. Pflug and W. Römisch, *Modeling, measuring and managing risk*, World Scientific, Singapore (2007).

[26] A. Prékopa, *Stochastic programming*, Kluwer, Dordrecht, 1995.

[27] F. Riedel, *Dynamic coherent risk measures*, Stochastic Processes and Their Applications **112** (2004), 185–200.

[28] R. Tyrrell Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970.

[29] R. Tyrrell Rockafellar, *Conjugate duality and optimization*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1974, Lectures given at the Johns Hopkins University, Baltimore, Md., June, 1973, Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 16.

[30] R. Tyrrell Rockafellar, *Convex analysis*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1997, Reprint of the 1970 original, Princeton Paperbacks.

[31] A. Ruszczyński, *Risk-averse dynamic programming for markov decision processes*, Mathematical Programming, to appear in 2010.

[32] A. Ruszczyński, *A regularized decomposition method for minimizing a sum of polyhedral functions*, Mathematical Programming **35** (1986), 309 – 333.

[33] A. Ruszczyński, *Decomposition methods*, in: Stochastic Programming, Handbooks Oper. Res. Management Sci., vol. 10, Elsevier, Amsterdam, 2003, pp. 141–211.

[34] A. Ruszczyński and A. Shapiro, *Optimization of risk measures*, in: Probabilistic and Randomized Methods for Design under Uncertainty (G. Calafiore and F. Dabbene, eds.), Springer, London, 2005.

[35] Andrzej Ruszczyński, *Nonlinear optimization*, Princeton University Press, Princeton, NJ, USA, 2006.

[36] Andrzej Ruszczyński and Alexander Shapiro, *Stochastic programming models*, Stochastic programming, Handbooks Oper. Res. Management Sci., vol. 10, Elsevier, Amsterdam, 2003, pp. 1–64.

[37] Andrzej Ruszczyński and Alexander Shapiro, *Conditional risk mappings*, Math. Oper. Res. **31** (2006), no. 3, 544–561.

[38] Andrzej Ruszczyński and Alexander Shapiro, *Optimization of convex risk functions*, Math. Oper. Res. **31** (2006), no. 3, 433–452.

[39] Andrzej Ruszczyński and Alexander Shapiro, *Corrigendum to: "optimization of convex risk functions," mathematics of operations research31 (2006) 433–452*, Math. Oper. Res. **32** (2007), no. 2, 496–496.

[40] G. Scandolo, *Risk measures in a dynamic setting*, PhD Thesis, Università degli Studi di Milano, Milan (2003).

[41] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński, *Lectures on stochastic programming: Modeling and theory*, MPS-SIAM series on optimization, no. 9, MPS-SIAM, Philadelphia, 2009.

[42] S. Simons, *Minimax and Monotonicity*, Springer, Berlin, 1998

[43] Giorgio Szegö, *Measures of risk*, Journal of Banking and Finance **26** (2002), no. 7, 1253 – 1272.

[44] S. Uryasev, R. T. Rockafellar and M. Zabarankin, *Deviation measures in risk analysis and optimization*, Finance and Stochastics **10** (2006), 51–74.

Curriculum Vitae

Ricardo A. Collado Soto

Education

2005-10 Ph. D. in Operations Research, Rutgers Center for Operations Resreach, Rutgers University.

1997-01 B. Sc. in Mathematics, University of Puerto Rico, Río Piedras Campus.

Academic Positions

2004-05 Instructor, Department of Mathematics, University of Puerto Rico, Río Piedras Campus.

2001-04 Research assistant, Department of Mathematics, University of Puerto Rico, Río Piedras Campus.

1998-04 Research assistant, Department of Mathematics, University of Puerto Rico, Río Piedras Campus.

1998-98 System administrator, Department of Mathematics, University of Puerto Rico, Río Piedras Campus.

1997-98 Teaching assistant, Department of Mathematics, University of Puerto Rico, Río Piedras Campus.

Publications

R. Collado, D. Papp and A. Ruszczyński, *Scenario decomposition of risk-averse multistage stochastic programming problems*, submitted for publication.

D. Papp, R. Collado, F. Alizadeh., *Nonnegativity and the Sum-of-Squares Property of Generalized Polynomials*, submitted for publication.

D. Papp, R. Collado, F. Alizadeh., *Extension of the semidefinite characterization of sum of squares functional systems to algebraic structures*, DIMACS technical report $\#2009 - 26$.

R. Collado, A. Kelmans, D. Krasner. *On convex polytopes in the plane "containing" and "avoiding" zero*, DIMACS technical report $\#2002 - 33$.