Math. Program., Ser. A 94: 1–19 (2002)

Digital Object Identifier (DOI) 10.1007/s10107-002-0313-2

Alexander Shapiro* · Tito Homem-de-Mello · Joocheol Kim

Conditioning of convex piecewise linear stochastic programs

Received: May 2000 / Accepted: May 2002-07-16
Published online: September 5, 2002 –

Abstract. In this paper we consider stochastic programming problems where the objective function is given as an expected value of a convex piecewise linear random function. With an optimal solution of such a problem we associate a condition number which characterizes well or ill conditioning of the problem. Using theory of Large Deviations we show that the sample size needed to calculate the optimal solution of such problem with a given probability is approximately proportional to the condition number.

Key words. stochastic programming – Monte Carlo simulation – large deviations theory – ill-conditioned problems

1. Introduction

Consider the stochastic programming problem

$$\operatorname*{Min}_{x\in S}\left\{f(x) := \mathbb{E}_P[h(x,\boldsymbol{\xi})]\right\}, \tag{1.1}$$

where $\boldsymbol{\xi}$ is a random vector having probability distribution P with support $\Xi \subset \mathbb{R}^d$, S is a nonempty closed subset of $\mathbb{R}^m$ and $h : \mathbb{R}^m \times \Xi \to \mathbb{R}$ is a real valued function (we use the bold face for the random vector $\boldsymbol{\xi}$ in order to distinguish it from its realization $\xi \in \mathbb{R}^d$). We discuss in this paper ill or well conditioning of an optimal solution x_0 of the above problem (1.1). We study the problem of conditioning of x_0 from the point of view of Monte Carlo sampling approximation approach. That is, suppose that an i.i.d. random sample $\boldsymbol{\xi}^1, ..., \boldsymbol{\xi}^N$, with the common distribution P, is generated and that the problem (1.1) is approximated by the problem

$$\operatorname*{Min}_{x\in S}\left\{\widehat{f}_N(x) := \frac{1}{N}\sum_{i=1}^{N} h(x,\boldsymbol{\xi}^i)\right\}. \tag{1.2}$$

We refer to (1.1) and (1.2) as the "true" (or expected value) and the sample average approximating (SSA) problems, respectively.

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

T. Homem-de-Mello: Department of Industrial, Welding and Systems Engineering, The Ohio State University, 1971 Neil Ave., Columbus, Ohio 43210-1271, USA, e-mail: homem-de-mello.1@osu.edu

J. Kim: School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205, USA

* The research of this author was supported, in part, by grant DMS-0073770 from the National Science Foundation.

It turns out that, for a certain class of problems, an optimal solution of the SAA problem (1.2) provides an *exact* optimal solution of the true problem (1.1) with probability one (w.p.1) for the sample size N large enough. In particular, this happens if the following assumptions hold:

(A1) For all $\xi \in \Xi$ the function $h(\cdot, \xi)$ is piecewise linear and convex.
(A2) The set S is polyhedral, i.e., is defined by a finite number of linear constraints.
(A3) The probability distribution P has a finite support, i.e., the set Ξ is finite.

Moreover, if the true problem (1.1) has unique optimal solution x_0, then probability of the event "$\widehat{x}_N = x_0$" approaches one exponentially fast as N tends to infinity. That is, there exists a constant $\beta > 0$ such that

$$\lim_{N\to\infty} \frac{1}{N} \log\big[1 - P(\widehat{x}_N = x_0)\big] = -\beta \tag{1.3}$$

(Shapiro and Homem-de-Mello [20], see also derivations below). By the event "$\widehat{x}_N = x_0$" we mean that the corresponding SAA problem has unique optimal solution equal to x_0. By "$\widehat{x}_N \neq x_0$" we denote the complement of that event.

In the subsequent analysis we assume that conditions (A1)–(A3) hold, and refer to problems satisfying conditions (A1)–(A3) as *convex piecewise linear*. An important class of such problems is given by two-stage linear stochastic programming problems with recourse and a finite number of scenarios (see, e.g., Birge and Louveaux [2] for a discussion of two-stage programming with recourse). As compared with other studies of exponential rates of convergence, based on the theory of Large Deviations (see, e.g., Kaniovski, King and Wets [10] and Dai, Chen and Birge [3] and references therein), the result (1.3) is different in that it asserts that $\widehat{x}_N$ is equal *exactly* to x_0 with probability approaching one exponentially fast. Of course, this is possible since we consider a specific class of problems satisfying assumptions (A1)–(A3).

The above is a qualitative result showing that one may not need a large sample in order to solve the true problem *exactly* with a high probability by solving the SAA problem. The required sample size N is, of course, problem dependent and may be difficult to estimate a priori. In some cases the optimal solution x_0 of the true problem is stable and a relatively small sample size N is needed in order to determine x_0 with a high probability by solving the corresponding SAA problem. It is natural to say that in such cases x_0 is well conditioned, as opposed to ill conditioned problems where a much larger sample is required. One may argue that in practical applications there is no need to solve the true problem exactly. Let us remark, however, that if the true problem has a large number of optimal or nearly optimal solutions (i.e., the problem is ill conditioned), then it may be difficult to validate a calculated solution for optimality. This is because in such cases the optimal value $\hat{v}_N$ of the SAA problem tends to give a downwards biased estimator of the optimal value v_0 of the true problem (see Mak, Morton and Wood [13] for a discussion of statistical lower bounds obtained via optimal values of SAA problems, Shapiro [21] and Kleywegt, Shapiro and Homem-de-Mello [11] for a discussion of the bias phenomenon for ill conditioned problems, and Linderoth, Shapiro and Wright [12] for additional numerical experiments).

From the above point of view, any problem (1.1) with multiple solutions is ill conditioned. In some cases the function $h(x, \xi)$ can be represented in the form

$$h(x, \xi) := c^T x + g(Ax, \xi),$$

where c is an m-dimensional column vector, A is an $n \times m$ matrix and $g : \mathbb{R}^n \times \Xi \to \mathbb{R}$ is a real valued function. For example, this is the case in two stage linear stochastic programming with recourse and nonrandom technology matrix A. If the matrix A has a deficient row rank less than m, then the function $h(\cdot, \xi)$ is constant on any affine subspace of $\mathbb{R}^m$ parallel to the linear space defined by the equations $Ax = 0$. This may result in nonuniqueness of optimal solutions of the corresponding true (expected value) problem. Nevertheless, such a problem can be "well-conditioned". Let us observe that the true and SAA problems (1.1) and (1.2) can be transformed into the following respective equivalent problems

$$\operatorname*{Min}_{\chi \in C} \left\{ Q(\chi) := \psi(\chi) + \mathbb{E}_P \left[g(\chi, \boldsymbol{\xi}) \right] \right\}, \tag{1.4}$$

$$\operatorname*{Min}_{\chi \in C} \left\{ \widehat{Q}_N(\chi) := \psi(\chi) + \frac{1}{N} \sum_{i=1}^{N} g(\chi, \boldsymbol{\xi}^i) \right\}, \tag{1.5}$$

where $C := A(S)$ and $\psi(\chi) := \inf\{c^T x : Ax = \chi,\ x \in S\}$. Note that the feasible set C of the above problem is convex if S is convex, and is polyhedral if S is polyhedral. Note also that the function $\psi(\chi)$ is convex piecewise linear if the set S is convex and polyhedral. Therefore, if the true problem (1.1) is convex piecewise linear, then the transformed true problem (1.4) is also convex piecewise linear. If the problem (1.4) has a unique optimal solution χ_0 which is well-conditioned, then it is natural to say that the true problem (1.1) is also well-conditioned with the same condition number.

Let us also mention that it is well known in almost every branch of numerical analysis that large problems tend to be ill conditioned, e.g., large linear programming problems tend to be degenerate, linear regression models with a large number of predictors tend to have the multicollinearity problem, etc. It is natural to assume that stochastic programming problems are no exceptions in this respect and large stochastic programming problems tend to be ill conditioned. However, our (admittedly limited) numerical experience suggests that in some cases stochastic problems with a finite but huge set Ξ are well conditioned. Of course, more numerical experiments are needed before a definite conclusion could be made.

In this paper we introduce a quantitative concept of the condition number associated with the optimal solution x_0 of the true problem. That condition number gives a characterization of ill (or well) conditioning of the problem from the point of view of Monte Carlo sample average approximation approach. It should be mentioned that the approach to conditioning discussed in this paper is stochastic in nature. From this point of view any deterministic problem with unique solution is well conditioned since in such case a sample of size $N = 1$ suffices to solve the problem exactly. Therefore, the impetus here is somewhat different from the one in a purely deterministic optimization (see Renegar [17] and Freund and Vera [6], and references therein, for a discussion of modern concepts of condition number in deterministic optimization).

We use the following notation and terminology throughout the paper. By $f'(x_0, d)$ and $h'_\xi(x_0, d)$ we denote the directional derivative of $f(\cdot)$ and $h(\cdot, \xi)$, respectively, at x_0 in the direction d. The tangent cone to a convex set S at a point $x \in S$ is denoted by $T_S(x)$, and by $S^{m-1} := \{x \in \mathbb{R}^m : \|x\| = 1\}$ we denote the unit sphere in the space $\mathbb{R}^m$. By $\mathbb{V}\mathrm{ar}[X]$ we denote the variance of the random variable X.

2. Condition number

We assume that conditions (A1)–(A3) hold, and that the true problem (1.1) has unique optimal solution x_0. It follows that the expected value function $f(x)$ is also convex piecewise linear. Moreover, we have then by the theory of linear programming that the optimal solution x_0 of the true problem is *sharp*, that is

$$f'(x_0, d) > 0, \quad \forall\, d \in T_S(x_0) \setminus \{0\}. \tag{2.1}$$

Furthermore, since it is assumed that the problem is convex piecewise linear the following property holds [20]:

(B) There exists a *finite* set $\{d_1, ..., d_\ell\} \subset T_S(x_0) \setminus \{0\}$ of directions, independent of the sample, such that if $\widehat{f}'_N(x_0, d_j) > 0$ for $j = 1, ..., \ell$, then $\widehat{x}_N = x_0$.

Let us also remark that because of the assumed piecewise linearity, we have here that if x_0 is a unique optimal solution of the SAA problem, then $\widehat{f}'_N(x_0, d) > 0$ for any $d \in T_S(x_0) \setminus \{0\}$.

Definition 1. *We call*

$$\kappa := \max_{j \in \{1,...,\ell\}} \frac{\mathbb{V}\mathrm{ar}\big[h'_{\boldsymbol{\xi}}(x_0, d_j)\big]}{\big[f'(x_0, d_j)\big]^2} \tag{2.2}$$

the condition number *of the true problem* (1.1).

The above definition is motivated by the following result which means that the sample size N required to achieve a given probability of the event "$\widehat{x}_N = x_0$" is roughly proportional to the condition number κ.

Theorem 1. *Suppose that the assumptions* (A1)–(A3) *are satisfied and that the true problem* (1.1) *has unique optimal solution* x_0. *Then the exponential rate* (1.3) *holds and the corresponding constant* β *is approximately equal to* $(2\kappa)^{-1}$, *where* κ *is given by* (2.2).

A formal derivation of the above theorem and exact meaning of the approximation $\beta \approx (2\kappa)^{-1}$ will be given in the remainder of this section.

It could be noticed that the above definition of the condition number κ depends on a choice of the set $\{d_1, ..., d_\ell\}$ satisfying property (B). This set is not uniquely defined, since by adding any $d \in T_S(x_0) \setminus \{0\}$ one still gets a set satisfying (B). We will see later that, for well conditioned problems, actually only a few directions are important.

Before giving a formal derivation of Theorem 1, let us make the following remarks. Under mild regularity conditions (e.g., [20]), and in particular if the true problem is convex piecewise linear, it follows that

$$\mathbb{E}_P\big[h'_{\boldsymbol{\xi}}(x_0, d)\big] = f'(x_0, d). \tag{2.3}$$

Thus, κ can be viewed as the largest *squared coefficient of variation* of random variables $h'_{\boldsymbol{\xi}}(x_0, d)$, $d \in \{d_1, ..., d_\ell\}$. Moreover, we have that if $\mathbb{V}\mathrm{ar}\left[h'_{\boldsymbol{\xi}}(x_0, d)\right] = 0$, then $h'_{\boldsymbol{\xi}}(x_0, d) = f'(x_0, d)$ for almost every $\boldsymbol{\xi}$. If this holds for every $d \in T_S(x_0)$, then $\kappa = 0$ and in that case $\widehat{x}_N = x_0$ for any sample.

We give now a derivation of Theorem 1. By property (B) we have that the event "$\widehat{x}_N \neq x_0$" is included in the union of the events "$\widehat{f}'_N(x_0, d_j) \leq 0$", $j = 1, ..., \ell$. For a direction $d \in \{d_1, ..., d_\ell\}$ consider the random variable

$$\eta(d, \boldsymbol{\xi}) := h'_{\boldsymbol{\xi}}(x_0, d),$$

its mean $\mu_d := \mathbb{E}\big[\eta(d, \boldsymbol{\xi})\big]$, moment generating function $M_d(t) := \mathbb{E}_P\big[e^{t\eta(d,\boldsymbol{\xi})}\big]$ and its rate function

$$I_d(s) := \sup_{t\in\mathbb{R}} \big[ts - \Lambda_d(t)\big], \tag{2.4}$$

where $\Lambda_d(t) := \log M_d(t)$. Note that $\mu_d = f'(x_0, d)$ and hence, by (2.1), $\mu_d > 0$. Since $\boldsymbol{\xi}$ is finite, we have that the moment generating function $M_d(t)$ is finite valued for all $t \in \mathbb{R}$. Moreover, $\widehat{f}'_N(x_0, d) = N^{-1} \sum_{i=1}^N \eta(d, \boldsymbol{\xi}^i)$, and hence it follows that

$$\frac{1}{N} \log\Big[P\big(\widehat{f}'_N(x_0, d) \leq 0\big)\Big] \leq -I_d(0), \tag{2.5}$$

and

$$\liminf_{N\to\infty} \frac{1}{N} \log\Big[P\big(\widehat{f}'_N(x_0, d) < 0\big)\Big] \geq -I_d(0). \tag{2.6}$$

The inequalities (2.5) and (2.6) correspond to the respective upper and lower bounds of Cramér's Large Deviation theorem. Note that the upper bound (2.5) is exact and holds for any $N = 1, ...,$ while the lower bound (2.6) is asymptotic.

Since the probability of the union of the events "$\widehat{f}'_N(x_0, d_j) \leq 0$", $j = 1, ..., \ell$, is less than or equal to the sum of the probabilities of these events, we obtain by (2.5) that

$$P\big(\widehat{x}_N \neq x_0\big) \leq \sum_{j=1}^{\ell} e^{-N I_{d_j}(0)} \leq \ell\, e^{-N\beta}, \tag{2.7}$$

where

$$\beta := \min\big\{I_{d_1}(0), ..., I_{d_\ell}(0)\big\}. \tag{2.8}$$

Moreover, if $\widehat{f}'_N(x_0, d) < 0$ for some $d \in \{d_1, ..., d_\ell\}$, then x_0 is not an optimal solution of the SAA problem. Therefore probability of the event "$\widehat{x}_N \neq x_0$" is greater than

or equal to the probability of each individual event "$\widehat{f}'_N(x_0, d_j) < 0$", $j \in \{1, ..., \ell\}$. Together with (2.6) this implies that

$$\liminf_{N\to\infty} \frac{1}{N} \log\Big[P\big(\widehat{x}_N \neq x_0\big)\Big] \geq -\beta. \tag{2.9}$$

It follows from (2.7) and (2.9) that

$$\lim_{N\to\infty} \frac{1}{N} \log\Big[P\big(\widehat{x}_N \neq x_0\big)\Big] = -\beta. \tag{2.10}$$

Of course, we also have that $P\big(\widehat{x}_N = x_0\big) = 1 - P\big(\widehat{x}_N \neq x_0\big)$, and hence (1.3) follows.

Let us estimate the constant β. Consider a direction $d \in \{d_1, ..., d_\ell\}$. Since the moment generating function $M_d(t)$ is finite, it is infinitely differentiable. Hence the function $\Lambda_d(t)$ is also infinitely differentiable and

$$\Lambda'_d(0) = \mathbb{E}\big[\eta(d, \boldsymbol{\xi})\big] = \mu_d \quad \text{and} \quad \Lambda''_d(0) = \mathbb{V}\text{ar}\big[\eta(d, \boldsymbol{\xi})\big].$$

Suppose further that the variance of $\eta(d, \omega)$ is not zero, and hence $\Lambda''_d(0) > 0$. For $s = \mu_d$ the maximum in the right hand side of (2.4) is attained at $t = 0$. It follows that $I_d(\mu_d) = -\Lambda_d(0) = 0$ and $I'_d(\mu_d) = 0$. Moreover, by the Implicit Function Theorem we have that

$$I''_d(\mu_d) = \frac{\partial^2 \phi(0, \mu_d)}{\partial s^2} - \left[\frac{\partial^2 \phi(0, \mu_d)}{\partial t\, \partial s}\right]^2 \left[\frac{\partial^2 \phi(0, \mu_d)}{\partial t^2}\right]^{-1},$$

where $\phi(t, s) := ts - \Lambda_d(t)$, and hence

$$I''_d(\mu_d) = \frac{1}{\Lambda''_d(0)} = \frac{1}{\mathbb{V}\text{ar}[\eta(d, \boldsymbol{\xi})]}.$$

Therefore, for "small" μ_d the second-order Taylor expansion of $I_d(s)$, at $s = \mu_d$, gives us

$$I_d(0) \approx \frac{\mu_d^2}{2\Lambda''_d(0)} = \frac{\big[\mathbb{E}\,\eta(d, \boldsymbol{\xi})\big]^2}{2\mathbb{V}\text{ar}\big[\eta(d, \boldsymbol{\xi})\big]} = \frac{\big[f'(x_0, d)\big]^2}{2\mathbb{V}\text{ar}\big[\eta(d, \boldsymbol{\xi})\big]}. \tag{2.11}$$

That is, for such d that $f'(x_0, d)$ is close to zero, $I_d(0)$ is approximately (up to the remainder of order $o(\mu_d^2)$) equal to $\frac{1}{2}[f'(x_0, d)]^2/\mathbb{V}\text{ar}[\eta(d, \boldsymbol{\xi})]$.

Therefore for problems where the minimal of the numbers μ_{d_j} is "small", we have that

$$\beta \approx \min_{j\in\{1,...,\ell\}} \frac{\big[f'(x_0, d_j)\big]^2}{2\mathbb{V}\text{ar}\big[\eta(d_j, \boldsymbol{\xi})\big]}. \tag{2.12}$$

The above derivations show that for convex piecewise linear problems $\beta \approx 1/(2\kappa)$, where the constant κ is defined in (2.2).

3. Estimation of sample sizes

The results in the previous section provide estimates for the constant β in (1.3), which in turn yields some information on how fast the probability $P(\widehat{x}_N = x_0)$ approaches one with increase of the sample size N. Note, however, that the upper bound (2.5), given by the Large Deviations theory, can be quite crude for "not too large" values of N. Therefore, the above Large Deviations type results have more of a *qualitative* rather than a quantitative value. One might then investigate sharper estimates for the corresponding probabilities. If such estimates can be obtained, then it will be possible to compute the sample size N required to make the probability of the event "$\widehat{x}_N \neq x_0$" smaller than a specified tolerance α.

Let us start by discussing some general results. Consider a sequence $X_1, X_2, \ldots$ of i.i.d. realizations of a (real valued) random variable X with finite mean μ and finite variance σ^2. The reader may think of X_i as the random variable $\eta(d, \boldsymbol{\xi}^i) = h'_{\boldsymbol{\xi}^i}(x_0, d)$, where d is a given direction and $\boldsymbol{\xi}^1$, ..., is the generated random sample. Suppose that for a given $\delta \geq 0$ we want to estimate the probability

$$p_N(\delta) := P\left(N^{-1}\sum_{i=1}^{N} X_i < \mu - \delta\right). \tag{3.1}$$

We have by the Central Limit Theorem (CLT) that $N^{-1/2}\sum_{i=1}^{N}(X_i - \mu)$ converges in distribution to normal $N(0, \sigma^2)$, and hence the probability $p_N(N^{-1/2}\delta)$ tends to $\Phi(-\delta/\sigma)$ as $N \to \infty$ (here $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution).

Of course, if the random variables X_i have a normal distribution, then their average is also normally distributed, and in that case $p_N(N^{-1/2}\delta) = \Phi(-\delta/\sigma)$, or equivalently $p_N(\delta) = \Phi(-\delta\sqrt{N}/\sigma)$. Note, however, that the CLT does not give a justification for the asymptotics $\Phi(-\delta\sqrt{N}/\sigma)$ of $p_N(\delta)$, as $N \to \infty$, for a general distribution. We have that $\Phi(-\delta\sqrt{N}/\sigma)$ approaches zero, as $N \to \infty$, at the exponential rate $\exp\left(-\frac{1}{2}N\delta^2/\sigma^2\right)$ which can be different from the corresponding exponential rate provided by the Large Deviations theory. It is interesting to note, however, that for ill-conditioned problems (where δ^2/σ^2 is "small") the exponential rate of convergence of $p_N(\delta)$ is well approximated by the one suggested by the CLT (see formula (3.7) below). Let us also note that in cases where the sample size N is not "too large" and $\delta\sqrt{N}/\sigma$ is well inside the interval $(0, 2)$, the value $\Phi(-\delta\sqrt{N}/\sigma)$ is not small. In such cases $\Phi(-\delta\sqrt{N}/\sigma)$ tends to give a better approximation of $p_N(\delta)$ than the one suggested by the exact asymptotics discussed below.

We discuss now the so-called *exact asymptotics* for the probabilities $p_N(\delta)$. That theory provides an estimate $J_N(\delta)$ of $p_N(\delta)$ in the sense that

$$\lim_{N\to\infty} \frac{p_N(\delta)}{J_N(\delta)} = 1.$$

Let $\Lambda(\cdot)$ and $I(\cdot)$ denote the logarithmic moment generating and the rate functions of X, respectively. We assume that the moment generating function of X, and hence $\Lambda(t)$, is

finite valued for all t in a neighborhood $\mathcal{N}$ of zero, which implies that the mean and variance of X are finite. This assumption also implies that $\Lambda(\cdot)$ is C^∞ on $\mathcal{N}$. The following proposition shows that $\Lambda(\cdot)$ is *strictly convex* on $\mathcal{N}$.

Proposition 1. *Let X be a real valued random variable with positive variance such that the moment generating function of X is finite valued for all t in an open convex neighborhood $\mathcal{N}$ of zero. Then $\Lambda(\cdot)$ is strictly convex on $\mathcal{N}$.*

Proof. As mentioned earlier, it follows from the assumption that the moment generating function of X is finite valued for all $t \in \mathcal{N}$ that $\Lambda(\cdot)$ is C^∞ on $\mathcal{N}$. By differentiating $\Lambda(t) = \log \mathbb{E}[e^{tX}]$ we obtain, for $t \in \mathcal{N}$,

$$\Lambda''(t) = \frac{\mathbb{E}\left[X^2 e^{tX}\right]\mathbb{E}\left[e^{tX}\right] - \left(\mathbb{E}\left[Xe^{tX}\right]\right)^2}{\left(\mathbb{E}\left[e^{tX}\right]\right)^2}.$$

Furthermore, by the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[Xe^{tX}\right] \le \mathbb{E}\left[|X|e^{tX}\right] \le \left(\mathbb{E}\left[X^2 e^{tX}\right]\right)^{1/2}\left(\mathbb{E}\left[e^{tX}\right]\right)^{1/2}. \tag{3.2}$$

Notice that the second inequality is strict if and only if there does not exist a constant $c > 0$ such that $X^2 e^{tX} = ce^{tX}$ w.p.1 (see, e.g., Royden [18, p.121]). This of course means that the second inequality in (3.2) is strict if and only if X^2 is not a.e. constant. Moreover, since it is assumed that $\mathbb{V}\mathrm{ar}(X) > 0$, we have that X is not a.e. constant. All together this implies that at least one inequality in (3.2) is strict, and hence we have that

$$\left(\mathbb{E}\left[Xe^{tX}\right]\right)^2 < \mathbb{E}\left[X^2 e^{tX}\right]\mathbb{E}\left[e^{tX}\right].$$

We obtain that $\Lambda''(t) > 0$ for all $t \in \mathcal{N}$, and hence $\Lambda(\cdot)$ is strictly convex on $\mathcal{N}$. □

In particular, if the set $\boldsymbol{\xi}$ is finite, then the moment generating function is finite valued for all $t \in \mathbb{R}$. In that case Proposition 1 shows that $\Lambda(\cdot)$ is strictly convex on $\mathbb{R}$.

Let $a \in \mathbb{R}$ be such that $\Lambda'(a) = \mu - \delta$. From Proposition 1 we have that $\Lambda(\cdot)$ is a strictly convex function on a neighborhood $\mathcal{N}$ of zero, and hence $\Lambda'(\cdot)$ is monotonically increasing on $\mathcal{N}$, with $\Lambda'(0) = \mu$ and $\Lambda''(0) = \sigma^2 > 0$. Therefore, for δ near zero, the solution a of the above equation exists and is unique, $a \le 0$ if $\delta \ge 0$, and $a \to 0$ as $\delta \to 0$. Moreover, $\Lambda''(a) > 0$. The estimate $J_N(\delta)$ is then given by

$$J_N(\delta) = \frac{Ce^{-NI(\mu-\delta)}}{\sqrt{a^2\Lambda''(a)2\pi N}} \tag{3.3}$$

([4, Thm. 3.7.4]). The constant C is equal to one if X has a non-lattice distribution. Otherwise, C can be calculated as follows. Let b be the largest number such that $(X - \mu + \delta)/b$ is an integer with probability one, i.e., b is the period of the distribution of $X - \mu + \delta$. If such b does not exist, then again we take $C = 1$. If such b exists (which is true for example if X is rational w.p.1 and $\mu - \delta$ is rational as well), then the constant C is given by $C = (ab)/(1 - e^{-ab})$. Note that $(ab)/(1 - e^{-ab})$ tends to one as $a \to 0$. Therefore, in any case we can take $C \approx 1$.

Let us now write the Taylor expansions of Λ and Λ' around zero. We have

$$\Lambda(t) \;=\; \Lambda(0) + \Lambda'(0)t + \tfrac{1}{2}\Lambda''(0)t^2 + o(t^2) \;=\; \mu t + \tfrac{1}{2}\sigma^2 t^2 + o(t^2), \tag{3.4}$$

$$\Lambda'(t) \;=\; \Lambda'(0) + \Lambda''(0)t + o(t) \;=\; \mu + \sigma^2 t + o(t), \tag{3.5}$$

and since $\Lambda'(a) = \mu - \delta$, we can approximate a (when δ, and hence a, is close to zero) by

$$a \approx -\frac{\delta}{\sigma^2}. \tag{3.6}$$

Moreover, we have that if $\Lambda'(t) = y$, then $I(y) = ty - \Lambda(t)$, and thus

$$I(\mu - \delta) = a(\mu - \delta) - \Lambda(a) = -a\delta - \tfrac{1}{2}\sigma^2 a^2 + o(a^2) \approx \frac{\delta^2}{2\sigma^2}.$$

Using also the approximation $\Lambda''(a) \approx \Lambda''(0) = \sigma^2$, we obtain that the estimate $J_N(\delta)$ of $p_N(\delta)$ can be approximated by

$$J_N(\delta) \;\approx\; \frac{\sigma}{\delta\sqrt{2\pi N}}\, e^{-N\delta^2/(2\sigma^2)}. \tag{3.7}$$

It is interesting to compare the above estimate of $p_N(\delta)$ with the corresponding Large Deviations bounds. In the present one-dimensional case the upper Large Deviations bound is a consequence of Chebyshev's inequality. Therefore, we have

$$p_N(\delta) \;\le\; e^{-NI(\mu-\delta)} \;\approx\; e^{-N\delta^2/(2\sigma^2)}. \tag{3.8}$$

Comparing the right hand sides of (3.7) and (3.8), we see that, while the exponential term is identical, the factor multiplying the exponential term is one in (3.8) and inversely proportional to $\sqrt{N}$ in (3.7). Therefore, the estimate $J_N(\delta)$ tends to be sharper.

Let us remark that the above estimates were computed using Taylor expansions (3.4) and (3.5). One can use, of course, higher order expansions, which will then provide more accurate estimates. This is developed in [8], to which we refer for details.

We now apply the above results to the estimation of the probability in the left hand side of (2.5). Let $\{d_1, \ldots, d_\ell\}$ be a set of directions satisfying property (B). For each $j = 1, \ldots, \ell$, denote

$$\mu_j := \mathbb{E}\big[h'_{\boldsymbol{\xi}}(x_0, d_j)\big] = f'(x_0, d_j) \;\text{ and }\; \sigma_j^2 := \mathbb{V}\mathrm{ar}\big[h'_{\boldsymbol{\xi}}(x_0, d_j)\big],$$

and let I_j denote the rate function of $h'_{\boldsymbol{\xi}}(x_0, d_j)$. Then $\mu_j > 0$ and $I_j(0)$ gives the corresponding exponential constant. Using $\delta = \mu_j$ in (3.7) we obtain

$$P\left(\widehat{f}'_N(x_0, d_j) \le 0\right) \;\approx\; \frac{1}{\sqrt{4\pi\beta_j N}}\, e^{-\beta_j N},$$

where $\beta_j := \mu_j^2/(2\sigma_j^2) \approx I_j(0)$. We have that

$$P\big(\widehat{x}_N \neq x_0\big) \leq \sum_{j=1}^{\ell} P\left(\widehat{f}_N'(x_0, d_j) \leq 0\right) \approx \sum_{j=1}^{\ell} \frac{1}{\sqrt{4\pi\beta_j N}} e^{-\beta_j N} \;\leq\; \frac{\ell}{\sqrt{4\pi\beta_0 N}}\, e^{-\beta_0 N}, \tag{3.9}$$

where $\beta_0 := \min_{1\leq j\leq \ell} \beta_j$. Note that $\beta_0 \approx \beta$, with the constant β defined in (2.8), and that β_0 is equal to $1/(2\kappa)$, where κ is the condition number defined in (2.2).

Also, we have that, for every $j \in \{1, ..., \ell\}$,

$$P\big(\widehat{x}_N \neq x_0\big) \;\geq\; P\big(\widehat{f}_N'(x_0, d_j) < 0\big) \;\approx\; \frac{1}{\sqrt{4\pi\beta_j N}}\, e^{-\beta_j N},$$

which in particular implies that

$$P\big(\widehat{x}_N \neq x_0\big) \;\gtrapprox\; \frac{1}{\sqrt{4\pi\beta_0 N}}\, e^{-\beta_0 N}. \tag{3.10}$$

The right sides of the inequalities (3.9) and (3.10) differ from each other by the factor ℓ. This illustrates again that the condition number κ characterizes the overall rate of convergence of $P\big(\widehat{x}_N \neq x_0\big)$ to zero.

We can now use the above results to obtain estimates of the sample size N which is needed to make $P\big(\widehat{x}_N \neq x_0\big)$ smaller than a specified tolerance α. A "sufficient" condition for N can be obtained by requiring the right hand side of (3.9) to be less than α (the quotes are due to the fact that the inequality in (3.9) is approximate). We get

$$2\beta_0 N + \log(2\beta_0 N) \;\geq\; \log\left(\frac{\ell^2}{2\pi\alpha^2}\right).$$

In order for N to satisfy the above inequality, it suffices that

$$N \;\geq\; \frac{1}{2\beta_0} \max\left\{1, \log\left(\frac{\ell^2}{2\pi\alpha^2}\right)\right\} \;=\; C_1\kappa, \tag{3.11}$$

where $C_1 := \max\left\{1, \log(\ell^2/(2\pi\alpha^2))\right\}$.

A more accurate estimate can be obtaining by solving the nonlinear equation

$$z + \log z = \log(\ell^2/(2\pi\alpha^2)). \tag{3.12}$$

By taking $z_0 = C_1$ as the initial point, this equation can be easily solved, say by Newton's method. Let C_2 denote the solution of equation (3.12). We can then estimate N by

$$N \;\geq\; C_2\kappa. \tag{3.13}$$

Of course, the constant (condition number) κ is unknown a priori. Note, however, that the above estimates can also be written (for $i = 1, 2$) as

$$N \geq \frac{C_i \mathbb{V}\mathrm{ar}\left[h'_{\boldsymbol{\xi}}(x_0, d)\right]}{[f'(x_0, d)]^2} \quad \text{for all } d \in \{d_1, ..., d_\ell\}. \tag{3.14}$$

Therefore, N can be estimated using a single direction such that $f'(x_0, d)$ is small. We discuss that in the next section.

The sample size estimate (3.14), while accurate, is difficult to compute in practice. Indeed, the expression on the right-hand side of (3.14) implies that both the mean and the variance of directional derivative of $h(\cdot, \boldsymbol{\xi})$ at x_0 in each direction d are known, which is hardly the case. Usually, those quantities can only be estimated. Even if a certain direction is given, the estimates of mean and variance will be typically noisy and thus the estimate will not be reliable.

A simpler method to obtain estimates for the sample size is based on the following two-stage procedure. Let N_0 be an initial sample size, whose value is determined by the user. Suppose the SAA problem (1.2) is solved R times (each time with a new stream of random numbers), and let $\widehat{x}_0$ denote the most frequent solution obtained, say, R_0 times. By taking $\widehat{x}_0$ as a candidate for the optimal solution, we can estimate the probability $P\big(\widehat{x}_N \neq x_0\big)$ above as $\widehat{\alpha}_0 := 1 - R_0/R$. Let C_0 denote the solution of the nonlinear equation

$$z + \log z = \log\left(\ell^2/(2\pi\widehat{\alpha}_0^2)\right). \tag{3.15}$$

Then, in parallel with inequality (3.13), we can estimate κ by

$$\widehat{\kappa} := \frac{N_0}{C_0}. \tag{3.16}$$

This estimate, in turn, can be substituted in inequality (3.13) to yield a new estimator for the sample size N that guarantees that the optimal solution will be obtained with a given probability at least $1 - \alpha$. Such procedure is, of course, heuristic; nevertheless, it requires little information about the system – in fact, only estimates for ℓ and α are needed. Notice that $\ell = 1$ gives the highest estimate for $\widehat{\kappa}$. In the next section we will see an example of application of this procedure.

4. Examples

We present now some examples to illustrate the ideas developed in the previous sections. Consider initially the following "median" problem. Let $\boldsymbol{\xi}$ be a (one dimensional) random variable, $S := \mathbb{R}$ and $h(x, \xi) := |x - \xi|$. Suppose that $\boldsymbol{\xi}$ has a discrete distribution with the odd number $r = 2k + 1$ of values equally spaced on the interval $[-1, 1]$, each having equal probability $1/r$. We have then that $x_0 = 0$ is the unique optimal solution of the true problem and for direction $d = 1$,

$$\mathbb{E}\big[h'_{\boldsymbol{\xi}}(x_0, d)\big] = r^{-1} \text{ and } \mathbb{V}\mathrm{ar}\big[h'_{\boldsymbol{\xi}}(x_0, d)\big] = 1 - r^{-2}.$$

Consequently the condition number is

$$\kappa = r^2 - 1 = 4k(k+1). \tag{4.1}$$

In that example the exact value of the exponential constant β is

$$\beta = \tfrac{1}{2}\log[r^2/(r^2-1)] \approx 1/(2r^2-1), \tag{4.2}$$

while the approximation $\beta \approx 1/(2\kappa)$ gives $1/(2\kappa) = 1/(2r^2-2)$ (see Shapiro and Homem-de-Mello [20] for a derivation).

Now let $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, ..., \boldsymbol{\xi}_m)$ be a random vector with independent components $\boldsymbol{\xi}_i$ each having the above discrete distribution, and let $h(x, \xi) := \sum_{i=1}^m |x_i - \xi_i|$. Then $x_0 = (0, ..., 0)$ is the optimal solution of the true problem with the same exponential constant and the condition number as in (4.1) and (4.2), respectively. This shows that for r not "too large", this separable problem is well conditioned, and hence a small sample suffices in order to solve it exactly with high probability (see Table 1).

We use this example to verify the accuracy of the estimates of the sample sizes given in (3.11) and (3.13). Let us fix $\alpha = 0.05$, i.e., we wish to obtain the true optimal solution with probability 0.95. Notice that both constants C_1 and C_2 in (3.11) and (3.13) depend on the number ℓ of directions; in this separable case, we have $\ell = 2m$. Table 1 below displays the values of N obtained with (3.11) (called N_1) and (3.13) (called N_2), as well as the corresponding probabilities that $\widehat{x}_N = x_0$, which for large N are very close to $1 - 2P(X \geq N/2)$, where X is a binomial random variable $B(N, q)$ with $q = (r-1)/(2r)$ (see [20]). Those probabilities are computed for various values of m and r, as the table shows. The last column displays the ratio N_1/N_2. Notice that the number of scenarios is given by r^m. Moreover, as remarked in [20], we can see that the sample size grows quadratically with r and logarithmically with m. Observe also that the probabilities corresponding to the more precise estimate N_2 are smaller than 0.95 for small r; this happens because, as remarked in section 3, the sample size estimates are more accurate when the underlying problem is ill-conditioned.

In the above example, the condition number κ was known. In general, however, κ can be difficult to compute, even for simple problems, and moreover it depends on the optimal solution x_0 which, of course, is not known a priori. In the next two examples below we use the following procedure to estimate κ at a given optimal solution x_0: first, we generate the corresponding Monte Carlo approximation problem with sample size N_0 to obtain an approximate solution $\widehat{x}_{N_0,1}$. We then independently replicate the experiment $T-1$ more times, hence obtaining T approximate solutions $\widehat{x}_{N_0,1}, \ldots, \widehat{x}_{N_0,T}$. Note that we are not interested here in the approximate objective values of the problem, but rather in the frequencies of the approximate solutions. Observe also that if the problem is ill-conditioned, then the most frequent approximate solution may not coincide with the true minimizer x_0. We exclude those $\widehat{x}_{N_0,i}$, $i = 1, \ldots, T$, which coincide with x_0, and find the most frequent approximate solution from the remaining $\widehat{x}_{N_0,i}$'s. Let the chosen solution be denoted by x_1. With x_0 and x_1, we can calculate the normalized direction $d := (x_1 - x_0)/\|x_1 - x_0\|$. Another possibility is to pick x_1 as the point $\widehat{x}_{N_0,i}$ whose objective function value is the closest to $f(x_0)$. Next, we fix ε to be a small number, say 0.01, and compute the objective values at x_0 and $x_0 + \varepsilon d$ exactly, i.e., by enumerating all possible scenarios. Of course, these small examples allow such computations; for larger

Table 1. Estimated sample sizes to attain probability 0.95 and exact probabilities $P(\widehat{x}_N = x_0)$ for the median problem

r	m	N_1	p_{N_1}	N_2	p_{N_2}	N_1/N_2
5	5	211	0.984	165	0.954	1.279
5	10	244	0.980	194	0.941	1.258
5	100	355	0.986	294	0.937	1.207
5	500	432	0.984	366	0.934	1.180
5	1000	465	0.987	398	0.934	1.168
11	5	1052	0.983	821	0.956	1.281
11	10	1218	0.984	968	0.950	1.258
11	100	1771	0.987	1470	0.949	1.205
11	500	2157	0.988	1830	0.948	1.179
11	1000	2323	0.989	1986	0.947	1.170
21	5	3854	0.984	3009	0.956	1.281
21	10	4464	0.985	3546	0.953	1.259
21	100	6491	0.988	5388	0.952	1.205
21	500	7907	0.989	6708	0.951	1.179
21	1000	8517	0.989	7282	0.951	1.170
31	5	8409	0.985	6564	0.955	1.281
31	10	9740	0.985	7737	0.956	1.259
31	100	14161	0.988	11756	0.953	1.205
31	500	17251	0.989	14636	0.952	1.179
31	1000	18582	0.989	15888	0.952	1.170
51	5	22773	0.985	17775	0.956	1.281
51	10	26378	0.985	20952	0.955	1.259
51	100	38351	0.988	31838	0.953	1.205
51	500	46720	0.989	39637	0.954	1.179
51	1000	50325	0.989	43028	0.953	1.170

problems, one can estimate those values by large samples. The directional derivative is then estimated by $[f(x_0 + \varepsilon d) - f(x_0)]/\varepsilon$.

We consider now the following two numerical examples. The first example is CEP1, which was used in [20] to illustrate the exponential rate of convergence to the optimal solution. The problem was originally described in [7]. The second problem is APL1P, which was described in [9] and also studied in [1].

The CEP1 problem has 8 decision variables with 5 constraints (plus lower bound constraints) on the first stage, and 15 decision variables with 7 constraints (plus lower bound constraints) on the second stage. The random variables appear only on the right hand side of the second stage. There are 3 independent and identically distributed random variables, each taking 6 possible values with equal probability, so the sample space has size $6^3 = 216$.

For the sake of verification, we solved the problem exactly by the Benders decomposition algorithm, and obtained the true minimizer x_0 of the problem, which in this case is unique. We then solved the corresponding Monte Carlo approximating problems with sample size $N_0 = 10$ for $T = 100$ replications. Using the procedure outlined above, we calculated $f'(x_0, d)$ and $\mathbb{V}\mathrm{ar}[h'_{\xi}(x_0, d)]$ for the direction $d := x_1 - x_0$, where x_1 is the second most often obtained solution. Table 2 below displays the results. The table also displays the value of N estimated with (3.13) that guarantees that the optimal solution will be obtained with probability at least 0.95. Note that this requires an estimate for ℓ. In this case we chose $\ell = 1$ due to the small number of decision variables. Note also that the estimate obtained for N ($N \geq 57$) is in agreement with the results obtained in [20] – in

that paper it was verified computationally that a sample size equal to 50 yields the optimal solution with probability 0.97. Larger values of ℓ, of course, yield larger sample sizes; for example, with $\ell = 6$ (which can be justified by the fact that there are 3 degrees of freedom for the first stage variables), we get the more conservative estimate $N \geq 111$.

The APL1P example is an electric power capacity expansion problem on a transportation network. The problem has two decision variables with 2 constraints (plus lower bound constraints) on the first stage, and 9 decision variables with 5 constraints (plus lower bound constraints) on the second stage. The random variables appear on both the right hand side and the technology matrix of the second stage. There are 5 independent random variables. The number of realizations per random variables yields a total of $4 \times 5 \times 4 \times 4 \times 4 = 1280$ scenarios. To estimate κ, we used the same procedure outlined above, with sample size $N_0 = 200$, $T = 100$ replications, and $d = x_1 - x_0$, where x_0 is an optimal solution and x_1 is an obtained solution whose objective function value is the closest to $f(x_0)$. As with the CEP1 problem, table 2 below displays the directional derivative in the direction d and its variance at the optimal solution, and the value of N estimated with (3.13) (and $\ell = 1$) that guarantees that the optimal solution will be obtained with probability at least 0.95. Note that the directional derivative is extremely small; this suggests that the problem is ill conditioned or even has multiple solutions, at least up to a certain precision. Hence, the estimate obtained for N is much larger than the total number of scenarios. That happened since the problem is small and ill conditioned. Of course, it makes sense to use Monte Carlo sampling techniques only for problems with a very large number of scenarios, so this example is given for illustration purposes only.

For the sake of verifying the alternative strategy for sample size estimation described at the end of section 3, we also list in Table 2 the estimate of κ obtained via (3.16), using the same respective number of samples and replications as above for problems CEP1 and APL1P. Those estimates for κ and N (which again correspond to 0.95 probability of optimality) are indicated by κ_{alt} and N_{alt}. As the table indicates, the estimate for problem CEP1 is reasonably close to the one obtained with (3.14) – and it is far easier to compute. For example, the sample size equal to 26 indicated by N_{alt} was verified in [20] to yield the optimal solution with probability between 0.905 and 0.958. The estimate for problem APL1P, in turn, is quite different from the one obtained with (3.14). Notice however that such estimate was based on an initial frequency of only 8% of occurrences of the optimal solution (more precisely, 8 occurrences out of 100 replications). In such cases, we would recommend using a larger sample size in the first stage of the two-stage estimation procedure, in order to obtain a more reliable estimate of the probability of obtaining the correct answer.

Table 2. Condition number and sample size estimates for the CEP1 and APL1P problems

	CEP1	APL1P
$f'(x_0, d)$	7.59	0.0005
$\mathbb{V}\text{ar}[h'_{\boldsymbol{\xi}}(x_0, d)]$	1068.3	19.5
κ	18.49	7.73×10^7
N	57	2.36×10^8
κ_{alt}	8.33	1248.4
N_{alt}	26	3797

Our final example is a stochastic vehicle allocation model in a single commodity network, described in Donohue and Birge [5] and also studied in Mak, Morton and Wood [13]. This minimization problem, called DB1 in the latter paper, has 5 decision variables and only one constraint (plus bound constraints) in the first stage, and 102 variables and 71 constraints (plus bound constraints) in the second stage. There are 46 independent and identically distributed random variables, which results in 4.5×10^{25} scenarios.

The size of the problem, of course, precludes its exact solution. To have a better understanding of the behavior of the model, we solved $M = 50$ independent replications of the SAA problem (1.2), each with the sample size $N = 250$ (notice that $N = 250$ is a tiny fraction of the total number of scenarios of the considered problem). We obtained the following solutions of the corresponding SAA problems:

$$\begin{aligned}
x_1 &= (11\ \ 13\ \ 8\ \ 12\ \ 7) \quad (24 \text{ times}) \\
x_2 &= (11\ \ 14\ \ 8\ \ 11\ \ 7) \quad (13 \text{ times}) \\
x_3 &= (11\ \ 13\ \ 9\ \ 11\ \ 7) \quad (8 \text{ times}) \\
x_4 &= (12\ \ 13\ \ 8\ \ 11\ \ 7) \quad (5 \text{ times})
\end{aligned}$$

The SAA problems were solved using an adapted version of the Stochastic Solutions model from the IBM Optimization Library.

We also solved the SAA problems with the same sample size and the same number of replications, but instead of using a standard Monte Carlo sampling we applied the *Latin Hypercube Sampling* (LHS) method. This sampling technique consists of dividing the (0, 1) interval into N subintervals of equal size (where N is the size of the sample) and picking one number randomly from each interval. The N obtained numbers are then randomly shuffled, and the resulting sequence is used to generate random variates from a given distribution, e.g., by means of the inverse transform method. The procedure is repeated for each of the components of the underlying random vector, yielding a stratified sample of size N of that vector. This technique, initially proposed by McKay, Conover and Beckman [14], has been thoroughly studied in terms of its theoretical properties and numerical efficacy; see, for instance, Stein [22] and Owen [16]. In the context of stochastic programming, Bailey, Jensen and Morton [1] have used the LHS strategy embedded in a response surface method.

The solution of 50 independent replications were:

$$\begin{aligned}
x_1 &= (11\ \ 13\ \ 8\ \ 12\ \ 7) \quad (24 \text{ times}) \\
x_2 &= (11\ \ 14\ \ 8\ \ 11\ \ 7) \quad (14 \text{ times}) \\
x_3 &= (11\ \ 13\ \ 9\ \ 11\ \ 7) \quad (7 \text{ times}) \\
x_4 &= (12\ \ 13\ \ 8\ \ 11\ \ 7) \quad (3 \text{ times}) \\
x_5 &= (10\ \ 14\ \ 8\ \ 12\ \ 7) \quad (2 \text{ times})
\end{aligned}$$

While the frequencies of each solution obtained with the LHS method were very similar to the standard Monte Carlo sampling, the variances of the point estimates were reduced about 200 times; therefore, we will present only the results obtained with LHS.

Table 3 below displays the results. The first row contains the average optimal value of the approximating problem (1.2) over the 50 replications, together with the half-width of a 95% confidence interval. Notice that such value constitutes an estimate for a *lower bound* of the original problem (see [13, 15] for discussions). The next four rows contain respectively the point estimates of the function values at the points x_1, x_2, x_3 and x_4 listed above, together with the half-width of 95% confidence intervals. These point estimates were calculated with the same random numbers used to solve the approximating problem. Thus, the estimate of $f(x_i)$ correspond to 50 batches of LHS samples of size 250 each, so the total sample size was 12,500. Notice that these point estimates all constitute estimates of upper bound to the optimal value of the original problem. Finally, the last row contains an estimate of the optimality gap obtained from the difference between the upper bound with x_1 minus the lower bound, together with the half-width of a 95% confidence interval. Notice that the gap is very small. Also, the obtained numbers agree with the results presented in Mak et al. [13], although more precise due to the use of a larger sample as well as the LHS strategy.

Notice that the ranking of x_1, x_2, x_3 and x_4 according to function values is the same as the ranking according to the frequency of each solution. However, the proximity of the those values among each other, and also among those values and the lower bound, suggest that those four points could, in principle, be accepted as optimal solutions. This can be made precise by testing whether the directional derivatives $f'(x_i; x_j - x_i)$ are zero or less.

Table 4 below displays the results of these tests. The second column displays the mean of each directional derivative, together with the half-width of a 95% confidence interval. The third and fourth columns correspond to respectively the hypothesis tests and the p-values of each test. All tests were performed after checking for normality of the data, to ensure validity of the tests. It is worthwhile mentioning that, unlike the function values estimates, the directional derivative estimates hardly benefit from the use of the LHS strategy – the variance reduction compared to standard Monte Carlo was minimal. This explains why the frequencies of solutions were about the same in both cases.

From table 4, we can see that both x_3 and x_4 are definitely worse points than x_1 and x_2 and thus can be discarded as candidates for the optimal solution. The results on the first row show that, with the sample size used, the hypothesis that x_2 is a better solution than x_1 can be rejected with the p-value of about 7%, and the 95% confidence interval for $f'(x_1, x_2 - x_1)$ contains zero. This suggests that we cannot use formula (3.14) to estimate the value of N that guarantees that the optimal solution will be obtained with probability 95%.

Table 3. Optimal value for approximating problem and point estimates

Optimal value of approx. problem	-17719.61 ± 3.17
Evaluation of point x_1	-17718.35 ± 3.12
Evaluation of point x_2	-17717.61 ± 3.28
Evaluation of point x_3	-17716.03 ± 3.37
Evaluation of point x_4	-17715.06 ± 3.43
Optimality gap	1.26 ± 0.49

Table 4. Estimates of directional derivatives and hypothesis tests

Derivative	mean $\pm\Delta$	Hypothesis	p-value
$f'(x_1, x_2 - x_1)$	0.74 ± 0.97	$H_0 : f'(x_1; x_2 - x_1) \leq 0$	0.0696
$f'(x_1, x_3 - x_1)$	2.31 ± 1.13	$H_0 : f'(x_1; x_3 - x_1) \leq 0$	0.0001
$f'(x_1; x_4 - x_1)$	3.29 ± 1.13	$H_0 : f'(x_1; x_4 - x_1) \leq 0$	0.0000
$f'(x_2, x_3 - x_2)$	1.58 ± 0.99	$H_0 : f'(x_2; x_3 - x_2) \leq 0$	0.0013
$f'(x_2, x_4 - x_2)$	2.55 ± 0.81	$H_0 : f'(x_2; x_4 - x_2) \leq 0$	0.0000

In order to overcome this difficulty, we shall use the two-stage estimation method presented at the end of section 3. Following the notation introduced in that discussion, let $N_0 = 250$ be the initial sample size and $R = 50$ the initial number of replications. Let $\widehat{x}_0 = x_1$ denote the most frequent solution obtained ($R_0 = 24$ times), which yields the estimate $\widehat{\alpha}_0 = 1 - R_0/R = 0.52$ for the probability of not obtaining the optimal solution. As an estimate for ℓ, we take $\ell_0 = 4$. The rationale for this choice is that, among x_1, x_2, x_3 and x_4, the first four coordinates vary along one direction and the fifth coordinate does not change. Next, let C_0 denote the solution of the nonlinear equation (3.15), which is $C_0 = 1.71$. From (3.16) we obtain the estimate $\widehat{\kappa} = 146.41$ and thus, using (3.13) we have that an estimate for the value of N that guarantees that the optimal solution will be obtained with a probability $1 - \alpha$ is

$$N \geq C_2 \widehat{\kappa},$$

where C_2 solves equation (3.12). For example, with $\alpha = 0.2$, $\alpha = 0.1$ and $\alpha = 0.05$ we obtain, respectively, $N \geq 446$, $N \geq 604$ and $N \geq 771$. Using $\ell_0 = 1$ instead of $\ell_0 = 4$ (i.e., a more conservative estimate) we get $N \geq 757$, $N \geq 1294$ and $N \geq 1920$ corresponding to the respective values of α. Notice that even the most conservative estimate ($N = 1920$) is far smaller than the number of scenarios. This is a strong indication that the underlying problem is, in fact, well-conditioned.

5. Conclusions

We have introduced in this paper the concept of *conditioning* of convex piecewise linear stochastic programs. In a well-conditioned problem, the solution of the Monte Carlo SAA problem coincides with the solution of the original problem with high probability, even for relatively small sample sizes. We also showed that conditioning of convex piecewise linear stochastic programs depends essentially on two factors: (i) how flat is the objective function around the optimal solution, and (ii) how much variability is inherent in the problem. In theory these factors can be quantified to determine the *condition number* of the problem.

On the numerical side, the introduced condition number can be used to estimate the sample size required for the solution of the Monte Carlo SAA problem to be equal to the solution of the original problem with a given probability. Since this estimate is typically difficult to compute in practice, we have provided another heuristic estimate, which is much simpler to calculate. The ideas introduced in the paper were illustrated through four examples with different characteristics. These numerical results indicate

that, indeed, well-conditioned problems can be solved quite accurately with relatively little effort.

Finally, another interesting aspect of the numerical examples presented in the paper was the use of the Latin Hypercube sampling technique to reduce the variance. It is well known that such methods can reduce the variance of point estimates very efficiently. We have observed this as well in our computations. However, it is still unclear what is the effect of such techniques in terms of the rate of convergence of solutions of the Monte Carlo SAA problems to solutions of the original problem. For example, for the DB1 problem described in section 4, the use of Latin Hypercube sampling did not improve the rate of convergence, but greatly reduced the variance of point estimates. For the APL1P problem, in turn, Latin Hypercube greatly sped up the rate of convergence (we did not present those numbers here since the issue is only marginally related to conditioning).

Acknowledgements. We thank two referees and the associate editor for their comments, which helped to improve the presentation of our results. We also thank David Morton from University of Texas for providing us the SMPS files for the DB1 problem.

References

1. T.G. Bailey, P. Jensen and D.P. Morton, "Response surface analysis of two-stage stochastic linear programming with recourse", *Naval Research Logistics* 46 (1999), 753–778.
2. J.R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer, New York, 1997.
3. L. Dai, C. H. Chen and J. R. Birge, "Convergence properties of two-stage stochastic programming", *J. Optim. Theory Appl.*, 106 (2000), 489–509.
4. A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd. ed., Springer-Verlag, New York, 1998.
5. C.J. Donohue and J.R. Birge, "An upper bound on the network recourse function", manuscript, University of Michigan, 1995.
6. R.M. Freund and J.R. Vera, "Some characterizations and properties of the "distance to ill-posedness" and the condition measure of a conic linear system", *Mathematical Programming*, 86 (1999), 225–260.
7. J.L. Higle and S. Sen, "Finite master programs in regularized stochastic decomposition", *Mathematical Programming*, 67 (1994), 143–168.
8. T. Homem-de-Mello, "Monte Carlo methods for discrete stochastic optimization", in *Stochastic Optimization: Algorithms and Applications* (S. Uryasev and P.M. Pardalos, eds.), 95–117, Kluwer Academic Publishers, 2000.
9. G. Infanger, "Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs," *Annals of Operations Research*, 39 (1992), 69–95.
10. Y.M. Kaniovski, A.J. King and R.J.-B. Wets, "Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems," *Annals of Operations Research*, 56 (1995), 189–208.
11. A. Kleywegt, A. Shapiro, and T. Homem-de-Mello, "The sample average approximation method for stochastic discrete optimization", *SIAM Journal on Optimization*, 12 (2001), 479–502.
12. J. Linderoth, A. Shapiro, and S. Wright, "The empirical behavior of sampling methods for stochastic programming", *Optimization Technical Report 02-01*, Computer Sciences Department, University of Wisconsin-Madison, 2002.
13. 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 (1999), 47–56.
14. M.D. McKay, R.J. Beckman and W.J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code", *Technometrics* 21 (1979), 239–245.
15. V.I. Norkin, G.Ch. Pflug and A. Ruszczyński, "A branch and bound method for stochastic global optimization", *Mathematical Programming*, 83 (1998), 425–450.
16. A.B. Owen, "Monte Carlo variance of scrambled equidistribution quadrature", *SIAM Journal of Numerical Analysis* 34 (1997), 1884–1910.
17. J. Renegar, "Condition numbers, the barrier method, and the conjugate-gradient method", *SIAM Journal on Optimization*, 6 (1996), 879–912.

18. H. Royden, *Real Analysis*, 3rd. ed., Macmillan, New York, 1988.
19. R.Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley, New York, NY, 1993.
20. A. Shapiro and T. Homem-de-Mello, "On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs", *SIAM Journal on Optimization*, 11 (2001), 70–86.
21. A. Shapiro, "Stochastic programming by Monte Carlo simulation methods", *The Stochastic Programming E-Print Series (SPEPS)*, available at the web site:
http://dochost.rz.hu-berlin.de/speps/contents00.html
22. M. Stein, "Large sample properties of simulations using Latin Hypercube Sampling", *Technometrics* 29 (1987), 143–151.