Math. Program., Ser. B (2008) 112:183–220
DOI 10.1007/s10107-006-0090-4

FULL LENGTH PAPER

Stochastic programming approach to optimization under uncertainty

Alexander Shapiro

Received: 31 January 2006 / Accepted: 16 September 2006 / Published online: 13 January 2007

Abstract In this paper we discuss computational complexity and risk averse approaches to two and multistage stochastic programming problems. We argue that two stage (say linear) stochastic programming problems can be solved with a reasonable accuracy by Monte Carlo sampling techniques while there are indications that complexity of multistage programs grows fast with increase of the number of stages. We discuss an extension of coherent risk measures to a multistage setting and, in particular, dynamic programming equations for such problems.

Keywords Two and multistage stochastic programming · Complexity · Monte Carlo sampling · Sample average approximation method · Coherent risk measures · Dynamic programming · Conditional risk mappings

Mathematics Subject Classification (2000) 90C15

1 Introduction

In many situations there is a need to make, hopefully an optimal, decision under conditions of uncertainty. Everybody would agree with this statement. There is a disagreement, however, with how to deal with such situations. Uncertainty can come in many different forms, and hence there are various ways how it can be modelled. In a mathematical approach one formulates an objective function $f : \mathbb{R}^n \to \mathbb{R}$ which should be optimized (say minimized) subject to specified constraints. That is, one formulates a mathematical programming problem:

A. Shapiro (✉)
School of Industrial and Systems Engineering, Georgia Institute of Technology,
Atlanta, GA 30332-0205, USA
e-mail: ashapiro@isye.gatech.edu

$$\operatorname*{Min}_{x\in X} f(x), \tag{1.1}$$

where the feasible set $X \subset \mathbb{R}^n$ is typically defined by a (finite or even infinite) number of constraints, say $X := \{x \in \mathbb{R}^n : g_i(x) \le 0,\ i \in I\}$ (the notation ":=" means "equal by definition"). Inevitably the objective and constraint functions depend on parameters, which we denote by vector $\xi \in \mathbb{R}^d$. That is, $f(x,\xi)$ and $g_i(x,\xi)$, $i \in I$, can be viewed as functions of the decision vector $x \in \mathbb{R}^n$ and parameter vector $\xi \in \mathbb{R}^d$.

Typically the parameter vector ξ is subject to an error (say a round-off error) or, even worse, is uncertain. In such cases fixing parameters to a nominal value $\xi = \xi^*$ and then solving the corresponding optimization problem, could lead to a poor solution. Let us note at this point that it is possible to reformulate problem (1.1) as an unconstrained problem by introducing penalties for violating the constraints. For example, we can define the extended real valued function $\bar{f}(x,\xi) := f(x,\xi)$ if $x \in \{x : g_i(x,\xi) \le 0,\ i \in I\}$ and $\bar{f}(x,\xi) := +\infty$ otherwise, and then rewrite the corresponding optimization problem as the unconstrained problem:

$$\operatorname*{Min}_{x\in\mathbb{R}^n} \bar{f}(x,\xi). \tag{1.2}$$

The reader should be warned that, although convenient from a mathematical point of view, such reformulation may disguise an essential difference between the objective and constraint functions. In some cases even small perturbations of the nominal values of the parameters may result in a severe infeasibility of an optimal solution x^* of the nominal problem. When satisfying the feasibility constraints is important, and often this is the case, such "nonrobustness" with respect to constraint violations can make the nominal solution useless, or even worse, misleading. This observation could be considered as a starting point for the robust approach to mathematical programming where the nominal problem is replaced by a "worst case" problem. We refer to Ben-Tal and Nemirovski [5] for a thorough discussion of the robust approach.

In this paper we consider an alternative approach of stochastic optimization. We view the uncertain parameter vector ξ as a *random* vector having probability distribution P supported on a (closed) set $\Xi \subset \mathbb{R}^d$. We then formulate the following stochastic programming problem:

$$\operatorname*{Min}_{x\in X} \mathbb{E}[f(x,\xi)], \tag{1.3}$$

where the expectation $\mathbb{E}[f(x,\xi)] = \int_\Xi f(x,\xi)dP(\xi)$ is taken with respect to the probability distribution P. For the moment we assume that the feasible set X is a well defined (deterministic) set. While making modelling in the form of optimization problem (1.3), we need to answer two basic questions:

(i) Whether the optimization problem (1.3) makes sense?
(ii) Could it be solved (numerically)?

These two questions cannot be separated since even if we are satisfied with the modelling part of the procedure, but the resulting optimization problem cannot be solved in a reasonable time with a reasonable accuracy, usefulness of such model could be questionable.

An answer to both questions is not obvious, and of course should depend on a class of considered problems. With respect to the question (i), two additional questions come to mind, namely:

- (i$'$) How do we know the probability distribution P?
- (i$''$) Why do we optimize the *expected* value of the objective function, i.e., why do we optimize on average?

Without specifying the probability distribution P, we even cannot formulate the problem mathematically. In some cases the relevant probability distribution can be estimated with a reasonable accuracy from an available historical data. However, in many cases it either could change with time or is based on a subjective judgement. The optimization of the expected value could be justified by an application of the Law of Large Numbers (LLN). That is, if we are supposed to solve the same problem, under the same probability distribution, many times, then formulation (1.3) gives a best possible solution *on average*. If, however, in the process we lose our investment and are forced out of business, it would not help that our decisions were optimal on average.

With respect to the second question (ii) of solving problem (1.3) numerically, let us observe that just evaluation of the objective function of that problem, at a considered point $x \in X$, requires calculation of the corresponding integral $\int_{\Xi} f(x, \xi) dP(\xi)$. Only in rather simple cases this integral can be written in a closed form. For continuous distributions, typically, this integral cannot be evaluated numerically with high accuracy already, say, for the number of random variables $d > 4$ (see, e.g., Niederreiter [36]).

The above discussion raises the question of whether the stochastic programming is a viable technique, moreover so in view of competing alternative approaches. In the remainder of this paper we will try to address the above questions. The reader could make his/her own conclusions whether the suggested answers are satisfactory.

2 Two-stage stochastic programming

In this section we discuss the two-stage stochastic programming approach. That is, it is assumed that two types of decision vectors $x \in \mathbb{R}^{n_1}$ and $y \in \mathbb{R}^{n_2}$ are involved. A decision about vector x has to be made "*here-and-now*" before a realization[1] of the corresponding random data vector ξ becomes known. After a realization of ξ becomes available, an optimal decision about y is made by solving the corresponding optimization problem:

$$\operatorname*{Min}_{y \in \mathcal{G}(x,\xi)} g(x, y, \xi), \tag{2.1}$$

[1] We denote by the same symbol ξ random vector and its particular realization, which one of these two meanings will be used in a particular situation will be clear from the context

where $\mathcal{G} : \mathbb{R}^{n_1} \times \Xi \rightrightarrows \mathbb{R}^{n_2}$ is a multifunction defining the corresponding feasible set and $g : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \Xi \to \mathbb{R}$. The second stage optimization problem (2.1) depends on the first stage decision vector x and data (parameter) vector ξ. At the first stage, one is supposed to optimize the expected value of the second stage problem, i.e., to solve the optimization problem:

$$\operatorname*{Min}_{x \in X} \big\{ f(x) := \mathbb{E}[F(x,\xi)] \big\}, \tag{2.2}$$

where X is a subset of $\mathbb{R}^{n_1}$ and $F(x,\xi)$ denotes the optimal value of problem (2.1). (We assume that the set X is nonempty and closed.) For example,[2] the above problem (2.1)–(2.2) becomes a two-stage stochastic *linear* program (with recourse) if

$$\begin{aligned} X &:= \{x : Ax + b \le 0\}, \; g(x,y,\xi) := \langle c,x \rangle + \langle q,y \rangle \text{ and} \\ \mathcal{G}(x,\xi) &:= \{y : Tx + Wy + h \le 0\}, \end{aligned} \tag{2.3}$$

and the data vector ξ is formed from some (all) elements of vectors q and h and matrices T and W. The concept of two-stage stochastic programming was introduced in Beale [4] and Dantzig [11], and was discussed in numerous publications (see, e.g., monographs [6,26,45,53]).

Let us remark that there is an additional difficulty here as compared with formulation (1.3). That is, the function $F(x,\xi)$ is not given explicitly and might be not finite valued. If for some $x \in X$ and $\xi \in \Xi$ the optimization problem (2.1) is unbounded from below, then $F(x,\xi) := -\infty$. This is a somewhat pathological situation meaning that for some feasible x and a possible realization of the data, one can improve the value of the second stage problem indefinitely. We assume that at the modelling stage, one makes sure that this does not happen, i.e., $F(x,\xi) > -\infty$ for all $(x,\xi) \in X \times \Xi$. It also might happen that the second stage problem is infeasible, i.e., $\mathcal{G}(x,\xi) = \emptyset$. In that case we define $F(x,\xi) := +\infty$. We have that if, for some $x \in X$, the second stage problem is infeasible with positive probability, then $\mathbb{E}[F(x,\xi)] = +\infty$ and such x cannot be a solution of the first stage problem. Therefore, de-facto the first stage problem (2.2) should be solved over such $x \in X$ that $F(x,\xi) < +\infty$ with probability one (w.p.1). It is said that the two-stage problem has *relatively complete recourse* if for every $x \in X$ the feasible set $\mathcal{G}(x,\xi)$, of the second-stage problem, is nonempty w.p.1, i.e., $\mathcal{G}(x,\xi) \neq \emptyset$ for almost every (a.e.) $\xi \in \Xi$. Of course, even in the case of relatively complete recourse it might happen that $\mathbb{E}[F(x,\xi)] = +\infty$, for some $x \in X$, if the probability distribution of $F(x,\xi)$ has sufficiently heavy tails.

A standard approach to solving the two stage problem (2.1)–(2.2) is by constructing scenarios. That is, one generates a finite number of points $\xi_k \in \Xi$, $k = 1,\dots,K$, called *scenarios*, and assigns to each ξ_k a positive weight p_k such that $\sum_{k=1}^K p_k = 1$. The generated set $\{\xi_1,\dots,\xi_K\}$ of scenarios, with the corresponding probabilities $p_1,\dots,p_K$, can be viewed as a representation of the

[2] By $\langle c,x \rangle$ we denote the standard scalar product of two vectors.

underlying probability distribution. With respect to this distribution we can write the expected value function $f(x) = \mathbb{E}[F(x,\xi)]$ as the finite summation $f(x) = \sum_{k=1}^{K} p_k F(x,\xi_k)$. By making one copy y_k of the second stage decision vector for every scenario ξ_k, i.e., by considering $y_k = y(\xi_k)$, $k = 1,\dots,K$, as a function of scenarios, we can write the two-stage problem (2.1)–(2.2) as one large optimization problem:

$$\begin{array}{cl} \operatorname*{Min}\limits_{x,y_1,\dots,y_K} & \sum_{k=1}^{K} p_k g(x,y_k,\xi_k) \\ \text{s.t.} & x \in X,\ y_k \in \mathcal{G}(x,\xi_k),\ k = 1,\dots,K. \end{array} \tag{2.4}$$

In particular, in the linear case (2.3) this becomes the linear programming problem:

$$\begin{array}{cl} \operatorname*{Min}\limits_{x,y_1,\dots,y_K} & \langle c,x\rangle + \sum_{k=1}^{K} p_k \langle q_k, y_k\rangle \\ \text{s.t.} & Ax + b \le 0,\ T_k x + W_k y_k + h_k \le 0,\ k = 1,\dots,K, \end{array} \tag{2.5}$$

where $\xi_k = (q_k, T_k, W_k, h_k)$, $k = 1,\dots,K$, are the corresponding scenarios. Over the years a considerable effort went into developing algorithms exploiting a specific structure of problems of the type (2.4), and especially linear problems (2.5) (see, e.g., [54] for a recent survey of such decomposition type algorithms).

If we view the generated scenarios (together with the corresponding weights/probabilities) as an approximation of the "true" probability distribution of the random data vector ξ, the natural question is whether by solving the corresponding problem (2.4) we can solve the "true" two-stage problem with a reasonable accuracy in a reasonable time. Modern computers coupled with good algorithms can solve linear programming problems of the form (2.5) with millions of variables and constraints. Yet, the number of scenarios needed to approximate the underlying probability distribution with a reasonable accuracy typically grows exponentially with increase of the number of random parameters. This poses serious doubts whether an answer to the question (ii), formulated in the Introduction, could be positive even for moderate size two stage linear stochastic programs. This is what we are going to discuss next.

3 Complexity of two-stage stochastic programs

Suppose for the moment that components ξ_i, $i = 1,\dots,d$, of the random data vector $\xi \in \mathbb{R}^d$ are independently distributed. Suppose, further, that we use r points for discretization of the (marginal) probability distribution of each component ξ_i. Then the resulting number of scenarios is $K = r^d$, i.e., it grows exponentially with increase of the number of random parameters. Already with, say, $r = 4$ and $d = 20$ we will have an astronomically large number of scenarios $4^{20} \approx 10^{12}$. In such situations it seems hopeless just to calculate with a high accuracy the value $f(x) = \mathbb{E}[F(x,\xi)]$ of the objective function at a given point $x \in X$, much less to solve the corresponding optimization problem.[3] And, indeed, it is shown in Dyer and Stougie [15] that, under the assumption that the stochastic parameters are independently distributed, two-stage linear stochastic programming problems are ♯P-hard.

Quite often in applications it does not make much sense to try to solve the corresponding stochastic problem with a high accuracy, say of order 10^{-3} or 10^{-4}, since the involved inaccuracies resulting from inexact modelling, distribution approximations etc. could be far bigger. Therefore, we approach now the problem from the point of view of Monte Carlo sampling techniques. Suppose that we can generate a random sample $\xi^1, \dots, \xi^N$ of N realizations of the random vector ξ, i.e., each ξ^j, $j = 1, \dots, N$, has the same probability distribution as ξ. While making the following theoretical analysis we assume that ξ^j, $j = 1, \dots, N$, are distributed independently,[4] i.e., the sample is iid. Consider the corresponding so-called sample average function $\hat{f}_N(x) := N^{-1} \sum_{j=1}^N F(x, \xi^j)$. The function $\hat{f}_N(x)$ depends on the generated random sample and therefore is random. For any fixed $x \in X$, we have that $\hat{f}_N(x)$ is an unbiased estimator of the expectation $f(x)$, i.e., $\mathbb{E}\big[\hat{f}_N(x)\big] = f(x)$, and by the LLN that $\hat{f}_N(x)$ converges to $f(x)$ w.p.1 as $N \to \infty$. This motivates to introduce the following, so-called *sample average approximation* (SAA), problem:

$$\operatorname*{Min}_{x \in X} \left\{ \hat{f}_N(x) := N^{-1} \sum_{j=1}^{N} F(x, \xi^j) \right\}. \tag{3.1}$$

Note that once the sample is generated, problem (3.1) becomes a problem of the form (2.4) with scenarios ξ^j, $j = 1, \dots, N$, each taken with equal probability $p_j = N^{-1}$, $j = 1, \dots, N$. Note also that the sample average approximation method is *not* an algorithm, one still has to solve the obtained optimization problem (3.1) by applying an appropriate numerical procedure. It is difficult to point out who was the first to suggest the SAA approach to solving stochastic problems. The idea of this method is quite simple and it was discovered and rediscovered by several authors under different names in different contexts.

It is possible to show that, under mild regularity conditions, the optimal value $\hat{v}_N$ and an optimal solution $\hat{x}_N$ of the SAA problem (3.1) converge w.p.1 as $N \to \infty$ to their counterparts[5] v^* and S^* of the "true" problem (2.2). However, the convergence could be slow. By the Central Limit Theorem (CLT), for

[3] Of course, in some very specific situations it is possible to calculate $\mathbb{E}[F(x, \xi)]$ in a closed form. Also if $F(x, \xi)$ is decomposable into the sum $\sum_{i=1}^d F_i(x, \xi_i)$, then $\mathbb{E}[F(x, \xi)] = \sum_{i=1}^d \mathbb{E}[F_i(x, \xi_i)]$ and hence the problem is reduced to calculations of one dimensional integrals. This happens in the case of the so-called simple recourse.

[4] In practical applications, in order to speed up the convergence, it is often advantageous to use Quasi-Monte Carlo techniques where the generated ξ^j are not independently distributed (cf., [24,28,42]).

[5] We denote by S^* the set of optimal solutions of the true problem (2.2). If $S^* = \{x^*\}$ is a singleton, then $\hat{x}_N$ converges to x^* w.p.1 under mild regularity conditions.

a fixed point $x \in X$, we have that $N^{1/2}\big[\hat{f}_N(x) - f(x)\big]$ converges in distribution to normal $\mathcal{N}(0, \sigma^2(x))$, where $\sigma^2(x) := \mathrm{Var}[F(x,\xi)]$. That is, $\hat{f}_N(x)$ converges to $f(x)$ at a rate of $O_p(N^{-1/2})$. It is possible to show that

$$\hat{v}_N = \min_{x \in S^*} \hat{f}_N(x) + o_p\left(N^{-1/2}\right), \tag{3.2}$$

provided that the set X is compact, [60]. In particular, if $S^* = \{x^*\}$ is singleton, then $\hat{v}_N$ converges to v^* at the same rate as $\hat{f}_N(x^*)$ converges to $f(x^*)$. It is also not difficult to verify that $v^* \geq \mathbb{E}[\hat{v}_N]$, i.e., $\hat{v}_N$ is a downward biased estimator of v^*. If $S^* = \{x^*\}$ is singleton, then the (negative) bias $\mathbb{E}[\hat{v}_N] - v^*$ tends to zero typically at a rate of $O(N^{-1})$, while if the true problem has more than one optimal solution, then this bias is typically of order $O(N^{-1/2})$ (see, e.g., [64] for a further discussion of statistical properties of the SAA estimators).

Although the rate of convergence of the SAA estimators can be enhanced, sometimes significantly, by various variance reduction techniques, the Monte Carlo approach does not allow to estimate the expectation $f(x)$ with a high accuracy. Therefore, it is somewhat surprising that the SAA approach could be quite efficient in solving a certain class of stochastic programming problems.

For $\varepsilon \geq 0$ we say that a point $\bar{x} \in X$ is an ε-optimal solution of problem (2.2) if $f(\bar{x}) \leq v^* + \varepsilon$, i.e., $\bar{x} \in X^*_\varepsilon$, where

$$X^*_\varepsilon := \big\{x \in X : f(x) \leq v^* + \varepsilon\big\} \quad \text{and} \quad v^* := \inf_{x \in X} f(x). \tag{3.3}$$

Note that the level set X^*_ε is nonempty for any $\varepsilon > 0$. Suppose that we solve the SAA problem (3.1) with an accuracy $\delta \in [0, \varepsilon)$. We ask now the question of how large should be the sample size N in order for a δ-optimal solution $\hat{x}_N$ of the SAA problem (3.1) to be an ε-optimal solution of the true problem (2.2). Since the sample is random, we could answer this question only with a certain confidence, say with probability at least $1 - \alpha$, where $\alpha \in (0, 1)$ is a chosen constant called significance level.

Let us make the following assumptions.

(A1) The expected value function $f(x)$ is well defined and finite valued for all $x \in X$.

(A2) There is a constant $\sigma > 0$ such that for any $x, x' \in X$, the moment generating function[6] $M_{x,x'}(t)$ of the random variable $Y_{x,x'} - \mathbb{E}[Y_{x,x'}]$, where $Y_{x,x'} := F(x,\xi) - F(x',\xi)$, satisfies:

$$M_{x,x'}(t) \leq \exp\left(\tfrac{1}{2}\sigma^2 t^2\right), \quad \forall t \in \mathbb{R}. \tag{3.4}$$

(A3) There exists a (measurable) function $\kappa : \Xi \to \mathbb{R}_+$ such that its moment generating function $M_\kappa(t)$ is finite valued for all t in a neighborhood of 0, and

[6] The moment generating function of a random variable Y is $M_Y(t) := \mathbb{E}[\exp(tY)]$.

$$\left|F(x,\xi) - F(x',\xi)\right| \le \kappa(\xi)\|x - x'\|, \ \ \forall x, x' \in X, \ \forall \xi \in \Xi. \tag{3.5}$$

The above assumption (A3) implies that the expectation $\mathbb{E}[\kappa(\xi)]$ is finite and the function $f(x)$ is Lipschitz continuous on X with Lipschitz constant $L = \mathbb{E}[\kappa(\xi)]$. It follows that the optimal value v^* of the true problem (2.2) is finite, provided the set X is bounded (recall that it was assumed that X is nonempty and closed). Moreover, by Cramér's Large Deviation Theorem we have that for any $L' > \mathbb{E}[\kappa(\xi)]$ there exists a positive constant $\beta = \beta(L')$ such that

$$\mathsf{Prob}\left(N^{-1}\sum_{j=1}^{N}\kappa(\xi^j) > L'\right) \le \exp(-N\beta). \tag{3.6}$$

By using theory of Large Deviations it is possible to prove the following result (cf., [27,64,68]).

Theorem 1 *Suppose that assumptions* (A1)–(A3) *hold, the set X has a finite diameter $D := \sup_{x,x' \in X} \|x - x'\|$ and let $\varepsilon > 0$, $\delta \in [0, \varepsilon)$, $\alpha \in (0,1)$, $L' > \mathbb{E}[\kappa(\xi)]$ and $\beta = \beta(L')$ be the corresponding constants. Then for the sample size N satisfying*[7]

$$N \ge \frac{8\sigma^2}{(\varepsilon - \delta)^2}\left[n \log\left(\frac{O(1)LD}{\varepsilon - \delta}\right) + \log\left(\frac{1}{\alpha}\right)\right] \bigvee \left[\beta^{-1}\log\left(\frac{2}{\alpha}\right)\right], \tag{3.7}$$

we have with probability at least $1 - \alpha$ that the following holds: "any δ-optimal solution of the SAA problem (3.1) *is an ε-optimal solution of the true problem* (2.2)".

Remark 1 If in condition (3.4) of assumption (A2), instead for all $t \in \mathbb{R}$, the corresponding inequality is satisfied for all t in a finite interval $[-a, a]$, where $a > 0$ is a given constant, then the estimate (3.7) still holds provided that $\varepsilon - \delta \le a\sigma^2$.

Remark 2 If the set X is finite, then, under assumptions (A1) and (A2), an estimate of the required sample size can be written as

$$N \ge \frac{2\sigma^2}{(\varepsilon - \delta)^2}\log\left(\frac{|X|}{\alpha}\right), \tag{3.8}$$

where $|X|$ denotes the cardinality (number of elements) of the set X (cf., [27]).

Remark 3 If the Lipschitz constant $\kappa(\xi)$ in (3.5) can be taken independent of ξ, i.e., $\kappa(\xi) \equiv L$, then $|F(x,\xi) - F(x',\xi)| \le LD$ for any $x, x' \in X$ and $\xi \in \Xi$. We

[7] By $O(1)$ we denote a generic constant independent of the data, and $a \vee b := \max\{a, b\}$.

have then that the assumption (A2) holds automatically and an estimate of the required sample size takes the form:

$$N \geq \left(\frac{O(1)LD}{\varepsilon - \delta} \right)^2 \left[n \log \left(\frac{O(1)LD}{\varepsilon - \delta} \right) + \log \left(\frac{1}{\alpha} \right) \right]. \tag{3.9}$$

If, for $\xi \in \Xi$, the function $F(\cdot, \xi)$ is Lipschitz continuous, then it is differentiable at every x except for x in a set Υ_ξ of Lebesgue measure zero and if, moreover, the set X is convex, the Lipschitz constant $\kappa(\xi)$ can be estimated by the maximum of $\|\nabla_x F(x, \xi)\|$ taken over $x \in X \setminus \Upsilon_\xi$. Consequently, we can take $L := \sup_{x \in X \setminus \Upsilon_\xi, \xi \in \Xi} \|\nabla_x F(x, \xi)\|$, provided that this constant is finite.

Remark 4 It was assumed in Theorem 1 that the set X has a finite diameter, i.e., that X is bounded. For convex problems this assumption can be relaxed. We say that the problem is *convex* if the set X is convex and the function $F(\cdot, \xi)$ is convex for every $\xi \in \Xi$ and real valued on a neighborhood of X. Then it follows that the expected value function $f(x)$ is also convex. Assume that the optimal value v^* of the true problem (2.2) is finite and for some $a > \varepsilon$ the level set X_a^*, defined in (3.3), has a finite diameter D_a^*. Note that the set X_ε^*, of ε-optimal solutions of the true problem (2.2), remains the same if the feasible set X is replaced by its subset X_a^*. Let N^* be an integer satisfying the inequality (3.7) with D replaced by D_a^*. Then by Theorem 1 we have that, with probability at least $1 - \alpha$, all δ-optimal solutions of the reduced SAA problem, where the set X is replaced by X_a^*, are ε-optimal solutions of the problem (2.2). Let us observe now that in this case the set of δ-optimal solutions of the reduced SAA problem coincides with the set of δ-optimal solutions of the original SAA problem. Indeed, suppose that the original SAA problem has a δ-optimal solution $x^* \in X \setminus X_a^*$. Let $\bar{x} \in \arg\min_{x \in X_a^*} \hat{f}_N(x)$, such a minimizer does exist since X_a^* is compact and $\hat{f}_N(x)$ is real valued convex and hence continuous. Then $\bar{x} \in X_\varepsilon^*$ and $\hat{f}_N(x^*) \leq \hat{f}_N(\bar{x}) + \delta$. By convexity of $\hat{f}_N(x)$ it follows that $\hat{f}_N(x) \leq \max\left\{\hat{f}_N(\bar{x}), \hat{f}_N(x^*)\right\}$ for all x on the segment joining $\bar{x}$ and x^*. This segment has a common point $\hat{x}$ with the set $X_a^* \setminus X_\varepsilon^*$. We obtain that $\hat{x} \in X_a^* \setminus X_\varepsilon^*$ is a δ-optimal solutions of the reduced SAA problem, a contradiction.

That is, with such sample size N^* we are guaranteed with probability at least $1 - \alpha$ that any δ-optimal solution of the SAA problem (3.1) is an ε-optimal solution of the true problem (2.2). Also assumptions (A2) and (A3) could be verified for x, x' in the set X_a^* only.

Remark 5 Suppose that the set S^*, of optimal solutions of the true problem, is nonempty and closed. Then it suffices in the assumption (A2) to verify condition (3.4) only for every $x \in X \setminus X_\varepsilon^*$ and some $x' \in S^*$, which may depend on x (cf., [64, p. 372]). For example, it suffices to verify (3.4) for every $x \in X \setminus X_\varepsilon^*$ and $x' \in \arg\min_{z \in S^*} \|x - z\|$. (Of course, if the set $X \setminus X_\varepsilon^*$ is empty, i.e., $X \subset X_\varepsilon^*$, then any point of X is an ε-optimal solution of the true problem.) If, moreover, $\kappa(\xi) \equiv L$, then for such x and x' we have $|F(x, \xi) - F(x', \xi)| \leq L\bar{D}$, where $\bar{D} := \sup_{x \in X \setminus X_\varepsilon^*} \operatorname{dist}(x, S^*)$. Suppose, further, that the problem is convex. Then (see Remark 4), for any $a > \varepsilon$, we can use X_a^* instead of X and to write the following estimate of the required sample size:

$$N \geq \left(\frac{O(1)L\bar{D}_{a,\varepsilon}}{\varepsilon - \delta}\right)^2 \left[n \log\left(\frac{O(1)LD_a^*}{\varepsilon - \delta}\right) + \log\left(\frac{1}{\alpha}\right)\right], \tag{3.10}$$

where D_a^* is the diameter of X_a^* and $\bar{D}_{a,\varepsilon} := \sup_{x \in X_a^* \setminus X_\varepsilon^*} \operatorname{dist}(x, S^*)$.

Remark 6 In some cases the convergence of optimal solutions of SAA problems is finite in the sense that w.p.1 for N large enough every optimal solution $\hat{x}_N$ of the SAA problem is an exact optimal solution of the true problem and, moreover, the probability of this event tends to one exponentially fast. This happens if the problem is convex and the true problem has sharp optimal solution $x^* \in X$, i.e., $f(x) \geq f(x^*) + c\|x - x^*\|$ for some $c > 0$ and all $x \in X$ (cf., [61,62,70]).

Remark 7 Suppose that $\kappa(\xi) \equiv L$, the problem is convex and the set S^*, of optimal solutions of the true problem, is nonempty and bounded. Then for any $a > 0$ and $\varepsilon \in (0, a)$, we can use the estimate (3.10). Suppose further that for some $\gamma \geq 1$, $c > 0$ and $\bar{a} > 0$, the following growth condition holds

$$f(x) \geq v^* + c\,[\operatorname{dist}(x, S^*)]^\gamma, \quad \forall\, x \in X_{\bar{a}}^*. \tag{3.11}$$

It follows from (3.11) that for any $a \leq \bar{a}$ and $x \in X_a^*$, the inequality $\operatorname{dist}(x, S^*) \leq (a/c)^{1/\gamma}$ holds. Consequently, for any $\varepsilon \in (0, \bar{a})$, by taking $a := \min\{2\varepsilon, \bar{a}\}$ and $\delta \in [0, \varepsilon/2]$ we obtain from (3.10) the following estimate of the required sample size

$$N \geq \left(\frac{O(1)L}{c^{1/\gamma}\varepsilon^{(\gamma-1)/\gamma}}\right)^2 \left[n \log\left(\frac{O(1)LD_a^*}{\varepsilon}\right) + \log\left(\frac{1}{\alpha}\right)\right]. \tag{3.12}$$

Note that if $S^* = \{x^*\}$ is a singleton, then it follows from (3.11) that $D_a^* \leq 2(a/c)^{1/\gamma}$. In particular, if $\gamma = 1$ and $S^* = \{x^*\}$ is a singleton, then D_a^* can be bounded by $4c^{-1}\varepsilon$ and hence we obtain the following estimate

$$N \geq O(1)c^{-2}L^2 \left[n \log\left(O(1)c^{-1}L\right) + \log\left(\alpha^{-1}\right)\right], \tag{3.13}$$

which does not depend on ε. This, of course, is in accordance with Remark 6. For $\gamma = 2$ condition (3.11) is called the "second-order" or "quadratic" growth condition. Under the quadratic growth condition the first term in the right hand side of (3.12) becomes of order $c^{-1}\varepsilon^{-1}L^2$.

Similar analysis can be performed without the condition $\kappa(\xi) \equiv L$, if instead we assume that the estimate (3.4) holds with σ^2 being proportional to $\|x - x'\|^2$. For example, if $F(x, \xi) - F(x', \xi)$ has a normal distribution, then σ^2 is equal to the variance of $F(x, \xi) - F(x', \xi)$, and hence by (3.5) we have that $\sigma^2 \leq \operatorname{Var}[\kappa(\xi)]\|x - x'\|^2$.

Typically, the above estimates of the sample size N are *far too conservative* for actual calculations. For practical applications there are techniques which allow to estimate (statistically) the error of a considered feasible solution $\bar{x}$ for a chosen sample size N (see [31,37]). The intrinsic value of these estimates is theoretical. In a sense, the above estimates of the sample size give an estimate of complexity of a considered class two-stage problems. For decomposition type algorithms the total number of iterations, required to solve the SAA problem, typically is independent of the sample size N (this is an empirical observation) and the computational effort at every iteration is proportional to N. Anyway size of the SAA problem (3.1), formulated in the form (2.4), grows linearly with increase of N. Consequently, e.g., in the case of linear two-stage problems, one can apply known complexity analysis to the obtained SAA problem.

Let us discuss now implications of the estimate (3.7), where we take, for example, $\delta \in [0, \varepsilon/2]$. The right hand side of (3.7) is proportional to σ^2/ε^2. Assumption (A2) requires for the probability distribution of the random variable $Y_{x,x'} := F(x,\xi) - F(x',\xi)$ to have sufficiently light tails. In particular, if $Y_{x,x'}$ has a normal distribution, then actually equality in (3.4) holds with σ^2 being the variance of $Y_{x,x'}$. In a sense, the constant σ^2 in (3.4) can be viewed as a bound reflecting variability of the random variables $Y_{x,x'}$, for $x, x' \in X$. Naturally, larger variability of the data should result in more difficulty in solving the problem. In order to see this, consider a simple case when the feasible set X consists of just two elements, i.e., $X = \{x_1, x_2\}$ with $f(x_2) - f(x_1) > \varepsilon > 0$. By solving the corresponding SAA problem we make the (correct) decision that x_1 is the ε-optimal solution if $\hat{f}_N(x_2) - \hat{f}_N(x_1) > 0$. If the random variable $F(x_2,\xi) - F(x_1,\xi)$ has a normal distribution with mean $\mu = f(x_2) - f(x_1)$ and variance σ^2, then $\hat{f}_N(x_2) - \hat{f}_N(x_1) \sim \mathcal{N}(\mu, \sigma^2/N)$ and the probability of the event "$\hat{f}_N(x_2) - \hat{f}_N(x_1) > 0$" (i.e., of the correct decision) is $\Phi(\mu\sqrt{N}/\sigma)$, where $\Phi(z)$ is the cumulative distribution function of $\mathcal{N}(0,1)$. We have that $\Phi(\varepsilon\sqrt{N}/\sigma) < \Phi(\mu\sqrt{N}/\sigma)$, and in order to make the probability of the incorrect decision less than α we have to take the sample size $N > z_\alpha^2\sigma^2/\varepsilon^2$, where $z_\alpha := \Phi^{-1}(1-\alpha)$. Even if $F(x_2,\xi) - F(x_1,\xi)$ is not normally distributed, the sample size of order σ^2/ε^2 could be justified asymptotically, say by applying the CLT. It also could be mentioned that if $F(x_2,\xi) - F(x_1,\xi)$ has a normal distribution (with known variance), then the uniformly most powerful test for testing $H_0 : \mu \le 0$ versus $H_a : \mu > 0$ is of the form: "reject H_0 if $\hat{f}_N(x_2) - \hat{f}_N(x_1)$ is bigger than a specified critical value" (this is a consequence of the Neyman–Pearson Lemma, see, e.g., [8, Sect. 8.3]). In other words, in such situations if we only have an access to a random sample, then solving the corresponding SAA problem is in a sense a best way to proceed.

The estimate (3.7) suggests complexity of order σ^2/ε^2 with respect to the desirable accuracy. This is in a sharp contrast with deterministic (convex) optimization where complexity usually is bounded in terms of $\log(\varepsilon^{-1})$. In view of the above discussion it should be not surprising that (even linear) two stage stochastic programs usually cannot be solved with a high accuracy. On the other hand, the estimate (3.7) depends *linearly* on the dimension n of the first stage decision vector. It also depends linearly on $\log(\alpha^{-1})$. This means that by increasing confidence, say, from 99 to 99.99% we need to increase the sample size by the factor of $\log 100 \approx 4.6$ at most.

This suggests that by using Monte Carlo sampling techniques one can solve two-stage stochastic programs with a reasonable accuracy, say with relative accuracy of 1% or 2%, in a reasonable time, provided that: (a) its variability is not too large, (b) it has relatively complete recourse, and (c) the corresponding SAA problem can be solved efficiently. And, indeed, this was verified in numerical experiments with two-stage problems having a linear second stage recourse (cf., [30,31,44,58,72]). Also it was demonstrated in theoretical studies and numerical experiments that Quasi-Monte Carlo techniques could significantly improve the accuracy of the SAA method (see, e.g., [36] for a general discussion of Quasi-Monte Carlo methods and [24,28,42] for stochastic programming applications).

The following example (taken from [67]) shows that the estimate (3.7) of the sample size cannot be significantly improved for the class of convex stochastic programs.

Example 1 Consider problem (2.2) with $F(x,\xi) := \|x\|^{2m} - 2m\langle\xi,x\rangle$, where m is a positive integer, and $X := \{x \in \mathbb{R}^n : \|x\| \le 1\}$. Suppose, further, that the random vector ξ has the normal distribution $\mathcal{N}(0,\sigma^2 I_n)$, where σ^2 is a positive constant and I_n is the $n \times n$ identity matrix, i.e., components ξ_i of ξ are independent and $\xi_i \sim \mathcal{N}(0,\sigma^2)$, $i = 1,\dots,n$. It follows that $f(x) = \|x\|^{2m}$, and hence for $\varepsilon \in [0,1]$ the set of ε-optimal solutions of the true problem (2.2) is $\{x : \|x\|^{2m} \le \varepsilon\}$. Now let $\xi^1,\dots,\xi^N$ be an iid random sample of ξ and $\bar{\xi}_N := (\xi^1 + \cdots + \xi^N)/N$. The corresponding sample average function is

$$\hat{f}_N(x) = \|x\|^{2m} - 2m\langle\bar{\xi}_N,x\rangle, \tag{3.14}$$

and the optimal solution $\hat{x}_N$ of the SAA problem is $\hat{x}_N = \|\bar{\xi}_N\|^{-r}\bar{\xi}_N$, where

$$r := \begin{cases} \frac{2m-2}{2m-1}, & \text{if} \quad \|\bar{\xi}_N\| \le 1, \\ 1, & \text{if} \quad \|\bar{\xi}_N\| > 1. \end{cases}$$

It follows that, for $\varepsilon \in (0,1)$, the optimal solution of the corresponding SAA problem is an ε-optimal solution of the true problem iff $\|\bar{\xi}_N\|^\nu \le \varepsilon$, where $\nu := \frac{2m}{2m-1}$. We have that $\bar{\xi}_N \sim \mathcal{N}(0,\sigma^2 N^{-1} I_n)$, and hence $N\|\bar{\xi}_N\|^2/\sigma^2$ has a chi-square distribution with n degrees of freedom. Consequently, the probability that $\|\bar{\xi}_N\|^\nu > \varepsilon$ is equal to the probability $\mathsf{Prob}\left(\chi_n^2 > N\varepsilon^{2/\nu}/\sigma^2\right)$. Moreover, $\mathbb{E}[\chi_n^2] = n$ and the probability $\mathsf{Prob}(\chi_n^2 > n)$ increases and tends to 1/2 as n increases. Consequently, for $\alpha \in (0,0.3)$ and $\varepsilon \in (0,1)$, for example, the sample size N should satisfy

$$N > \frac{n\sigma^2}{\varepsilon^{2/\nu}} \tag{3.15}$$

in order to have the property: "with probability $1-\alpha$ an (exact) optimal solution of the SAA problem is an ε-optimal solution of the true problem". Compared with (3.7), the lower bound (3.15) also grows linearly in n and is proportional to $\sigma^2/\varepsilon^{2/\nu}$. It remains to note that the constant ν decreases to one as m increases.

Note that in this example the growth condition (3.11) holds with $\gamma = 2m$, and the power constant of ε in the estimate (3.15) is in accordance with the estimate (3.12).

Of course, in this example the "true" optimal solution is $\bar{x} = 0$, and one does not need sampling in order to solve this problem. Note, however, that the sample average function $\hat{f}_N(x)$ here depends on the random sample only through the data average vector $\bar{\xi}_N$. Therefore, any numerical procedure based on averaging a generated random sample, will need a sample of size N satisfying the estimate (3.15) in order to produce an ε-optimal solution. □

It follows from assumption (A1) that for any $x \in X$ the optimal value $F(x, \xi)$ of the second stage problem is finite for a.e. $\xi \in \Xi$, i.e., that the considered two-stage problem has relatively complete recourse. This assumption was essential in the above analysis. Of course, one can make sure at the modelling stage that the relatively complete recourse holds. For example, suppose that the feasible set of the second stage problem is given in the form

$$\mathcal{G}(x, \xi) := \left\{ y \in \mathbb{R}^{n_2} : G(x, y, \xi) \in C \right\}, \tag{3.16}$$

where $G : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \Xi \to \mathbb{R}^m$ and $C \subset \mathbb{R}^m$ is a closed convex cone with a nonempty interior. Let e be an interior point of C and $\pi > 0$ be a constant, and consider the following modification of the second stage problem (2.1):

$$\begin{array}{cl} \underset{y \in \mathbb{R}^{n_2}, t \in \mathbb{R}}{\text{Min}} & g(x, y, \xi) + \pi t \\ \text{s.t} & G(x, y, \xi) + te \in C, \; t \geq 0. \end{array} \tag{3.17}$$

This modification makes the second stage problem always feasible (just take $t > 0$ large enough such that $e + t^{-1} G(x, y, \xi) \in C$). By taking the penalty parameter π sufficiently large, one may hope that solving the modified two-stage problem, with the second stage problem (2.1) replaced by (3.17), will lead to "nearly feasible and nearly optimal" solution of the original two-stage problem (2.1)–(2.2). Note, however, that the variance of the optimal value $F_\pi(x, \xi)$ of the second stage problem (3.17) grows with increase of π in a way more or less proportional[8] to π. Therefore, for large values of the penalty parameter π one may need an unrealistically large sample size N in order to solve the corresponding modified problem with a reasonable accuracy.

In general, the considered approach of two-stage stochastic programming with recourse is not suitable to handle situations where certain events can happen with very small probabilities but with huge costs. It does not make much

[8] If a random variable Y can take two values, say a and b with respective probabilities $1 - p$ and p, then for small $p > 0$ and large b such that the expected value $\mathbb{E}[Y] = (1 - p)a + pb$ remains constant, the variance of Y grows asymptotically proportionally to b.

sense to mix such catastrophic events with regular events trying to optimize the cost on average.

4 Risk averse approach

In this section we discuss questions (i′) and (i″) posed in the Introduction. Suppose that we do not know the corresponding probability distribution P exactly, but we could reasonably identify a relevant family $\mathcal{A}$ of probability distributions. Then we can reformulate problem (2.2) as the following minimax stochastic program:

$$\operatorname*{Min}_{x\in X}\left\{f(x) := \sup_{P\in\mathcal{A}} \mathbb{E}_P[F(x,\xi)]\right\}, \tag{4.1}$$

by hedging against a worst possible distribution (e.g., [13,19,22,63,75]). For example, it could be easier to evaluate some moments $\mathbb{E}_P[\psi_i(\xi)]$, $i = 1,\dots,m$, of ξ, than its complete distribution. This corresponds to the so-called Problem of Moments, where the set $\mathcal{A}$ is defined by specifying equality and/or inequality type constraints on these moments. Usually such moment constraints lead to extreme (worst case) distributions having a finite support of at most $m+1$ points, and from a practical point of view could be too loose. For some other approaches to minimax stochastic programming see, e.g., [66,69].

As far as question (i″) is concerned one can try to reach a compromise between optimizing the objective on average and at the same time reducing its variability. That is, consider the following mean-risk averse problem:

$$\operatorname*{Min}_{x\in X}\left\{f(x) := \rho[F(x,\xi)]\right\}, \tag{4.2}$$

where $\rho[Z] := \mathbb{E}[Z] + \lambda\mathbb{D}[Z]$, $\lambda \geq 0$ is a weight parameter and $\mathbb{D}[Z]$ is a measure of dispersion (variability) of random variable $Z = Z(\xi)$ (cf., [38]). It seems natural to use variance $\mathrm{Var}[Z]$ or standard deviation $\sqrt{\mathrm{Var}[Z]}$ as the dispersion measure $\mathbb{D}[Z]$. Such choice of the dispersion measure was suggested by Markowitz [32], more than 50 years ago, and was extensively used for portfolio selections. As the following example shows, however, there is a certain problem with using the corresponding risk measure for stochastic programming.

Example 2 Suppose that the space $\Xi = \{\xi_1,\xi_2\}$ consists of two points with associated probabilities p and $1-p$, for some $p \in (0,1)$. Consider dispersion measure $\mathbb{D}[Z]$, defined on the space of functions (random variables) $Z : \Xi \to \mathbb{R}$, either of the form $\mathbb{D}[Z] := \sqrt{\mathrm{Var}[Z]}$ or $\mathbb{D}[Z] := \mathrm{Var}[Z]$, and the corresponding $\rho[Z] := \mathbb{E}[Z]+\lambda\mathbb{D}[Z]$. Consider functions $Z_1, Z_2 : \Xi \to \mathbb{R}$ defined $Z_1(\xi_1) = -a$ and $Z_1(\xi_2) = 0$, where a is some positive number, and $Z_2(\xi_1) = Z_2(\xi_2) = 0$. Now, for $\mathbb{D}[Z] := \sqrt{\mathrm{Var}[Z]}$, we have that $\rho[Z_2] = 0$ and $\rho[Z_1] = -pa+\lambda a\sqrt{p(1-p)}$. It follows that for any $\lambda > 0$ and $p < (1+\lambda^{-2})^{-1}$ we have that $\rho[Z_1] > \rho[Z_2]$. Similarly, for $\mathbb{D}[Z] := \mathrm{Var}[Z]$ we have that $\rho[Z_1] > \rho[Z_2]$ if $a > \lambda^{-1}$ and

$p < [1-(\lambda a)^{-1}]^{-1}$. That is, although Z_2 dominates Z_1 in the sense that $Z_2(\xi) \geq Z_1(\xi)$ for *every* possible realization of $\xi \in \Xi$, we have here that $\rho[Z_1] > \rho[Z_2]$.

Consider now an optimization problem of the form (4.2) with

$$X := \left\{x = (x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 = 1,\ x_1 \geq 0,\ x_2 \geq 0\right\}$$

and $F(x,\xi) := x_1 Z_1(\xi) + x_2 Z_2(\xi)$. Let $\bar{x} := (1,0)$ and $x^* := (0,1)$. Note that the set X is formed by vectors $t\bar{x} + (1-t)x^*$, $t \in [0,1]$. We have here that $F(x,\xi) = x_1 Z_1(\xi)$, and hence $F(\bar{x},\xi)$ is dominated by $F(x,\xi)$ for any $x \in X$. And yet $\bar{x}$ is not an optimal solution of the corresponding optimization (minimization) problem since $\rho[F(\bar{x},\xi)] = \rho[Z_1]$ is greater than $\rho[F(x^*,\xi)] = \rho[Z_2]$.

It turns our that there is a duality relation between the minimax (4.1) and risk averse (4.2) formulations of stochastic programs. We view now risk measure $\rho[Z]$ as a mapping assigning to a (measurable) function $Z : \Xi \to \mathbb{R}$ a real number. For technical reasons we need to define a space of functions $Z(\xi)$ for which $\rho[Z]$ is defined. We assume that there is a reference probability measure (distribution) P on Ξ and for $p \in [1, +\infty)$ consider the space $\mathcal{Z} := L_p(\Xi, \mathcal{F}, P)$ of random variables $Z(\xi)$ having finite p-th order moments.[9] Consider the following conditions (axioms) associated with a risk measure (function) $\rho : \mathcal{Z} \to \mathbb{R}$.

(C1) *Convexity*:

$$\begin{aligned}\rho(\alpha Z_1 + (1-\alpha)Z_2) &\leq \alpha\rho(Z_1) \\ &+(1-\alpha)\rho(Z_2) \quad \text{for all } Z_1, Z_2 \in \mathcal{Z} \text{ and } \alpha \in [0,1].\end{aligned}$$

(C2) *Monotonicity*:[10] If $Z_1, Z_2 \in \mathcal{Z}$ and $Z_2 \succeq Z_1$, then $\rho(Z_2) \geq \rho(Z_1)$.

(C3) *Translation Equivariance*: If $a \in \mathbb{R}$ and $Z \in \mathcal{Z}$, then $\rho(Z + a) = \rho(Z) + a$.

(C4) *Positive homogeneity*: If $\alpha > 0$ and $Z \in \mathcal{Z}$, then $\rho(\alpha Z) = \alpha\rho(Z)$.

The above axioms were introduced and risk measures satisfying these axioms were called *coherent risk measures* in the pioneering paper by Artzner et al [2] (for a discussion of a relation between axiomatics of risk and dispersion measures see [50,51]).

Recall that with Banach space $\mathcal{Z} := L_p(\Xi, \mathcal{F}, P)$ is associated its dual space $\mathcal{Z}^* := L_q(\Xi, \mathcal{F}, P)$, where $q \in (1, +\infty]$ is such that $1/p + 1/q = 1$, with the corresponding scalar product

$$\langle \zeta, Z \rangle := \int_{\Xi} Z(\xi)\zeta(\xi) dP(\xi), \quad Z \in \mathcal{Z},\ \zeta \in \mathcal{Z}^*.$$

[9] Here $(\Xi, \mathcal{F}, P)$ is viewed as a probability space. If Ξ is a subset of $\mathbb{R}^d$, then we assume that the sigma algebra $\mathcal{F}$ is formed by Borel subsets of Ξ. The space $L_p(\Xi, \mathcal{F}, P)$ consists from classes of $\mathcal{F}$-measurable functions $Z : \Xi \to \mathbb{R}$ such that $\mathbb{E}_P|Z|^p < +\infty$, which can differ from each other on a set of P-measure zero.

[10] The notation $Z_2 \succeq Z_1$ means that $Z_2(\xi) \geq Z_1(\xi)$ for all $\xi \in \Xi$. Since we deal here with $Z_1(\xi)$ and $Z_2(\xi)$ viewed as random variables defined on the probability space $(\Xi, \mathcal{F}, P)$, we can relax this to requiring $Z_2(\xi) \geq Z_1(\xi)$ to hold for a.e. $\xi \in \Xi$.

Note that $\zeta(\xi)dP(\xi)$ can be considered as a probability measure (distribution) on $(\Xi,\mathcal{F})$, provided that $\zeta(\cdot)$ is a *probability density* function, i.e., $\zeta(\xi) \geq 0$ for all $\xi \in \Xi$ and $\int_\Xi \zeta dP = 1$. In that case we write $\langle\zeta,Z\rangle = \mathbb{E}_\zeta[Z]$, viewing this scalar product as the expectation of Z with respect to the probability density ζ. The following duality result is a consequence of the Fenchel–Moreau Theorem. With various degrees of generality it was obtained in [2,10,12,21,43,50,56].

Theorem 2 *Let $\mathcal{Z} := L_p(\Xi,\mathcal{F},P)$, $\mathcal{Z}^* := L_q(\Xi,\mathcal{F},P)$, with $p \in [1,+\infty)$, and $\rho : \mathcal{Z} \to \mathbb{R}$ be a risk measure. Then ρ is a coherent risk measure* (*i.e., satisfy conditions* (C1)–(C4)) *if and only if there exists a convex set $\mathcal{A} \subset \mathcal{Z}^*$ of probability density functions such that*

$$\rho[Z] = \sup_{\zeta\in\mathcal{A}} \mathbb{E}_\zeta[Z], \quad \forall Z \in \mathcal{Z}. \tag{4.3}$$

The representation (4.3) shows that for coherent risk measures, in a sense, formulations (4.1) and (4.2) are dual to each other.

Let us make the following observations. It is possible to show that convexity assumption (C1) and monotonicity assumption (C2) imply that the (real valued) function ρ is continuous in the norm topology of $L_p(\Xi,\mathcal{F},P)$ (cf., [56]). The monotonicity assumption (condition (C2)) is important in several respects. If it does not hold, then we may end up, while solving (4.2), in a situation where $F(x_2,\xi) \geq F(x_1,\xi)$ for all possible realizations of ξ, and yet we prefer decision x_2 to x_1 (see Example 2). Also, conditions (C1) and (C2) imply that if $F(\cdot,\xi)$ is convex for every $\xi \in \Xi$, then the corresponding composite function $f(x) = \rho[F(x,\xi)]$ is also convex. For this, convexity preserving, property to hold the condition of monotonicity is essential.

Remark 8 We can define the space $\mathcal{Z}$ to be the space of all bounded measurable functions $Z : \Xi \to \mathbb{R}$, and to pair this space with the space $\mathcal{Z}^*$ of all signed finite Borel measures on Ξ with the corresponding scalar product $\langle\mu,Z\rangle := \int_\Xi Z(\xi)d\mu(\xi)$, $\mu \in \mathcal{Z}^*$, $Z \in \mathcal{Z}$. Then the result of Theorem 2 holds with $\mathcal{A} \subset \mathcal{Z}^*$ being a (convex) set of probability measures and the expectation $\mathbb{E}_\mu[Z]$, in the right hand side of (4.3), is taken with respect to probability measure $\mu \in \mathcal{A}$. Suppose, further, that the set $\Xi \subset \mathbb{R}^d$ is compact and let us take $\mathcal{A}$ to be the set of *all* probability measures on Ξ. Then the maximum of $\mathbb{E}_\mu[Z]$, over $\mu \in \mathcal{A}$, is attained at a measure of mass one at a point $a \in \Xi$. That is, in that case the representation (4.3) takes the form $\rho[Z] = \sup_{a\in\Xi} Z(a)$, and problem (4.2) can be written as the min-max optimization problem:

$$\operatorname*{Min}_{x\in X} \left\{ f(x) := \sup_{a\in\Xi} [F(x,a)] \right\}, \tag{4.4}$$

In case the set $\Xi := \{\xi_1,\dots,\xi_K\}$ is finite, we can view a function $Z : \Xi \to \mathbb{R}$ as a vector $(Z(\xi_1),\dots,Z(\xi_K)) \in \mathbb{R}^K$. Moreover, if Ξ is equipped with sigma algebra $\mathcal{F}$ of all subsets of Ξ, then we can identify $L_p(\Xi,\mathcal{F},P)$ with $\mathbb{R}^K$, equipped with the corresponding ℓ_p-norm. Let $F(x,\xi)$ be the optimal value of the second stage problem (2.1). Suppose that $F(x,\xi)$ is finite for every $x \in X$ and $\xi \in \Xi$, i.e., the problem has relatively complete recourse. Then the function $\rho[F(x,\xi)]$ is well defined and problem (4.2) can be considered as a two-stage risk averse stochastic problem. By making one copy of the second stage decision vector for every scenario ξ_k (compare with (2.4)), we can write this two-stage problem in the form:

$$\begin{array}{cl} \underset{x,y_1,\dots,y_K}{\text{Min}} & \rho\left[\left(g(x,y_1,\xi_1),\dots,g(x,y_K,\xi_K)\right)\right] \\ \text{s.t.} & x \in X,\ y_k \in \mathcal{G}(x,\xi_k),\ k=1,\dots,K. \end{array} \tag{4.5}$$

In particular, in the case of linear second stage problem (2.3), the above formulation of two-stage risk averse problem takes the form:

$$\begin{array}{cl} \underset{x,y_1,\dots,y_K}{\text{Min}} & \langle c,x\rangle + \rho\left[\left(\langle q_1,y_1\rangle,\dots,\langle q_K,y_K\rangle\right)\right] \\ \text{s.t.} & Ax+b \le 0,\ T_kx+W_ky_k+h_k \le 0,,\ k=1,\dots,K. \end{array} \tag{4.6}$$

We would like to emphasize that the monotonicity condition (C2) is essential in verification of the equivalence of formulations (4.2) and (4.5) of the corresponding two-stage problem.

Consider the mean-deviation risk function $\rho[Z] := \mathbb{E}[Z] + \lambda\sqrt{\mathrm{Var}[Z]}$. Here $\lambda \ge 0$ is a weight parameter and all expectations are taken with respect to the reference distribution P. It is natural to assume here existence of second order moments, i.e., that $Z \in L_2(\Xi,\mathcal{F},P)$. It turns out that this risk measure satisfies conditions (C1),C(3) and (C4), but for $\lambda > 0$ not condition (C2) (see Example 2). This in turn may result in suboptimality of solutions of the associated two-stage programs and that the corresponding formulation (4.5) is not equivalent (even in the linear case (4.6)) to the original problem (4.2) (cf., Takriti and Ahmed [71]).

A class of coherent risk measures is given by mean-semideviation risk functions:

$$\rho[Z] := \mathbb{E}[Z] + \lambda\left(\mathbb{E}\left\{\left[Z-\mathbb{E}(Z)\right]_+^p\right\}\right)^{1/p}, \tag{4.7}$$

where $p \in [1,+\infty)$, $[z]_+^p := (\max\{z,0\})^p$ and all expectations are taken with respect to the reference distribution P. For any $\lambda \ge 0$ these risk functions satisfy conditions (C1),C(3),C(4), and for $\lambda \in [0,1]$ also the monotonicity condition (C2), i.e., for $\lambda \in [0,1]$ these are coherent risk measures. Another important class of coherent risk measures is

$$\rho[Z] := \mathbb{E}[Z] + \inf_{t\in\mathbb{R}} \mathbb{E}\left\{a_1[t-Z]_+ + a_2[Z-t]_+\right\}, \tag{4.8}$$

where $a_1 \in [0,1]$ and $a_2 \ge 0$ are constants. It is natural to use for these risk measures the space $\mathcal{Z} := L_1(\Xi,\mathcal{F},P)$ together with its dual space $\mathcal{Z}^* = L_\infty(\Xi,\mathcal{F},P)$. The dual representation (4.3) then holds and the corresponding set $\mathcal{A} \subset \mathcal{Z}^*$ can be written in the form:

$$\mathcal{A} = \left\{\zeta \in \mathcal{Z}^* : 1-a_1 \le \zeta(\xi) \le 1+a_2,\ \text{a.e. } \xi \in \Xi,\quad \mathbb{E}[\zeta]=1\right\} \tag{4.9}$$

(again all expectations here are taken with respect to the reference measure P). For $a_1 \in [0,1]$ and $a_2 \geq 0$ all densities in the right hand side of (4.9) are nonnegative, and the risk function defined in (4.8) is a coherent risk measure. We can write this risk measure in the form

$$\rho[Z] = (1 - a_1)\mathbb{E}[Z] + a_1 CV@R_\kappa[Z],$$

where $\kappa := a_2/(a_1 + a_2)$ and

$$CV@R_\kappa[Z] := \inf_{t \in \mathbb{R}} \left\{ t + \frac{1}{1-\kappa} \mathbb{E}\big([Z - t]_+\big) \right\} \tag{4.10}$$

is the so-called Conditional Value at Risk function (see Rockafellar and Uryasev [49]). For many other examples of risk functions satisfying some/all conditions (C1)–(C4), their dual representations and their subdifferentials we may refer to [55].

Now the question is how the computational complexity of the risk averse problem (4.2) is compared with complexity of the expected value problem (2.2). For risk functions of the form (4.7) or (4.8) the SAA method can be applied to problem (4.2) in a straightforward way with similar sample size estimates. That is, for a generated sample $\xi^1, \ldots, \xi^N \sim P$, replace the reference distribution P by its empirical (sample) approximation[11] $\hat{P}_N := N^{-1} \sum_{j=1}^N \Delta(\xi^j)$, i.e., replace the corresponding expectations by their sample averages. In that respect the risk measure (4.8) is especially convenient. For this risk measure the corresponding problem (4.2) takes the form:

$$\operatorname*{Min}_{x \in X, t \in \mathbb{R}} \left\{ f(x) := \mathbb{E}\big[F(x,\xi) + a_1[t - F(x,\xi)]_+ + a_2[F(x,\xi) - t]_+\big] \right\}, \tag{4.11}$$

and for a finite (not too large) number of scenarios can be numerically solved by decomposition techniques (see, e.g., [66] for details and numerical experiments).

Example 3 (Newsvendor Problem) A newsvendor has to decide about quantity x of newspapers which he purchases from a distributor at the beginning of a day at the cost of c per unit. He can sell a newspaper at the price s per unit and unsold newspapers can be returned to the vendor at the price of r per unit. It is assumed that $0 \leq r < c < s$. If the demand D for the newspapers, at a particular day, turns out to be greater than or equal to the order quantity x, then he makes the profit $sx - cx = (s - c)x$, while if D is less than x, his profit is $sD + r(x - D) - cx = (r - c)x + (s - r)D$. Thus the profit is a function of x and D and is given by

$$F(x, D) = [(s-c)x]\,\delta(D - x) + [(r-c)x + (s-r)D]\,\delta(x - D),$$

[11] By $\Delta(\xi)$ we denote probability measure (distribution) of mass one at the point ξ.

where $\delta(t) = 0$ if $t < 0$, and $\delta(t) = 1$ if $t \geq 0$. The objective of the newsvendor is to maximize his profit. Viewing the demand D as uncertain, the risk averse formulation (4.2) of the corresponding optimization problem can be written here as follows:

$$\operatorname*{Min}_{x \geq 0} \left\{ f(x) := \rho[-F(x,D)] \right\}. \tag{4.12}$$

Note that in order to formulate this as a minimization, rather than maximization, problem we used negative of the profit function. Let G be a reference distribution of the demand D supported on $\Xi := \mathbb{R}_+$. We can view G as a cumulative distribution function (cdf) supported on $\mathbb{R}_+$, i.e., $G(t) = 0$ for any $t < 0$. Note that $F(x,D)$ is piecewise linear in D. Therefore we can use the space $\mathcal{Z} := L_1(\mathbb{R}_+, \mathcal{F}, G)$. Furthermore, assuming that $\rho : \mathcal{Z} \to \mathbb{R}$ is a coherent risk measure, we have by Theorem 2 that there is a set $\mathcal{A} \subset \mathcal{Z}^* = L_\infty(\mathbb{R}_+, \mathcal{F}, G)$ such that the representation (4.3) holds.

Using integration by parts it is not difficult to verify that

$$\mathbb{E}_G[-F(x,D)] = (c-s)x + (s-r) \int_0^x G(t)dt.$$

It is also possible to show that with every coherent risk measure $\rho : \mathcal{Z} \to \mathbb{R}$ is associated a cdf $\bar{G}(t)$, supported on $\mathbb{R}_+$ and independent of the parameters r, c, s, such that

$$\rho[-F(x,D)] = (c-s)x + (s-r) \int_0^x \bar{G}(t)dt \tag{4.13}$$

(cf., [1,63]). In particular, for the (coherent) risk measure (4.8) it is possible to show, by using its dual representation with the set $\mathcal{A}$ given in (4.9), that $\bar{G}(t) = \min\{(1+a_2)G(t), 1\}$. Consequently, for this risk measure an optimal solution of problem (4.12) is given by the quantile $\bar{x} = G^{-1}(\gamma/(1+a_2))$, where $\gamma := (s-c)/(s-r)$. Naturally, with increase of the uncertainty parameter a_2, the corresponding (conservative) decision $\bar{x}$ is monotonically decreasing. □

5 Multistage stochastic programming

We discuss now stochastic programming in a dynamic setting when decisions are made in several, say T, stages depending on information available at a current stage $t = 1, \ldots, T$. In a generic form a T-stage stochastic programming problem can be written in the following nested formulation

$$\operatorname*{Min}_{x_1 \in \mathcal{G}_1} F_1(x_1) + \mathbb{E}\left[\inf_{x_2 \in \mathcal{G}_2(x_1,\xi_2)} F_2(x_2,\xi_2) + \mathbb{E}\left[\cdots + \mathbb{E}\left[\inf_{x_T \in \mathcal{G}_T(x_{T-1},\xi_T)} F_T(x_T,\xi_T) \right]\right]\right], \tag{5.1}$$

driven by the random data process $\xi_2,\dots,\xi_T$. Here $x_t \in \mathbb{R}^{n_t}$, $t=1,\dots,T$, are decision variables, $F_t : \mathbb{R}^{n_t} \times \mathbb{R}^{d_t} \to \mathbb{R}$ are continuous functions and $\mathcal{G}_t : \mathbb{R}^{n_{t-1}} \times \mathbb{R}^{d_t} \rightrightarrows \mathbb{R}^{n_t}$, $t=2,\dots,T$, are measurable closed valued multifunctions, the function $F_1 : \mathbb{R}^{n_1} \to \mathbb{R}$ and the set $\mathcal{G}_1 \subset \mathbb{R}^{n_1}$ are deterministic. For example, in the linear case: $F_t(x_t,\xi_t) := \langle c_t, x_t\rangle$, $\mathcal{G}_1 := \{x_1 : A_1x_1 = b_1,\ x_1 \geq 0\}$,

$$\mathcal{G}_t(x_{t-1},\xi_t) := \left\{x_t : B_tx_{t-1} + A_tx_t = b_t,\ x_t \geq 0\right\},\ t=2,\dots,T,$$

$\xi_1 := (c_1, A_1, b_1)$ is known at the first stage (and hence is nonrandom), and $\xi_t := (c_t, B_t, A_t, b_t) \in \mathbb{R}^{d_t}$, $t=2,\dots,T$, are data vectors some (all) elements of which can be random.

There are several equivalent ways how this formulation can be made precise. One approach is to consider decision variables $x_t = x_t(\xi_{[t]})$, $t=1,\dots,T$, as functions of the data process $\xi_{[t]} := (\xi_1,\dots,\xi_t)$ up to time t. Such a sequence of (measurable) mappings $x_t(\xi_{[t]})$, $t=1,\dots,T$, is called an *implementable policy* (recall that ξ_1 is deterministic). An implementable policy is said to be feasible if it satisfies the feasibility constraints, i.e.,

$$x_t(\xi_{[t]}) \in \mathcal{G}_t(x_{t-1}(\xi_{[t-1]}),\xi_t),\ t=2,\dots,T,\ \text{w.p.1.} \tag{5.2}$$

We can formulate the multistage problem (5.1) in the form

$$\begin{array}{cl} \underset{x_1,x_2(\cdot),\dots,x_T(\cdot)}{\text{Min}} & \mathbb{E}\left[F_1(x_1) + F_2(x_2(\xi_{[2]}),\xi_2) + \cdots + F_T\left(x_T(\xi_{[T]}),\xi_T\right)\right] \\ \text{s.t.} & x_1 \in \mathcal{G}_1,\ x_t(\xi_{[t]}) \in \mathcal{G}_t(x_{t-1}(\xi_{[t-1]}),\xi_t), \quad t=2,\dots,T. \end{array} \tag{5.3}$$

Note that optimization in (5.3) is performed over implementable and feasible policies.

Another possible way is to write the corresponding dynamic programming equations. That is, consider the last stage problem

$$\underset{x_T \in \mathcal{G}_T(x_{T-1},\xi_T)}{\text{Min}} F_T(x_T,\xi_T). \tag{5.4}$$

The optimal value of this problem, denoted $Q_T(x_{T-1},\xi_T)$, depends on the decision vector x_{T-1} and data ξ_T. At stage $t=2,\dots,T-1$, we write the problem:

$$\begin{array}{l} \underset{x_t}{\text{Min}}\ F_t(x_t,\xi_t) + \mathbb{E}\left\{Q_{t+1}\left(x_t,\xi_{[t+1]}\right)\middle|\xi_{[t]}\right\} \\ \text{s.t.}\ \ x_t \in \mathcal{G}_t(x_{t-1},\xi_t), \end{array} \tag{5.5}$$

where $\mathbb{E}\left[\,\cdot\,|\xi_{[t]}\right]$ denotes conditional expectation. Its optimal value depends on the decision x_{t-1} at the previous stage and realization of the data process $\xi_{[t]}$, and denoted $Q_t\left(x_{t-1},\xi_{[t]}\right)$. The idea is to calculate the (so-called *cost-to-go* or *value*) functions $Q_t\left(x_{t-1},\xi_{[t]})\right)$, recursively, going backward in time. At the first stage we finally need to solve the problem:

$$\underset{x_1 \in \mathcal{G}_1}{\text{Min}}\ F_1(x_1) + \mathbb{E}\left[Q_2\left(x_1,\xi_2\right)\right]. \tag{5.6}$$

The corresponding dynamic programming equations are

$$Q_t\left(x_{t-1}, \xi_{[t]}\right) = \inf_{x_t \in \mathcal{G}_t(x_{t-1}, \xi_t)} \left\{ F_t(x_t, \xi_t) + \mathcal{Q}_{t+1}\left(x_t, \xi_{[t]}\right) \right\}, \tag{5.7}$$

where $\mathcal{Q}_{t+1}\left(x_t, \xi_{[t]}\right) := \mathbb{E}\left\{ Q_{t+1}\left(x_t, \xi_{[t+1]}\right) \middle| \xi_{[t]} \right\}$. We have that an implementable policy $\bar{x}_t(\xi_{[t]})$ is optimal iff

$$\bar{x}_t(\xi_{[t]}) \in \arg\min_{x_t \in \mathcal{G}_t(x_{t-1}, \xi_t)} \left\{ F_t(x_t, \xi_t) + \mathcal{Q}_{t+1}\left(x_t, \xi_{[t]}\right) \right\}, \quad t = 1, \ldots, T, \ \text{ w.p.1.} \tag{5.8}$$

If the random process is *Markovian* (i.e., the conditional distribution of ξ_{t+1} given $\xi_{[t]} = (\xi_1, \ldots, \xi_t)$ is the same as the conditional distribution of ξ_{t+1} given ξ_t), then $Q_t\left(x_{t-1}, \xi_t\right)$ is a function of x_{t-1} and ξ_t. We say that the process $\xi_1, \ldots, \xi_T$ is *between stages independent* if ξ_{t+1} is independent of $\xi_{[t]}$ for $t = 1, \ldots, T-1$. In that case $\mathbb{E}\left[Q_{t+1}\left(x_t, \xi_{t+1}\right) \middle| \xi_t\right] = \mathcal{Q}_{t+1}(x_t)$ does not depend on ξ_t.

In some specific cases it is possible to solve these dynamic programming equations either analytically or numerically. However, more often than not it is quite impossible to solve these equations as they are. Let us consider the following example of portfolio selection. This example is sufficiently simple so to some extent it can be analyzed analytically. We will also use this example later to demonstrate some ideas and difficulties associated with multistage stochastic programming.

Example 4 (Portfolio Selection) Suppose that we want to invest an amount of W_0 in n assets, x_i, $i = 1, \ldots, n$, in each. Suppose, further, that we can rebalance our portfolio at several, say T, periods of time. That is, at the beginning we choose values $x_{i0} \geq 0$ of our assets subject to the budget constraint $\sum_{i=1}^{n} x_{i0} = W_0$. At the period $t = 1, \ldots, T$, our wealth is $W_t = \sum_{i=1}^{n} \xi_{it} x_{i,t-1}$, where $\xi_{it} = (1 + R_{it})$ and R_{it} is the return of the i-th asset at the period t. Our objective is to maximize the expected utility

$$\operatorname{Max} \mathbb{E}\left[U(W_T)\right] \tag{5.9}$$

at the end of the considered period, subject to the balance constraints

$$\sum_{i=1}^{n} x_{it} = W_t \quad \text{and} \quad x_t \geq 0, \ t = 0, \ldots, T-1. \tag{5.10}$$

We use notation $x_t := (x_{1t}, \ldots, x_{nt})$ and $\xi_t := (\xi_{1t}, \ldots, \xi_{nt})$, and as before $\xi_{[t]} := (\xi_1, \ldots, \xi_t)$ for the history of the data process up to time t. The values of the decision vector x_t, chosen at stage t, may depend on the information $\xi_{[t]}$ available up to time t, but not on the future observations. The decision process has the form

$$\text{decision}(x_0) \rightsquigarrow \text{observation}(\xi_1) \rightsquigarrow \text{decision}(x_1)$$
$$\rightsquigarrow \cdots \rightsquigarrow \text{observation}(\xi_T) \rightsquigarrow \text{decision}(x_T).$$

In order to derive dynamic programming equations consider the last stage $t = T-1$. At that stage we have to solve the problem

$$\begin{array}{ll}\underset{x_{T-1}\ge 0}{\text{Max}} & \mathbb{E}\left[U\left(\sum_{i=1}^n \xi_{i,T}x_{i,T-1}\right)\big|\,\xi_{[T-1]}\right] \\ \text{s.t.} & \sum_{i=1}^n x_{i,T-1} = W_{T-1}.\end{array} \tag{5.11}$$

Its optimal value $Q_{T-1}\left(W_{T-1},\xi_{[T-1]}\right)$ is a function of W_{T-1} and $\xi_{[T-1]}$. At stage $t = T-2,\dots,0$ we need to solve

$$\begin{array}{ll}\underset{x_t\ge 0}{\text{Max}} & \mathbb{E}\left[Q_{t+1}\left(\sum_{i=1}^n \xi_{i,t+1}x_{it},\xi_{[t+1]}\right)\big|\,\xi_{[t]}\right] \\ \text{s.t.} & \sum_{i=1}^n x_{it} = W_t.\end{array} \tag{5.12}$$

Its optimal value is $Q_t\left(W_t,\xi_{[t]}\right)$. Note that if the process ξ_t is between stages independent, then the value function $Q_t(W_t)$ is independent of $\xi_{[t]}$ and is a function of one variable W_t.

Consider the logarithmic utility function $U(z) := \log z$. Then, for $W_{T-1} > 0$,

$$Q_{T-1}\left(W_{T-1},\xi_{[T-1]}\right) = \log W_{T-1} + Q_{T-1}\left(1,\xi_{[T-1]}\right),$$

and by induction

$$Q_1\left(W_1,\xi_1\right) = \log W_1 + \mathbb{E}\left[Q_1\left(1,\xi_1\right)\right] + \sum_{t=2}^{T-1}\mathbb{E}\left[Q_t\left(1,\xi_{[t]}\right)\big|\,\xi_{[t-1]}\right]. \tag{5.13}$$

Consequently, the first stage optimal solution is obtained by solving the problem:

$$\underset{x_0\ge 0}{\text{Max}}\,\mathbb{E}\left[\log\left(\sum_{i=1}^n \xi_{i1}x_{i0}\right)\right] \text{ s.t. } \sum_{i=1}^n x_{i0} = W_0. \tag{5.14}$$

That is, the first stage optimal solution can be obtained in a *myopic* way by solving the (static) problem (5.14). The optimal value v^* of the corresponding multistage problem is

$$v^* = Q_0(W_0) + \mathbb{E}\left[Q_1\left(1,\xi_1\right)\right] + \sum_{t=2}^{T-1}\mathbb{E}\left[Q_t\left(1,\xi_{[t]}\right)\big|\,\xi_{[t-1]}\right]. \tag{5.15}$$

Consider now the power utility function $U(z) := z^\gamma$, where $\gamma \le 1$, and suppose that the random process ξ_t is between stages independent. Then $Q_{T-1}\left(W_{T-1}\right) = W_{T-1}^\gamma Q_{T-1}(1)$, and by induction $Q_1(W_1) = W_1^\gamma \prod_{t=1}^{T-1} Q_t(1)$. Consequently, the first stage optimal solution is obtained in a *myopic* way by solving the problem:

$$\underset{x_0\ge 0}{\text{Max}}\,\mathbb{E}\left[\left(\sum_{i=1}^n \xi_{i1}x_{i0}\right)^\gamma\right] \text{ s.t. } \sum_{i=1}^n x_{i0} = W_0. \tag{5.16}$$

The optimal value of the corresponding multistage problem is

$$v^* = W_0^\gamma \prod_{t=0}^{T-1} Q_t(1)\,. \tag{5.17}$$

We see that in the above cases one needs to solve just a (static) one-stage stochastic program in order to find the first stage optimal solutions. Of course, such myopic behavior of multistage programs is rather exceptional. For instance, introduction of transaction costs into this model destroys this myopic property. □

6 Complexity of multistage stochastic programs

A standard approach to solving multistage stochastic programs is by a discretization formulated in a form of *scenario tree*. That is, at period $t = 1$ we have one root node associated with the (deterministic) value of ξ_1. At period $t = 2$ we have as many nodes as many different realizations of ξ_2 are considered. Each of them is connected with the root node by an arc. For each node i at period $t = 2$ (which corresponds to a particular realization ξ_2^i of ξ_2) we create as many nodes at period $t = 3$ as different values of ξ_3 may follow ξ_2^i, and we connect them with the node i, etc. Generally, nodes at period t correspond to possible values of ξ_t that may occur. Each node ξ_t^i at period t is connected to a unique node at period $t - 1$, called its *ancestor* node, and is also connected to several nodes at period $t + 1$, called its *children*. With every arc of the tree, connecting a node ξ_t^i with its child node ξ_{t+1}^{ij}, is associated (conditional) probability $p_{ij} > 0$ such that $\sum_j p_{ij} = 1$. A *scenario* is a path starting at the root node and ending at a node of the last period T, i.e., each scenario represents a particular realization $\xi_{[T]} = (\xi_1, \dots, \xi_T)$ of the considered process. The probability of a scenario is given by the product of the conditional probabilities p_{ij} corresponding to the arcs of its path. Once such a scenario tree is constructed, the obtained multistage stochastic program can be written as a one large (deterministic) optimization problem of the form (5.3) with a finite number of decision variables $x_t(\xi_{[t]})$.

If one views a constructed scenario tree (with a manageable number of scenarios) as an accurate representation of reality, then there is no principle difference between the numerical complexity of two and multi-stage stochastic programming. Yes, it is more difficult to solve multi than two-stage (say linear) stochastic programs with a comparable number of scenarios, but the difference is not that dramatic. A considerable effort went into development of efficient algorithms for solving (mainly linear) multistage stochastic programs by utilizing their particular (decomposable) structure (see, e.g., [54] for a recent survey).

On the other hand, if the number of scenarios is astronomically large, then in both two and multi-stage cases the corresponding deterministic optimization problems are unsolvable. However, we argued in Sect. 3 that some classes of two-stage stochastic programs can be solved with a reasonable accuracy by Monte Carlo sampling techniques. The corresponding estimate (3.7) of the required sample size does not depend on the number of scenarios which can be even infinite. The SAA method can be also applied to multistage stochastic programs. That is, a scenario tree is constructed by sampling in the following way. First, a random sample ξ_2^i, $i = 1, \dots, N_1$, of N_1 realizations of the random vector ξ_2 is generated. These realizations are viewed as nodes at the second period, each taken with probability $1/N_1$. For each realization ξ_2^i, $i = 1, \dots, N_1$, a random sample ξ_3^{ij}, $j = 1, \dots, N_2$, of N_2 realizations[12] of the random vector ξ_3 are generated in accordance with the conditional distribution of ξ_3 given $\xi_2 = \xi_2^i$. And so on for the later stages. In that way a scenario tree is generated with the total number of scenarios $N = \prod_{t=1}^{T-1} N_t$, each with equal probability $1/N$. We refer to this process of generating scenario trees as *conditional sampling*.

After a (random) scenario tree is generated by conditional sampling, the obtained multistage problem is solved by an appropriate algorithm. It is possible to show that, under mild regularity conditions, the optimal value and a first stage optimal solution of such SAA problem converge to their true counterparts w.p.1 as the sample sizes N_t, $t = 1, \dots, T-1$, tend (simultaneously) to infinity (cf., [39,41,65]). Note, however, that although is a step in a right direction, such consistency result by itself does not justify this method since a scenario tree required to solve the corresponding "true" problem with a reasonable accuracy could be far too large to handle numerically.

It is possible to show that an analogue of the estimate (3.9) of the sample size holds for 3-stage problems. That is, under certain regularity conditions, for $T = 3$, $\varepsilon > 0$, $\alpha \in (0,1)$ and appropriate constants $L_1, L_2, L_3, D_1, D_2, \sigma_1^2, \sigma_2^2$, analogous to constants used in the estimate (3.7) for two stage programs, and the sample sizes N_1 and N_2 satisfying

$$\begin{aligned} O(1)\Bigg[\left(\frac{L_1 D_1}{\varepsilon}\right)^{n_1} \exp\left\{-\frac{O(1)N_1\varepsilon^2}{\sigma_1^2}\right\} \\ + \left(\frac{L_3 D_1}{\varepsilon}\right)^{n_1} \left(\frac{L_2 D_2}{\varepsilon}\right)^{n_2} \exp\left\{-\frac{O(1)N_2\varepsilon^2}{\sigma_2^2}\right\}\Bigg] \le \alpha, \end{aligned} \tag{6.1}$$

we have that any first stage $\varepsilon/2$-optimal solution of the SAA problem is an ε-optimal solution of the corresponding true problem with probability at least $1 - \alpha$ (see [67] for details). In particular, suppose that $N_1 = N_2$ and take $L := \max\{L_1, L_2, L_3\}$, $D := \max\{D_1, D_2\}$ and $\sigma^2 := \max\{\sigma_1^2, \sigma_2^2\}$. Then the above estimate of the required sample size $N_1 = N_2$ takes the form:

$$N_1 \ge \frac{O(1)\sigma^2}{\varepsilon^2}\left[(n_1 + n_2)\log\left(\frac{O(1)LD}{\varepsilon}\right) + \log\left(\frac{1}{\alpha}\right)\right]. \tag{6.2}$$

[12] It is also possible to consider a sampling scheme where a different number, say N_2^i, of random realizations of ξ_3, conditional on $\xi_2 = \xi_2^i$, is generated. We use the same sample size N_2 in order to simplify the presentation.

Note that, similar to the analysis of two stage programming in Sect. 3, the above sample size estimates were derived under the assumption of relatively complete recourse (it was also assumed in [67] that the random process ξ_t is between stages independent).

The sample size estimate (6.2) looks similar to the estimate (3.9) for two stage programs. Note, however, that the total number of scenarios of the corresponding SAA tree (generated by conditional sampling) is $N_1 N_2 = N_1^2$. This analysis can be extended to T-stage stochastic programming with a conclusion that the corresponding number of scenarios is growing *exponentially* with increase of the number of stages. Consequently, the deterministic formulation of the constructed (by conditional sampling) SAA problem becomes far too large for a numerical solution with increase of the number of stages. Therefore, one is forced to take progressively smaller sample sizes N_t for later stages hoping that it will have a little effect on the first stage solution.

The above analysis indicates a dramatic difference between complexity of two and multi-stage stochastic programming. It should be clearly stated, however, that this does not prove in a rigorous way that multistage (even linear) stochastic programming problems are computationally intractable for large, say $T \geq 5$, number of stages. Little is known about complexity of multistage stochastic programming and the topic requires a further investigation. An essential difference between two and multi-stage programs was also observed, e.g., in stability analysis of stochastic programs (see [23]).

Example 5 (Portfolio Selection continued) Consider the problem of portfolio selection discussed in Example 4. Suppose that the random process $\xi_1, \ldots, \xi_T$ is between stages independent and we use power utility function $U(z) := z^\gamma$. It was shown in Example 4 that in this case the problem is myopic, and hence in order to find an optimal first stage solution we only need to solve the one-stage stochastic problem (5.16). But suppose that we don't know this and would like to estimate the optimal value v^* of the corresponding T-stage problem by using the SAA method. To this end we employ conditional sampling with the corresponding sample sizes N_t at stages $t = 1, \ldots, T$.

In accordance with (5.17) we have that the optimal value $\hat{v}_N$ of this SAA problem can be written in the form

$$\hat{v}_N = W_0^\gamma \prod_{t=0}^{T-1} \hat{Q}_{N_{t+1}}(1), \tag{6.3}$$

where $\hat{Q}_{N_{t+1}}(W_t)$ is the optimal value of the SAA counterpart of the corresponding "true" problem. Note that since the conditional sample is generated here in the "between stages independent way", the random variables $\hat{Q}_{N_{t+1}}(1)$ are mutually independent. Therefore

$$\mathbb{E}\left[\hat{v}_N\right] = W_0^\gamma \prod_{t=0}^{T-1} \mathbb{E}\left[\hat{Q}_{N_{t+1}}(1)\right] = v^* \prod_{t=0}^{T-1} (1 + \beta_t), \tag{6.4}$$

where

$$\beta_t := \frac{\mathbb{E}\left[\hat{Q}_{N_{t+1}}(1)\right] - Q_t(1)}{Q_t(1)}$$

is the relative bias of the optimal value of the corresponding t-th stage SAA problem. This indicates that in this example the bias of the optimal value of the SAA problem grows *exponentially* with increase of the number of stages, and hence the SAA estimate $\hat{v}_N$ could be considerably bigger than v^*, and hence far too optimistic, for large number of stages. Some numerical experiments seem to confirm that the bias in this problem grows fast with increase of the number of stages (cf., [7]). □

7 Risk averse multistage programming

In this section we discuss possible extensions of the coherent risk measures, discussed in Sect. 4, to multistage programming. Several approached were suggested in the recent literature for extending risk averse approach to a dynamical setting (e.g., [3,10,17,25,48]). We follow below the approach of conditional risk mappings developed in Ruszczyński and Shapiro [57]. For the sake of simplicity and in order to avoid technical complications, we assume that the underlying process has a discrete distribution with a finite support and can be represented by a (finite) scenario tree. Note that at this moment we do not assume any probability distribution on the considered scenario tree.

Let us denote by Ω_t the set of all nodes at stage $t = 1, \dots, T$, and $K_t := |\Omega_t|$ be the cardinality of Ω_t. With the set Ω_T we associate sigma algebra $\mathcal{F}_T$ of *all* its subsets. The set Ω_T can be represented as union of disjoint sets $C_1, \dots, C_{K_{T-1}}$, with each C_k being the set of children of a node at stage $T-1$. Let $\mathcal{F}_{T-1}$ be the subalgebra of $\mathcal{F}_T$ generated by sets $C_1, \dots, C_{K_{T-1}}$, i.e., these sets form the set of elementary events of $\mathcal{F}_{T-1}$. By this construction there is a one-to-one correspondence between elementary events of $\mathcal{F}_{T-1}$ and the set Ω_{T-1} of nodes at stage $T-1$. By continuing this process we construct a sequence of sigma algebras (called filtration) $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_T$. Note that $\mathcal{F}_1$ corresponds to the unique root node and hence $\mathcal{F}_1 = \{\emptyset, \Omega_T\}$.

Consider a node $a \in \Omega_t$. We denote by $C_a \subset \Omega_{t+1}$ the set of all children nodes of a. Since there is a one-to-one correspondence between nodes of Ω_t and elementary events of the sigma algebra $\mathcal{F}_t$, we can identify a with an elementary event of $\mathcal{F}_t$. We have that the sets C_a, $a \in \Omega_t$, are disjoint and $\Omega_{t+1} = \cup_{a \in \Omega_t} C_a$. By taking all children of every node of C_a at later stages, we eventually can identify with C_a a subset of Ω_T. With every node a at stage t we associate a risk function:

$$\rho^a : \mathbb{R}^{|C_a|} \to \mathbb{R}, \quad a \in \Omega_t. \tag{7.1}$$

For example, we can use (coherent) risk functions of the form (4.7) or (4.8), say

$$\rho^a[Z] := \mathbb{E}[Z] + \lambda_a \mathbb{E}\big[Z - \mathbb{E}[Z]\big]_+, \quad Z \in \mathbb{R}^{|C_a|}, \tag{7.2}$$

where[13] $\lambda_a \in [0,1]$ and the expectations are taken with respect to a chosen probability distribution p^a on the set C_a.

We can write $\mathbb{R}^{K_{t+1}}$ as the Cartesian product of the spaces $\mathbb{R}^{|C_a|}$, $a \in \Omega_t$. That is, $\mathbb{R}^{K_{t+1}} = \mathbb{R}^{|C_{a_1}|} \times \cdots \times \mathbb{R}^{|C_{a_{K_t}}|}$, where $\{a_1, \ldots, a_{K_t}\} = \Omega_t$. Define the mappings

$$\rho_{t+1} := (\rho^{a_1}, \ldots, \rho^{a_{K_t}}) : \mathbb{R}^{K_{t+1}} \to \mathbb{R}^{K_t}, \quad t = 1, \ldots, T-1, \tag{7.3}$$

associated with risk functions ρ^a. Recall that the set Ω_{t+1} of nodes at stage $t+1$ is identified with the set of elementary events of sigma algebra $\mathcal{F}_{t+1}$, and its sigma subalgebra $\mathcal{F}_t$ is generated by sets C_a, $a \in \Omega_t$.

We denote by $\mathcal{Z}_T$ the set of all functions $Z : \Omega_T \to \mathbb{R}$. We can identify every such function with a vector of the space $\mathbb{R}^{K_T}$, i.e., the set $\mathcal{Z}_T$ can be identified with the space $\mathbb{R}^{K_T}$. We have that a function $Z : \Omega_T \to \mathbb{R}$ is $\mathcal{F}_{T-1}$-measurable iff it is constant on every set C_a, $a \in \Omega_{T-1}$. We denote by $\mathcal{Z}_{T-1}$ the subset of $\mathcal{Z}_T$ formed by $\mathcal{F}_{T-1}$-measurable functions. The set $\mathcal{Z}_{T-1}$ can be identified with $\mathbb{R}^{K_{T-1}}$. And so on, we can construct a sequence $\mathcal{Z}_t$, $t = 1, \ldots, T$, of spaces of $\mathcal{F}_t$-measurable functions $Z : \Omega_T \to \mathbb{R}$ such that $\mathcal{Z}_1 \subset \cdots \subset \mathcal{Z}_T$ and each $\mathcal{Z}_t$ can be identified with the space $\mathbb{R}^{K_t}$. Recall that $K_1 = 1$, and hence $\mathcal{Z}_1$ can be identified with $\mathbb{R}$. We view the mapping ρ_{t+1}, defined in (7.3), as a mapping from the space $\mathcal{Z}_{t+1}$ into the space $\mathcal{Z}_t$. Conversely, with any mapping $\rho_{t+1} : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ we can associate a set of risk functions of the form (7.1).

We say that a mapping $\rho_{t+1} : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ is a *conditional risk mapping* if it satisfies the following conditions (cf. [57]):

(C*1) *Convexity*:

$$\rho_{t+1}(\alpha Z_1 + (1-\alpha) Z_2) \le \alpha \rho_{t+1}(Z_1) + (1-\alpha)\rho_{t+1}(Z_2),$$
$$\forall Z_1, Z_2 \in \mathcal{Z}_{t+1},\ \forall \alpha \in [0,1].$$

(C*2) *Monotonicity*: If $Z_1, Z_2 \in \mathcal{Z}_{t+1}$ and $Z_2 \ge Z_1$, then $\rho_{t+1}(Z_2) \ge \rho_{t+1}(Z_1)$.

(C*3) *Translation Equivariance*: If $Z' \in \mathcal{Z}_t$ and $Z \in \mathcal{Z}_{t+1}$, then $\rho_{t+1}(Z + Z') = \rho_{t+1}(Z) + Z'$.

(C*4) *Positive homogeneity*: If $\alpha > 0$ and $Z \in \mathcal{Z}_{t+1}$, then $\rho_{t+1}(\alpha Z) = \alpha \rho_{t+1}(Z)$.

It is straightforward to see that conditions (C*1), (C*2) and (C*4) hold iff the corresponding conditions (C1), (C2) and (C4), defined in Sect. 4, hold for every risk function ρ^a associated with ρ_{t+1}. Also by construction (7.3) of ρ_{t+1}, we have that condition (C*3) holds iff condition (C3) holds for all ρ^a. That is, ρ_{t+1} is a *conditional risk mapping* iff every corresponding risk function ρ^a is a *coherent risk measure.*

[13] Note that we do not have here measurability problems since all considered sets are finite, and use the space $\mathcal{Z} := \mathbb{R}^{|C_a|}$ which can be viewed as the space of all functions $Z : C_a \to \mathbb{R}$.

By Theorem 2 with each coherent risk function ρ^a, $a \in \Omega_t$, is associated a set $\mathcal{A}(a)$ of probability measures (vectors) such that

$$\rho^a(Z) = \max_{p \in \mathcal{A}(a)} \mathbb{E}_p[Z]. \tag{7.4}$$

Here $Z \in \mathbb{R}^{K_{t+1}}$ is a vector corresponding to function $Z : \Omega_{t+1} \to \mathbb{R}$,

$$\mathbb{E}_p[Z] = \langle p, Z \rangle = \sum_{k=1}^{K_{t+1}} p_k Z_k$$

and $\mathcal{A}(a) = \mathcal{A}_{t+1}(a)$ is a closed convex set of probability vectors[14] $p \in \mathbb{R}^{K_{t+1}}$ such that $p_k = 0$ if $k \in \Omega_{t+1} \setminus C_a$, i.e., all probability measures of $\mathcal{A}_{t+1}(a)$ are supported on the set C_a. We can now represent the corresponding conditional risk mapping ρ_{t+1} as a maximum of conditional expectations as follows. For an *arbitrary* probability distribution $\nu = (\nu_a)_{a \in \Omega_t}$ on Ω_t, define:

$$\mathcal{D}_{t+1} := \left\{ \mu = \sum_{a \in \Omega_t} \nu_a p^a : p^a \in \mathcal{A}_{t+1}(a) \right\}. \tag{7.5}$$

It is not difficult to see that $\mathcal{D}_{t+1} \subset \mathbb{R}^{K_{t+1}}$ is a convex set of probability vectors. Moreover, since each $\mathcal{A}_{t+1}(a)$ is compact, the set $\mathcal{D}_{t+1}$ is also compact and hence is closed. For any $\mu = \sum_{a \in \Omega_t} \nu_a p^a \in \mathcal{D}_{t+1}$ and $Z \in \mathbb{R}^{K_{t+1}}$ we have[15]

$$\mathbb{E}_\mu \left[Z | \mathcal{F}_t \right](a) = \langle p^a, Z \rangle = \mathbb{E}_{p^a}[Z], \quad a \in \Omega_t. \tag{7.6}$$

It follows then by (7.4) that

$$\rho_{t+1}(Z) = \max_{\mu \in \mathcal{D}_{t+1}} \mathbb{E}_\mu \left[Z | \mathcal{F}_t \right], \tag{7.7}$$

where the maximum in the right hand side of (7.7) is taken pointwise in $a \in \Omega_t$. Note also that any distribution of $\mathcal{D}_{t+1}$ agrees with the distribution ν on Ω_t (for an extension of the representation (7.7) of conditional risk mappings to general, not necessarily finitely supported, distributions see [57]).

For a sequence $\rho_{t+1} : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$, $t = 1, \dots, T-1$, of conditional risk mappings consider the following risk averse analogue formulation of the multistage program (5.1):

[14] A vector $p \in \mathbb{R}^{K_t}$ is said to be a probability vector if all its components p_k, $k = 1, \dots, K_t$, are nonnegative and $\sum_{k=1}^{K_t} p_k = 1$. Such a probability vector can be considered as a probability distribution on Ω_t.

[15] Recall that the space $\mathcal{Z}_{t+1}$ can be identified with $\mathbb{R}^{K_{t+1}}$.

$$\begin{aligned}&\operatorname*{Min}_{x_1\in\mathcal{G}_1} F_1(x_1)+\rho_2\Bigg[\inf_{x_2\in\mathcal{G}_2(x_1,\omega)} F_2(x_2,\omega)+\cdots+\rho_{T-1}\Bigg[\inf_{x_{T-1}\in\mathcal{G}_T(x_{T-2},\omega)}\\&\quad\times\Big\{F_{T-1}(x_{T-1},\omega)+\rho_T\big[\inf_{x_T\in\mathcal{G}_T(x_{T-1},\omega)} F_T(x_T,\omega)\big]\Big\}\Bigg]\Bigg].\end{aligned}\tag{7.8}$$

Here $\Omega := \Omega_T$, the objective functions $F_t : \mathbb{R}^{n_{t-1}} \times \Omega \to \mathbb{R}$ are real valued functions and $\mathcal{G}_t : \mathbb{R}^{n_{t-1}} \times \Omega \rightrightarrows \mathbb{R}^{n_t}$, $t = 2, \dots, T$, are multifunctions such that $F_t(x_t, \cdot)$ and $\mathcal{G}_t(x_{t-1}, \cdot)$ are $\mathcal{F}_t$-measurable for all x_t and x_{t-1}.

There are several ways how the above nested formulation (7.8) can be formalized, we proceed as follows (see [57] for details). In a ways similar to (5.3) we can write problem (7.8) in the form

$$\begin{array}{ll}\operatorname*{Min}\limits_{x_1,x_2(\cdot),\dots,x_T(\cdot)} & \tilde{\rho}\big[F_1(x_1)+F_2(x_2(\omega),\omega)+\cdots+F_T\,(x_T(\omega),\omega)\,\big]\\ \text{s.t.} & x_1\in\mathcal{G}_1,\ x_t(\omega)\in\mathcal{G}_t(x_{t-1}(\omega),\omega),\quad t=2,\dots,T.\end{array}\tag{7.9}$$

Here $\tilde{\rho} := \rho_2 \circ \cdots \rho_T$ is the composite risk function, i.e., for $Z_t \in \mathcal{Z}_t, t = 1, \dots, T$,

$$\tilde{\rho}(Z_1+\cdots+Z_T)=Z_1+\rho_2\left[Z_2+\cdots+\rho_{T-1}\left[Z_{T-1}+\rho_T[Z_T]\right]\right].\tag{7.10}$$

Recall that $\mathcal{Z}_1$ is identified with $\mathbb{R}$, and hence Z_1 is a real number. The optimization in (7.9) is performed over functions $x_t : \Omega \to \mathbb{R}, t = 1, \dots, T$, satisfying the corresponding constraints, which imply that each $x_t(\omega)$ is $\mathcal{F}_t$-measurable and hence each $F_t(x_t(\omega), \omega)$ is $\mathcal{F}_t$-measurable.

An alternative approach to formalizing nested formulation (7.8) is to write dynamic programming equations. That is, for the last period T we have

$$Q_T(x_{T-1},\omega):=\inf_{x_T\in\mathcal{G}_T(x_{T-1},\omega)} F_T(x_T,\omega),\tag{7.11}$$

$$\mathcal{Q}_T(x_{T-1},\omega):=\rho_T[Q_T(x_{T-1},\omega)],\tag{7.12}$$

and for $t = T - 1, \dots, 2$,

$$\mathcal{Q}_t(x_{t-1},\omega):=\rho_t\left[Q_t(x_{t-1},\omega)\right],\tag{7.13}$$

where

$$Q_t(x_{t-1},\omega):=\inf_{x_t\in\mathcal{G}_t(x_{t-1},\omega)}\left\{F_t(x_t,\omega)+\mathcal{Q}_{t+1}(x_t,\omega)\right\}.\tag{7.14}$$

Of course, equations (7.13) and (7.14) can be combined into one equation:[16]

$$Q_t(x_{t-1},\omega)=\inf_{x_t\in\mathcal{G}_t(x_{t-1},\omega)}\left\{F_t(x_t,\omega)+\rho_{t+1}\left[Q_{t+1}(x_t,\omega)\right]\right\}.\tag{7.15}$$

[16] With some abuse of the notation we write $Q_{t+1}(x_t, \omega)$ for the value of $Q_{t+1}(x_t)$ at $\omega \in \Omega$, and $\rho_{t+1}\left[Q_{t+1}(x_t, \omega)\right]$ for $\rho_{t+1}\left[Q_{t+1}(x_t)\right](\omega)$.

Finally, at the first stage we solve the problem

$$\underset{x_1 \in \mathcal{G}_1}{\text{Min}}\ \rho_2[Q_2(x_1,\omega)]. \tag{7.16}$$

It is important to emphasize that in the above development of the dynamic programming equations the monotonicity condition (C*2) plays a crucial role, because only then we can move the optimization under the risk operation.

Remark 9 By using representation (7.7), we can write the dynamic programming equations (7.15) in the form

$$Q_t(x_{t-1},\omega) = \inf_{x_t \in \mathcal{G}_t(x_{t-1},\omega)} \left\{ F_t(x_t,\omega) + \sup_{\mu \in \mathcal{D}_{t+1}} \mathbb{E}_\mu \left[Q_{t+1}(x_t) \middle| \mathcal{F}_t \right](\omega) \right\}. \tag{7.17}$$

Note that the left and right hand side functions in (7.17) are $\mathcal{F}_t$-measurable, and hence this equation can be written in terms of $a \in \Omega_t$ instead of $\omega \in \Omega$. Recall that every $\mu \in \mathcal{D}_{t+1}$ is representable in the form $\mu = \sum_{a \in \Omega_t} \nu_a p^a$ (see (7.5)) and that

$$\mathbb{E}_\mu \left[Q_{t+1}(x_t) \middle| \mathcal{F}_t \right](a) = \mathbb{E}_{p^a}[Q_{t+1}(x_t)],\ \ a \in \Omega_t. \tag{7.18}$$

We say that the problem is *convex* if the functions $F_t(\cdot,\omega)$, $Q_t(\cdot,\omega)$ and the sets $\mathcal{G}_t(x_{t-1},\omega)$ are convex. If the problem is convex, then (since the set $\mathcal{D}_{t+1}$ is convex compact) the 'inf' and 'sup' operators in the right hand side of (7.17) can be interchanged to obtain a dual problem, and for a given x_{t-1} and every $a \in \Omega_t$ the dual problem has an optimal solution $\bar{p}^a \in \mathcal{A}_{t+1}(a)$. Consequently, for $\bar{\mu}_{t+1} := \sum_{a \in \Omega_t} \nu_a \bar{p}^a$ we have that an optimal solution of the original problem and the corresponding cost-to-go functions satisfy the following dynamic programming equations:

$$Q_t(x_{t-1},\omega) = \inf_{x_t \in \mathcal{G}_t(x_{t-1},\omega)} \left\{ F_t(x_t,\omega) + \mathbb{E}_{\bar{\mu}_{t+1}} \left[Q_{t+1}(x_t) | \mathcal{F}_t \right](\omega) \right\}. \tag{7.19}$$

Moreover, it is possible to choose the "worst case" distributions $\bar{\mu}_{t+1}$ in a consistent way, i.e., such that each $\bar{\mu}_{t+1}$ coincides with $\bar{\mu}_t$ on $\mathcal{F}_t$ (cf. [57]). That is, consider the first-stage problem (7.16). We have that (recall that at the first stage there is only one node, $\mathcal{F}_1 = \{\emptyset, \Omega\}$ and $\mathcal{D}_2 = \mathcal{A}_2$)

$$\rho_2[Q_2(x_1)] = \sup_{\mu \in \mathcal{D}_2} \mathbb{E}_\mu[Q_2(x_1)|\mathcal{F}_1] = \sup_{\mu \in \mathcal{D}_2} \mathbb{E}_\mu[Q_2(x_1)]. \tag{7.20}$$

By convexity and since $\mathcal{D}_2$ is compact, we have that there is $\bar{\mu}_2 \in \mathcal{D}_2$ (an optimal solution of the dual problem) such that the optimal value of the first stage problem is equal to the optimal value and the set of optimal solutions of the first stage problem is contained in the set of optimal solutions of the problem

$$\underset{x_1 \in \mathcal{G}_1}{\operatorname{Min}} \mathbb{E}_{\bar{\mu}_2}[Q_2(x_1)]. \tag{7.21}$$

Let $\bar{x}_1$ be an optimal solution of the first stage problem. Then we can choose $\bar{\mu}_3 \in \mathcal{D}_3$, of the form $\bar{\mu}_3 := \sum_{a \in \Omega_2} \nu_a \bar{p}^a$, such that Eq. (7.19) holds with $t = 2$ and $x_1 = \bar{x}_1$. Moreover, we can take the probability measure $\nu = (\nu_a)_{a \in \Omega_2}$ to be the same as $\bar{\mu}_2$, and hence to ensure that $\bar{\mu}_3$ coincides with $\bar{\mu}_2$ on $\mathcal{F}_2$. Next, for every node $a \in \Omega_2$ choose a corresponding (second-stage) optimal solution and repeat the construction to produce an appropriate $\bar{\mu}_4 \in \mathcal{D}_4$, and so on for later stages. In that way, assuming existence of optimal solutions, we can construct a probability distribution $\bar{\mu}_2, \ldots, \bar{\mu}_T$ on the considered scenario tree such that the obtained multistage problem, of the regular form (5.1), has the same cost-to-go (value) functions as the original problem (7.8) and has an optimal solution which also is an optimal solution of the problem (7.8) (in that sense the obtained multistage problem, driven by dynamic programming equations (7.19), is "almost equivalent" to the original problem). In particular, it may happen (see Remark 11) that each distribution $\bar{\mu}_t$ is degenerate, i.e., is a distribution of mass one at a point $\bar{a}_t \in \Omega_t$ (in that case consistency of these distributions means that $\bar{a}_{t+1}$ is a child node of $\bar{a}_t$, $t = 2, \ldots, T-1$). If this happens, then the corresponding probability distribution on the considered scenario tree degenerates into distribution of mass one at the sample (scenario) path $\bar{a}_2, \ldots, \bar{a}_T$, and the obtained multistage problem becomes deterministic. In that case the original problem (7.8) has a constant (i.e., independent of realizations of the uncertain data) optimal policy. Of course, this may happen only in rather specific cases.

If the risk mappings ρ_{t+1} are taken to be conditional expectations, i.e., for $t = 1, \ldots, T-1$ the set $\mathcal{D}_{t+1} = \{\bar{\mu}_{t+1}\}$ in (7.7) is a singleton, then the composite function $\tilde{\rho}$ becomes an expectation operator. In that case problems (7.8) and (7.9) are equivalent to respective problems (5.1) and (5.3) discussed in Sect. 5, and (7.11)–(7.15) become standard dynamic programming equations (of the form (7.19)). For general conditional risk mappings the corresponding composite risk function $\tilde{\rho}$ can be quite complicated.

A natural way for constructing risk averse multistage programs is the following. Consider a multistage stochastic program of the form (5.1) driven by random process ξ_t, $t = 2, \ldots, T$. Assume that this process has a discrete distribution with a finite support and hence can be represented by a scenario tree. We refer to the probability distribution on this tree defined by the considered process as the reference distribution. At each node $a \in \Omega_t$ of this scenario tree we can define a risk function of the form (4.7) or (4.8) with respect to the reference distribution, for example we can use mean-absolute semideviation risk measures (7.2). For such constructions the analysis simplifies considerably if we assume the between stages independence condition, i.e., that random vectors $\xi_t, t = 2, \ldots, T$, are independently distributed. Under this condition of between stages independence we have that the functions $\mathcal{Q}_t(x_{t-1})$ are independent of the random data and the objective function in (7.9) can be written as

$$F_1(x_1) + \rho_2\left[F_2(x_2(\xi_2), \xi_2)\right] + \cdots + \rho_T\left[F_T\left(x_T(\xi_T), \xi_T\right)\right]. \tag{7.22}$$

The condition of between stages independence can be formalized as follows.

Consider a sequence ρ_{t+1} of conditional risk mappings and a sequence $Z_{t+1} \in \mathcal{Z}_{t+1}$, $t = 1, \ldots, T-1$. We say that the *between stages independence condition* holds if:

(i1) For $t = 1, \ldots, T-1$ the corresponding risk functions ρ^a, in (7.1), do not depend on the node $a \in \Omega_t$, i.e., $\rho^a = \rho^{a'}$ for any $a, a' \in \Omega_t$.
(i2) Each Z_{t+1} (viewed as a function $Z_{t+1} : \Omega_{t+1} \to \mathbb{R}$) restricted to C_a does not depend on the node $a \in \Omega_t$.

The above condition (i1) can be equivalently formulated as that the corresponding sets $\mathcal{A}_{t+1}(a) \equiv \mathcal{A}_{t+1}$ do not depend on $a \in \Omega_t$. Of course, this implies that each set C_a has the same cardinality for every $a \in \Omega_t$. For example, risk functions defined in (7.2) satisfy condition (i1) if for every $t = 1, \ldots, T-1$, probability vectors p^a and coefficients λ_a are the same for every $a \in \Omega_t$. If, moreover, the sequence $Z_1, \ldots, Z_T$, considered as a sequence of random variables with respect to the probability distribution imposed by vectors p^a, is independently distributed, then the between stages independence condition follows.

If the between stages independence condition holds, then $\rho_{t+1}[Z_{t+1}]$ is constant for every Z_{t+1}, $t = 1, \ldots, T-1$. Consequently, under the between stages independence condition we have (compare with (7.10)):

$$\tilde{\rho}(Z_1 + \cdots + Z_T) = Z_1 + \rho_2[Z_2] + \cdots + \rho_T[Z_T]. \tag{7.23}$$

We say that the between stages independence condition holds for the multistage problem (7.8) if the corresponding conditional risk mappings ρ_{t+1}, and $Z_{t+1}(\omega) := Q_{t+1}(x_t, \omega)$, $t = 1, \ldots, T-1$, satisfy the conditions (i1) and (i2) for any $x_t \in \mathbb{R}^{n_t}$. For example, this holds if ξ_t, $t = 2, \ldots, T$, is an independently distributed sequence of random vectors (i.e., this process satisfies the condition of between stages independence), $F_t(x_t, \xi_t)$ and $\mathcal{G}_t(x_{t-1}, \xi_t)$ are functions of this process and the corresponding risk functions ρ^a are defined in a form based on expectations taken with respect to the considered distribution of the process ξ_t, e.g., in the form (7.2). Under the between stages independence condition we have that the functions $\mathcal{Q}_t(x_{t-1})$, defined in (7.12), are independent of ω, and the problem (7.9) can be written in the form

$$\begin{array}{cl} \underset{x_1, x_2(\cdot), \ldots, x_T(\cdot)}{\text{Min}} & F_1(x_1) + \rho_2\left[F_2(x_2(\omega), \omega)\right] + \cdots + \rho_T\left[F_T\left(x_T(\omega), \omega\right)\right] \\ \text{s.t.} & x_1 \in \mathcal{G}_1, \; x_t(\omega) \in \mathcal{G}_t(x_{t-1}(\omega), \omega), \; t = 2, \ldots, T. \end{array} \tag{7.24}$$

Remark 10 It should be noted that the constructions of this section are made under the implicit assumption that the cost-to-go functions $Q_t(x_t, \omega)$ are *finite valued*, and in particular under the assumption of relatively complete recourse.

Remark 11 Let us define, for every node $a \in \Omega_t$, $t = 1, \dots, T-1$, the corresponding set $\mathcal{A}(a) = \mathcal{A}_{t+1}(a)$ to be the set of *all* probability measures (vectors) on the set C_a (recall that $C_a \subset \Omega_{t+1}$ is the set of children nodes of a, and that all probability measures of $\mathcal{A}_{t+1}(a)$ are supported on C_a). Then the maximum in the right hand side of (7.4) is attained at a measure of mass one at a point of the set C_a (compare with Remark 8). Consequently (see (7.18)), for such choice of the sets $\mathcal{A}_{t+1}(a)$, the dynamic programming equations (7.17) can be written as

$$Q_t(x_{t-1}, a) = \inf_{x_t \in \mathcal{G}_t(x_{t-1}, a)} \left\{ F_t(x_t, a) + \max_{\omega \in C_a} Q_{t+1}(x_t, \omega) \right\}, \quad a \in \Omega_t. \tag{7.25}$$

It is interesting to note (see Remark 9) that if the problem is convex, then it is possible to construct a probability distribution (on the considered scenario tree), defined by a sequence $\bar{\mu}_t$, $t = 2, \dots, T$, of consistent probability distributions, such that the obtained multistage program, of the regular form (5.1), is "almost equivalent" to the "min-max" formulation (7.25). In some cases the corresponding distribution can be degenerate in the sense that each $\bar{\mu}_t$ is a distribution of mass one at a point $\bar{a}_t \in \Omega_t$, i.e., $\bar{\mu}_t = \Delta(\bar{a}_t)$, and that $\bar{a}_{t+1}$ is a child node of node $\bar{a}_t$. In order to see when this can happen let us consider the construction of Remark 9. Arguing by induction, suppose that we already have a sequence of nodes $\bar{a}_\tau \in \Omega_\tau$, $\tau = 2, \dots, t$, such that each $\bar{a}_{\tau+1}$ is a child node of $\bar{a}_\tau$ and $\bar{\mu}_\tau = \Delta(\bar{a}_\tau)$. Suppose that for $a = \bar{a}_t$ the right hand side of (7.25) has an optimal solution $\bar{x}_t$. Then for $a = \bar{a}_t$, the dual of the min-max problem in the right hand side of (7.25), which is obtained by interchanging the 'min' and 'max' operators, has the same optimal value iff this min-max problem has a saddle point $(\bar{x}_t, \bar{a}_{t+1}) \in \mathcal{G}_t(x_{t-1}, \bar{a}_t) \times C_{\bar{a}_t}$. Only in that case we can take $\bar{\mu}_{t+1} := \Delta(\bar{a}_{t+1})$ and continue the process of constructing the corresponding sample path.

In the present setting the between stages independent condition can be formulated as that for $t = 1, \dots, T-1$, the set of children nodes of every $a \in \Omega_t$ is the same (i.e., does not depend on $a \in \Omega_t$). That is, in the between stages independent case we can view the corresponding scenario tree as defined by a sequence $\mathcal{C}_t$, $t = 2, \dots, T$, of (finite) sets such that the set of nodes at stage t can be identified with $\mathcal{C}_t$, and the set of children nodes of every node $a \in \mathcal{C}_{t-1}$ coincides with $\mathcal{C}_t$. Then dynamic programming equations (7.25) can be written as

$$Q_t(x_{t-1}, a) = \inf_{x_t \in \mathcal{G}_t(x_{t-1}, a)} \left\{ F_t(x_t, a) + \max_{\omega \in \mathcal{C}_{t+1}} Q_{t+1}(x_t, \omega) \right\}, \quad a \in \mathcal{C}_t. \tag{7.26}$$

Example 6 (Portfolio Selection continued) Consider again the portfolio selection problem discussed in Example 4. Suppose that the process ξ_t is between stages independent, and our objective now is to maximize $\sum_{t=1}^{T} \rho_t(W_t)$ subject to the balance constraints, where $-\rho_t$, $t = 1, \dots, T$, are chosen coherent risk measures. (Note that here we have to solve a maximization, rather than minimization, problem. Therefore, we use coherent risk measures with negative sign.) Then we have that $Q_T(W_{T-1}) = Q_T(1)W_{T-1}$, $Q_{T-1}(W_{T-2})$

$= Q_T(1)Q_{T-1}(1)W_{T-2}$, etc, where $Q_t(W_{t-1})$ is the optimal value of:

$$\begin{array}{ll} \underset{W,x_{t-1}}{\text{Max}} & \rho_t(W) \\ \text{s.t.} & W = \sum_{i=1}^n \xi_{it} x_{i,t-1}, \\ & \sum_{i=1}^n x_{i,t-1} = W_{t-1},\ x_{i,t-1} \geq 0,\ i = 1,\dots,n. \end{array} \tag{7.27}$$

At the first stage the optimal solution is obtained by solving the above problem (7.27) for $t = 1$. That is, under the assumption of between stages independence, the optimal policy is myopic in the sense that it involves solutions of single-stage models.

It is interesting to note that, in this example, at each stage t risk aversion is controlled by the corresponding risk measure (function) ρ_t alone. In particular, *first* stage optimal solutions are obtained by solving an optimization problem based on risk function ρ_1, and are independent of a choice of the following risk functions ρ_t, $t \geq 2$. □

8 Concluding remarks

In the previous sections we discussed some recent advances in our understanding of stochastic programming. Of course, there are many questions open for discussion which require a further investigation. It was already mentioned at the end of Sect. 3 that the considered approach of stochastic programming with recourse is not well suited to handle catastrophic events. How can one evaluate a cost of a collapsing bridge or a power blackout in a big city? It does not seem to be a good idea to satisfy an electricity demand on average. Ideally one would like to make sure that catastrophic events never happen. This, however, could be impossible to maintain under all possible circumstances or the involved costs could be unrealistically large. Therefore, it could be reasonable to approach this problem by enforcing constraints which make probability of catastrophic events to happen very small. This leads to the concept of *chance* or *probabilistic* constraints. Chance constraints were introduced in Charnes et al. [9] (see Prékopa [45,46] for a thorough discussion of chance constraints in stochastic optimization). Chance constraints are difficult to handle, both numerically and from the modelling point of view (see, e.g., [18] for a worst case (minimax) type approach to chance constraints). In that respect we would like to mention a tractable convex approximations approach initiated by Nemirovski [33], and developed further in [34,35].

It was assumed so far that the involved probability distributions are independent of our decisions. In principle, it is also possible to make the corresponding probability distribution P_x dependent on the decision vector x. In that case the objective function, say of the problem (2.2), takes the form

$$f(x) := \mathbb{E}_{P_x}[F(x,\xi)] = \int_{\Xi} F(x,\xi)dP_x(\xi). \tag{8.1}$$

By choosing a reference distribution $\bar{P}$, it is possible to rewrite this objective function as

$$f(x) = \mathbb{E}_{\bar{P}}[F(x,\xi)L(x,\xi)] = \int_{\Xi} F(x,\xi)L(x,\xi)d\bar{P}(\xi), \tag{8.2}$$

where $L(x,\xi) := (dP_x/d\bar{P})(\xi)$ is the so-called Likelihood Ratio (LR) function. Here $dP_x/d\bar{P}$ is the Radon-Nikodym derivative, i.e., $L_x(\cdot) = L(x,\cdot)$ is the density of P_x with respect to $\bar{P}$, assuming of course that such density exists. For example, if the space $\Xi = \{\xi_1,\dots,\xi_K\}$ is finite and $P_x = (p_{x,1},\dots,p_{x,K})$ and $\bar{P} = (\bar{p}_1,\dots,\bar{p}_K)$ are respective probability distributions on Ξ, then the corresponding LR function is $L(x,\xi_k) = p_{x,k}/\bar{p}_k$, $k = 1,\dots,K$.

By making "change-of-variables" (8.2), we represent the objective function $f(x)$ as expectation of $R(x,\xi) := F(x,\xi)L(x,\xi)$ with respect to a fixed (independent of x) distribution, and consequently can apply the standard methodology. There are several problems with this approach, however. The obtained objective function $R(\cdot,\xi)$ often is nonconvex, irrespective whether $F(\cdot,\xi)$ is convex or not. Another serious problem is that this method is very sensitive to a choice of the reference measure $\bar{P}$. Note that although the expected value function remains the same, the variance of $R(x,\xi)$ depends on the distribution $\bar{P}$. Unless the reference distribution is carefully controlled, this variance could become very large making Monte Carlo sampling calculations based on representation (8.2) infeasible (cf. [52]). This should be not surprising since, for example in the case of finite support, the LR is given by the ratio $p_{x,k}/\bar{p}_k$ of two small numbers and could be numerically very unstable. One can try to turn this into an advantage by choosing the reference distribution in such a way as to reduce the corresponding variance, this is the basic idea of the so-called *importance sampling* method. Note, however, that a good choice of the reference distribution for one value of x could be disastrous for other values, and it is difficult to control this process in an iterative optimization routine.

It was assumed in Sect. 2 that the second stage problem (2.1) is an optimization problem subject to constraints defining the feasible set $\mathcal{G}(x,\xi)$. It is possible to consider situations where this feasible set is defined by equilibrium constraints, say of the form

$$\mathcal{G}(x,\xi) := \big\{y \in \mathbb{R}^{n_2} : 0 \in H(x,y,\xi) + \mathcal{N}_C(y)\big\}, \tag{8.3}$$

where $H : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \Xi \to \mathbb{R}^{n_2}$, C is a convex closed subset of $\mathbb{R}^{n_2}$ and $\mathcal{N}_C(y)$ denotes the normal cone to C at $y \in \mathbb{R}^{n_2}$ (by definition $\mathcal{N}_C(y) = \emptyset$ if $y \notin C$). For such defined set $\mathcal{G}(x,\xi)$, problem (2.1) belongs to the class of so-called Mathematical Programming with Equilibrium Constraints (MPEC) problems. In that case the corresponding two-stage program (2.1)–(2.2) can be viewed as a (two-stage) Stochastic MPEC (SMPEC) problem. Such two-stage SMPEC problems were considered by Patriksson and Wynter [40] (see also [16,29,74]). It is also possible, at least theoretically, to extend the SMPEC to a multistage setting. From theoretical point of view it is more or less straightforward to apply the SAA method to SMPEC problems with similar estimates of required sample sizes (cf. [70]). Note, however, that MPEC problems typically are nonconvex and nonsmooth and difficult to solve. Although some significant progress was made recently in understanding of theoretical and numerical properties of MPEC problems (cf., [20,47,59]), it remains to be shown that, say, the SAA method is numerically viable for solving realistic SMPEC problems.

Acknowledgements Supported in part by the National Science Foundation awards DMS-0510324 and DMI-0619977.

References

1. Ahmed, S., Cakmak, U., Shapiro, A.: Coherent risk measures in inventory problems. Eur. J. Oper. Res. (in press, 2007)
2. Artzner, P., Delbaen, F., Eber, J.-M., Heath, D.: Coherent measures of risk. Math. Financ. **9**, 203–228 (1999)
3. Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., Ku, H.: Coherent multiperiod risk measurement. Manuscript, ETH Zürich (2003)
4. Beale, E.M.L.: On minimizing a convex function subject to linear inequalities. J. R. Stat. Soc. B **17**, 173–184 (1955)
5. Ben-Tal, A., Nemirovski, A.: Selected topics in robust convex optimization. Math. Prog. B, this issue
6. Birge, J.R., Louveaux, F.V.: Introduction to Stochastic Programming. Springer-Verlag, New York (1997)
7. Blomvall, J., Shapiro, A.: Solving multistage asset investment problems by Monte Carlo based optimization. Math. Prog. B **108**, 571–595 (2007)
8. Casella, G., Berger, R.: Statistical Inference. 2nd Edn, Duxbury (2001)
9. Charnes, A., Cooper, W.W., Symonds, G.H.: Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. Manage. Sci. **4**, 235–263 (1958)
10. Cheridito, P., Delbaen, F., Kupper, M.: Coherent and convex risk measures for bounded càdlàg processes. Stochas. Processes Appl. **112**, 1–22 (2004)
11. Dantzig, G.B.: Linear programming under uncertainty. Manage. Sci. **1**, 197–206 (1955)
12. Delbaen, F.: Coherent risk measures on general probability spaces. Essays in Honour of Dieter Sondermann. Springer, Heidelberg (2002)
13. Dupačová, J.: Minimax approach to stochastic linear programming and the moment problem. Optimierung, Sstochastik Und Mathematische Methoden der Wirtschaftswissenschaften **58**, 466–467 (1978)
14. Dupačová, J.: The minimax approach to stochastic programming and an illustrative application. Stochastics **20**, 73–88 (1987)
15. Dyer, M., Stougie, L.: Computational complexity of stochastic programming problems. Math. Prog. **106**(3), 423–432 (2006)
16. Evgrafov, A., Patriksson, M.: On the existence of solutions to stochastic mathematical programs with equilibrium constraints. J. Optim. Theory Appl. **121**, 67–76 (2004)
17. Eichhorn, A., Römisch, W.: Polyhedral risk measures in stochastic programming. SIAM J. Optim. **16**, 69–95 (2005)
18. Erdoğan, E., Iyengar, G.: Ambiguous chance constrained problems and robust optimization. Math. Prog. **107**, 37–61 (2006)
19. Ermoliev, Y., Gaivoronski, A., Nedeva, C.: Stochastic optimization problems with partially known distribution functions. SIAM J. Control Optim. **23**, 697–716 (1985)
20. Fletcher, R., Leyffer, S.: Numerical experience with solving MPECs as NLPs, University of Dundee Report NA210 (2002)
21. Föllmer, H., Schied, A.: Convex measures of risk and trading constraints. Financ. Stochas. **6**, 429–447 (2002)

22. Gaivoronski, A.: A numerical method for solving stochastic programming problems with moment constraints on a distribution function. Ann. Oper. Res. **31**, 347–370 (1991)
23. Heitsch, H., Römisch, W., Strugarek, C.: Stability of multistage stochastic programs. SIAM J. Optim. (in press 2007)
24. Homem-de-Mello, T.: On rates of convergence for stochastic optimization problems under non-i.i.d. sampling, Manuscript, Dept. of Industrial Engineering and Management Sciences, Northwestern University (2006)
25. Iyengar, G.: Robust dynamic programming Math. Oper. Res. **30**, 1–21 (2005)
26. Kall, P.: Stochastic Linear Programming. Springer-Verlag, Berlin (1976)
27. Kleywegt, A.J., Shapiro, A., Homem-de-Mello, T.: The sample average approximation method for stochastic discrete optimization. SIAM J. Optim. **12**, 479–502 (2001)
28. Koivu, M.: Variance reduction in sample approximations of stochastic programs. Math. Prog. **103**, 463–485 (2005)
29. Lin, G.H., Fukushima, M.: A Class of stochastic mathematical programs with complementarity constraints: reformulations and algorithms. J. Indus. Manage. Optim. **1**, 99–122 (2005)
30. Linderoth, J., Shapiro, A., Wright, S.: The empirical behavior of sampling methods for stochastic programming. Ann. Oper. Res. **142**, 215–241 (2006)
31. Mak, W.K., Morton, D.P., Wood, R.K.: Monte Carlo bounding techniques for determining solution quality in stochastic programs. Oper. Res. Lett. **24**, 47–56 (1999)
32. Markowitz, H.M.: Portfolio selection. J. Financ. **7**, 77–91 (1952)
33. Nemirovski, A.: On tractable approximations of randomly perturbed convex constraints. Proceedings of the 42nd IEEE Conference on Decision and Control Maui, Hawaii USA, December 2003, 2419–2422
34. Nemirovski, A., Shapiro, A.: Scenario approximations of chance constraints. In: Calafiore, G., Dabbene, F. (eds.), Probabilistic and Randomized Methods for Design under Uncertainty, pp. 3–48, Springer, London (2005)
35. Nemirovski, A., Shapiro, A.: Convex approximations of chance constrained programs. SIAM J. Optim. (in press 2007)
36. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods, SIAM, Philadelphia (1992)
37. Norkin, V.I., Pflug, G.Ch., Ruszczyński, A.: A branch and bound method for stochastic global optimization. Math. Prog. **83**, 425–450 (1998)
38. Ogryczak, W., Ruszczyński, A.: From stochastic dominance to mean–risk models: semideviations as risk measures. Eur. J. Oper. Res. **116**, 33–50 (1999)
39. Olsen, P.: Discretization of multistage stochastic programming problems. Math. Prog. Study **6**, 111–124 (1976)
40. Patriksson, M., Wynter, L.: Stochastic mathematical programs with equilibrium constraints. Oper. Res. Lett. **25**, 159–167 (2000)
41. Pennanen, T.: Epi-convergent discretizations of multistage stochastic programs. Math. Oper. Res. **30**, 245–256 (2005)
42. Pennanen, T., Koivu, M.: Epi-convergent discretizations of stochastic programs via integration quadratures. Numerische Mathematik **100**, 141–163 (2005)
43. Pflug, G.Ch.: Subdifferential representations of risk measures. Math. Prog. **108**, 339–354 (2007)
44. Plambeck, E.L., Fu, B.R., Robinson, S.M., Suri, R.: Sample-path optimization of convex stochastic performance functions. Math. Prog. B **75**, 137–176 (1996)
45. Prékopa, A.: Stochastic Programming, Kluwer, Dordrecht, Boston (1995)
46. Prékopa, A.: Probabilistic programming. In: Ruszczyński, A., Shapiro, A., (eds.) Stochastic Programming, Handbook in OR & MS, vol. 10, North-Holland Publishing Company, Amsterdam (2003)
47. Ralph, D., Wright, S.J.: Some properties of regularization and penalization schemes for MPECs. Optim. Methods Software **19**, 527–556 (2004)
48. Riedel, F.: Dynamic coherent risk measures. Stochas. Processes Appl. **112**, 185–200 (2004)
49. Rockafellar, R.T., Uryasev, S.P.: Optimization of conditional value-at-risk. J. Risk **2**, 21–41 (2000)
50. Rockafellar, R.T., Uryasev, S., Zabarankin, M.: Generalized deviations in risk analysis. Financ. Stochast. **10**, 51–74 (2006)

51. Rockafellar, R.T., Uryasev, S., Zabarankin, M.: Optimality conditions in portfolio analysis with generalized deviation measures. Math. Prog. (in press 2006)
52. Rubinstein, R.Y., Shapiro, A.: Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method. Wiley, New York (1993)
53. Ruszczyński, A., Shapiro, A. (Eds.): Stochastic Programming, Handbook in OR & MS, vol. 10, North-Holland Publishing Company, Amsterdam (2003)
54. Ruszczyński, A.: Decomposition methods. In: Ruszczyński, A., Shapiro, A. (Eds.) Stochastic Programming, Handbook in OR & MS, vol. 10, North-Holland Publishing Company, Amsterdam (2003)
55. Ruszczyński, A., Shapiro, A.: Optimization of risk measures. In: Calafiore, G., Dabbene, F. (Eds.) Probabilistic and Randomized Methods for Design under Uncertainty, pp. 117–158, Springer, London (2005)
56. Ruszczyński, A., Shapiro, A.: Optimization of convex risk functions. Math. Oper. Res. **31**, 433–452 (2006)
57. Ruszczyński, A., Shapiro, A.: Conditional risk mappings. Math. Oper. Res. **31**, 544–561 (2006)
58. Santoso, T., Ahmed, S., Goetschalckx, M., Shapiro, A.: A stochastic programming approach for supply chain network design under uncertainty. Eur. J. Oper. Res. **167**, 96–115 (2005)
59. Scholtes, S.: Convergence properties of a regularization scheme for mathematical programs with complementarity constraints. SIAM J. Optim. **11**, 918–936 (2001)
60. Shapiro, A.: Asymptotic analysis of stochastic programs. Ann. Oper. Res. **30**, 169–186 (1991)
61. Shapiro, A., Homem-de-Mello, T.: On rate of convergence of Monte Carlo approximations of stochastic programs. SIAM J. Optim. **11**, 70–86 (2000)
62. Shapiro, A., Homem-de-Mello, T., Kim, J.C.: Conditioning of stochastic programs. Math. Prog. **94**, 1–19 (2002)
63. Shapiro, A., Kleywegt, A.: Minimax analysis of stochastic programs. Optim. Method Software **17**, 523–542 (2002)
64. Shapiro, A.: Monte Carlo sampling methods. In: Ruszczyński, A., Shapiro, A. (Eds.) Stochastic Programming, Handbook in OR & MS, Vol. 10, North-Holland Publishing Company, Amsterdam (2003)
65. Shapiro, A.: Inference of statistical bounds for multistage stochastic programming problems. Math. Method Oper. Res. **58**, 57–68 (2003)
66. Shapiro, A., Ahmed, S.: On a class of minimax stochastic programs. SIAM J. Optim. **14**, 1237–1249 (2004)
67. Shapiro, A.: On complexity of multistage stochastic programs. Oper. Res. Lett. **34**, 1–8 (2006)
68. Shapiro, A., Nemirovski, A.: On complexity of stochastic programming problems. In: Jeyakumar, V., Rubinov, A.M. (Eds.), Continuous Optimization: Current Trends and Applications, pp. 111–144, Springer, Heidelberg (2005)
69. Shapiro, A.: Worst-case distribution analysis of stochastic programs. Math. Program. Ser. B **107**, 91–96 (2006)
70. Shapiro, A., Xu, H.: Stochastic methemtical programs with equilibrium constraints, modeling and sample average approxiamtion. E-print available at: http://www.optimization-online.org, 2005
71. Takriti, S., Ahmed, S.: On Robust optimization of two-stage systems. Math. Program. **99**, 109–126 (2004)
72. Verweij, B., Ahmed, S., Kleywegt, A.J., Nemhauser, G., Shapiro, A.: The sample average approximation method applied to stochastic routing problems: a computational study. Comput. Optim. Appl. **24**, 289–333 (2003)
73. Wets, R.J.-B.: Programming under uncertainty: the equivalent convex program. SIAM J. Appl. Math. **14**, 89–105 (1966)
74. Xu, H.: An implicit programming approach for a class of stochastic mathematical programs with equilibrium constraints. SIAM J. Optim. **16**, 670–696 (2006)
75. Žáčková, J.: On minimax solutions of stochastic linear programming problems. Čas. Pěst. Mal. **91**, 423–430 (1966)