SIAM J. OPTIM.
Vol. 12, No. 2, pp. 479–502

THE SAMPLE AVERAGE APPROXIMATION METHOD FOR STOCHASTIC DISCRETE OPTIMIZATION*

ANTON J. KLEYWEGT†, ALEXANDER SHAPIRO†, AND TITO HOMEM-DE-MELLO‡

Abstract. In this paper we study a Monte Carlo simulation–based approach to stochastic discrete optimization problems. The basic idea of such methods is that a random sample is generated and the expected value function is approximated by the corresponding sample average function. The obtained sample average optimization problem is solved, and the procedure is repeated several times until a stopping criterion is satisfied. We discuss convergence rates, stopping rules, and computational complexity of this procedure and present a numerical example for the stochastic knapsack problem.

Key words. stochastic programming, discrete optimization, Monte Carlo sampling, law of large numbers, large deviations theory, sample average approximation, stopping rules, stochastic knapsack problem

AMS subject classifications. 90C10, 90C15

PII. S1052623499363220

1. Introduction.

In this paper we consider optimization problems of the form

$$\min_{x\in\mathcal{S}} \{g(x) := \mathbb{E}_P G(x,W)\}\,. \tag{1.1}$$

Here W is a random vector having probability distribution P, $\mathcal{S}$ is a *finite* set (e.g., $\mathcal{S}$ can be a finite subset of $\mathbb{R}^n$ with integer coordinates), $G(x,w)$ is a real valued function of two (vector) variables x and w, and $\mathbb{E}_P G(x,W) = \int G(x,w)P(dw)$ is the corresponding expected value. We assume that the expected value function $g(x)$ is well defined, i.e., for every $x \in \mathcal{S}$ the function $G(x,\cdot)$ is measurable and $\mathbb{E}_P\{|G(x,W)|\} < \infty$.

We are particularly interested in problems with the following characteristics:

1. The expected value function $g(x) := \mathbb{E}_P G(x,W)$ cannot be written in a closed form, and/or its values cannot be easily calculated.
2. The function $G(x,w)$ is easily computable for given x and w.
3. The set $\mathcal{S}$ of feasible solutions, although finite, is very large, so that enumeration approaches are not feasible. For instance, in the example presented in section 4, $\mathcal{S} = \{0,1\}^k$ and hence $|\mathcal{S}| = 2^k$; i.e., the size of the feasible set grows exponentially with the number of variables.

It is well known that many discrete optimization problems are hard to solve. Another difficulty here is that the objective function $g(x)$ can be complicated and/or difficult to compute even approximately. Therefore stochastic discrete optimization problems are difficult indeed and little progress in solving such problems numerically has been reported so far. There is an extensive literature addressing stochastic discrete optimization problems in which the number of feasible solutions is sufficiently small to

*Received by the editors November 1, 1999; accepted for publication (in revised form) May 14, 2001; published electronically December 14, 2001.
http://www.siam.org/journals/siopt/12-2/36322.html
†School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205 (Anton.Kleywegt@isye.gatech.edu, Alexander.Shapiro@isye.gatech.edu). The first author's work was supported by the National Science Foundation under grant DMI-9875400. The second author's work was supported by the National Science Foundation under grant DMS-0073770.
‡Department of Industrial, Welding and Systems Engineering, The Ohio State University, Columbus, OH 43210-1271 (homem-de-mello.1@osu.edu).

allow estimation of $g(x)$ for each solution x. Examples of this literature are Hochberg and Tamhane [12]; Bechhofer, Santner, and Goldsman [2]; Futschik and Pflug [7, 8]; and Nelson et al. [17]. Another approach that has been studied consists of modifying the well-known simulated annealing method in order to account for the fact that the objective function values are not known exactly. Work on this topic includes Gelfand and Mitter [9], Alrefaei and Andradóttir [1], Fox and Heine [6], Gutjahr and Pflug [10], and Homem-de-Mello [13]. A discussion of two-stage stochastic integer programming problems with recourse can be found in Birge and Louveaux [3]. A branch and bound approach to solving stochastic integer programming problems was suggested by Norkin, Ermoliev, and Ruszczyński [18] and Norkin, Pflug, and Ruszczyński [19]. Schultz, Stougie, and Van der Vlerk [20] suggested an algebraic approach to solving stochastic programs with integer recourse by using a framework of Gröbner basis reductions.

In this paper we study a Monte Carlo simulation–based approach to stochastic discrete optimization problems. The basic idea is simple indeed—a random sample of W is generated and the expected value function is approximated by the corresponding sample average function. The obtained sample average optimization problem is solved, and the procedure is repeated several times until a stopping criterion is satisfied. The idea of using sample average approximations for solving stochastic programs is a natural one and was used by various authors over the years. Such an approach was used in the context of a stochastic knapsack problem in a recent paper of Morton and Wood [16].

The organization of this paper is as follows. In the next section we discuss a statistical inference of the sample average approximation method. In particular, we show that with probability approaching 1 exponentially fast with increase of the sample size, an optimal solution of the sample average approximation problem provides an exact optimal solution of the "true" problem (1.1). In section 3 we outline an algorithm design for the sample average approximation approach to solving (1.1), and in particular we discuss various stopping rules. In section 4 we present a numerical example of the sample average approximation method applied to a stochastic knapsack problem, and section 5 gives conclusions.

2. Convergence results. As mentioned in the introduction, we are interested in solving stochastic discrete optimization problems of the form (1.1). Let $W^1, \ldots, W^N$ be an independently and identically distributed (i.i.d.) random sample of N realizations of the random vector W. Consider the sample average function

$$\hat{g}_N(x) := \frac{1}{N} \sum_{j=1}^{N} G(x, W^j)$$

and the associated problem

$$\min_{x \in \mathcal{S}} \ \hat{g}_N(x). \tag{2.1}$$

We refer to (1.1) and (2.1) as the "true" (or expected value) and sample average approximation (SAA) problems, respectively. Note that $\mathbb{E}[\hat{g}_N(x)] = g(x)$.

Since the feasible set $\mathcal{S}$ is finite, problems (1.1) and (2.1) have nonempty sets of optimal solutions, denoted $\mathcal{S}^*$ and $\hat{\mathcal{S}}_N$, respectively. Let v^* and $\hat{v}_N$ denote the optimal values,

$$v^* := \min_{x \in \mathcal{S}} g(x) \quad \text{and} \quad \hat{v}_N := \min_{x \in \mathcal{S}} \hat{g}_N(x),$$

of the respective problems. We also consider sets of ε-optimal solutions. That is, for $\varepsilon \geq 0$, we say that $\bar{x}$ is an ε-optimal solution of (1.1) if $\bar{x} \in \mathcal{S}$ and $g(\bar{x}) \leq v^* + \varepsilon$. The sets of all ε-optimal solutions of (1.1) and (2.1) are denoted by $\mathcal{S}^\varepsilon$ and $\hat{\mathcal{S}}_N^\varepsilon$, respectively. Clearly for $\varepsilon = 0$ set $\mathcal{S}^\varepsilon$ coincides with $\mathcal{S}^*$, and $\hat{\mathcal{S}}_N^\varepsilon$ coincides with $\hat{\mathcal{S}}_N$.

2.1. Convergence of objective values and solutions. The following proposition establishes convergence with probability one (w.p.1) of the above statistical estimators. By the statement "an event happens w.p.1 for N large enough" we mean that for P—almost every realization $\omega = \{W^1, W^2, \ldots\}$ of the random sequence—there exists an integer $N(\omega)$ such that the considered event happens for all samples $\{W^1, \ldots, W^n\}$ from ω with $n \geq N(\omega)$. Note that in such a statement the integer $N(\omega)$ depends on the sequence ω of realizations and therefore is random.

PROPOSITION 2.1. *The following two properties hold:* (i) $\hat{v}_N \to v^*$ *w.p.1 as* $N \to \infty$, *and* (ii) *for any* $\varepsilon \geq 0$ *the event* $\{\hat{\mathcal{S}}_N^\varepsilon \subset \mathcal{S}^\varepsilon\}$ *happens w.p.1 for* N *large enough.*

Proof. It follows from the (strong) law of large numbers that for any $x \in \mathcal{S}$, $\hat{g}_N(x)$ converges to $g(x)$ w.p.1 as $N \to \infty$. Since the set $\mathcal{S}$ is finite and the union of a finite number of sets each of measure zero also has measure zero, it follows that, w.p.1, $\hat{g}_N(x)$ converges to $g(x)$ uniformly in $x \in \mathcal{S}$. That is,

$$\delta_N := \max_{x \in \mathcal{S}} |\hat{g}_N(x) - g(x)| \to 0, \quad \text{w.p.1 as} \quad N \to \infty. \tag{2.2}$$

Since $|\hat{v}_N - v^*| \leq \delta_N$, it follows that, w.p.1, $\hat{v}_N \to v^*$ as $N \to \infty$.

For a given $\varepsilon \geq 0$ consider the number

$$\rho(\varepsilon) := \min_{x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon} g(x) - v^* - \varepsilon. \tag{2.3}$$

Since for any $x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon$ it holds that $g(x) > v^* + \varepsilon$ and the set $\mathcal{S}$ is finite, it follows that $\rho(\varepsilon) > 0$.

Let N be large enough such that $\delta_N < \rho(\varepsilon)/2$. Then $\hat{v}_N < v^* + \rho(\varepsilon)/2$, and for any $x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon$ it holds that $\hat{g}_N(x) > v^* + \varepsilon + \rho(\varepsilon)/2$. It follows that if $x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon$, then $\hat{g}_N(x) > \hat{v}_N + \varepsilon$ and hence x does not belong to the set $\hat{\mathcal{S}}_N^\varepsilon$. The inclusion $\hat{\mathcal{S}}_N^\varepsilon \subset \mathcal{S}^\varepsilon$ follows, which completes the proof. □

Note that if δ is a number such that $0 \leq \delta \leq \varepsilon$, then $\mathcal{S}^\delta \subset \mathcal{S}^\varepsilon$ and $\hat{\mathcal{S}}_N^\delta \subset \hat{\mathcal{S}}_N^\varepsilon$. Consequently it follows by the above proposition that for any $\delta \in [0, \varepsilon]$ the event $\{\hat{\mathcal{S}}_N^\delta \subset \mathcal{S}^\varepsilon\}$ happens w.p.1 for N large enough. It also follows that if $\mathcal{S}^\varepsilon = \{x^*\}$ is a singleton, then $\hat{\mathcal{S}}_N^\varepsilon = \{x^*\}$ w.p.1 for N large enough. In particular, if the true problem (1.1) has a unique optimal solution x^*, then w.p.1 for sufficiently large N the approximating problem (2.1) has a unique optimal solution $\hat{x}_N$ and $\hat{x}_N = x^*$. Also consider the set $A := \{g(x) - v^* : x \in \mathcal{S}\}$. The set A is a subset of the set $\mathbb{R}_+$ of nonnegative numbers and $|A| \leq |\mathcal{S}|$, and hence A is finite. It follows from the above analysis that for any $\varepsilon \in \mathbb{R}_+ \setminus A$ the event $\{\hat{\mathcal{S}}_N^\varepsilon = \mathcal{S}^\varepsilon\}$ happens w.p.1 for N large enough.

2.2. Convergence rates. The above results do not say anything about the rates of convergence of $\hat{v}_N$ and $\hat{\mathcal{S}}_N^\delta$ to their true counterparts. In this section we investigate such rates of convergence. By using the theory of large deviations (LD), we show that, under mild regularity conditions and $\delta \in [0, \varepsilon]$, the probability of the event $\{\hat{\mathcal{S}}_N^\delta \subset \mathcal{S}^\varepsilon\}$ approaches 1 exponentially fast as $N \to \infty$. Next we briefly outline some background of the LD theory.

Consider a random (real valued) variable X having mean $\mu := \mathbb{E}[X]$. Its moment-generating function $M(t) := \mathbb{E}[e^{tX}]$ is viewed as an extended valued function, i.e., it can take value $+\infty$. It holds that $M(t) > 0$ for all $t \in \mathbb{R}$, $M(0) = 1$, and the domain $\{t : M(t) < +\infty\}$ of the moment-generating function is an interval containing zero. The conjugate function

$$I(z) := \sup_{t \in \mathbb{R}} \{tz - \Lambda(t)\}, \tag{2.4}$$

of the logarithmic moment-generating function $\Lambda(t) := \log M(t)$, is called the (LD) *rate* function of X. It is possible to show that both functions $\Lambda(\cdot)$ and $I(\cdot)$ are convex.

Consider an i.i.d. sequence $X_1, \ldots, X_N$ of replications of the random variable X, and let $Z_N := N^{-1} \sum_{i=1}^N X_i$ be the corresponding sample average. Then for any real numbers a and $t \geq 0$ it holds that $P(Z_N \geq a) = P(e^{tZ_N} \geq e^{ta})$, and hence it follows from Chebyshev's inequality that

$$P(Z_N \geq a) \;\leq\; e^{-ta}\mathbb{E}\left[e^{tZ_N}\right] \;=\; e^{-ta}[M(t/N)]^N.$$

By taking the logarithm of both sides of the above inequality, changing variables $t' := t/N$, and minimizing over $t' \geq 0$, it follows for $a \geq \mu$ that

$$\frac{1}{N} \log\left[P(Z_N \geq a)\right] \;\leq\; -I(a). \tag{2.5}$$

Note that for $a \geq \mu$ it suffices to take the supremum in the definition (2.4) of $I(a)$ for $t \geq 0$, and therefore this constraint is omitted. Inequality (2.5) corresponds to the upper bound of Cramér's LD theorem.

The constant $I(a)$ in (2.5) gives, in a sense, the best possible exponential rate at which the probability $P(Z_N \geq a)$ converges to zero for $a > \mu$. This follows from the lower bound

$$\liminf_{N \to \infty} \frac{1}{N} \log\left[P(Z_N \geq a)\right] \;\geq\; -I(a) \tag{2.6}$$

of Cramér's LD theorem. A simple sufficient condition for (2.6) to hold is that the moment-generating function $M(t)$ is finite valued for all $t \in \mathbb{R}$. For a thorough discussion of the LD theory, the interested reader is referred to Dembo and Zeitouni [5].

The rate function $I(z)$ has the following properties: The function $I(z)$ is convex and attains its minimum at $z = \mu$, and $I(\mu) = 0$. Moreover, suppose that the moment-generating function $M(t)$ is finite valued for all t in a neighborhood of $t = 0$. Then X has finite moments, and it follows by the dominated convergence theorem that $M(t)$, and hence the function $\Lambda(t)$, are infinitely differentiable at $t = 0$, and $\Lambda'(0) = \mu$. Consequently for $a > \mu$ the derivative of $\psi(t) := ta - \Lambda(t)$ at $t = 0$ is greater than zero, and hence $\psi(t) > 0$ for $t > 0$ small enough. In that case it follows that $I(a) > 0$. Also, $I'(\mu) = 0$ and $I''(\mu) = \sigma^{-2}$, and hence by Taylor's expansion

$$I(a) = \frac{(a-\mu)^2}{2\sigma^2} + o(|a-\mu|^2). \tag{2.7}$$

Consequently, for a close to μ one can approximate $I(a)$ by $(a-\mu)^2/(2\sigma^2)$. Moreover, for any $\tilde{\epsilon} > 0$ there is a neighborhood $\mathcal{N}$ of μ such that

$$I(a) \geq \frac{(a-\mu)^2}{(2+\tilde{\epsilon})\sigma^2} \quad \forall\, a \in \mathcal{N}. \tag{2.8}$$

In particular, one can take $\tilde{\epsilon} = 1$.

Now we return to problems (1.1) and (2.1). Consider numbers $\varepsilon \geq 0$, $\delta \in [0, \varepsilon]$, and the event $\{\hat{\mathcal{S}}_N^\delta \subset \mathcal{S}^\varepsilon\}$. It holds that

$$\left\{\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon\right\} = \bigcup_{x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon} \bigcap_{y \in \mathcal{S}} \{\hat{g}_N(x) \leq \hat{g}_N(y) + \delta\}, \tag{2.9}$$

left: there exists x not in S^\epsilon such that is delta optimal for the approximate function

and hence

$$P\left(\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon\right) \leq \sum_{x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon} P\left(\bigcap_{y \in \mathcal{S}} \{\hat{g}_N(x) \leq \hat{g}_N(y) + \delta\}\right). \tag{2.10}$$

Consider a mapping $u : \mathcal{S} \setminus \mathcal{S}^\varepsilon \mapsto \mathcal{S}$. It follows from (2.10) that

$$P\left(\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon\right) \leq \sum_{x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon} P\Big(\hat{g}_N(x) - \hat{g}_N(u(x)) \leq \delta\Big). \tag{2.11}$$

We assume that the mapping $u(x)$ is chosen in such a way that for some $\varepsilon^* > \varepsilon$

$$g(u(x)) \leq g(x) - \varepsilon^* \quad \text{for all} \quad x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon. \tag{2.12}$$

Note that if $u(\cdot)$ is a mapping from $\mathcal{S} \setminus \mathcal{S}^\varepsilon$ into the set $\mathcal{S}^*$, i.e., $u(x) \in \mathcal{S}^*$ for all $x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon$, then (2.12) holds with

$$\varepsilon^* := \min_{x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon} g(x) - v^*, \tag{2.13}$$

and that $\varepsilon^* > \varepsilon$ since the set $\mathcal{S}$ is finite. Therefore a mapping $u(\cdot)$ that satisfies condition (2.12) always exists.

For each $x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon$, let

$$H(x, w) := G(u(x), w) - G(x, w).$$

Note that $\mathbb{E}[H(x, W)] = g(u(x)) - g(x)$, and hence $\mathbb{E}[H(x, W)] \leq -\varepsilon^*$. Let $W^1, \ldots, W^N$ be an i.i.d. random sample of N realizations of the random vector W, and consider the sample average function

$$\hat{h}_N(x) := \frac{1}{N} \sum_{j=1}^N H(x, W^j) = \hat{g}_N(u(x)) - \hat{g}_N(x).$$

It follows from (2.11) that

$$P\left(\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon\right) \leq \sum_{x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon} P\left(\hat{h}_N(x) \geq -\delta\right). \tag{2.14}$$

Let $I_x(\cdot)$ denote the LD rate function of $H(x, W)$. Inequality (2.14) together with (2.5) implies that

$$P\left(\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon\right) \leq \sum_{x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon} e^{-N I_x(-\delta)}. \tag{2.15}$$

It is important to note that the above inequality (2.15) is not asymptotic and is valid for any random sample of size N.

Assumption (A). For every $x \in \mathcal{S}$ the moment-generating function of the random variable $H(x, W)$ is finite valued in a neighborhood of 0.

The above assumption (A) holds, for example, if $H(x, W)$ is a bounded random variable, or if $H(x, \cdot)$ grows at most linearly and W has a distribution from the exponential family.

PROPOSITION 2.2. *Let ε and δ be nonnegative numbers such that $\delta \leq \varepsilon$. Then*

$$P\left(\hat{\mathcal{S}}_N^{\delta} \not\subset \mathcal{S}^{\varepsilon}\right) \leq |\mathcal{S} \setminus \mathcal{S}^{\varepsilon}| e^{-N\gamma(\delta,\varepsilon)}, \tag{2.16}$$

where

$$\gamma(\delta, \varepsilon) := \min_{x \in \mathcal{S} \setminus \mathcal{S}^{\varepsilon}} I_x(-\delta). \tag{2.17}$$

Moreover, if Assumption (A) *holds, then $\gamma(\delta, \varepsilon) > 0$.*

Proof. Inequality (2.16) is an immediate consequence of inequality (2.15). It holds that $-\delta > -\varepsilon^* \geq \mathbb{E}[H(x, W)]$, and hence it follows by Assumption (A) that $I_x(-\delta) > 0$ for every $x \in \mathcal{S} \setminus \mathcal{S}^{\varepsilon}$. This implies that $\gamma(\delta, \varepsilon) > 0$. □

The following asymptotic result is an immediate consequence of inequality (2.16),

$$\limsup_{N \to \infty} \frac{1}{N} \log\left[1 - P(\hat{\mathcal{S}}_N^{\delta} \subset \mathcal{S}^{\varepsilon})\right] \leq -\gamma(\delta, \varepsilon). \tag{2.18}$$

Inequality (2.18) means that the probability of the event $\{\hat{\mathcal{S}}_N^{\delta} \subset \mathcal{S}^{\varepsilon}\}$ approaches 1 *exponentially fast* as $N \to \infty$. This suggests that Monte Carlo sampling, combined with an efficient method for solving the deterministic SAA problem, can efficiently solve the type of problems under study, provided that the constant $\gamma(\delta, \varepsilon)$ is not "too small."

It follows from (2.7) that

$$I_x(-\delta) \approx \frac{(-\delta - \mathbb{E}[H(x, W)])^2}{2\sigma_x^2} \geq \frac{(\varepsilon^* - \delta)^2}{2\sigma_x^2}, \tag{2.19}$$

where ε^* is defined in (2.13) and

$$\sigma_x^2 := \mathrm{Var}[H(x, W)] = \mathrm{Var}[G(u(x), W) - G(x, W)].$$

Therefore the constant $\gamma(\delta, \varepsilon)$, given in (2.17), can be approximated by

$$\gamma(\delta, \varepsilon) \approx \min_{x \in \mathcal{S} \setminus \mathcal{S}^{\varepsilon}} \frac{(-\delta - \mathbb{E}[H(x, W)])^2}{2\sigma_x^2} \geq \frac{(\varepsilon^* - \delta)^2}{2\sigma_{\max}^2} > \frac{(\varepsilon - \delta)^2}{2\sigma_{\max}^2}, \tag{2.20}$$

where

$$\sigma_{\max}^2 := \max_{x \in \mathcal{S} \setminus \mathcal{S}^{\varepsilon}} \mathrm{Var}[G(u(x), W) - G(x, W)]. \tag{2.21}$$

A result similar to the one of Proposition 2.2 was derived in [14] by using slightly different arguments. The LD rate functions of the random variables $G(x, W)$ were used there, which resulted in estimates of the exponential constant similar to the estimate (2.20) but with σ_x^2 replaced by the variance of $G(x, W)$. Due to a positive correlation between $G(x, W)$ and $G(u(x), W)$, the variance of $G(x, W) - G(u(x), W)$ tends to be smaller than the variance of $G(x, W)$, thereby providing a smaller upper

bound on $P(\hat{\mathcal{S}}_N^\delta \not\subset \mathcal{S}^\varepsilon)$, especially when $u(x)$ is chosen to minimize $\mathrm{Var}[G(x,W) - G(u(x),W)]/[g(x) - g(u(x))]^2$. This suggests that the estimate given in (2.20) could be more accurate than the one obtained in [14].

To illustrate some implications of the bound (2.16) for issues of the complexity of solving stochastic problems, let us fix a significance level $\alpha \in (0,1)$, and estimate the sample size N which is needed for the probability $P(\hat{\mathcal{S}}_N^\delta \subset \mathcal{S}^\varepsilon)$ to be at least $1-\alpha$. By requiring that the right-hand side of (2.16) be less than or equal to α, we obtain that

$$N \geq \frac{1}{\gamma(\delta,\varepsilon)} \log\left(\frac{|\mathcal{S} \setminus \mathcal{S}^\varepsilon|}{\alpha}\right). \tag{2.22}$$

Moreover, it follows from (2.8) and (2.17) that $\gamma(\delta,\varepsilon) \geq (\varepsilon-\delta)^2/(3\sigma_{\max}^2)$ for all $\varepsilon \geq 0$ sufficiently small. Therefore it holds that for all $\varepsilon > 0$ small enough and $\delta \in [0,\varepsilon)$, a sufficient condition for (2.22) is that

$$N \geq \frac{3\sigma_{\max}^2}{(\varepsilon-\delta)^2} \log\left(\frac{|\mathcal{S}|}{\alpha}\right). \tag{2.23}$$

It appears that the bound (2.23) may be too conservative for practical estimates of the required sample sizes (see the discussion in section 4.2). However, the estimate (2.23) has interesting consequences for complexity issues. A key characteristic of (2.23) is that N depends only *logarithmically* both on the size of the feasible set $\mathcal{S}$ and on the tolerance probability α. An important implication of such behavior is the following. Suppose that (i) the size of the feasible set $\mathcal{S}$ grows at most exponentially in the length of the problem input, (ii) the variance $\sigma_{\max}^2$ grows polynomially in the length of the problem input, and (iii) the complexity of finding a δ-optimal solution for (2.1) grows polynomially in the length of the problem input and the sample size N. Then a solution can be generated in time that grows polynomially in the length of the problem input such that, with probability at least $1-\alpha$, the solution is ε-optimal for (1.1). A careful analysis of these issues is beyond the scope of this paper, and requires further investigation.

Now suppose for a moment that the true problem has unique optimal solution x^*, i.e., $\mathcal{S}^* = \{x^*\}$ is a singleton, and consider the event that the SAA problem (2.1) has unique optimal solution $\hat{x}_N$ and $\hat{x}_N = x^*$. We denote that event by $\{\hat{x}_N = x^*\}$. Furthermore, consider the mapping $u : \mathcal{S} \setminus \mathcal{S}^\varepsilon \mapsto \{x^*\}$, i.e., $u(x) \equiv x^*$, and the corresponding constant $\gamma^* := \gamma(0,0)$. That is,

$$\gamma^* = \min_{x \in \mathcal{S} \setminus \{x^*\}} I_x(0), \tag{2.24}$$

with $I_x(\cdot)$ being the LD rate function of $G(x^*,W) - G(x,W)$. Note that $\mathbb{E}[G(x^*,W) - G(x,W)] = g(x^*) - g(x)$, and hence $\mathbb{E}[G(x^*,W) - G(x,W)] < 0$ for every $x \in \mathcal{S} \setminus \{x^*\}$. Therefore, if Assumption (A) holds, i.e., the moment-generating function of $G(x^*,W) - G(x,W)$ is finite valued in a neighborhood of 0, then $\gamma^* > 0$.

PROPOSITION 2.3. *Suppose that the true problem has unique optimal solution x^* and the moment-generating function of each random variable $G(x^*,W) - G(x,W)$, $x \in \mathcal{S} \setminus \{x^*\}$, is finite valued on $\mathbb{R}$. Then*

$$\lim_{N\to\infty} \frac{1}{N} \log\left[1 - P(\hat{x}_N = x^*)\right] = -\gamma^*. \tag{2.25}$$

Proof. It follows from (2.18) that

$$\limsup_{N\to\infty} \frac{1}{N} \log\left[1 - P(\hat{x}_N = x^*)\right] \;\leq\; -\gamma^*. \tag{2.26}$$

Consider the complement of the event $\{\hat{x}_N = x^*\}$, which is denoted $\{\hat{x}_N \neq x^*\}$. The event $\{\hat{x}_N \neq x^*\}$ is equal to the union of the events $\{\hat{g}_N(x) \leq \hat{g}_N(x^*)\}$, $x \in \mathcal{S} \setminus \{x^*\}$. Therefore, for any $x \in \mathcal{S} \setminus \{x^*\}$,

$$P\left(\hat{x}_N \neq x^*\right) \;\geq\; P\left(\hat{g}_N(x) \leq \hat{g}_N(x^*)\right).$$

By using the lower bound (2.6) of Cramér's LD theorem, it follows that the inequality

$$\liminf_{N\to\infty} \frac{1}{N} \log\left[1 - P(\hat{x}_N = x^*)\right] \;\geq\; -I_x(0) \tag{2.27}$$

holds for every $x \in \mathcal{S} \setminus \{x^*\}$. Inequalities (2.26) and (2.27) imply (2.25). □

Suppose that $\mathcal{S}^* = \{x^*\}$ and consider the number

$$\kappa \;:=\; \max_{x\in\mathcal{S}\setminus\{x^*\}} \frac{\mathrm{Var}[G(x,W) - G(x^*,W)]}{[g(x) - g(x^*)]^2}. \tag{2.28}$$

It follows from (2.7) and (2.24) that $\kappa \approx 1/(2\gamma^*)$. One can view κ as a *condition number* of the true problem. That is, the sample size required for the event $\{\hat{x}_N = x^*\}$ to happen with a given probability is roughly proportional to κ. The number defined in (2.28) can be viewed as a discrete version of the condition number introduced in [22] for piecewise linear continuous problems.

For a problem with a large feasible set $\mathcal{S}$, the number $\min_{x\in\mathcal{S}\setminus\{x^*\}} g(x) - g(x^*)$, although positive if $\mathcal{S}^* = \{x^*\}$, tends to be small. Therefore the sample size required to calculate the exact optimal solution x^* with a high probability could be very large, even if the optimal solution x^* is unique. For ill-conditioned problems it makes sense to search for approximate (ε-optimal) solutions of the true problem. In that respect the bound (2.16) is more informative since the corresponding constant $\gamma(\delta,\varepsilon)$ is guaranteed to be at least of the order $(\varepsilon - \delta)^2/(2\sigma^2_{\max})$.

It is also insightful to note the behavior of the condition number κ for a discrete optimization problem with linear objective function $G(x,W) := \sum_{i=1}^k W_i x_i$ and feasible set $\mathcal{S}$ given by the vertices of the unit hypercube in $\mathbb{R}^k$, i.e., $\mathcal{S} := \{0,1\}^k$. In that case the corresponding true optimization problem is

$$\min_{x\in\{0,1\}^k} \left\{ g(x) = \sum_{i=1}^k \bar{w}_i x_i \right\},$$

where $\bar{w}_i := E[W_i]$. Suppose that $\bar{w}_i > 0$ for all $i \in \{1,\ldots,k\}$, and hence the origin is the unique optimal solution of the true problem, i.e., $\mathcal{S}^* = \{0\}$. Let

$$\vartheta_i^2 \;:=\; \frac{\mathrm{Var}[W_i]}{(E[W_i])^2}$$

denote the squared coefficient of variation of W_i, and let

$$\rho_{ij} \;:=\; \frac{\mathrm{Cov}[W_i, W_j]}{\sqrt{\mathrm{Var}[W_i]}\sqrt{\mathrm{Var}[W_j]}}$$

denote the correlation coefficient between W_i and W_j. It follows that for any $x \in \{0,1\}^k \setminus \{0\}$,

$$\frac{\mathrm{Var}\left[\sum_{i=1}^k W_i x_i\right]}{\left[\sum_{i=1}^k \bar{w}_i x_i\right]^2} = \frac{\sum_{i=1}^k \sum_{j=1}^k \rho_{ij}\vartheta_i \bar{w}_i x_i \vartheta_j \bar{w}_j x_j}{\sum_{i=1}^k \sum_{j=1}^k \bar{w}_i x_i \bar{w}_j x_j} \leq \max_{i\in\{1,\ldots,k\}} \vartheta_i^2.$$

Thus

$$\kappa = \max_{x\in\{0,1\}^k\setminus\{0\}} \frac{\mathrm{Var}\left[\sum_{i=1}^k W_i x_i\right]}{\left[\sum_{i=1}^k \bar{w}_i x_i\right]^2} = \max_{i\in\{1,\ldots,k\}} \vartheta_i^2.$$

The last equality follows because the maximum is attained by setting $x_i = 1$ for the index i for which W_i has the maximum squared coefficient of variation ϑ_i^2, and setting $x_j = 0$ for the remaining variables. Thus, in this example the condition number κ is equal to the maximum squared coefficient of variation of the W_i's.

2.3. Asymptotics of sample objective values. Next we discuss the asymptotics of the SAA optimal objective value $\hat{v}_N$. For any subset $\mathcal{S}'$ of $\mathcal{S}$ the inequality $\hat{v}_N \leq \min_{x\in\mathcal{S}'} \hat{g}_N(x)$ holds. In particular, by taking $\mathcal{S}' = \mathcal{S}^*$, it follows that $\hat{v}_N \leq \min_{x\in\mathcal{S}^*} \hat{g}_N(x)$, and hence

$$\mathbb{E}[\hat{v}_N] \leq \mathbb{E}\left\{\min_{x\in\mathcal{S}^*} \hat{g}_N(x)\right\} \leq \min_{x\in\mathcal{S}^*} \mathbb{E}[\hat{g}_N(x)] = v^*.$$

That is, the estimator $\hat{v}_N$ has a negative bias (cf. Norkin, Pflug, and Ruszczyński [19] and Mak, Morton, and Wood [15]).

It follows from Proposition 2.1 that w.p.1, for N sufficiently large, the set $\hat{\mathcal{S}}_N$ of optimal solutions of the SAA problem is included in $\mathcal{S}^*$. In that case it holds that

$$\hat{v}_N = \min_{x\in\hat{\mathcal{S}}_N} \hat{g}_N(x) \geq \min_{x\in\mathcal{S}^*} \hat{g}_N(x).$$

Since the opposite inequality always holds, it follows that, w.p.1, $\hat{v}_N - \min_{x\in\mathcal{S}^*} \hat{g}_N(x) = 0$ for N large enough. Multiplying both sides of this equation by $\sqrt{N}$ it follows that, w.p.1, $\sqrt{N}\,[\hat{v}_N - \min_{x\in\mathcal{S}^*} \hat{g}_N(x)] = 0$ for N large enough, and hence

$$\lim_{N\to\infty} \sqrt{N}\left[\hat{v}_N - \min_{x\in\mathcal{S}^*} \hat{g}_N(x)\right] = 0 \quad \text{w.p.1.} \tag{2.29}$$

not rigorous

Since convergence w.p.1 implies convergence in probability, it follows from (2.29) that $\sqrt{N}\,[\hat{v}_N - \min_{x\in\mathcal{S}^*} \hat{g}_N(x)]$ converges in probability to zero, i.e.,

$$\hat{v}_N = \min_{x\in\mathcal{S}^*} \hat{g}_N(x) + o_p(N^{-1/2}).$$

Furthermore, since $v^* = g(x)$ for any $x \in \mathcal{S}^*$, it follows that

$$\sqrt{N}\left[\min_{x\in\mathcal{S}^*} \hat{g}_N(x) - v^*\right] = \sqrt{N} \min_{x\in\mathcal{S}^*} [\hat{g}_N(x) - v^*] = \min_{x\in\mathcal{S}^*} \left\{\sqrt{N}\,[\hat{g}_N(x) - g(x)]\right\}.$$

Suppose that for every $x \in \mathcal{S}$ the variance

$$\sigma^2(x) := \mathrm{Var}[G(x,W)] \tag{2.30}$$

exists. Then it follows by the central limit theorem (CLT) that, for any $x \in \mathcal{S}$, $\sqrt{N}[\hat{g}_N(x) - g(x)]$ converges in distribution to a normally distributed variable $Z(x)$ with zero mean and variance $\sigma^2(x)$. Moreover, again by the CLT, random variables $Z(x)$ have the same covariance function as $G(x, W)$, i.e., the covariance between $Z(x)$ and $Z(x')$ is equal to the covariance between $G(x, W)$ and $G(x', W)$ for any $x, x' \in \mathcal{S}$. Hence the following result is obtained (it is similar to an asymptotic result for stochastic programs with continuous decision variables which was derived in [21]). We use "$\Rightarrow$" to denote convergence in distribution.

PROPOSITION 2.4. *Suppose that variances $\sigma^2(x)$, defined in* (2.30)*, exist for every $x \in \mathcal{S}^*$. Then*

$$\sqrt{N}(\hat{v}_N - v^*) \;\Rightarrow\; \min_{x \in \mathcal{S}^*} Z(x), \tag{2.31}$$

where $Z(x)$ are normally distributed random variables with zero mean and the covariance function given by the corresponding covariance function of $G(x, W)$. In particular, if $\mathcal{S}^* = \{x^*\}$ is a singleton, then

$$\sqrt{N}(\hat{v}_N - v^*) \;\Rightarrow\; N(0, \sigma^2(x^*)). \tag{2.32}$$

Although for any given x the mean (expected value) of $Z(x)$ is zero, the expected value of the minimum of $Z(x)$ over a subset $\mathcal{S}'$ of $\mathcal{S}$ can be negative and tends to be smaller for a larger set $\mathcal{S}'$. Therefore, it follows from (2.31) that for ill-conditioned problems, where the set of optimal or nearly optimal solutions is large, the estimate $\hat{v}_N$ of v^* tends to be heavily biased. Note that convergence in distribution does not necessarily imply convergence of the corresponding means. Under mild additional conditions it follows from (2.31) that $\sqrt{N}[\mathbb{E}(\hat{v}_N) - v^*] \to \mathbb{E}[\min_{x \in \mathcal{S}^*} Z(x)]$.

3. Algorithm design. In the previous section we established a number of convergence results for the SAA method. The results describe how the optimal value $\hat{v}_N$ and the set $\hat{\mathcal{S}}_N^\varepsilon$ of ε-optimal solutions of the SAA problem converge to their true counterparts v^* and $\mathcal{S}^\varepsilon$, respectively, as the sample size N increases. These results provide some theoretical justification for the proposed method. When designing an algorithm for solving stochastic discrete optimization problems, many additional issues have to be addressed. Some of these issues are discussed in this section.

3.1. Selection of the sample size. In an algorithm, a finite sample size N or a sequence of finite sample sizes has to be chosen, and the algorithm has to stop after a finite amount of time. An important question is how these choices should be made. Estimate (2.23) gives a bound on the sample size required to find an ε-optimal solution with probability at least $1-\alpha$. This estimate has two shortcomings for computational purposes. First, for many problems it is not easy to compute the estimate, because $\sigma^2_{\max}$ and in some problems also $|\mathcal{S}|$ may be hard to compute. Second, as demonstrated in section 4.2, the bound may be far too conservative to obtain a practical estimate of the required sample size. To choose N, several trade-offs should be taken into account. With larger N, the objective function of the SAA problem tends to be a more accurate estimate of the true objective function, an optimal solution of the SAA problem tends to be a better solution, and the corresponding bounds on the optimality gap, discussed later, tend to be tighter. However, depending on the SAA problem (2.1) and the method used for solving the SAA problem, the computational complexity for solving the SAA problem increases at least linearly, and often exponentially, in the sample size N. Thus, in the choice of sample size N, the trade-off between the quality

of an optimal solution of the SAA problem and the bounds on the optimality gap, on the one hand, and computational effort, on the other hand, should be taken into account. Also, the choice of sample size N may be adjusted dynamically, depending on the results of preliminary computations. This issue is addressed in more detail later.

Typically, estimating the objective value $g(x)$ of a feasible solution $x \in \mathcal{S}$ by the sample average $\hat{g}_N(x)$ requires much less computational effort than solving the SAA problem (for the same sample size N). Thus, although computational complexity considerations motivate one to choose a relatively small sample size N for the SAA problem, it makes sense to choose a larger sample size N' to obtain an accurate estimate $\hat{g}_{N'}(\hat{x}_N)$ of the objective value $g(\hat{x}_N)$ of an optimal solution $\hat{x}_N$ of the SAA problem. A measure of the accuracy of a sample average estimate $\hat{g}_{N'}(\hat{x}_N)$ of $g(\hat{x}_N)$ is given by the corresponding sample variance $S^2_{N'}(\hat{x}_N)/N'$, which can be calculated from the same sample of size N'. Again the choice of N' involves a trade-off between computational effort and accuracy, measured by $S^2_{N'}(\hat{x}_N)/N'$.

3.2. Replication. If the computational complexity of solving the SAA problem increases faster than linearly in the sample size N, it may be more efficient to choose a smaller sample size N and to generate and solve several SAA problems with i.i.d. samples, that is, to replicate generating and solving SAA problems.

With such an approach, several issues have to be addressed. One question is whether there is a guarantee that an optimal (or ε-optimal) solution for the true problem will be produced if a sufficient number of SAA problems, based on independent samples of size N, are solved. One can view such a procedure as Bernoulli trials with probability of success $p = p(N)$. Here "success" means that a calculated optimal solution $\hat{x}_N$ of the SAA problem is an optimal solution of the true problem. It follows from Proposition 2.1 that this probability p tends to 1 as $N \to \infty$, and, moreover, by Proposition 2.2 it tends to 1 exponentially fast if Assumption (A) holds. However, for a finite N the probability p can be small or even zero. The probability of producing an optimal solution of the true problem at least once in M trials is $1 - (1-p)^M$, and this probability tends to one as $M \to \infty$, provided p is positive. Thus a relevant question is whether there is a guarantee that p is positive for a given sample size N. The following example shows that the sample size N required for p to be positive is problem-specific, cannot be bounded by a function that depends only on the number of feasible solutions, and can be arbitrarily large.

Example. Suppose that $\mathcal{S} := \{-1, 0, 1\}$, that W can take two values w_1 and w_2 with respective probabilities $1-\gamma$ and γ, and that $G(-1, w_1) := -1$, $G(0, w_1) := 0$, $G(1, w_1) := 2$, and $G(-1, w_2) := 2k$, $G(0, w_2) := 0$, $G(1, w_2) := -k$, where k is an arbitrary positive number. Let $\gamma = 1/(k+1)$. Then $g(x) = (1-\gamma)G(x, w_1) + \gamma G(x, w_2)$, and thus $g(-1) = k/(k+1)$, $g(0) = 0$, and $g(1) = k/(k+1)$. Therefore $x^* = 0$ is the unique optimal solution of the true problem. If the sample does not contain any observations w_2, then $\hat{x}_N = -1 \neq x^*$. Suppose the sample contains at least one observation w_2. Then $\hat{g}_N(1) \leq [2(N-1) - k]/N$. Thus $\hat{g}_N(1) < 0 = \hat{g}_N(0)$ if $N \leq k/2$, and $\hat{x}_N = 1 \neq x^*$. Thus a sample of size $N > k/2$ at least is required, in order for $x^* = 0$ to be an optimal solution of the SAA problem. (Note that $\mathrm{Var}[G(-1, W) - G(0, W)]$ and $\mathrm{Var}[G(1, W) - G(0, W)]$ are $\Theta(k)$, which causes the problem to become harder as k increases.)

Another issue that has to be addressed is the choice of the number M of replications. In a manner similar to the choice of sample size N, the number M of replications may be chosen dynamically. One approach to doing this is discussed next. For sim-

plicity of presentation, suppose that each SAA replication produces one candidate solution, which can be an optimal (ε-optimal) solution of the SAA problem. Let $\hat{x}_N^m$ denote the candidate solution produced by the mth SAA replication (trial). The optimality gap $g(\hat{x}_N^m) - v^*$ can be estimated, as described in the next section. If a stopping criterion based on the optimality gap estimate is satisfied, then no more replications are performed. Otherwise, additional SAA replications with the same sample size N are performed, or the sample size N is increased. The following argument provides a simple guideline as to whether an additional SAA replication with the same sample size N is likely to provide a better solution than the best solution found so far.

Note that, by construction, the random variables $g(\hat{x}_N^m)$, $m = 1, \ldots$, are i.i.d., and their common probability distribution has a finite support because the set $\mathcal{S}$ is finite. Suppose that M replications with sample size N have been performed so far. If the probability distribution of $g(\hat{x}_N)$ were continuous, then the probability that the $(M+1)$th SAA replication with the same sample size would produce a better solution than the best of the solutions produced by the M replications so far would be equal to $1/(M+1)$. Because in fact the distribution of $g(\hat{x}_N)$ is discrete, this probability is less than or equal to $1/(M+1)$. Thus, when $1/(M+1)$ becomes sufficiently small, additional SAA replications with the same sample size are not likely to be worth the effort, and either the sample size N should be increased or the procedure should be stopped.

3.3. Performance bounds. To assist in making stopping decisions, as well as for other performance evaluation purposes, one would like to compute the optimality gap $g(\hat{x}) - v^*$ for a given solution $\hat{x} \in \mathcal{S}$. Unfortunately, the very reason for the approach described in this paper implies that both terms of the optimality gap are hard to compute. As before,

$$\hat{g}_{N'}(\hat{x}) \;:=\; \frac{1}{N'} \sum_{j=1}^{N'} G(\hat{x}, W^j)$$

is an unbiased estimator of $g(\hat{x})$, and the variance of $\hat{g}_{N'}(\hat{x})$ is estimated by $S_{N'}^2(\hat{x})/N'$, where $S_{N'}^2(\hat{x})$ is the sample variance of $G(\hat{x}, W^j)$, based on the sample of size N'.

An estimator of v^* is given by

$$\bar{v}_N^M \;:=\; \frac{1}{M} \sum_{m=1}^{M} \hat{v}_N^m,$$

where $\hat{v}_N^m$ denotes the optimal objective value of the mth SAA replication. Note that $\mathbb{E}[\bar{v}_N^M] = \mathbb{E}[\hat{v}_N]$, and hence the estimator $\bar{v}_N^M$ has the same negative bias as $\hat{v}_N$. Proposition 2.4 indicates that this bias tends to be bigger for ill-conditioned problems with larger sets of optimal, or nearly optimal, solutions. Consider the corresponding estimator $\hat{g}_{N'}(\hat{x}) - \bar{v}_N^M$ of the optimality gap $g(\hat{x}) - v^*$, at the point $\hat{x}$. Since

$$\mathbb{E}\left[\hat{g}_{N'}(\hat{x}) - \bar{v}_N^M\right] \;=\; g(\hat{x}) - \mathbb{E}[\hat{v}_N] \;\geq\; g(\hat{x}) - v^*, \tag{3.1}$$

it follows that on average the above estimator overestimates the optimality gap $g(\hat{x}) - v^*$. It is possible to show (Norkin, Pflug, and Ruszczyński [19], and Mak, Morton, and Wood [15]) that the bias $v^* - \mathbb{E}[\hat{v}_N]$ is monotonically decreasing in the sample size N.

The variance of $\bar{v}_N^M$ is estimated by

$$\frac{S_M^2}{M} \;=\; \frac{1}{M(M-1)} \sum_{m=1}^{M} \left(\hat{v}_N^m - \bar{v}_N^M\right)^2. \tag{3.2}$$

If the M samples, of size N, and the evaluation sample, of size N', are independent, then the variance of the optimality gap estimator $\hat{g}_{N'}(\hat{x}) - \bar{v}_N^M$ can be estimated by $S_{N'}^2(\hat{x})/N' + S_M^2/M$.

An estimator of the optimality gap $g(\hat{x}) - v^*$ with possibly smaller variance is $\bar{g}_N^M(\hat{x}) - \bar{v}_N^M$, where

$$\bar{g}_N^M(\hat{x}) \;:=\; \frac{1}{M} \sum_{m=1}^{M} \hat{g}_N^m(\hat{x})$$

and $\hat{g}_N^m(\hat{x})$ is the sample average objective value at $\hat{x}$ of the mth SAA sample of size N,

$$\hat{g}_N^m(\hat{x}) \;:=\; \frac{1}{N} \sum_{j=1}^{N} G(\hat{x}, W^{mj}).$$

The variance of $\bar{g}_N^M(\hat{x}) - \bar{v}_N^M$ is estimated by

$$\frac{\bar{S}_M^2}{M} \;=\; \frac{1}{M(M-1)} \sum_{m=1}^{M} \left[\left(\hat{g}_N^m(\hat{x}) - \hat{v}_N^m\right) - \left(\bar{g}_N^M(\hat{x}) - \bar{v}_N^M\right)\right]^2.$$

Which estimator of the optimality gap has the least variance depends on the correlation between $\hat{g}_N^m(\hat{x})$ and $\hat{v}_N^m$, as well as on the sample sizes N, N', and M. For many applications, one would expect positive correlation between $\hat{g}_N^m(\hat{x})$ and $\hat{v}_N^m$. The additional computational effort to compute $\hat{g}_N^m(\hat{x})$ for $m = 1, \ldots, M$ should also be taken into account when evaluating any such variance reduction. Either way, the CLT can be applied to the optimality gap estimators $\hat{g}_{N'}(\hat{x}) - \bar{v}_N^M$ and $\bar{g}_N^M(\hat{x}) - \bar{v}_N^M$, so that the accuracy of an optimality gap estimator can be taken into account by adding a multiple z_α of its estimated standard deviation to the gap estimator. Here $z_\alpha := \Phi^{-1}(1-\alpha)$, where $\Phi(z)$ is the cumulative distribution function of the standard normal distribution. For example, if $\hat{x} \in \mathcal{S}$ denotes the candidate solution with the best value of $\hat{g}_{N'}(\hat{x})$ found after M replications, then an optimality gap estimator taking accuracy into account is given by either

$$\hat{g}_{N'}(\hat{x}) - \bar{v}_N^M + z_\alpha \left(\frac{S_{N'}^2(\hat{x})}{N'} + \frac{S_M^2}{M}\right)^{1/2}$$

or

$$\bar{g}_N^M(\hat{x}) - \bar{v}_N^M + z_\alpha \frac{\bar{S}_M}{\sqrt{M}}.$$

For algorithm control, it is useful to separate an optimality gap estimator into its components. For example,

$$\begin{aligned} &\hat{g}_{N'}(\hat{x}) - \bar{v}_N^M + z_\alpha \left(\frac{S_{N'}^2(\hat{x})}{N'} + \frac{S_M^2}{M}\right)^{1/2} \\ &\quad= \left(\hat{g}_{N'}(\hat{x}) - g(\hat{x})\right) + \left(g(\hat{x}) - v^*\right) + \left(v^* - \bar{v}_N^M\right) + z_\alpha \left(\frac{S_{N'}^2(\hat{x})}{N'} + \frac{S_M^2}{M}\right)^{1/2}. \end{aligned} \tag{3.3}$$

In the four terms on the right-hand side of the above equation, the first term has expected value zero; the second term is the true optimality gap; the third term is the bias term, which has positive expected value decreasing in the sample size N; and the fourth term is the accuracy term, which is decreasing in the number M of replications and the sample size N'. Thus a disadvantage of these optimality gap estimators is that the gap estimator may be large if M, N, or N' is small, even if $\hat{x}$ is an optimal solution, i.e., $g(\hat{x}) - v^* = 0$.

3.4. Postprocessing, screening, and selection. Suppose a decision has been made to stop, for example when the optimality gap estimator has become small enough. At this stage the candidate solution $\hat{x} \in \mathcal{S}$ with the best value of $\hat{g}_{N'}(\hat{x})$ can be selected as the chosen solution. However, it may be worthwhile to perform a more detailed evaluation of the candidate solutions produced during the replications. There are several statistical screening and selection methods for selecting subsets of solutions or a single solution, among a (reasonably small) finite set of solutions, using samples of the objective values of the solutions. Many of these methods are described in Hochberg and Tamhane [12] and Bechhofer, Santner, and Goldsman [2]. In the numerical tests described in section 4, a combined procedure was used, as described in Nelson et al. [17]. During the first stage of the combined procedure, a subset $\mathcal{S}''$ of the candidate solutions $\mathcal{S}' := \{\hat{x}_N^1, \ldots, \hat{x}_N^M\}$ is chosen (called screening) for further evaluation, based on its sample average values $\hat{g}_{N'}(\hat{x}_N^m)$. During the second stage, sample sizes $N'' \geq N'$ are determined for more detailed evaluation, based on the sample variances $S^2_{N'}(\hat{x}_N^m)$. Then $N'' - N'$ additional observations are generated, and the candidate solution $\hat{x} \in \mathcal{S}''$ with the best value of $\hat{g}_{N''}(\hat{x})$ is selected as the chosen solution. The combined procedure guarantees that the chosen solution $\hat{x}$ has objective value $g(\hat{x})$ within a specified tolerance δ of the best value $\min_{\hat{x}_N^m \in \mathcal{S}'} g(\hat{x}_N^m)$ over all candidate solutions $\hat{x}_N^m$ with probability at least equal to specified confidence level $1 - \alpha$.

3.5. Algorithm. Next we state a proposed algorithm for the type of stochastic discrete optimization problem studied in this paper.

SAA Algorithm for Stochastic Discrete Optimization.

1. Choose initial sample sizes N and N', a decision rule for determining the number M of SAA replications (possibly involving a maximum number M' of SAA replications with the same sample size, such that $1/(M'+1)$ is sufficiently small), a decision rule for increasing the sample sizes N and N' if needed, and tolerance ε.
2. For $m = 1, \ldots, M$, do steps 2.1 through 2.3.
   - 2.1 Generate a sample of size N and solve the SAA problem (2.1) with objective value $\hat{v}_N^m$ and ε-optimal solution $\hat{x}_N^m$.
   - 2.2 Estimate the optimality gap $g(\hat{x}_N^m) - v^*$ and the variance of the gap estimator.
   - 2.3 If the optimality gap and the variance of the gap estimator are sufficiently small, go to step 4.
3. If the optimality gap or the variance of the gap estimator is too large, increase the sample sizes N and/or N', and return to step 2.
4. Choose the best solution $\hat{x}$ among all candidate solutions $\hat{x}_N^m$ produced, using a screening and selection procedure. Stop.

4. Numerical tests. In this section we describe an application of the SAA method to an optimization problem. The purposes of these tests are to investigate

the viability of the SAA approach, as well as to study the effects of problem parameters, such as the number of decision variables and the condition number κ, on the performance of the method.

4.1. Resource allocation problem. We apply the method to the following resource allocation problem. A decision maker has to choose a subset of k known alternative projects to take on. For this purpose a known quantity q of relatively low-cost resource is available to be allocated. Any additional amount of resource required can be obtained at a known incremental cost of c per unit of resource. The amount W_i of resource required by each project i is not known at the time the decision has to be made, but we assume that the decision maker has an estimate of the probability distribution of $W = (W_1, \ldots, W_k)$. Each project i has an expected net reward (expected revenue minus expected resource use times the lower cost) of r_i. Thus the optimization problem can be formulated as follows:

$$\max_{x \in \{0,1\}^k} \left\{ \sum_{i=1}^{k} r_i x_i - c\,\mathbb{E}\left[\sum_{i=1}^{k} W_i x_i - q \right]^+ \right\}, \tag{4.1}$$

where $[x]^+ := \max\{x, 0\}$. This problem can also be described as a knapsack problem, where a subset of k items has to be chosen, given a knapsack of size q into which to fit the items. The size W_i of each item i is random, and a per unit penalty of c has to be paid for exceeding the capacity of the knapsack. For this reason the problem is called the *static stochastic knapsack problem* (SSKP).

This problem was chosen for several reasons. First, expected value terms similar to that in the objective function of (4.1) occur in many interesting stochastic optimization problems. One such example is airline crew scheduling. An airline crew schedule is made up of crew pairings, where each crew pairing consists of a number of consecutive days (duties) of flying by a crew. Let $\{p_1, \ldots, p_k\}$ denote the set of pairings that can be chosen from. Then a crew schedule can be denoted by the decision vector $x \in \{0,1\}^k$, where $x_i = 1$ means that pairing p_i is flown. The cost $C_i(x)$ of a crew pairing p_i is given by

$$C_i(x) \;=\; \max\left\{ \sum_{d \in p_i} b_d(x), f t_i(x), g n_i \right\},$$

where $b_d(x)$ denotes the cost of duty d in pairing p_i, $t_i(x)$ denotes the total time duration of pairing p_i, n_i denotes the number of duties in pairing p_i, and f and g are constants determined by contracts. Even ignoring airline recovery actions such as cancellations and rerouting, $b_d(x)$ and $t_i(x)$ are random variables. The optimization problem is then

$$\min_{x \in \mathcal{X} \subset \{0,1\}^k} \sum_{i=1}^{k} \mathbb{E}[C_i(x)] x_i,$$

where $\mathcal{X}$ denotes the set of feasible crew schedules. Thus the objective function of the crew pairing problem can be written in a form similar to that of the objective function of (4.1).

Another example is a stochastic shortest path problem, where travel times are random and a penalty is incurred for arriving late at the destination. In this case,

the cost $C(x)$ of a path x is given by

$$C(x) \;=\; \sum_{(i,j)\in x} b_{ij} + c\left[\sum_{(i,j)\in x} t_{ij} - q\right]^+,$$

where b_{ij} is the cost of traversing arc (i,j), t_{ij} is the time of traversing arc (i,j), q is the available time to travel to the destination, and c is the penalty per unit time late. The optimization problem is then

$$\min_{x\in\mathcal{X}} \mathbb{E}[C(x)],$$

where $\mathcal{X}$ denotes the set of feasible paths in the network from the specified origin to the specified destination.

A second reason for choosing the SSKP is that objective functions with terms such as $\mathbb{E}[\sum_{i=1}^k W_i x_i - q]^+$ are interesting for the following reason. For many stochastic optimization problems good solutions can be obtained by replacing the random variables W by their means and then solving the resulting deterministic optimization problem $\max_x G(x, E[W])$, called the expected value problem (Birge and Louveaux [3]). It is easy to see that this may not be the case if the objective contains an expected value term as in (4.1). For a given solution x, this term may be very large but may become small if $W_1, \ldots, W_k$ are replaced by their means. In such a case, the obtained expected value problem may produce very bad solutions for the corresponding stochastic optimization problem.

The SSKP was also chosen because it is of interest by itself. One application is the decision faced by a contractor who can take on several contracts, such as an electricity supplier who can supply power to several groups of customers or a building contractor who can bid on several construction projects. The amount of work that will be required by each contract is unknown at the time the contracting decision has to be made. The contractor has the capacity to do work at a certain rate at relatively low cost, for example to generate electricity at a low-cost nuclear power plant. However, if the amount of work required exceeds the capacity, additional capacity has to be obtained at high cost, for example additional electricity can be generated at high-cost oil or natural gas–fired power plants. Norkin, Ermoliev, and Ruszczyński [18] also give several interesting applications of stochastic discrete optimization problems.

Note that the SAA problem of the SSKP can be formulated as the following integer linear program:

$$\begin{array}{lll} \max_{x,z} & \sum_{i=1}^k r_i x_i - \frac{c}{N}\sum_{j=1}^N z_j & \\ \text{subject to} & z_j \;\geq\; \sum_{i=1}^k W_i^j x_i - q, & j = 1,\ldots,N, \\ & x_i \;\in\; \{0,1\}, & i = 1,\ldots,k, \\ & z_j \;\geq\; 0, & j = 1,\ldots,N. \end{array} \tag{4.2}$$

This problem can be solved with the branch and bound method, using the linear programming relaxation to provide upper bounds.

4.2. Numerical results. We present results for two sets of instances of the SSKP. The first set of instances has 10 decision variables, and the second set has 20 decision variables each. For each set we present one instance (called instances 10D and 20D, respectively) that was designed to be hard (large condition number κ), and one randomly generated instance (called instances 10R and 20R, respectively).

TABLE 4.1
Condition numbers κ, optimal values v^*, and values $g(\bar{x})$ of optimal solutions $\bar{x}$ of expected value problems $\max_x G(x, E[W])$, for instances presented.

Instance	Condition number κ	Optimal value v^*	Expected value $g(\bar{x})$
10D	107000	42.7	26.2
10R	410	46.3	28.2
20D	954000	96.5	75.9
20R	233	130.3	109.0

Table 4.1 shows the condition numbers, the optimal values v^*, and the values $g(\bar{x})$ of the optimal solutions $\bar{x}$ of the associated expected value problems $\max_x G(x, E[W])$ for the four instances.

For all instances of the SSKP, the size variables W_i are independent normally distributed, for ease of evaluation of the results produced by the SAA method, as described in the next paragraph. For the randomly generated instances, the rewards r_i were generated from the uniform $(10, 20)$ distribution, the mean sizes μ_i were generated from the uniform $(20, 30)$ distribution, and the size standard deviations σ_i were generated from the uniform $(5, 15)$ distribution. For all instances, the per unit penalty $c = 4$.

If $W_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, k$, are independent normally distributed random variables, then the objective function of (4.1) can be written in closed form. That is, the random variable $Z(x) := \sum_{i=1}^{k} W_i x_i - q$ is normally distributed with mean $\mu(x) = \sum_{i=1}^{k} \mu_i x_i - q$ and variance $\sigma(x)^2 = \sum_{i=1}^{k} \sigma_i^2 x_i^2$. It is also easy to show, since $Z(x) \sim N(\mu(x), \sigma(x)^2)$, that

$$\mathbb{E}[Z(x)]^+ \;=\; \mu(x)\Phi\left(\frac{\mu(x)}{\sigma(x)}\right) + \frac{\sigma(x)}{\sqrt{2\pi}} \exp\left(\frac{-\mu(x)^2}{2\sigma(x)^2}\right),$$

where Φ denotes the standard normal cumulative distribution function. Thus, it follows that

$$g(x) \;=\; \sum_{i=1}^{k} r_i x_i - c\left[\mu(x)\Phi\left(\frac{\mu(x)}{\sigma(x)}\right) + \frac{\sigma(x)}{\sqrt{2\pi}} \exp\left(-\frac{\mu(x)^2}{2\sigma(x)^2}\right)\right]. \tag{4.3}$$

The benefit of such a closed form expression is that the objective value $g(x)$ can be computed quickly and accurately, which is useful for solving small instances of the problem by enumeration or branch and bound (cf. Cohn and Barnhart [4]) and for evaluation of solutions produced by the SAA Algorithm. Good numerical approximations are available for computing $\Phi(x)$, such as *Applied Statistics* Algorithm AS66 (Hill [11]). The SAA Algorithm was executed without the benefit of a closed form expression for $g(x)$, as would be the case for most probability distributions; (4.3) was used only to evaluate the solutions produced by the SAA Algorithm.

The first numerical experiment was conducted to observe how the exponential convergence rate established in Proposition 2.2 applies in the case of the SSKP, and to investigate how the convergence rate is affected by the number of decision variables and the condition number κ. Figures 4.1 and 4.2 show the estimated probability that an SAA optimal solution $\hat{x}_N$ has objective value $g(\hat{x}_N)$ within relative tolerance d of the optimal value v^*, i.e., $\hat{P}[v^* - g(\hat{x}_N) \leq d\, v^*]$, as a function of the sample size N, for different values of d. The experiment was conducted by generating $M = 1000$ independent SAA replications for each sample size N, computing SAA optimal

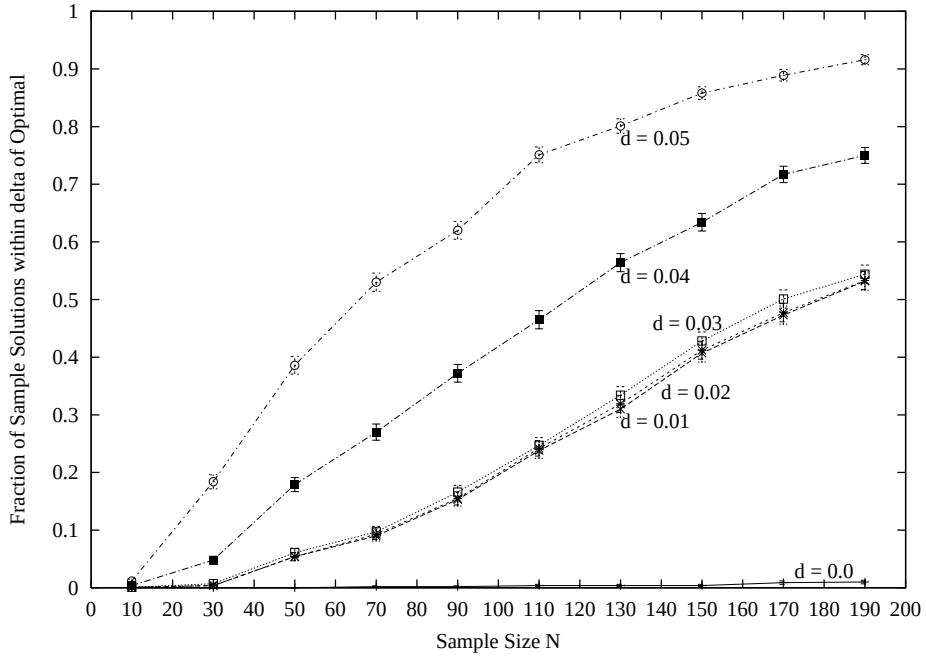

FIG. 4.1. *Probability of* SAA *optimal solution* $\hat{x}_N$ *having objective value* $g(\hat{x}_N)$ *within relative tolerance* d *of the optimal value* v^*, $\hat{P}[v^* - g(\hat{x}_N) \leq d\,v^*]$, *as a function of sample size* N *for different values of* d, *for instance* 20D.

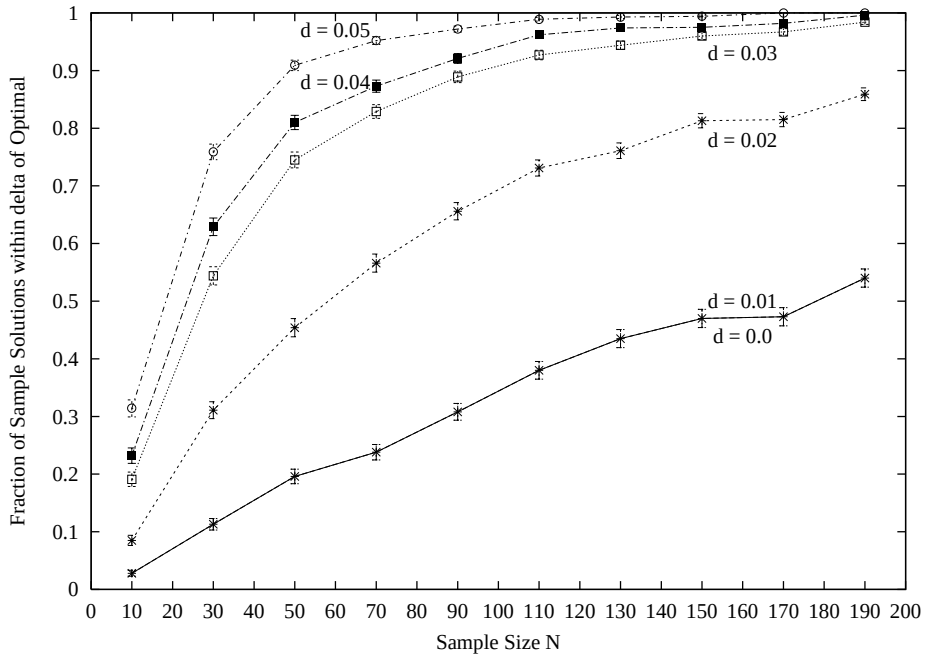

FIG. 4.2. *Probability of* SAA *optimal solution* $\hat{x}_N$ *having objective value* $g(\hat{x}_N)$ *within relative tolerance* d *of the optimal value* v^*, $\hat{P}[v^* - g(\hat{x}_N) \leq d\,v^*]$, *as a function of sample size* N *for different values of* d, *for instance* 20R.

solutions $\hat{x}_N^m$, $m = 1, \ldots, M$, and their objective values $g(\hat{x}_N^m)$ using (4.3), and then counting the number M_d of times that $v^* - g(\hat{x}_N^m) \leq d\, v^*$. Then the probability was estimated by $\hat{P}[v^* - g(\hat{x}_N) \leq d\, v^*] = M_d/M$, and the variance of this estimator was estimated by

$$\widehat{\mathrm{Var}}[\hat{P}] \;=\; \frac{M_d(1 - M_d/M)}{M(M-1)}.$$

The figures also show error bars of length $2(\widehat{\mathrm{Var}}[\hat{P}])^{1/2}$ on each side of the point estimate M_d/M.

One noticeable effect is that the probability that an SAA replication generates an optimal solution ($d = 0$) increases much more slowly with increase in the sample size N for the harder instances (10D and 20D) with poor condition numbers κ than for the randomly generated instances with better condition numbers. However, the probability that an SAA replication generates a reasonably good solution (e.g., $d = 0.05$) increases quite quickly with increase in the sample size N for both the harder instances and for the randomly generated instances.

The second numerical experiment demonstrates how the objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ change as the sample size N increases, and how this change is affected by the number of decision variables and the condition number κ. In this experiment, the maximum number of SAA replications with the same sample size N was chosen as $M' = 50$. Additionally, after $M'' = 20$ replications with the same sample size N, the variance $S^2_{M''}$ of $\hat{v}_N^m$ was computed as in (3.2), because it is an important term in the optimality gap estimator (3.3). If $S^2_{M''}$ was too large, it indicated that the optimality gap estimate would be too large and that the sample size N should be increased. Otherwise, if $S^2_{M''}$ was not too large, then SAA replications were performed with the same sample size N until M' SAA replications had occurred. If the optimality gap estimate was greater than a specified tolerance, then the sample size N was increased and the procedure was repeated. Otherwise, if the optimality gap estimate was less than a specified tolerance, then a screening and selection procedure was applied to all the candidate solutions $\hat{x}_N^m$ generated, and the best solution among these was chosen.

Figures 4.3 and 4.4 show the objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ produced during the course of the algorithm. There were several noticeable effects. First, good and often optimal solutions were produced early in the execution of the algorithm, but the sample size N had to be increased several times thereafter before the optimality gap estimate became sufficiently small for stopping, without any improvement in the quality of the generated solutions. Second, for the randomly generated instances a larger proportion of the SAA optimal solutions $\hat{x}_N^m$ were optimal or had objective values close to optimal, and optimal solutions were produced with smaller sample sizes N than were required for the harder instances. For example, for the harder instance with 10 decision variables (instance 10D), the optimal solution was first produced after $m = 6$ replications with sample size $N = 120$; and for instance 10R, the optimal solution was first produced after $m = 2$ replications with sample size $N = 20$. Also, for the harder instance with 20 decision variables (instance 20D), the optimal solution was not produced in any of the 270 total number of replications (but the second-best solution was produced 3 times); and for instance 20R, the optimal solution was first produced after $m = 15$ replications with sample size $N = 50$. Third,

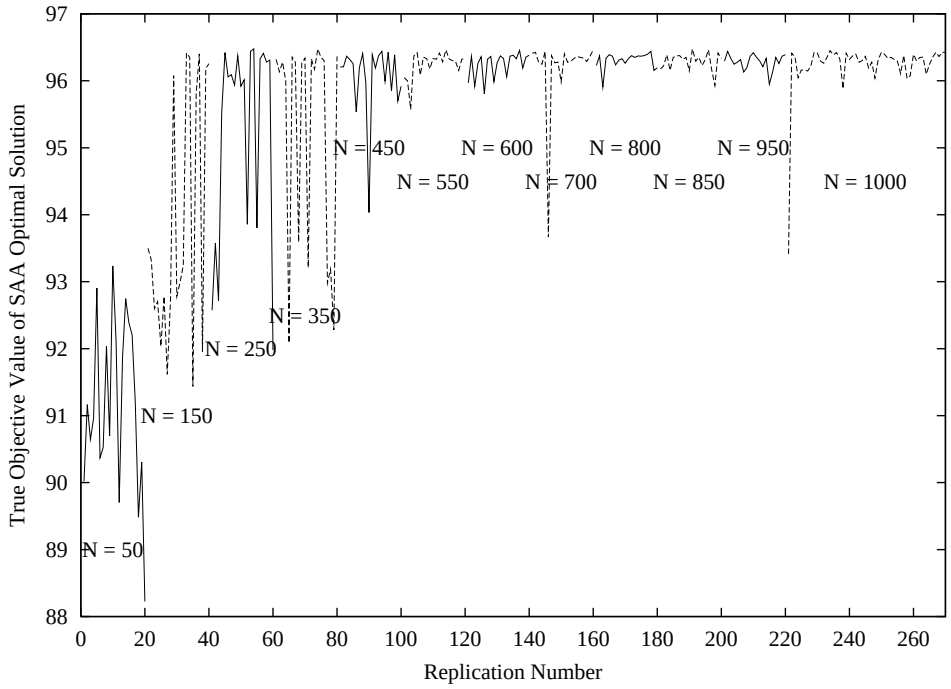

FIG. 4.3. *Improvement of objective values* $g(\hat{x}_N^m)$ *of* SAA *optimal solutions* $\hat{x}_N^m$ *as the sample size* N *increases, for instance* 20D.

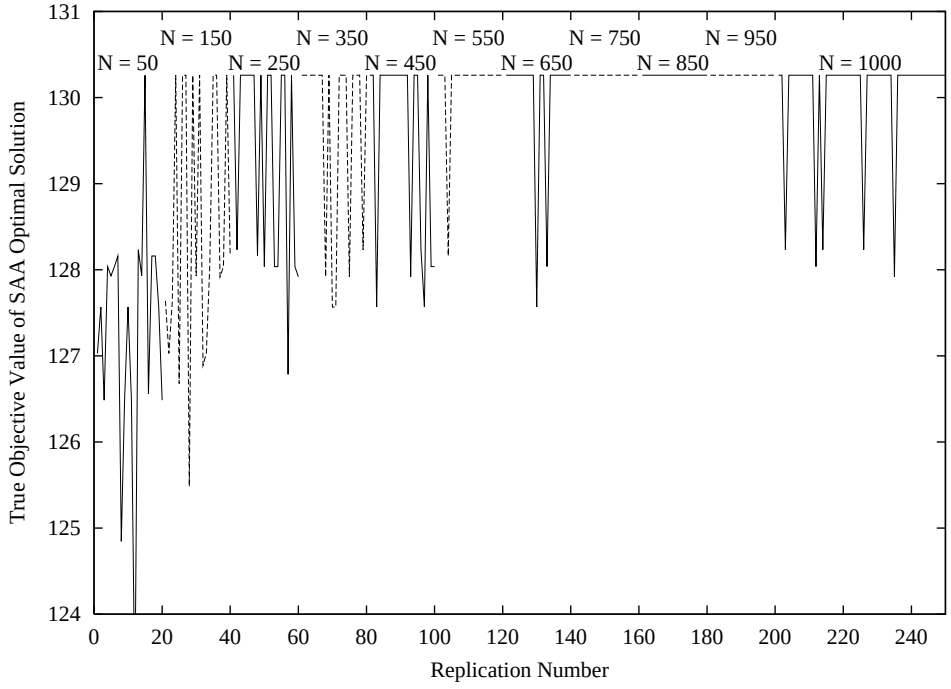

FIG. 4.4. *Improvement of objective values* $g(\hat{x}_N^m)$ *of* SAA *optimal solutions* $\hat{x}_N^m$ *as the sample size* N *increases, for instance* 20R.

for each of the instances, the expected value problem $\max_x G(x, E[W])$ was solved, with its optimal solution denoted by $\bar{x}$. The objective value $g(\bar{x})$ of each $\bar{x}$ is shown in Table 4.1. It is interesting to note that even with small sample sizes N, every solution $\hat{x}_N^m$ produced had a better objective value $g(\hat{x}_N^m)$ than $g(\bar{x})$.

As mentioned above, in the second numerical experiment it was noticed that often the optimality gap estimate is large, even if an optimal solution has been found, i.e., $v^* - g(\hat{x}) = 0$. (This is also a common problem in deterministic discrete optimization.) Consider the components of the optimality gap estimator $\hat{g}_{N'}(\hat{x}) - \bar{v}_N^M$ given in (3.3). The first component $g(\hat{x}) - \hat{g}_{N'}(\hat{x})$ can be made small with relatively little computational effort by choosing N' sufficiently large. The second component, the true optimality gap $v^* - g(\hat{x})$, is often small after only a few replications m with a small sample size N. The fourth component $z_\alpha(S_{N'}^2(\hat{x})/N' + S_M^2/M)^{1/2}$ can also be made small with relatively little computational effort by choosing N' and M sufficiently large. The major part of the problem seems to be caused by the third term $\bar{v}_N^M - v^*$ and by the fact that $\mathbb{E}[\bar{v}_N^M] - v^* \geq 0$, as identified in (3.1). It was also mentioned that the bias $\mathbb{E}[\bar{v}_N^M] - v^*$ decreases as the sample size N increases. However, the second numerical experiment indicated that a significant bias can persist even if the sample size N is increased far beyond the sample size needed for the SAA method to produce an optimal solution.

The third numerical experiment demonstrates the effect of the number of decision variables and the condition number κ on the bias in the optimality gap estimator. Figures 4.5 and 4.6 show how the relative bias $\bar{v}_N^M/v^*$ of the optimality gap estimate changes as the sample size N increases, for different instances. The most noticeable effect is that the bias decreases much more slowly for the harder instances than for the randomly generated instances as the sample size N increases. This is in accordance with the asymptotic result (2.31) of Proposition 2.4.

Two estimators of the optimality gap $v^* - g(\hat{x})$ were discussed in section 3.3, namely, $\bar{v}_N^M - \hat{g}_{N'}(\hat{x})$ and $\bar{v}_N^M - \bar{g}_N^M(\hat{x})$. It was mentioned that the second estimator may have smaller variance than the first, especially if there is positive correlation between $\hat{g}_N^m(\hat{x})$ and $\hat{v}_N^m$. It was also pointed out that the second estimator requires additional computational effort, because after $\hat{x}$ is produced by solving the SAA problem for one sample, the second estimator requires the computation of $\hat{g}_N^m(\hat{x})$ for all the remaining samples $m = 1, \ldots, M$. The fourth numerical experiment compares the optimality gap estimates and their variances. Sample sizes of $N = 50$ and $N' = 2000$ were used, and $M = 50$ replications were performed.

Table 4.2 shows the optimality gap estimates $\bar{v}_N^M - \hat{g}_{N'}(\hat{x})$ and $\bar{v}_N^M - \bar{g}_N^M(\hat{x})$, with their variances $\widehat{\mathrm{Var}}[\bar{v}_N^M - \hat{g}_{N'}(\hat{x})] = S_{N'}^2(\hat{x})/N' + S_M^2/M$ and $\widehat{\mathrm{Var}}[\bar{v}_N^M - \bar{g}_N^M(\hat{x})] = \bar{S}_M^2/M$, respectively; the correlation $\widehat{\mathrm{Cor}}[\bar{v}_N^M, \bar{g}_N^M(\hat{x})]$; and the computation times of the gap estimates. In each case, the bias $\bar{v}_N^M - v^*$ formed the major part of the optimality gap estimate; the standard deviations of the gap estimators were small compared with the bias. There was positive correlation between $\bar{v}_N^M$ and $\bar{g}_N^M(\hat{x})$, and the second gap estimator had smaller variances, but this benefit is obtained at the expense of relatively large additional computational effort.

In section 2.2, an estimate $N \approx 3\sigma_{\max}^2 \log(|\mathcal{S}|/\alpha)/(\varepsilon - \delta)^2$ of the required sample size was derived. For the instances presented here, using $\varepsilon = 0.5$, $\delta = 0$, and $\alpha = 0.01$, these estimates were of the order of 10^6 and thus much larger than the sample sizes that were actually required for the specified accuracy. The sample size estimates using $\sigma_{\max}^2$ were smaller than the sample size estimates using $\max_{x \in \mathcal{S}} \mathrm{Var}[G(x, W)]$ by a factor of approximately 10.

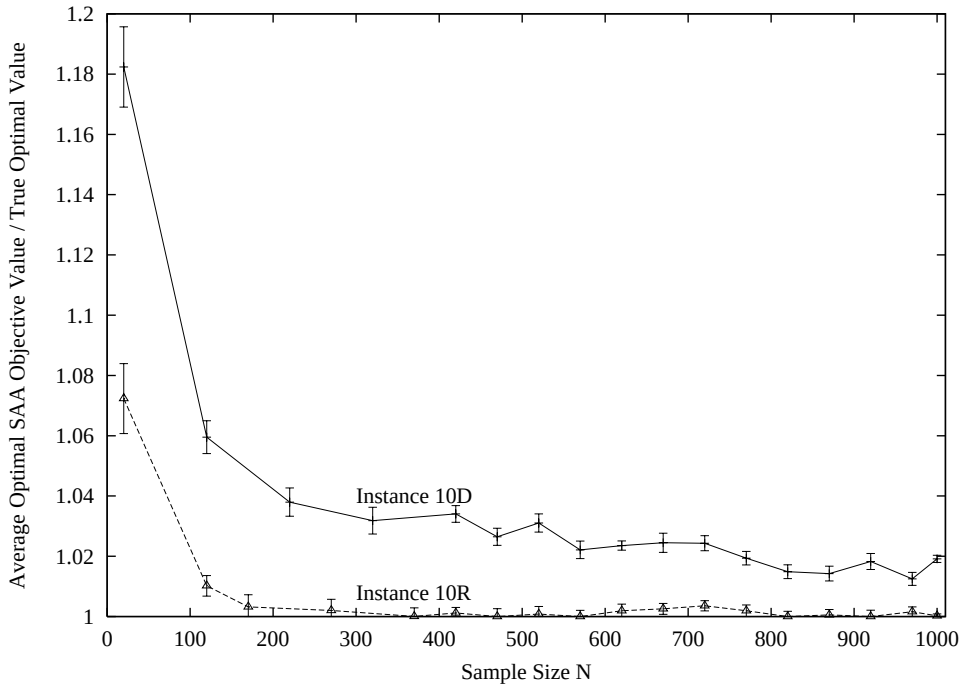

FIG. 4.5. *Relative bias* $\bar{v}_N^M/v^*$ *of the optimality gap estimator as a function of the sample size* N*, for instances* 10D *and* 10R*, with* 10 *decision variables.*

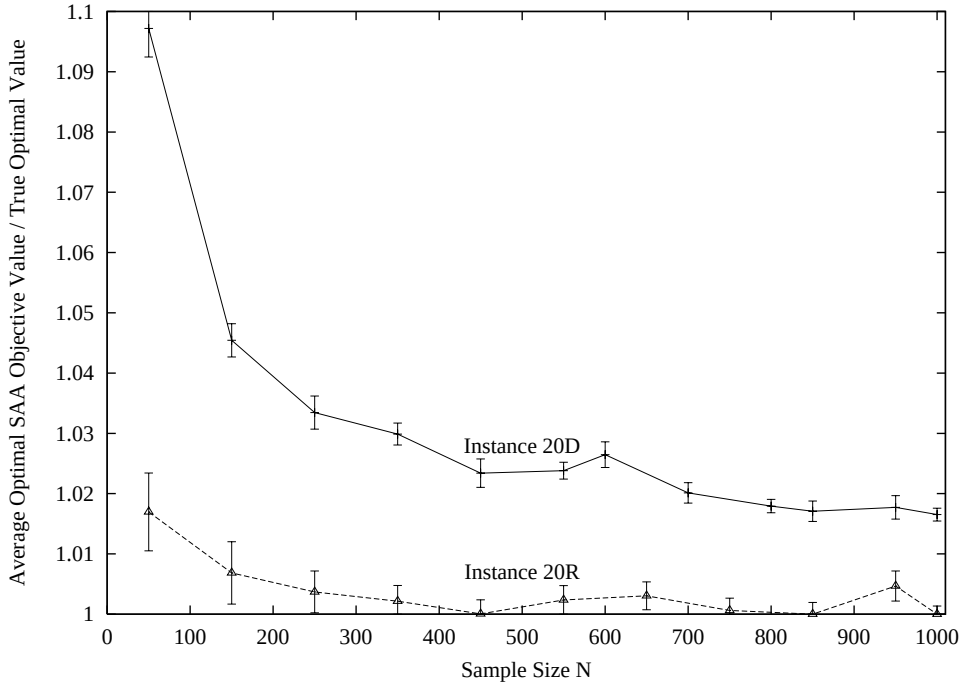

FIG. 4.6. *Relative bias* $\bar{v}_N^M/v^*$ *of the optimality gap estimate as a function of the sample size* N*, for instances* 20D *and* 20R*, with* 20 *decision variables.*

TABLE 4.2
Optimality gap estimates $\bar{v}_N^M - \hat{g}_{N'}(\hat{x})$ and $\bar{v}_N^M - \bar{g}_N^M(\hat{x})$, with their variances and computation times.

Instance	Opt. gap $v^* - g(\hat{x})$	Estimate $\bar{v}_N^M - \hat{g}_{N'}(\hat{x})$	$\widehat{\mathrm{Var}}[\bar{v}_N^M - \hat{g}_{N'}(\hat{x})]$ $= S^2_{N'}(\hat{x})/N' + S^2_M/M$	CPU time
10D	0	3.46	0.200	0.02
10R	0	1.14	0.115	0.01
20D	0.148	8.46	0.649	0.02
20R	0	3.34	1.06	0.02

Instance	Opt. gap $v^* - g(\hat{x})$	Estimate $\bar{v}_N^M - \bar{g}_N^M(\hat{x})$	$\widehat{\mathrm{Var}}[\bar{v}_N^M - \bar{g}_N^M(\hat{x})]$ $= \bar{S}^2_M/M$	Correlation $\widehat{\mathrm{Cor}}[\bar{v}_N^M, \bar{g}_N^M(\hat{x})]$	CPU time
10D	0	3.72	0.121	0.203	0.24
10R	0	1.29	0.035	0.438	0.24
20D	0.148	9.80	0.434	0.726	0.49
20R	0	3.36	0.166	0.844	0.47

Several variance reduction techniques can be used. Compared with simple random sampling, Latin hypercube sampling reduced the variances by factors varying between 1.02 and 2.9 and increased the computation time by a factor of approximately 1.2. Also, to estimate $g(x)$ for any solution $x \in \mathcal{S}$, it is natural to use $\sum_{i=1}^k W_i x_i$ as a control variate, because $\sum_{i=1}^k W_i x_i$ should be correlated with $[\sum_{i=1}^k W_i x_i - q]^+$, and the mean of $\sum_{i=1}^k W_i x_i$ is easy to compute. Using this control variate reduced the variances of the estimators of $g(x)$ by factors between 2.0 and 3.0 and increased the computation time by a factor of approximately 2.0.

5. Conclusion. We proposed a sample average approximation method for solving stochastic discrete optimization problems, and we studied some theoretical as well as practical issues important for the performance of this method. It was shown that the probability that a replication of the SAA method produces an optimal solution increases at an exponential rate in the sample size N. It was found that this convergence rate depends on the conditioning of the problem, which in turn tends to become poorer with an increase in the number of decision variables. It was also shown that the sample size required for a specified accuracy increases proportional to the logarithm of the number of feasible solutions. It was found that for many instances the SAA method produces good and often optimal solutions with only a few replications and a small sample size. However, the optimality gap estimator considered here was in each case too weak to indicate that a good solution had been found. Consequently the sample size had to be increased substantially before the optimality gap estimator indicated that the solutions were good. Thus, a more efficient optimality gap estimator can make a substantial contribution toward improving the performance guarantees of the SAA method during execution of the algorithm. The SAA method has the advantage of ease of use in combination with existing techniques for solving deterministic optimization problems.

The proposed method involves solving several replications of the SAA problem (2.1), and possibly increasing the sample size several times. An important issue is the behavior of the computational complexity of the SAA problem (2.1) as a function of the sample size. Current research aims at investigating this behavior for particular classes of problems.

REFERENCES

[1] M. H. Alrefaei and S. Andradóttir, *A simulated annealing algorithm with constant temperature for discrete stochastic optimization*, Management Science, 45 (1999), pp. 748–764.

[2] R. E. Bechhofer, T. J. Santner, and D. M. Goldsman, *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons*, John Wiley, New York, NY, 1995.

[3] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, Springer Ser. Oper. Res., Springer-Verlag, New York, NY, 1997.

[4] A. Cohn and C. Barnhart, *The stochastic knapsack problem with random weights: A heuristic approach to robust transportation planning*, in Proceedings of the Triennial Symposium on Transportation Analysis (TRISTAN III), San Juan, PR, 1998.

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

[6] B. L. Fox and G. W. Heine, *Probabilistic search with overrides*, Ann. Appl. Probab., 5 (1995), pp. 1087–1094.

[7] A. Futschik and G. C. Pflug, *Confidence sets for discrete stochastic optimization*, Ann. Oper. Res., 56 (1995), pp. 95–108.

[8] A. Futschik and G. C. Pflug, *Optimal allocation of simulation experiments in discrete stochastic optimization and approximative algorithms*, European J. Oper. Res., 101 (1997), pp. 245–260.

[9] S. B. Gelfand and S. K. Mitter, *Simulated annealing with noisy or imprecise energy measurements*, J. Optim. Theory Appl., 62 (1989), pp. 49–62.

[10] W. Gutjahr and G. C. Pflug, *Simulated annealing for noisy cost functions*, J. Global Optim., 8 (1996), pp. 1–13.

[11] I. D. Hill, *Algorithm AS66: The normal integral*, Applied Statistics, 22 (1973), pp. 424–427.

[12] Y. Hochberg and A. Tamhane, *Multiple Comparison Procedures*, John Wiley, New York, NY, 1987.

[13] T. Homem-de-Mello, *Variable-Sample Methods and Simulated Annealing for Discrete Stochastic Optimization*, manuscript, Department of Industrial, Welding and Systems Engineering, The Ohio State University, Columbus, OH, 1999.

[14] T. Homem-de-Mello, *Monte Carlo methods for discrete stochastic optimization*, in Stochastic Optimization: Algorithms and Applications, S. Uryasev and P. M. Pardalos, eds., Kluwer Academic Publishers, Norwell, MA, 2000, pp. 95–117.

[15] W. K. Mak, D. P. Morton, and R. K. Wood, *Monte Carlo bounding techniques for determining solution quality in stochastic programs*, Oper. Res. Lett., 24 (1999), pp. 47–56.

[16] D. P. Morton and R. K. Wood, *On a stochastic knapsack problem and generalizations*, in Advances in Computational and Stochastic Optimization, Logic Programming, and Heuristic Search: Interfaces in Computer Science and Operations Research, D. L. Woodruff, ed., Kluwer Academic Publishers, Dordrecht, the Netherlands, 1998, pp. 149–168.

[17] B. L. Nelson, J. Swann, D. M. Goldsman, and W. Song, *Simple procedures for selecting the best simulated system when the number of alternatives is large,* Oper. Res., to appear.

[18] V. I. Norkin, Y. M. Ermoliev, and A. Ruszczyński, *On optimal allocation of indivisibles under uncertainty*, Oper. Res., 46 (1998), pp. 381–395.

[19] V. I. Norkin, G. C. Pflug, and A. Ruszczyński, *A branch and bound method for stochastic global optimization*, Math. Programming, 83 (1998), pp. 425–450.

[20] R. Schultz, L. Stougie, and M. H. Van der Vlerk, *Solving stochastic programs with integer recourse by enumeration: A framework using Gröbner basis reductions*, Math. Programming, 83 (1998), pp. 229–252.

[21] A. Shapiro, *Asymptotic analysis of stochastic programs*, Ann. Oper. Res., 30 (1991), pp. 169–186.

[22] A. Shapiro, T. Homem-de-Mello, and J. C. Kim, *Conditioning of Convex Piecewise Linear Stochastic Programs*, manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 2000.