MONTE CARLO SAMPLING APPROACH TO STOCHASTIC PROGRAMMING *

A. Shapiro[1]

Abstract. Various stochastic programming problems can be formulated as problems of optimization of an expected value function. Quite often the corresponding expectation function cannot be computed exactly and should be approximated, say by Monte Carlo sampling methods. In fact, in many practical applications, Monte Carlo simulation is the only reasonable way of estimating the expectation function. We discuss converges properties of the sample average approximation (SAA) approach to stochastic programming. We argue that the SAA method is easily implementable and can be surprisingly efficient for some classes of stochastic programming problems.

1. Introduction

Consider the following stochastic programming problem

$$\underset{x \in X}{\text{Min}} \; \left\{ f(x) := \mathbb{E}_P[F(x, \boldsymbol{\xi})] \right\}. \tag{1}$$

Here $F(x, \xi)$ is a 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 ξ we denote their realizations). The random vector $\boldsymbol{\xi}$ represents the uncertain parameters (data) of the problem. It is assumed that the probability distribution P is known and the corresponding expected value function $f(x)$ is well defined. We refer to (1) as the "true" (or expected value) problem.

Suppose that P has a finite support, i.e., $\boldsymbol{\xi}$ can take values $\xi_1, ..., \xi_K$ (called *scenarios*) with respective probabilities $p_1, ..., p_K$. In that case

$$\mathbb{E}_P[F(x, \boldsymbol{\xi})] = \sum_{k=1}^{K} p_k F(x, \xi_k). \tag{2}$$

In two-stage linear stochastic programming with recourse, the feasible set X can be written in the form $X := \{x : Ax = b, \; x \geq 0\}$ and $F(x, \xi)$ is given by the optimal value of the second stage program:

$$\underset{y \geq 0}{\text{Min}} \, c^T x + q^T y \;\; \text{s.t.} \;\; Tx + Wy = h, \tag{3}$$

with $\xi := (q, h, T, W)$.

If the number of scenarios is finite, then the two stage linear stochastic program can be formulated as one (large) linear program:

$$
\begin{array}{ll}
\text{Min}_{x,y_1,...,y_K} & c^T x + \sum_{k=1}^{K} p_k q_k^T y_k \\
\text{subject to} & Ax = b, \\
& T_k x + W_k y_k = h_k, \quad k = 1, ..., K, \\
& x \geq 0, \ y_k \geq 0, \quad k = 1, ..., K.
\end{array}
$$

However, typically the number K of scenarios grows *exponentially* with the dimension of the data $\boldsymbol{\xi}$. For example, if d random components of $\boldsymbol{\xi}$ are independent, each having just 3 possible realizations, then the total number of scenarios $K = 3^d$. No computer in a foreseeable future will be able to handle calculations involving 3^{100} scenarios.

In view of the exponential growth of the number of scenarios, with increase of the number of random parameters, the natural question is: *how difficult is to solve the above two-stage problem*? And, moreover, what can be said about *multi-stage problems*?

Example 1.1 (financial planning). Suppose that we want to invest an amount of W_0 in n assets, x_i, $i = 1, ..., n$, in each. That is,

$$W_0 = \sum_{i=1}^{n} x_i. \tag{4}$$

After one period of time our wealth becomes

$$W_1 = \sum_{i=1}^{n} \xi_i x_i, \tag{5}$$

where $\xi_i = 1 + R_i$ and R_i is the return of the i-th asset. We would like to maximize W_1 by making an "optimal" distribution of our initial wealth. Of course, we have to make a decision about x_i *before* a realization of the returns R_i (of ξ_i) become known. Suppose that we have an idea, may be from historical data, about probability distribution of $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_n)$. Then we may think about maximizing W_1 on *average*. That is, we would like to maximize the expected value $\mathbb{E}[W_1]$ of our wealth subject to the budget constraint (4) and "no borrowing" constraints $x_i \geq 0$. This leads to the optimization problem

$$\text{Max}_{x \geq 0} \ \mathbb{E}[W_1] \quad \text{subject to} \ \sum_{i=1}^{n} x_i = W_0. \tag{6}$$

We have that

$$\mathbb{E}[W_1] = \mathbb{E}\left[\sum_{i=1}^{n} \boldsymbol{\xi}_i x_i\right] = \sum_{i=1}^{n} \mu_i x_i,$$

where $\mu_i = \mathbb{E}[\boldsymbol{\xi}_i]$. Consequently, problem (6) has the simple optimal solution of investing everything into the asset with the maximal expected return.

Suppose now that we have a target wealth of τ. If W_1 falls short of τ we are penalized by $r(W_1 - \tau)$, and if W_1 exceeds τ we are rewarded by $q(W_1 - \tau)$, with $r > q$. This leads to the following utility function

$$
U(w) := \begin{cases} r(w - \tau), & \text{if} \quad w \leq \tau, \\ q(w - \tau), & \text{if} \quad w \geq \tau, \end{cases} \tag{7}
$$

and to the optimization problem

$$\text{Max}_{x \geq 0} \ \mathbb{E}\left[F(x, \boldsymbol{\xi})\right] \quad \text{s.t.} \ \sum_{i=1}^{n} x_i = W_0, \tag{8}$$

where $F(x, \xi) := U\left(\sum_{i=1}^{n} \xi_i x_i\right)$. The concept of utility function is well known in economics and game theory (see, e.g., Levy [10] and references therein).

Let us observe that the above $F(x, \xi)$ can be represented as the optimal value of the problem

$$
\begin{array}{ll}
\text{Max}_{y \geq 0, z \geq 0} & qy - rz \\
\text{subject to} & \sum_{i=1}^{n} \xi_i x_i = \tau + y + z.
\end{array} \tag{9}
$$

Therefore we can view our decision process in two stages. At the first stage we have to make a decision about x_i *before* realizations of the random variables $\boldsymbol{\xi}_i$ become known. At the second stage, after ξ_i become known, we optimize our behavior by solving problem (9). Of course, in this example the second stage problem (9) has a closed form simple solution.

Consider now the deterministic counterpart

$$\text{Max}_{x \geq 0} \ F(x, \mu) \quad \text{s.t.} \ \textstyle\sum_{i=1}^n x_i = W_0, \tag{10}$$

of the stochastic problem (8). That is, random variables $\boldsymbol{\xi}_i$ in (8) are replaced by their means $\mu_i = \mathbb{E}[\boldsymbol{\xi}_i]$. Let us note that for any x, the function $F(x, \cdot)$ is concave. Therefore, by Jensen's inequality, $F(x, \mu) \geq \mathbb{E}[F(x, \boldsymbol{\xi})]$. It follows that the optimal value of problem (10) is bigger than the optimal value of problem (8). This should be not surprising that the deterministic problem (10) may give a too "optimistic" optimal value since it does not take into account a possible variability of the random data $\boldsymbol{\xi}$.

Suppose now that we can rebalance our portfolio at several, say T, periods of time. That is, at the beginning we choose values x_{i0} of our assets subject to the budget constraint

$$\textstyle\sum_{i=1}^n x_{i0} = W_0. \tag{11}$$

At the period $t = 1, ..., T$, our wealth is

$$W_t = \textstyle\sum_{i=1}^n \xi_{it} x_{i,t-1}, \tag{12}$$

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

$$\text{Max} \, \mathbb{E}\left[U(W_T)\right] \tag{13}$$

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

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

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

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

This becomes a multi-stage stochastic programming problem. We refer to [1] and [16] for a further discussion of two and multi-stage stochastic programming problems.

2. Monte Carlo sampling approach

Let $\xi^1, ..., \xi^N$ be a generated (say, iid) random sample drawn from P. For a given $x \in X$, the sample average

$$\hat{f}_N(x) := N^{-1} \sum_{j=1}^N F(x, \xi^j) \tag{14}$$

can be used as an unbiased estimator of the expected value $f(x) = \mathbb{E}_P[F(x, \boldsymbol{\xi})]$. By the Law of Large Numbers, we have that $\hat{f}_N(x)$ converges to $f(x)$ with probability one (w.p.1) as $N \to \infty$. However, the convergence is of

order $O_p(N^{-1/2})$ and is notoriously slow. In order to improve the accuracy by one digit the sample size should be increased 100 times. By the Central Limit Theorem

$$N^{1/2}\left[\hat{f}_N(x) - f(x)\right] \Rightarrow N(0, \sigma^2(x)),$$

where $\sigma^2(x) := \mathbb{V}ar[F(x, \boldsymbol{\xi})]$ and " $\Rightarrow$ " denotes convergence in distribution. The good news is that the rate of convergence does not depend on the number of scenarios, only on the variance $\sigma^2(x)$. The accuracy can be improved by variance reduction techniques. However, the rate of the square root of N (of Monte Carlo sampling estimation) cannot be changed.

Two basic philosophies of the Monte Carlo approach are the *interior* and *exterior* sampling. In interior sampling methods, sampling is performed inside a chosen algorithm with new (independent) or updated samples generated in the process of iterations, e.g., Higle and Sen (stochastic decomposition method) [5], Infanger (statistical L-shape method) [7], Norkin, Pflug and Ruszczynski (stochastic branch and bound method) [13].

In the exterior sampling approach the true problem is approximated by the sample average approximation problem:

(SAA) $$\underset{x \in X}{\text{Min}}\left\{\hat{f}_N(x) := N^{-1}\sum_{j=1}^{N}F(x, \xi^j)\right\}.$$

Once the sample $\xi^1, ..., \xi^N \sim P$ is generated, the SAA problem becomes a deterministic optimization problem and can be solved by an appropriate algorithm.

It is difficult to point out an exact origin of this method. Variants of this approach were suggested by a number of authors under different names (e.g., Rubinstein and Shapiro (stochastic counterpart method) [15], Plambeck, Fu, Robinson and Suri (sample-path optimization) [14], Mak, Morton and Wood [12], Kleywegt, Shapiro and Homem-De-Mello [9]). Advantages of the SAA method are: (i) ease of numerical implementation, often one can use existing software, (ii) good convergence properties, (iii) well developed statistical inference: validation and error analysis, stopping rules, (iv) easily amenable to variance reduction techniques, (v) ideal for parallel computations. Numerical experiments with the SAA method can be also found in Linderoth, Shapiro and Wright [11] and Verweij, Ahmed, Kleywegt, Nemhauser and Shapiro [24].

The SAA approach can be related to the idea of *common random numbers generation* which is well known in the area of simulation. Suppose that the feasible set X consists of two points, $X = \{x_1, x_2\}$. Note that for the purpose of optimization we are not interested in an accurate estimation of the expectations $f(x_1)$ and $f(x_2)$, but rather in their difference $f(x_1) - f(x_2)$. Now if the same sample is used in the estimators $\hat{f}_N(x_1)$ and $\hat{f}_N(x_2)$, then the variance of $N^{1/2}\left[\hat{f}_N(x_1) - \hat{f}_N(x_2)\right]$ is $\mathbb{V}ar[F(x_1, \boldsymbol{\xi})] + \mathbb{V}ar[F(x_2, \boldsymbol{\xi})] - 2\,\mathbb{C}ov[F(x_1, \boldsymbol{\xi}), F(x_2, \boldsymbol{\xi})]$. This variance can be much smaller than $\mathbb{V}ar[F(x_1, \boldsymbol{\xi})] + \mathbb{V}ar[F(x_2, \boldsymbol{\xi})]$, when independent samples are used.

We will use the following notation: v^0 is the optimal value of the true problem, S^0 is the optimal solutions set of the true problem, S^ε is the set of ε-optimal solutions of the true problem, $\hat{v}_N$ is the optimal value of the SAA problem, $\hat{S}_N^\varepsilon$ is the set of ε-optimal solutions of the SAA problem, $\hat{x}_N$ is an optimal solution of the SAA problem.

There is a vast literature on statistical properties of the SAA estimators $\hat{v}_N$ and $\hat{x}_N$. By the Law of Large Numbers, $\hat{f}_N(x)$ converge (pointwise) to $f(x)$ w.p.1. Under mild additional conditions, this implies that $\hat{v}_N \to v^0$ and $\text{dist}(\hat{x}_N, S^0) \to 0$ w.p.1, as $N \to \infty$. In particular, $\hat{x}_N \to x^0$ w.p.1 if the true optimal set $S^0 = \{x^0\}$ is a singleton. That is, in the terminology of theoretical statistics, the estimators $\hat{v}_N$ and $\hat{x}_N$ are consistent. Since the pioneering work of Wald [25] on consistency of Maximum Likelihood estimators, there have been numerous publications on that subject. In a recent literature on stochastic programming the consistency of statistical estimators is often treated in a framework of the epi-convergence theory (see Dupačová and Wets [4] and references therein).

There are Central Limit Theorem type results about asymptotics of $\hat{v}_N$ and $\hat{x}_N$. We have that

$$\hat{v}_N = \min_{x \in S^0} \hat{f}_N(x) + o_p(N^{-1/2}), \tag{15}$$

Shapiro [18]. In particular, if $S^0 = \{x^0\}$, then

$$N^{1/2}[\hat{v}_N - v^0] \Rightarrow N(0, \sigma^2(x^0)). \tag{16}$$

These results suggest that the optimal value of the SAA problem converges at a rate of $O_p(N^{-1/2})$. In particular, if $S^0 = \{x^0\}$, then $\hat{v}_N$ converges to v^0 at the same rate as $\hat{f}_N(x^0)$ converges to $f(x^0)$.

If $S^0 = \{x^0\}$ is a singleton, then under certain regularity conditions, $N^{1/2}(\hat{x}_N - x^0)$ converges in distribution to a limiting (often normal) distribution (asymptotic normality of M-estimators, Huber [6]). Regularity conditions which usually assumed in order to ensure such asymptotics are that the expected value function $f(x)$ is smooth (twice differentiable) at x^0 and the Hessian matrix $\nabla^2 f(x^0)$ is positive definite (cf., King and Rockafellar [8], Shapiro [17]). This typically happens if the probability distribution P is *continuous*. In such cases, and if x^0 is an interior point of the feasible set X, $\hat{x}_N$ converges to x^0 at the same rate as the stochastic approximation iterates calculated with the optimal step sizes (Shapiro [19]). Some results about asymptotics of the SAA optimal solution set $\hat{S}_N$ are available also in cases where S^0 is not a singleton (Dentcheva and Römisch [3]).

2.1. Complexity issues

Using the theory of Large Deviations (see, e.g., Dembo and Zeitouni [2]) it is possible to give the following estimates. Suppose, for a moment, that the feasible set X is *finite*. Consider a mapping $u : X \setminus S^\varepsilon \to S^0$, and

$$H(x, \xi) := F(u(x), \xi) - F(x, \xi).$$

Suppose that for every $x \in X$ the moment generating function of $H(x, \boldsymbol{\xi})$ is finite valued in a neighborhood of zero. Let ε and δ be nonnegative numbers such that $\delta \leq \varepsilon$. Then there is $\gamma(\delta, \varepsilon) > 0$ such that

$$P\left(\hat{S}_N^\delta \not\subset S^\varepsilon\right) \leq |X| e^{-N\gamma(\delta, \varepsilon)}. \tag{17}$$

The constant $\gamma(\delta, \varepsilon)$ can be estimated

$$\gamma(\delta, \varepsilon) \geq \frac{(\varepsilon^* - \delta)^2}{3\sigma^2} > \frac{(\varepsilon - \delta)^2}{3\sigma^2}, \tag{18}$$

where

$$\varepsilon^* := \min_{x \in X \setminus S^\varepsilon} f(x) - v^0$$

and

$$\sigma_{\max}^2 := \max_{x \in X \setminus S^\varepsilon} \mathbb{V}ar[H(x, \boldsymbol{\xi})]$$

(Kleywegt, Shapiro and Homem-de-Mello [9]). Note that $\varepsilon^* > \varepsilon$. This gives the following estimate of the sample size N which guarantees that $\mathbb{P}\left(\hat{S}_N^\delta \subset S^\varepsilon\right) \geq 1 - \alpha$, for a given $\alpha \in (0, 1)$:

$$N \geq \frac{3\sigma_{\max}^2}{(\varepsilon - \delta)^2} \log\left(\frac{|X|}{\alpha}\right). \tag{19}$$

The above estimate (19) typically is too conservative to be used for practical calculations. However, it indicates that the required sample size grows as a logarithm of $|X|$.

Now let X be a bounded subset of $\mathbb{R}^n$. Then for a given $\nu > 0$, consider a finite subset X_ν of X such that for any $x \in X$ there is $x' \in X_\nu$ satisfying $\|x - x'\| \leq \nu$. If D is the diameter of the set X, then such set X_ν can be constructed with $|X_\nu| \leq \left(\frac{D}{\nu}\right)^n$. Reducing the feasible set X to its subset X_ν, we obtain the following estimate of the required sample size to solve the reduced problem

$$N \geq \frac{3\sigma_{\max}^2}{(\varepsilon - \delta)^2}\left[n\log\left(\frac{D}{\nu}\right) - \log\alpha\right].$$

Suppose, further, that $f(x)$ is Lipschitz continuous modulus L. By taking $\nu := (\varepsilon - \delta)/(2L)$ we obtain the following estimate of the required sample size to solve the the true problem

$$N \geq \frac{12\sigma_{\max}^2}{(\varepsilon - \delta)^2}\left[n\log\left(\frac{2DL}{(\varepsilon - \delta)^2}\right) - \log\alpha\right]. \tag{20}$$

This suggests a *linear* growth of the required sample size with the dimensionality n of the first stage problem. We refer to [16, Chapter 6] for a detail discussion of the above derivations.

Suppose now that the true problem is *convex piecewise linear*. That is: (i) the distribution P has a finite support, (ii) for almost every ξ the function $F(\cdot, \xi)$ is piecewise linear and convex, (iii) the feasible set X is polyhedral (i.e., is defined by a finite number of linear constraints). Suppose also that the optimal solutions set S^0, of the true problem, is nonempty and bounded. Then the following holds.

(1) W.p.1 for N large enough, $\hat{x}_N$ is an *exact* optimal solution of the true problem. More precisely, w.p.1 for N large enough, the set $\hat{S}_N$ of optimal solutions of the SAA problem is nonempty and forms a face of the (polyhedral) set S^0.

(2) Probability of the event $\{\hat{S}_N \subset S^0\}$ tends to one *exponentially fast*. That is, there exists a constant $\gamma > 0$ such that

$$\lim_{N\to\infty}\frac{1}{N}\log\left[1 - P(\hat{S}_N \subset S^0)\right] = -\gamma, \tag{21}$$

(Shapiro and Homem-de-Mello [22]).

Suppose that the problem is *convex piecewise linear*, and let x^0 be unique optimal solution of the true problem. Then

$$f'(x^0, d) > 0, \quad \forall\, d \in T_X(x^0) \setminus \{0\}. \tag{22}$$

(Here $f'(x^0, d)$ denotes the directional derivative of the function $f(\cdot)$ at x^0 in the direction d, and $T_X(x^0)$ denotes the tangent cone to X at x^0.) Furthermore, there exists a *finite* set $\{d_1, ..., d_\ell\} \subset T_X(x_0)$ of nonzero directions, independent of the sample, such that if $\hat{f}'_N(x^0, d_j) > 0$ for $j = 1, ..., \ell$, then $\hat{x}_N = x^0$. We call

$$\kappa := \max_{j\in\{1,...,\ell\}}\frac{\mathbb{V}ar\left[F'_\xi(x^0, d_j)\right]}{\left[f'(x^0, d_j)\right]^2} \tag{23}$$

the *condition number* of the true problem. Recall that $\mathbb{E}\left[F'_\xi(x^0, d)\right] = f'(x^0, d)$.

For convex piecewise linear problems with unique optimal solution, the exponential constant γ, defined in the estimate (21), is approximately equal to $(2\kappa)^{-1}$. This means that the sample size N required to achieve a given probability of the event "$\hat{x}_N = x^0$" is roughly proportional to the condition number κ. More accurately, for large N and κ,

$$P(\hat{x}_N \neq x^0) \approx \frac{Ce^{-N/(2\kappa)}}{\sqrt{4\pi N/(2\kappa)}}, \tag{24}$$

where C is a positive constant close to one and independent of the sample (Shapiro, Homem de Mello and Kim [23]).

2.2. **Validation analysis**

How one can evaluate quality of a given solution $\hat{x} \in X$? There are two basic approaches to validation of a given solution: (i) Evaluate the gap $f(\hat{x}) - v^0$. (ii) Verify the KKT optimality conditions at $\hat{x}$. The following statistical test, based on estimation of $f(\hat{x}) - v^0$, was suggested by Norkin, Pflug and Ruszczynski [13] and developed further in Mak, Morton and Wood [12].

(1) Estimate $f(\hat{x})$ by the sample average $\hat{f}_{N'}(\hat{x})$, using sample of a large size N'.
(2) Solve the SAA problem M times, using M independent samples, each of size N. Let $\hat{v}_N^{(1)}, ..., \hat{v}_N^{(M)}$ be the optimal values of the corresponding SAA problems. Estimate $\mathbb{E}[\hat{v}_N]$ by the average $M^{-1} \sum_{j=1}^{M} \hat{v}_N^{(j)}$.

Note that

$$\mathbb{E}\left[\hat{f}_{N'}(\hat{x}) - M^{-1} \sum_{j=1}^{M} \hat{v}_N^{(j)}\right] = \left(f(\hat{x}) - v^0\right) + \left(v^0 - \mathbb{E}[\hat{v}_N]\right),$$

and that $v^0 - \mathbb{E}[\hat{v}_N] > 0$. The bias $v^0 - \mathbb{E}[\hat{v}_N]$ is positive and (under mild regularity conditions)

$$\lim_{N \to \infty} N^{1/2} \left(v^0 - \mathbb{E}[\hat{v}_N]\right) = \mathbb{E}\left[\max_{x \in S^0} Y(x)\right],$$

where $(Y(x_1), ..., Y(x_k))$ has a multivariate normal distribution with zero mean vector and covariance matrix given by the covariance matrix of the random vector $(F(x_1, \boldsymbol{\xi}), ..., F(x_k, \boldsymbol{\xi}))$ (Shapiro [18]). For ill-conditioned problems this bias is of order $O(N^{-1/2})$ and can be large if the ε-optimal solution set S^ε is large for some small $\varepsilon \geq 0$.

It is also possible to give the following common random numbers variant of the above procedure (Mak, Morton and Wood [12]). Generate a sample (of size N) and calculate the gap

$$\hat{f}_N(\hat{x}) - \inf_{x \in X} \hat{f}_N(x).$$

Repeat this procedure M times (with independent samples), and calculate the average of the above gaps. This procedure works well for well conditioned problems, but does not improve the bias problem.

Another approach to validation of a given solution $\hat{x} \in X$ is based on verification of first-order (KKT) optimality conditions. Let the feasible set be defined by constraints

$$X := \{x \in \mathbb{R}^n : c_i(x) = 0, \ i \in I, \ c_i(x) \leq 0, \ i \in J\},$$

where $c_i(x)$ are continuously differentiable functions. Suppose that the probability distribution of $\boldsymbol{\xi}$ is continuous. Then, under mild regularity conditions (e.g., [16, Chapter 2]), $F(\cdot, \boldsymbol{\xi})$ is differentiable at $\hat{x}$ w.p.1 and

$$\nabla f(\hat{x}) = \mathbb{E}_P\left[\nabla_x F(\hat{x}, \boldsymbol{\xi})\right]. \tag{25}$$

The KKT-optimality conditions at an optimal solution $x^0 \in S^0$ can be written as follows:

$$-\nabla f(x^0) \in C(x^0), \tag{26}$$

where

$$C(x) := \left\{y = \sum_{i \in I \cup J(x)} \lambda_i \nabla c_i(x), \ \lambda_i \geq 0, \ i \in J(x)\right\},$$

and $J(x) := \{i : c_i(x) = 0, \ i \in J\}$. The idea of the KKT test is to estimate the distance

$$\delta(\hat{x}) := \text{dist}\left(-\nabla f(\hat{x}), C(\hat{x})\right),$$

by using the sample estimator

$$\hat{\delta}_N(\hat{x}) := \operatorname{dist}\left(-\nabla \hat{f}_N(\hat{x}), C(\hat{x})\right).$$

The covariance matrix of $\nabla \hat{f}_N(\hat{x})$ can be estimated (from the same sample), and hence a confidence region for $\nabla f(\hat{x})$ can be constructed. This allows to construct a statistical test for validation of the KKT conditions (see Shapiro and Homem-de-Mello [21] for details).

3. Multistage stochastic programming

Consider the following (nested) formulation of a multi-stage linear stochastic programming problem

$$\operatorname*{Min}_{\substack{A_1 x_1 = b_1 \\ x_1 \geq 0}} c_1 x_1 + \mathbb{E}\Big\{ \operatorname*{Min}_{\substack{B_2 x_1 + A_2 x_2 = b_2 \\ x_2 \geq 0}} c_2 x_2 + \cdots$$

$$\cdots + \mathbb{E}\Big[\operatorname*{Min}_{\substack{B_T x_{T-1} + A_T x_T = b_T \\ x_T \geq 0}} c_T x_T \Big]\Big\}.$$

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

$$\operatorname*{Min}_{\substack{B_2 x_1 + A_2 x_2 = b_2 \\ x_2 \geq 0}} c_2 x_2 +$$

$$\mathbb{E}\Big\{ \cdots + \mathbb{E}\Big[\operatorname*{min}_{\substack{B_T x_{T-1} + A_T x_T = b_T \\ x_T \geq 0}} c_T x_T \Big]\Big\},$$

then we can write the T–stage problem in the following form of two-stage programming problem

$$\operatorname*{Min}_{x_1} c_1 x_1 + \mathbb{E}\big[Q_2(x_1, \boldsymbol{\xi}_2)\big] \text{ s.t. } A_1 x_1 = b_1, \ x_1 \geq 0.$$

Note, however, that if $T \geq 3$, then the above problem in itself is a stochastic programming problem. Consequently, if the number of involved scenarios is very large, or infinite, then the optimal value $Q_2(x_1, \xi_2)$ may be calculated only approximately, say by sampling.

If we relax the nonanticipativity constraints starting from the second stage of the above T–stage problem, we obtain the following two-stage program

$$\operatorname*{Min}_{x_1} c_1 x_1 + \mathbb{E}\Big[Q_2(x_1, \boldsymbol{\xi}_{[2,T]})\Big] \text{ s.t. } A_1 x_1 = b_1, \ x_1 \geq 0,$$

where $\xi_{[2,T]} := (\xi_2, ..., \xi_T)$ and $Q_2(x_1, \xi_{[2,T]})$ is the optimal value of the corresponding linear program. For example, for $T = 3$, the value $Q_2(x_1, \xi_{[2,T]})$ is given by the optimal value of the problem

$$\begin{aligned} \operatorname*{Min}_{x_2, x_3} \quad & c_2 x_2 + c_3 x_3 \\ \text{subject to} \quad & B_2 x_1 + A_2 x_2 = b_2, \\ & B_3 x_2 + A_3 x_3 = b_3, \\ & x_2 \geq 0, \ x_3 \geq 0. \end{aligned}$$

Since the above two-stage problem is obtained by a relaxation of the nonanticipativity constraints, its optimal value is smaller than the optimal value of the corresponding T–stage problem. Note that direct sampling of scenarios solves consistently (i.e., in the limit) the relaxed two-stage problem and not the original T–stage problem.

If the number of scenarios is finite, then linear programming formulations of the T–stage and its two-stage relaxation problems are comparable in size. However, the two-stage problem can be solved efficiently by sampling even if the number of scenarios is very large, while the corresponding multistage problem may be intractable. In order to solve the multistage problem consistently by sampling one needs to employ a conditional sampling scheme. However, the number of required scenarios in a corresponding sample average approximation, may grow exponentially fast with the number of stages. We refer to Shapiro [20] for a further discussion of this problem.

References

[1] J.R. Birge and F.V. Louveaux, *Introduction to Stochastic Programming*, Springer-Verlag, New York, 1997.

[2] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Springer-Verlag, New York, 1998.

[3] J. Dupačová and R.J.B. Wets, Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems, *The Annals of Statistics*, 16:1517–1549, 1988.

[4] D. Dentcheva and W. Römisch, Differential stability of two-stage stochastic programs, *SIAM Journal on Optimization*, 11:87–112, 2000.

[5] J. L. Higle and S. Sen, *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming*, Kluwer Academic Publishers, Dordrecht, 1996.

[6] P. J. Huber, The behavior of maximum likelihood estimates under nonstandard conditions, *Proc. Fifth Berkeley Symp.Math. Statist. Probab.*, vol. 1, 221-233, Univ. California Press, 1967.

[7] G. Infanger, *Planning Under Uncertainty: Solving Large Scale Stochastic Linear Programs*, Boyd and Fraser Publishing Co., Danvers, MA, 1994.

[8] A.J. King and R.T. Rockafellar, Asymptotic theory for solutions in statistical estimation and stochastic programming, *Mathematics of Operations Research*, 18:148-162, 1993.

[9] A. J. Kleywegt, A. Shapiro, and T. Homem-De-Mello, The sample average approximation method for stochastic discrete optimization, *SIAM Journal of Optimization*, 12:479–502, 2001.

[10] H. Levy, Stochastic dominance and expected utility: survey and analysis, *Management Science*, 38:555–593, 1992.

[11] J. Linderoth, A. Shapiro, and S. Wright, The empirical behavior of sampling methods for stochastic programming, *Annals of Operations Research*, to appear.

[12] W. K. Mak, D. P. Morton, and R. K. Wood, Monte Carlo bounding techniques for determining solution quality in stochastic programs, *Operations Research Letters*, 24:47–56, 1999.

[13] V. I. Norkin, G. Ch. Pflug, and A. Ruszczyński, A branch and bound method for stochastic global optimization, *Mathematical Programming*, 83:425–450, 1998.

[14] E. L. Plambeck, B. R. Fu, S. M. Robinson, and R. Suri, Sample-path optimization of convex stochastic performance functions, *Mathematical Programming, Series B*, 75:137–176, 1996.

[15] R. Y. Rubinstein and A. Shapiro, Optimization of static simulation models by the score function method, *Mathematics and Computers in Simulation*, 32:373–392, 1990.

[16] A. Ruszczyński and A. Shapiro, Eds., *Stochastic Programming*, in: Handbooks in OR & MS, Vol. 10, North-Holland Publishing Company, 2003 (in press).

[17] A. Shapiro, Asymptotic properties of statistical estimators in stochastic programming, *Annals of Statistics*, 17:841-858, 1989.

[18] A. Shapiro, Asymptotic analysis of stochastic programs, *Annals of Operations Research*, 30:169–186, 1991.

[19] A. Shapiro, Simulation-based optimization: Convergence analysis and statistical inference, *Stochastic Models*, 12:425–454, 1996.

[20] A. Shapiro, Statistical inference of multistage stochastic programming problems, *E-print available at* http://www.optimization-online.org, 2002.

[21] A. Shapiro and T. Homem-de-Mello, A simulation-based approach to two-stage stochastic programming with recourse, *Mathematical Programming*, 81:301–325, 1998.

[22] A. Shapiro and T. Homem-de-Mello, On rate of convergence of Monte Carlo approximations of stochastic programs, *SIAM Journal on Optimization*, 11:70–86, 2000.

[23] A. Shapiro, T. Homem de Mello, and J.C. Kim, Conditioning of stochastic programs, *Mathematical Programming*, 94:1–19, 2002.

[24] B. Verweij, S. Ahmed, A. J. Kleywegt, G. Nemhauser, and A. Shapiro, The sample average approximation method applied to stochastic routing problems: a computational study, *Computational Optimization and Applications*, 24:289–333, 2003.

[25] A. Wald, Note on the consistency of the maximum likelihood estimates, *Annals Mathematical Statistics*, 20:595–601, 1949.