Math Meth Oper Res (2003) 58:57–68
DOI 10.1007/s001860300280

Mathematical Methods
of Operations Research


# Inference of statistical bounds for multistage stochastic programming problems

**Alexander Shapiro***

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205, USA

Manuscript received: January 2002 / Final version received: January 2003

**Abstract.** We discuss in this paper statistical inference of sample average approximations of multistage stochastic programming problems. We show that any random sampling scheme provides a valid statistical lower bound for the optimal (minimum) value of the true problem. However, in order for such lower bound to be consistent one needs to employ the conditional sampling procedure. We also indicate that fixing a feasible first-stage solution and then solving the sampling approximation of the corresponding $(T-1)$-stage problem, does not give a valid statistical upper bound for the optimal value of the true problem.

**Key words:** Stochastic programming, Multistage stochastic programs with recourse, Monte Carlo sampling, Statistical bounds, Consistent estimators

## 1 Introduction

It is well known that even a crude discretization of the distribution of the random parameters involved in a stochastic programming problem results in exponential growth of the number of scenarios. This, in turn, precludes calculation of the corresponding expected values since the number of scenarios is just too large. Therefore, that way or another, realistic stochastic programming problems could be only solved by some sort of sampling which drastically reduces the size of the set of considered scenarios. One possible approach to such a reduction is based on the Monte Carlo sampling techniques. That is, the "true" (expected value) optimization problem is approximated by a "manageable" problem based on a randomly generated

*Supported, in part, by the National Science Foundation under grant DMS-0073770.

sample from the entire scenarios population. In order to have an idea about the accuracy of such an approximation one needs some type of inference describing statistical properties of the calculated estimates. And, indeed, for *two-stage* stochastic programming problems with recourse such statistical inference is quite well developed.

Much less is known about multistage stochastic programming problems with recourse (see, e.g., [2] for a discussion of multistage stochastic programming). In order to see where the difficulty is in extending the theory from two to multistage programming let us discuss the following abstract framework of stochastic programming. Consider the expected value optimization problem

$$\operatorname*{Min}_{x\in X}\big\{f(x) := \mathbb{E}\big[F(x,\boldsymbol{\xi})\big]\big\}. \tag{1.1}$$

Here $F(x,\xi)$ is a real valued (or, more generally, extended real valued) function of two vector variables $x \in \mathbb{R}^n$ and $\xi \in \mathbb{R}^d$, $X$ is a given subset of $\mathbb{R}^n$, and the expectation is taken with respect to the probability distribution $P$ of the random data vector $\boldsymbol{\xi}$ (by bold script, like $\boldsymbol{\xi}$, we denote random vectors, while by $\xi$ we denote their realizations). The distribution $P$ is supposed to be known. We denote by $\Xi \subset \mathbb{R}^d$ the support of the probability distribution of $\boldsymbol{\xi}$. We assume that for any considered point $x \in \mathbb{R}^n$ the expected value $\mathbb{E}\big[F(x,\boldsymbol{\xi})\big]$ is well defined, i.e., the function $F(x,\cdot)$ is measurable and either $\mathbb{E}\big[F(x,\boldsymbol{\xi})_+\big] < +\infty$ or $\mathbb{E}\big[(-F(x,\boldsymbol{\xi}))_+\big] < +\infty$, where $a_+ := \max\{a,0\}$.

In the case of two-stage programming, $F(x,\xi)$ can be viewed as the optimal value of the second stage optimization problem. By generating a sample $\xi^1,\ldots,\xi^N$, of $N$ replications of the random vector $\boldsymbol{\xi}$, one can construct the following, so-called sample average approximation (SAA), problem

$$\operatorname*{Min}_{x\in X}\left\{\hat{f}_N(x) := \frac{1}{N}\sum_{i=1}^{N} F\big(x,\xi^i\big)\right\}. \tag{1.2}$$

We can consider the generated sample from two points of view, namely as the sequence $\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^N$ of random vectors, or as its realization $\xi^1,\ldots,\xi^N$. The generated sample need not be i.i.d., i.e., random vectors $\boldsymbol{\xi}^i$ need not be (stochastically) independent of each other. We only assume that the marginal probability distribution of each $\boldsymbol{\xi}^i$ is $P$, and that the (strong) Law of Large Numbers (LLN) holds pointwise, i.e., for any $x \in X$ we have that $\hat{f}_N(x) \to f(x)$ with probability one (w.p.1) as $N \to \infty$. Of course, if the sample is i.i.d., then the LLN holds provided that the expected value $f(x)$ is well defined. There are important cases where the LLN holds even if the sample is not i.i.d. (see, e.g., [1]).

It is implicitly assumed in the above construction that for any $x \in X$ and $\xi \in \Xi$, one can efficiently calculate the value and derivatives of the objective function $F(x,\xi)$, and hence to solve the SAA problem (1.2) by an appropriate (deterministic) optimization algorithm. Statistical properties of the optimal value $\hat{v}_N$ and an optimal solution $\hat{x}_N$ of the SAA problem (1.2) have been thoroughly investigated, we may refer to [11, Chapter 6], for example, for a flavor of these results.

We have that for any fixed $x \in X$, $\mathbb{E}\big[F(x,\boldsymbol{\xi}^i)\big] = f(x)$ and hence $\mathbb{E}\big[\hat{f}_N(x)\big] = f(x)$, i.e., $\hat{f}_N(x)$ is an unbiased estimator of $f(x)$. However, it is not difficult to verify, and is well known, that

$$\inf_{x\in X} \mathbb{E}\left[\hat{f}_N(x)\right] \geq \mathbb{E}\left[\inf_{x\in X} \hat{f}_N(x)\right], \tag{1.3}$$

i.e., $v^* \geq \mathbb{E}[\hat{v}_N]$ where $v^*$ denotes the optimal value of the true problem (1.1). That is, $\hat{v}_N$ is a biased estimator of $v^*$.

We say that an estimator $\tilde{v}_N$ is a *valid* statistical *lower* bound of the true optimal value $v^*$ if $v^* \geq \mathbb{E}[\tilde{v}_N]$, and that $\tilde{v}_N$ is *consistent* if $\tilde{v}_N$ tends to $v^*$ w.p.1 as $N \to \infty$. By (1.3) we have that $\hat{v}_N$ is a valid statistical lower bound of $v^*$. It is also possible to show that, under mild regularity conditions, $\hat{v}_N$ is consistent (see, e.g., [3], [9] and references therein, and section 3 below), and the bias $v^* - \mathbb{E}[\hat{v}_N]$ converges to zero as $N \to \infty$ at a rate of $O(N^{-1/2})$, [12].

Since for any $\bar{x} \in X$ we have that $\mathbb{E}[\hat{f}_N(\bar{x})] = f(\bar{x}) \geq v^*$, we can view $\hat{f}_N(\bar{x})$ as a valid statistical *upper* bound of $v^*$. By the LLN it is also consistent if $f(\bar{x}) = v^*$, i.e., $\bar{x}$ is an optimal solution of the true problem (1.1). Such statistical bounds were suggested by Norkin, Pflug and Ruszczyński [7], and developed further in Mak, Morton and Wood [6], and turned out to be very useful for numerical validation of two-stage stochastic programs. For numerical results and experiments with these statistical bounds we refer to [6], [5], [14].

The goal of this paper is two fold, namely to point out some difficulties in extending this methodology to multistage programming and to provide some initial results. In the next section we discuss, in particular, why straightforward sampling in the multistage case does not produce consistent lower bounds and why it is difficult to construct computationally manageable and tight statistical upper bounds. In section 3 we show that under certain regularity conditions consistent lower bounds can be obtained by conditional sampling.

## 2 Multistage sampling bounds

Consider the following $T$-stage linear stochastic programming problem with recourse

$$\operatorname*{Min}_{x_1\in X_1} c_1x_1 + \mathbb{E}\left[\min_{x_2\in X_2(x_1,\xi_2)} c_2x_2 + \mathbb{E}\left[\cdots + \mathbb{E}\left[\min_{x_T\in X_T(x_{T-1},\xi_T)} c_Tx_T\right]\right]\right] \tag{2.1}$$

driven by the random data process $\xi_2, \ldots, \xi_T$. Here $x_t \in \mathbb{R}^{n_t}$, $t = 1, \ldots, T$, are decision variables, $X_1 := \{x_1 : A_1x_1 = b_1,\ x_1 \geq 0\}$,

$$X_t(x_{t-1}, \xi_t) := \{x_t : B_tx_{t-1} + A_tx_t = b_t,\ x_t \geq 0\}, \quad t = 2, \ldots, T,$$

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

If we denote by $Q_2(x_1, \xi_2)$ the optimal value of the $(T-1)$-stage problem:

$$\operatorname*{Min}_{x_2\in X_2(x_1,\xi_2)} c_2x_2 + \mathbb{E}\left[\cdots + \mathbb{E}\left[\min_{x_T\in X_T(x_{T-1},\xi_T)} c_Tx_T\right]\right], \tag{2.2}$$

then we can write the $T$-stage problem (2.1) in the following form of two-stage programming problem

$$\operatorname*{Min}_{x_1 \in X_1} c_1 x_1 + \mathbb{E}\big[Q_2(x_1, \boldsymbol{\xi}_2)\big]. \tag{2.3}$$

Note, however, that if $T > 2$, then problem (2.2) in itself is a stochastic programming problem. Consequently, if the number of scenarios involved in (2.2) is very large, or infinite, then the optimal value $Q_2(x_1, \xi_2)$ can be calculated only approximately, say by sampling.

Let us note that if we relax the nonanticipativity constraints starting from the second stage of the multistage problem (2.1), we obtain the following two-stage program

$$\operatorname*{Min}_{x_1 \in X_1} c_1 x_1 + \mathbb{E}\big[Q(x_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T)\big], \tag{2.4}$$

where $Q(x_1, \xi_2, \ldots, \xi_T)$ is the optimal value of the problem:

$$\operatorname*{Min}_{x_2, \ldots, x_T} \sum_{t=2}^{T} c_t x_t \quad \text{subject to } B_t x_{t-1} + A_t x_t = b_t, \quad x_t \geq 0, \ \ t = 2, \ldots, T, \tag{2.5}$$

and the expectation is taken with respect to the (joint) distribution of the random vector $\boldsymbol{\xi} := (\boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T)$. Since problem (2.4) is obtained by relaxing the nonanticipativity constraints, its optimal value is less than or equal to the optimal value of the multistage problem (2.1). For $T \geq 3$, only in rather exceptional cases the optimal values of (2.4) and (2.1) are the same.

For given $x_1$ and $\xi_2$, the corresponding expected value(s) can be estimated by generating random samples and solving the obtained SAA problems. Let us observe that in case $T > 2$, it follows from (1.3) that for any estimator $\widehat{Q}_2(x_1, \xi_2)$ of $Q_2(x_1, \xi_2)$ obtained in that way the following relation holds

$$Q_2(x_1, \xi_2) \geq \mathbb{E}\Big[\widehat{Q}_2(x_1, \boldsymbol{\xi}_2) \big| \boldsymbol{\xi}_2 = \xi_2\Big] \tag{2.6}$$

for every feasible $x_1$ and $\xi_2$. That is, for $T \geq 3$ any SAA estimator of $Q_2(x_1, \xi_2)$ is biased downwards.

There are several ways how one can sample from the random process $\boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T$. One possible approach is to view $(\boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T)$ as a random vector and to sample from its (joint) distribution. That is, to generate a random sample $\xi_2^i, \ldots, \xi_T^i$, $i = 1, \ldots, N$, of $N$ scenarios and to solve the obtained SAA problem

$$\operatorname*{Min}_{x_1 \in X_1} c_1 x_1 + \frac{1}{N} \sum_{i=1}^{N} Q(x_1, \xi_2^i, \ldots, \xi_T^i). \tag{2.7}$$

We have here that the optimal value of the above problem (2.7) provides a valid (and consistent) statistical lower bound for the two-stage relaxation (2.4). Therefore, if the optimal value of (2.4) is strictly smaller than the optimal value of (2.1) (which is typical), then such straightforward sampling gives a valid, but not consistent, statistical lower bound for the multistage problem (2.1).

Let us observe that if the number of scenarios of the multistage problem (2.1) is finite, then by generating a sufficiently large scenario sample $\xi_2^i, \ldots, \xi_T^i$, $i = 1, \ldots, N$, one can reconstruct the tree scenario structure of the true problem (2.1). That is, some of the generated second stage vectors $\xi_2^i$ can

be equal to each other, and so on. Note, however, that if the number of scenarios is very large, then even for $T = 3$ the sample size required to reconstruct the corresponding tree structure with a reasonable accuracy can be comparable with the total number of scenarios.

The above discussion also shows that a valid *upper* statistical bound cannot be obtained by a straightforward sampling. In order to compute such an upper bound one needs to construct an implementable and feasible policy. Recall that a sequence of mappings $x_t(\cdot)$, $t = 1, \ldots, T$, is called an *implementable* policy if $x_1(\cdot) = x_1 \in \mathbb{R}^{n_1}$ and each $x_t(\cdot) \in \mathbb{R}^{n_t}$, $t = 2, \ldots, T$, is a function of $x_1$ and the history $\xi_{[1,t]} := (\xi_1, \ldots, \xi_t)$ of the process up to time $t$. An implementable policy is feasible if it satisfies, w.p.1., the corresponding feasibility constraints at each stage $t = 1, \ldots, T$. Given any implementable and feasible policy $x_1, x_2(x_1, \xi_{[1,2]}), \ldots, x_T(x_1, \xi_{[1,T]})$, the expectation

$$\mathbb{E}\big[c_1 x_1 + \boldsymbol{c}_2 x_2(x_1, \boldsymbol{\xi}_{[1,2]}) + \cdots + \boldsymbol{c}_T x_T(x_1, \boldsymbol{\xi}_{[1,T]})\big] \tag{2.8}$$

provides an upper bound for the optimal value of the true multistage problem. The above expectation can be estimated by the average

$$\frac{1}{N}\sum_{i=1}^{N}\Big[c_1 x_1 + c_2^i x_2(x_1, \xi^i_{[1,2]}) + \cdots + c_T^i x_T(x_1, \xi^i_{[1,T]})\Big] \tag{2.9}$$

for a generated sample $\xi_2^i, \ldots, \xi_T^i$ of $N$ realizations of the random process $\boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T$. Although (2.9) gives a valid statistical upper bound, its quality depends on the chosen policy.

In order to improve *lower* bounds one needs to increase the sample size at every stage conditionally on the scenarios generated at the previous stage. This is a standard practice of deterministic procedures for scenario trees construction (see [4], [8]). In the next section we investigate such conditional constructions applied to random sampling. Let us also remark that the above analysis can be extended to nonlinear multistage stochastic programming problems as well.

## 3 Conditional sampling of multistage programs

In this section we study statistical properties of the following random sample construction. First we generate a random sample

$$\xi_2^i = \big(c_2^i, B_2^i, A_2^i, b_2^i\big), \quad i = 1, \ldots, N_1, \tag{3.1}$$

of $N_1$ replications of the random vector $\boldsymbol{\xi}_2$. Then for every $i \in \{1, \ldots, N_1\}$, we generate a random sample

$$\xi_3^{ij} = \big(c_3^{ij}, B_3^{ij}, A_3^{ij}, b_3^{ij}\big), \quad j = 1, \ldots, N_2, \tag{3.2}$$

from the conditional distribution of $\boldsymbol{\xi}_3$ given the event $\boldsymbol{\xi}_2 = \xi_2^i$. And so on for the later stages. We refer to the above sampling scheme as the *conditional sampling*. The sample size associated with each node of a stage $t \in \{2, \ldots, T-1\}$ need not be the same, we assumed it constant for the sake of simplicity.

In order to simplify the presentation we discuss now the case of the linear multistage program (2.1) with $T = 3$. As we shall see, an extension from two

to three stage programming is not a trivial one. In fact, it already demonstrates the main difficulties of extending the analysis from two to multistage programs.

With the generated sample (3.1)–(3.2) is associated the following SAA three-stage program

$$\operatorname*{Min}_{x_1 \in X_1} c_1 x_1 + \frac{1}{N_1} \sum_{i=1}^{N_1} \widehat{Q}_{2,N_2}(x_1, \xi_2^i), \tag{3.3}$$

where $\widehat{Q}_{2,N_2}(x_1, \xi_2^i)$ is the optimal value of

$$\operatorname*{Min}_{x_2 \in X_2(x_1, \xi_2^i)} c_2^i x_2 + \frac{1}{N_2} \sum_{j=1}^{N_2} Q_3(x_2, \xi_3^{ij}), \tag{3.4}$$

with

$$Q_3(x_2, \xi_3) := \inf_{x_3 \in X_3(x_2, \xi_3)} c_3 x_3.$$

The constructed three-stage stochastic programming problem (3.3)–(3.4) has $N = N_1 N_2$ scenarios, each with equal probability $1/N$. It can be noted that for any fixed $j \in \{1, \ldots, N_2\}$ in the above conditional sampling, the corresponding sample $(\xi_2^i, \xi_3^{ij})$, $i = 1, \ldots, N_1$, is a random sample from the distribution of the random vector $(\boldsymbol{\xi}_2, \boldsymbol{\xi}_3)$. Therefore, if $N_2 = 1$, then the above conditional sampling becomes the same as the sampling used in construction of the SAA problem (2.7). Note also that at this stage we do not specify how the conditional samples $\xi_3^{ij}$ are generated. For example, we do not necessarily assume that for different $i, k \in \{1, \ldots, N_1\}$ the corresponding random samples $\boldsymbol{\xi}_3^{ij}$ and $\boldsymbol{\xi}_3^{kj}$, $j = 1, \ldots, N_2$, are independent of each other conditional on $\boldsymbol{\xi}_2^i$ and $\boldsymbol{\xi}_2^k$, respectively. By $\hat{v}_{N_1,N_2}$ and $\hat{S}_{N_1,N_2}$ we denote the optimal value and the set of optimal solutions, respectively, of the problem (3.3).

As it was discussed in the previous section, we have that

$$\inf_{x_1 \in X_1} \left\{ c_1 x_1 + \mathbb{E}\big[Q_2(x_1, \boldsymbol{\xi}_2)\big] \right\} \geq \mathbb{E}\left[ \inf_{x_1 \in X_1} \left\{ c_1 x_1 + \frac{1}{N_1} \sum_{i=1}^{N_1} Q_2(x_1, \boldsymbol{\xi}_2^i) \right\} \right] \tag{3.5}$$

and

$$Q_2(x_1, \xi_2^i) \geq \mathbb{E}\Big[ \widehat{Q}_{2,N_2}(x_1, \boldsymbol{\xi}_2) \big| \boldsymbol{\xi}_2 = \xi_2^i \Big] \tag{3.6}$$

It follows from (3.5) and (3.6) that $\hat{v}_{N_1,N_2}$ gives a valid statistical lower bound for the optimal value $v^*$ of the corresponding (true) three-stage stochastic programming problem. We show now that, under certain regularity conditions, $\hat{v}_{N_1,N_2} \to v^*$ w.p.1 as $N_1 \to \infty$ and $N_2 \to \infty$, i.e., that $\hat{v}_{N_1,N_2}$ is a consistent estimator of $v^*$.

We will need the following results. It is said that a set $V \subset \mathbb{R}^n$ is a neighborhood of the set $X$ if the set $V$ is open and the topological closure of $X$ is contained in $V$. For sets $A, B \subset \mathbb{R}^n$ we denote by $\operatorname{dist}(x, A) := \inf_{x' \in A} \|x - x'\|$ the distance from $x \in \mathbb{R}^n$ to $A$, and by

$$\mathbb{D}(A, B) := \sup_{x \in A} \operatorname{dist}(x, B)$$

the deviation of the set $A$ from the set $B$. Recall the functions $f(x)$ and $\hat{f}_N(x)$ defined in (1.1) and (1.2), respectively.

**Proposition 3.1.** *Suppose that*: (i) *for every* $\xi \in \Xi$ *the function* $F(\cdot, \xi)$ *is convex*, (ii) *the set* $X$ *is convex and compact*, (iii) *the expected value function* $f(x)$ *is finite valued and the* (*strong*) *LLN holds pointwise for every* $x$ *in a neighborhood of the set* $X$. *Then* $\hat{f}_N(\cdot)$ *converges to* $f(\cdot)$ *w.p.1 uniformly on* $X$, *i.e.*,

$$\sup_{x \in X} \left| \hat{f}_N(x) - f(x) \right| \to 0 \qquad \text{w.p.1 as } N \to \infty. \tag{3.7}$$

The above is an easy consequence of the following result from finite dimensional convex analysis. Let $V$ be a neighborhood of $X$ on which $f(\cdot)$ is convex and finite, $D$ be a dense subset of $V$, and $f_N(\cdot)$ be a sequence of (deterministic) convex functions such that $f_N(x)$ converges to $f(x)$, as $N \to \infty$, for every $x \in D$. Then the convergence of $f_N$ to $f$ is uniform on the compact set $X$, [10].

Now we can proceed as follows. We can view $\hat{f}_N(\cdot) = \hat{f}_N(\cdot, \omega)$ as a sequence of random functions defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. By the LLN we have that for any $x \in V$, $\hat{f}_N(x)$ converges to $f(x)$ w.p.1 as $N \to \infty$. This means that there exists a set $\Upsilon_x \in \mathcal{F}$ of $\mathbb{P}$-measure zero such that for any $\omega \in \Omega \setminus \Upsilon_x$, $\hat{f}_N(x, \omega)$ tends to $f(x)$ as $N \to \infty$. This also implies that $f(\cdot)$ is convex. Let $D$ be a dense and countable subset of $V$. Consider the set $\Upsilon := \cup_{x \in D} \Upsilon_x$. Since the set $D$ is countable and $\mathbb{P}(\Upsilon_x) = 0$ for every $x \in D$, we have that $\mathbb{P}(\Upsilon) = 0$. We also have that for any $\omega \in \Omega \setminus \Upsilon$, $\hat{f}_N(x, \omega)$ converges to $f(x)$, as $N \to \infty$, pointwise on $D$. Consequently, for any $\omega \in \Omega \setminus \Upsilon$, this convergence is uniform in $x \in X$, and hence (3.7) holds.

By using the uniform convergence result of the above proposition it is possible to show that, under mild regularity conditions, the optimal value $\hat{v}_N$ and the set $\hat{S}_N$ of optimal solutions of the SAA problem (1.2) are consistent estimators of their true counterparts. By employing tools of epi-convergence (rather than uniform convergence) such results are given in [3], [9], for example. Therefore, the result of the following proposition should be not surprising. We give its proof for the sake of completeness and since some of the techniques used in the proof are employed later.

**Proposition 3.2.** *Suppose that*: (i) *for every* $\xi \in \Xi$ *the function* $F(\cdot, \xi)$ *is convex*, (ii) *the set* $X$ *is closed and convex*, (iii) *the set* $S$ *of optimal solutions of* (1.1) *is nonempty and bounded*, (iv) *the expected value function* $f(x)$ *is finite valued and the* (*strong*) *LLN holds for every* $x$ *in a neighborhood of the set* $S$. *Then* $\hat{v}_N \to v^*$ *and* $\mathbb{D}(\hat{S}_N, S) \to 0$ *w.p.1 as* $N \to \infty$.

*Proof.* It follows from the assumption (i) that the expected value function $f(x)$ is convex. Together with (ii) this implies that the set $S$ is convex. By the assumptions (iii) and (iv), there exists a convex compact set $C$ such that $S$ is contained in the interior of $C$ and $f(x)$ is finite valued, and hence is continuous, on a neighborhood of $C$. It follows that the set $S$ is closed, and hence is compact. By Proposition 3.1 we have that $\hat{f}_N(x)$ converges w.p.1 to $f(x)$ uniformly in $x \in C$. Consider the set $\widetilde{S}_N$ of minimizers of $\hat{f}_N(x)$ over $X \cap C$. Since $X \cap C$ is nonempty and compact and $\hat{f}_N(x)$ is continuous on a neighborhood of $C$, the set $\widetilde{S}_N$ is nonempty. By standard arguments it follows then that $\mathbb{D}(\widetilde{S}_N, S) \to 0$ w.p.1 as $N \to \infty$. Because of the convexity assumptions, any minimizer of $\hat{f}_N(x)$ over $X \cap C$ which lies inside the interior of $C$, is also an optimal solution of the SAA problem (1.2). Therefore, w.p.1 for $N$ large

enough we have that $\widetilde{S}_N = \widehat{S}_N$. Consequently we can restrict both optimization problems (1.1) and (1.2) to the compact set $X \cap C$, and hence the assertions of the above proposition follow. $\square$

A few remarks are now in order. It follows from the convexity assumption (i) that the expected value function $f(x)$ is convex. Moreover, since by the assumption (iv), the function $f(x)$ is finite valued on an open set, it follows that $f(x)$ is proper, i.e., $f(x) > -\infty$ for all $x \in \mathbb{R}^n$. We assume in the above proposition that $f(x)$ is finite valued only on a neighborhood of the set $S$, and it may happen that $f(x) = +\infty$ for some $x \in X$. It was possible to push the proof through since in the considered convex case local optimality implies global optimality. In the case of two-stage programming we have that $f(x) = +\infty$ for some $x \in X$ if the associated second stage problem is infeasible with a positive probability $p$. In that case the corresponding second stage SAA problem will also be infeasible, and hence $\hat{f}_N(x) = +\infty$, w.p.1 for $N$ large enough. Of course, if $p$ is very small, then the required sample size for that event to happen could be very large.

Now let us discuss consistency of $\hat{v}_{N_1,N_2}$. Consider the expected value function

$$Q_1(x_1) := \mathbb{E}[Q_2(x_1, \boldsymbol{\xi}_2)].$$

We make the following assumptions.

**(A1)** The set $S$ of optimal solutions of the first stage problem (2.3) is nonempty and bounded.

**(A2)** The expected value function $Q_1(x_1)$ is finite valued for all $x_1 \in V$, where $V$ is a neighborhood of $S$.

**(A3)** The (strong) LLN holds pointwise, i.e., for any $x_1 \in V$ the following holds

$$\frac{1}{N_1}\sum_{i=1}^{N_1} Q_2(x_1, \boldsymbol{\xi}_2^i) \to Q_1(x_1) \quad \text{w.p.1 as } N_1 \to \infty. \tag{3.8}$$

It follows that the set $S$ is compact, and by Proposition 3.1 that the convergence in (3.8) is uniform in $x_1$ on any compact subset of $V$.

We make similar assumptions about the second stage problem. We denote by $\Xi_2 \subset \mathbb{R}^{d_2}$ the support of the probability distribution of $\boldsymbol{\xi}_2$.

**(A4)** There exists a bounded set $W \subset \mathbb{R}^{n_2}$ such that for any $x_1 \in V$ and $\xi_2 \in \Xi_2$, the set of optimal solutions of the second stage problem (2.7) is nonempty and is contained in the interior of $W$.

**(A5)** The conditional expectation $\mathbb{E}\big[Q_3(x_2, \boldsymbol{\xi}_3)\big|\boldsymbol{\xi}_2 = \xi_2\big]$ is finite valued for all $x_2 \in W$ and $\xi_2 \in \Xi_2$.

We also need the following LLN holding uniformly with respect to the distribution of the random vector $\boldsymbol{\xi}_2$. Recall that the random sample $\boldsymbol{\xi}_3^{ij}, j = 1, \ldots, N_2$, is derived from the conditional distribution of $\boldsymbol{\xi}_3$ given $\boldsymbol{\xi}_2 = \xi_2^i$. We can view $\boldsymbol{\xi}_3^{ij} = \xi_3^{ij}(\omega)$ as defined on a measurable space $(\Omega, \mathcal{F})$ equipped with a probability measure $\mathbb{P}_{\xi_2^i}$.

**(A6)** For every $x_2 \in W$ there exists an $\mathcal{F}$-measurable set $\Upsilon_{x_2} \subset \Omega$ such that for any $\xi_2^i \in \Xi_2$ it follows that $\mathbb{P}_{\xi_2^i}(\Upsilon_{x_2}) = 0$ and for any $\omega \in \Omega \setminus \Upsilon_{x_2}$ the limit

$$\lim_{N_2\to\infty}\left[\frac{1}{N_2}\sum_{j=1}^{N_2} Q_3\big(x_2,\xi_3^{ij}(\omega)\big)\right] = \mathbb{E}\big[Q_3(x_2,\boldsymbol{\xi}_3)|\boldsymbol{\xi}_2=\xi_2^i\big] \tag{3.9}$$

holds.

If the random vectors $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$ are independent of each other, then the probability distribution of $\boldsymbol{\xi}_3$ is independent of the event $\boldsymbol{\xi}_2=\xi_2$. In that case assumption (A6) is just the pointwise LLN specified in the following assumption (A7). Another case where assumption (A6) is reduced to the pointwise LLN of assumption (A7) is when the support $\Xi_2$ of $\boldsymbol{\xi}_2$ is finite.

**(A7)** For any $x_2\in W$ and $\xi_2^i\in\Xi_2$ the following (strong) LLN holds

$$\frac{1}{N_2}\sum_{j=1}^{N_2} Q_3\big(x_2,\boldsymbol{\xi}_3^{ij}\big) \to \mathbb{E}\big[Q_3(x_2,\boldsymbol{\xi}_3)|\boldsymbol{\xi}_2=\xi_2^i\big] \quad \text{w.p.1 as } N_2\to\infty. \tag{3.10}$$

Assumption (A7) holds, in particular, if the sample $\boldsymbol{\xi}_3^{ij}$, $j=1,\ldots,N_2$, conditional on $\boldsymbol{\xi}_2^i$, is i.i.d.

**Proposition 3.3.** *Suppose that assumptions* (A1)–(A6) *hold. Then* $\hat{v}_{N_1,N_2}\to v^*$ *and* $\mathbb{D}(\widehat{S}_{N_1,N_2},S)\to 0$ *w.p.*1 *as* $N_1\to\infty$ *and* $N_2\to\infty$.

*Proof.* In a way similar to the proof of Proposition 3.1, it can be shown that assumptions (A5) and (A6) imply that the convergence is uniform in $x_2$ on any compact subset of $W$. That is, for any compact set $C\subset W$ there exists an $\mathcal{F}$-measurable set $\Upsilon\subset\Omega$ such that for any $\xi_2^i\in\Xi_2$ it follows that $\mathbb{P}_{\xi_2^i}(\Upsilon)=0$ and for any $\omega\in\Omega\setminus\Upsilon$ the limit

$$\lim_{N_2\to\infty}\left\{\sup_{x_2\in C}\left|\frac{1}{N_2}\sum_{j=1}^{N_2} Q_3\big(x_2,\xi_3^{ij}(\omega)\big)-\mathbb{E}\big[Q_3(x_2,\boldsymbol{\xi}_3)|\boldsymbol{\xi}_2=\xi_2^i\big]\right|\right\}=0 \tag{3.11}$$

holds. Moreover, by assumption (A4) we can choose $C$ in such a way that for any $x_1\in V$ and $\xi_2\in\Xi_2$, the set of optimal solutions of the second stage problem (2.7) is nonempty and is contained in the interior of $C$. It follows then that $\widehat{Q}_{2,N_2}(x_1,\xi_2^i)$ converges w.p.1 to $Q_2(x_1,\xi_2^i)$, as $N_2\to\infty$, uniformly in $x_1\in V$ and $\xi_2^i\in\Xi_2$. Since the convergence in (3.8) is uniform in $x_1$ on any compact subset of $V$, it follows that $N_1^{-1}\sum_{i=1}^{N_1}\widehat{Q}_{2,N_2}(x_1,\boldsymbol{\xi}_2^i)$ converges w.p.1 to $Q_1(x_1)$, as $N_1\to\infty$ and $N_2\to\infty$, uniformly in $x_1$ on any compact subset of $V$. The assertions then follow. □

Let us observe that it was possible to construct in the above proof such set $\Upsilon$, independent of $\xi_2^i$, because of the assumption (A6). It is straightforward to generalize this assumption to $T$-stage problems with $T\geq 3$, and hence to extend the above consistency results.

Note again that under the assumption that the random vectors $\boldsymbol{\xi}_2$ and $\boldsymbol{\xi}_3$ are independent, the sample $\boldsymbol{\xi}_3^{ij}$ does not depend on the probability distribution of $\boldsymbol{\xi}_2$. Therefore, in that case we can generate a random sample $\boldsymbol{\xi}_3^j$, $j=1,\ldots,N_2$, of $N_2$ replications of $\boldsymbol{\xi}_3$, independent of the second stage sample $\boldsymbol{\xi}_2^i$, and to take $\boldsymbol{\xi}_3^{ij}:=\boldsymbol{\xi}_3^j$ for all $i$ and $j$. The consistency results of Proposition

3.3 then hold. Generating sample in that way simplifies the problem (3.4), since then $Q_3(x_2, \xi_3^{ij}) = Q_3(x_2, \xi_3^j)$ is independent of $i$.

Suppose now that the total number of scenarios of the considered (true) three-stage problem is finite. Then the expected value function $Q_1(x_1)$ is convex piecewise linear. Suppose further that $Q_1(x_1)$ is finite for all $x_1$ in a neighborhood of the optimal solutions set $S$. Then the set $S$ is a polyhedron. Consider the (random) function:

$$\tilde{f}_{N_1,N_2}(x_1) := c_1 x_1 + \frac{1}{N_1}\sum_{i=1}^{N_1} \widehat{Q}_{2,N_2}(x_1, \xi_2^i). \tag{3.12}$$

Note that $\widehat{S}_{N_1,N_2}$ is the set of minimizers of $\tilde{f}_{N_1,N_2}(x_1)$ over $X_1$. Consider the event:

$(\mathcal{E})$ "The set $\widehat{S}_{N_1,N_2}$ is nonempty and forms a face of the set $S$".

Of course, it follows from the above event that $\widehat{S}_{N_1,N_2}$ is a subset of $S$. We can view $\tilde{f}_{N_1,N_2}(x_1) = \tilde{f}_{N_1,N_2}(x_1, \omega)$ as a sequence of random functions defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. By saying that the event $(\mathcal{E})$ happens w.p.1 for $N_1$ and $N_2$ large enough we mean that for $\mathbb{P}$-almost every $\omega \in \Omega$ there exists an integer $M = M(\omega)$ such that for all $N_1 \geq M$ and $N_2 \geq M$ the event $(\mathcal{E})$ happens. The following result, about finite convergence of the set of optimal solutions of the SAA three-stage program, is an extension of Theorem 2.3 in [13].

**Proposition 3.4.** *Suppose that the number of scenarios of the considered (true) three-stage problem is finite and assumptions* (A1)–(A5) *and* (A7) *hold. Then the event* $(\mathcal{E})$ *happens w.p.1 for* $N_1$ *and* $N_2$ *large enough.*

*Proof.* Since the number scenarios is finite we have that the function $Q_1(x_1)$ is convex piecewise linear. Since the set $X_1$ is a polyhedron and because of the assumptions (A1)–(A2), it follows then that the set $S$ is a polyhedron. Consider the functions $f(x_1) := c_1 x_1 + Q_1(x_1)$ and

$$\hat{f}_{N_1}(x_1) := c_1 x_1 + \frac{1}{N_1}\sum_{i=1}^{N_1} Q_2(x_1, \xi_2^i),$$

and the function $\tilde{f}_{N_1,N_2}(x_1)$ defined in (3.12). By using the polyhedral structure of the problem it is possible to show, in the same way as in the proof of Lemma 2.4 in [13], the following: there exist two finite sets $\Delta \subset S$ and $\Theta \subset (X_1 \cap V) \setminus S$ such that $\Delta$ forms the set of extreme points of $S$ and if the following condition holds

$$\tilde{f}_{N_1,N_2}(x) < \tilde{f}_{N_1,N_2}(z) \quad \text{for any } x \in \Delta \text{ and } z \in \Theta, \tag{3.13}$$

then the set $\widehat{S}_{N_1,N_2}$ is nonempty and forms a face of the set $S$, i.e., the event $(\mathcal{E})$ happens.

Since the sets $\Delta$ and $\Theta$ are finite, we have that there exists $\varepsilon > 0$ such that $f(z) - f(x) \geq \varepsilon$ for all $x \in \Delta$ and $z \in \Theta$. Also since the number of scenarios is finite, the LLN specified in assumption (A3) holds uniformly in $x_1 \in \Delta \cup \Theta$. Therefore we have that $\hat{f}_{N_1}(z) - \hat{f}_{N_1}(x) \geq \varepsilon/2$ for all $x \in \Delta$ and $z \in \Theta$ w.p.1 for $N_1$ large enough. Also by assumptions (A4) and (A7) we have that the inequality

$$\max_{x_1 \in \Delta \cup \Theta} \left| \tilde{f}_{N_1,N_2}(x_1) - \hat{f}_{N_1}(x_1) \right| < \varepsilon/4 \tag{3.14}$$

holds w.p.1 for $N_1$ and $N_2$ large enough. It follows then that (3.13) holds w.p.1 for $N_1$ and $N_2$ large enough, and hence the proof is complete. $\square$

Unfortunately, it is not clear whether it is possible to extend a result from [13] to show an exponential rate of convergence of the probability of the event $(\mathcal{E})$ to one. This is because (see (3.6)) we have here that $\widehat{Q}_{2,N_2}(x_1, \xi_2^i)$ is a biased estimator of $Q_2(x_1, \xi_2^i)$.

## 4 Conclusions

We showed that by generating a sample from the random process $\boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T$ governing a considered multistage program and solving the obtained SAA problems, one obtains a valid statistical lower bound of the true optimal value $v^*$. So validity of such statistical lower bound holds for any random sample. However, in order to construct a consistent statistical lower bound one needs to employ the conditional sampling scheme. Unfortunately, the number of scenarios in conditional sampling grows fast with the number $T$ of stages. That is, if we generate $N_1$ scenarios at the second stage, $N_2$ scenarios at the third stage conditional on every second stage scenario and etc., then the total number of scenarios is $N = \prod_{t=1}^{T-1} N_t$. If the random vectors $\boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T$ are independent of each other, then one can generate independent random samples, of sizes $N_1, \ldots, N_{T-1}$, from the respective random vectors $\boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_T$, and to employ these sample in the corresponding conditional sampling scheme. This simplifies the constructed (sample) $T$-stage problem, although the total number of scenarios $N = \prod_{t=1}^{T-1} N_t$ remains the same.

It was also demonstrated that fixing a feasible first stage solution and then constructing a SAA estimate for the obtained $(T-1)$-stage problem, does not give a valid statistical upper bound of $v^*$ for any considered sampling scheme if $T \geq 3$. In order to compute a valid statistical upper bound one needs to construct an implementable and feasible policy. However, for $T \geq 3$, it could be difficult to construct such a policy which will provide a tight and numerically feasible upper bound.

## References

[1] Asmussen S (1987) Applied Probability and Queues. John Wiley & Sons, New York, NY
[2] Birge JR, Louveaux F (1997) Introduction to Stochastic Programming. Springer Series in Operations Research. Springer-Verlag, New York, NY
[3] Dupačová J, Wets RJ-B (1988) Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. The Annals of Statistics 16:1517–1549
[4] Dupačová J, Consigli G, Wallace SW (2001) Scenarios for multistage stochastic programs. Annals of Operations Research 100:25–53
[5] Linderoth J, Shapiro A, Wright S (2002) The empirical behavior of sampling methods for stochastic programming. Optimization Technical Report 02-01, Computer Sciences Department, University of Wisconsin-Madison
[6] Mak WK, Morton DP, Wood RK (1999) Monte Carlo bounding techniques for determining solution quality in stochastic programs. Operations Research Letters 24:47–56
[7] Norkin VI, Pflug GCh, Ruszczyński A (1998) A branch and bound method for stochastic global optimization. Mathematical Programming 83:425–450

[8] Pflug GCh (2001) Scenario tree generation for multiperiod financial optimization by optimal discretization. Mathematical Programming, Ser. B 89:251–271
[9] Robinson SM (1996) Analysis of sample-path optimization. Mathematics of Operations Research 21:513–528
[10] Rockafellar RT (1970) Convex Analysis. Princeton University Press, Princeton, New Jersey
[11] Rubinstein RY, Shapiro A (1993) Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method. John Wiley & Sons, Chichester, England
[12] Shapiro A (1991) Asymptotic analysis of stochastic programs. Annals of Operations Research 30:169–186
[13] Shapiro A, Homem-de-Mello T (2000) On rate of convergence of Monte Carlo approximations of stochastic programs. SIAM Journal on Optimization 11:70–86
[14] Verweij B, Ahmed S, Kleywegt AJ, Nemhauser G, Shapiro A (2003) The sample average approximation method applied to stochastic routing problems: a computational study. Computational Optimization and Applications 24:289–333