THE SAMPLE AVERAGE APPROXIMATION METHOD FOR STOCHASTIC DISCRETE OPTIMIZATION

ANTON J. KLEYWEGT[†‡] AND ALEXANDER SHAPIRO[†§]

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 consequently 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 and stopping rules of this procedure and present a numerical example of 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

1. Introduction.

In this paper we consider optimization problems of the form

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

Here W is a random vector having probability distribution P, $G(x,w)$ is a real valued function, and $\mathcal{S}$ is a *finite* set, for example $\mathcal{S}$ can be a finite subset of $\mathbb{R}^k$ with integer coordinates. 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 P-measurable and $\mathbb{E}_P\{|G(x,W)|\} < \infty$. We are particularly interested in problems for which the expected value function $g(x) \equiv \mathbb{E}_P G(x,W)$ cannot be written in a closed form and/or its values cannot be easily calculated, while $G(x,w)$ is easily computable for given x and w.

It is well known that many discrete optimization problems are hard to solve. Here on top of this we have additional difficulties since 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. A discussion of two stage stochastic integer programming problems with recourse can be found in Birge and Louveaux [2]. A branch and bound approach to solving stochastic integer programming problems was suggested by Norkin, Pflug and Ruszczynski [9]. Schultz, Stougie, and Van der Vlerk [10] 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 consequently 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 [7].

[†] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205.

[‡] Supported by the National Science Foundation under grant DMI-9875400.

[§] Supported by the National Science Foundation under grant DMI-9713878.

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 one 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 $W^1, ..., W^N$ be an i.i.d. random sample of N realizations of the random vector W. Consider the sample average function

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

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,

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

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.

In the following proposition we show 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. By the strong Law of Large Numbers we have 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, w.p.1,

$$\delta_N \equiv \max_{x\in\mathcal{S}} |\hat{g}_N(x) - g(x)| \to 0 \quad \text{as} \quad N \to \infty. \tag{2.2}$$

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

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

$$\alpha(\varepsilon) \;\equiv\; \min_{x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon} g(x) - v^* - \varepsilon. \tag{2.3}$$

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 $\alpha(\varepsilon) > 0$.

Let N be large enough such that $\delta_N < \alpha(\varepsilon)/2$. Then $\hat{v}_N < v^* + \alpha(\varepsilon)/2$, and for any $x \in \mathcal{S} \setminus \mathcal{S}^\varepsilon$ it holds that $\hat{g}_N(x) > v^* + \varepsilon + \alpha(\varepsilon)/2$. 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 $\hat{\mathcal{S}}^\varepsilon_N$. The inclusion $\hat{\mathcal{S}}^\varepsilon_N \subset \mathcal{S}^\varepsilon$ follows, which completes the proof. □

It follows that if, for some $\varepsilon \ge 0$, $\mathcal{S}^\varepsilon = \{x^*\}$ is a singleton, then w.p.1, $\hat{\mathcal{S}}^\varepsilon_N = \{x^*\}$ 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 $\hat{x}_N = x^*$.

In the next section, and in section 4, it is demonstrated that $\alpha(\varepsilon)$, defined in (2.3), is an important measure of the well-conditioning of a stochastic discrete optimization problem.

2.2. Convergence Rates.

The above results do not say anything about the rates of convergence of $\hat{v}_N$ and $\hat{\mathcal{S}}^\varepsilon_N$ 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, the probability of the event $\{\hat{\mathcal{S}}^\varepsilon_N \subset \mathcal{S}^\varepsilon\}$ approaches one exponentially fast as $N \to \infty$. Next we briefly outline some background of the LD theory.

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

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

where $M(t) \equiv \mathbb{E}\{e^{tX}\}$ is the moment generating function of X. By taking the logarithm of both sides of the above inequality, changing variables $t' \equiv t/N$ and minimizing over $t' > 0$, we obtain

$$\frac{1}{N} \log\left[P(Z_N \ge a)\right] \;\le\; -I(a), \tag{2.4}$$

where

$$I(z) \;\equiv\; \sup_{t \ge 0}\{tz - \Lambda(t)\}$$

is the conjugate of the logarithmic moment generating function $\Lambda(t) \equiv \log M(t)$. In LD theory, $I(z)$ is called the large deviations rate function, and the inequality (2.4) corresponds to the upper bound of Cramér's LD theorem.

Although we do not need this in the subsequent analysis, it could be mentioned that the constant $I(a)$ in (2.4) gives, in a sense, the best possible exponential rate at which the probability $P(Z_N \ge a)$ converges to zero. This follows from the corresponding lower bound of Cramér's LD theorem. For a thorough discussion of the LD theory, an interested reader is referred to Dembo and Zeitouni [4].

The rate function $I(z)$ has the following properties. Suppose that the random variable X has mean μ. Then the function $I(z)$ is convex, attains its minimum at $z = \mu$, and $I(\mu) = 0$. Moreover, suppose that the moment generating function $M(t)$, of X, is finite valued for all t in a neighborhood of $t = 0$. Then 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 $\psi(t) \equiv ta - \Lambda(t)$ at $t = 0$ is greater than zero, and hence $\psi(t) > 0$ for $t > 0$ small enough. It follows that in that case $I(a) > 0$.

Now we return to the problems (1.1) and (2.1). Consider $\varepsilon \geq 0$ and the numbers δ_{N} and $\alpha(\varepsilon)$ defined in (2.2) and (2.3), respectively. Then it holds that if $\delta_{N} < \alpha(\varepsilon)/2$, then $\{\hat{\mathcal{S}}^{\varepsilon}_{N} \subset \mathcal{S}^{\varepsilon}\}$. Since the complement of the event $\{\delta_{N} < \alpha(\varepsilon)/2\}$ is given by the union of the events $|\hat{g}_{N}(x) - g(x)| \geq \alpha(\varepsilon)/2$ over all $x \in \mathcal{S}$, and the probability of that union is less than or equal to the sum of the corresponding probabilities, it follows that

$$1 - P\left(\hat{\mathcal{S}}^{\varepsilon}_{N} \subset \mathcal{S}^{\varepsilon}\right) \;\leq\; \sum_{x \in \mathcal{S}} P\left\{|\hat{g}_{N}(x) - g(x)| \geq \alpha(\varepsilon)/2\right\}.$$

We make the following assumption.

Assumption A For any $x \in \mathcal{S}$, the moment generating function $M(t)$ of the random variable $G(x, W)$ is finite valued in a neighborhood of $t = 0$.

Under Assumption A, it follows from the LD upper bound (2.4) that for any $x \in \mathcal{S}$ there are constants $\gamma_x > 0$ and $\gamma'_x > 0$ such that

$$P\left\{|\hat{g}_{N}(x) - g(x)| \geq \alpha(\varepsilon)/2\right\} \;\leq\; e^{-N\gamma_x} + e^{-N\gamma'_x}.$$

Namely, the constants γ_x and γ'_x are given by values of the rate functions of $G(x, W)$ and $-G(x, W)$ at $g(x) + \alpha(\varepsilon)/2$ and $-g(x) + \alpha(\varepsilon)/2$, respectively. Since the set $\mathcal{S}$ is finite, by taking $\gamma \equiv \min_{x \in \mathcal{S}}\{\gamma_x, \gamma'_x\}$, the following result is obtained (it is similar to an asymptotic result for piecewise linear continuous problems derived in [12]).

Proposition 2.2. *Suppose that Assumption A holds. Then there exists a constant $\gamma > 0$ such that the following inequality holds:*

$$\limsup_{N \to \infty} \frac{1}{N} \log\left[1 - P(\hat{\mathcal{S}}^{\varepsilon}_{N} \subset \mathcal{S}^{\varepsilon})\right] \;\leq\; -\gamma. \tag{2.5}$$

The inequality (2.5) means that the probability of the event $\{\hat{\mathcal{S}}^{\varepsilon}_{N} \subset \mathcal{S}^{\varepsilon}\}$ approaches one exponentially fast as $N \to \infty$. Unfortunately it appears that the corresponding constant γ, giving the exponential rate of convergence, cannot be calculated (or even estimated) a priori, i.e., before the problem is solved. Therefore the above result is more of theoretical value. Let us mention at this point that the above constant γ depends, through the corresponding rate functions, on the number $\alpha(\varepsilon)$. Clearly, if $\alpha(\varepsilon)$ is "small", then an accurate approximation would be required in order to find an ε-optimal solution of the true problem. Therefore, in a sense, $\alpha(\varepsilon)$ characterizes a well conditioning of the set $\mathcal{S}^{\varepsilon}$.

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 $\mathcal{S}' = \mathcal{S}^*$ we obtain 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. Mak, Morton, and Wood [6]).

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 we have 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 $\sqrt{N}$ we obtain 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

$$\lim_{N\to\infty} \sqrt{N}\left[\hat{v}_N - \min_{x\in\mathcal{S}^*} \hat{g}_N(x)\right] = 0 \quad w.p.1. \tag{2.6}$$

Since convergence w.p.1 implies convergence in probability, it follows from (2.6) 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

$$\sigma^2(x) \;\equiv\; \mathrm{Var}\{G(x,W)\} \tag{2.7}$$

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 $Y(x)$ with zero mean and variance $\sigma^2(x)$. Moreover, again by the CLT, random variables $Y(x)$ have the same autocovariance function as $G(x,W)$, i.e., the covariance between $Y(x)$ and $Y(x')$ 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 [11]). We use "⇒" to denote convergence in distribution.

Proposition 2.3. *Suppose that variances $\sigma^2(x)$, defined in* (2.7), *exist for every $x \in \mathcal{S}^*$. Then*

$$\sqrt{N}(\hat{v}_N - v^*) \;\Rightarrow\; \min_{x\in\mathcal{S}^*} Y(x), \tag{2.8}$$

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

$$\sqrt{N}(\hat{v}_N - v^*) \;\Rightarrow\; N(0, \sigma^2(x^*)). \tag{2.9}$$

3. Algorithm Design.

In the previous section we established a number of convergence results for the sample average approximation method. The results describe how the optimal value $\hat{v}_N$ and the ε-optimal solutions set $\hat{\mathcal{S}}_N^\varepsilon$ 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. Of course, in a real application the sample size N cannot go to infinity. A finite sample size N 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. To choose N, several trade-offs should be taken into account. We have that the objective value and an optimal solution of the SAA problem tend to be better for larger N, 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 motivates 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. We 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 one as $N \to \infty$, and moreover by Proposition 2.2 it tends to one exponentially fast if Assumption A holds. However, for a finite N the probability p can be small or even zero. We have that after M trials the probability of producing an exact optimal solution of the true problem at least once is $1-(1-p)^M$, and that 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 a required sample size N is problem specific, does not depend in general on the number of feasible solutions, and can be arbitrarily large.

Example. Suppose that $\mathcal{S} \equiv \{-1,0,1\}$, that W can take two values w_1 and w_2 with respective probabilities $1-\gamma$ and γ, and that $G(-1,w_1) \equiv -1$, $G(0,w_1) \equiv 0$, $G(1,w_1) \equiv 2$, and $G(-1,w_2) \equiv 2k$, $G(0,w_2) \equiv 0$, $G(1,w_2) \equiv -k$, 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. □

Another issue that has to be addressed, is the choice of the number M of replications. 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 simplicity of presentation, suppose 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 m-th SAA replication (trial). Some candidate solutions are better than others. Using the larger sample with size N', an accurate estimate $\hat{g}_{N'}(\hat{x}_N^m)$ of the objective value $g(\hat{x}_N^m)$ of each candidate solution $\hat{x}_N^m$ is obtained. The estimate $\hat{g}_{N'}(\hat{x}_N^m)$ can be compared with the objective value estimates $\hat{g}_{N'}(\hat{x}_N^{m'})$ of previously produced candidate solutions $\hat{x}_N^{m'}$, $m' < m$, to determine if the newest solution $\hat{x}_N^m$ appears to be better than all previously produced candidate solutions $\hat{x}_N^{m'}$. If for several successive SAA replications the candidate solutions are worse than the best candidate solution produced so far, it indicates that another SAA replication is not very likely to produce a better solution, using the same sample size N. At that stage it seems that the best recourse for the algorithm is to either increase the sample size N, or to accept the best solution found so far. This decision is discussed in more detail later.

Thus a relevant question is how many successive SAA replications should be performed without improvement in the best candidate solution found so far, before the decision is made to either increase the sample size N or to stop. The following decision rule may provide some guidelines. Suppose that m successive SAA replications have been performed without improvement in the best candidate solution found so far. One can compute a Bayesian estimate of the probability ρ that another SAA replication will find a better candidate solution. Since one does not have any prior information about ρ, it makes sense to assume that the prior distribution of ρ is uniform on the interval $[0,1]$. Let Z denote the number of SAA replications until a better candidate solution than the best candidate solution found so far is produced. Then by Bayes' formula we have

$$\begin{aligned} P(\rho \in dx \mid Z > m) &= \frac{P(Z > m \mid \rho \in dx)\, dx}{\int_0^1 P(Z > m \mid \rho \in dx)\, dx} \\ &= \frac{(1-x)^m dx}{\int_0^1 (1-x)^m dx} \quad = \quad (m+1)(1-x)^m dx. \end{aligned}$$

That is, the posterior probability density function of ρ is $\pi(x \mid m) = (m+1)(1-x)^m$. For example, with $m = 50$ we have that $\rho \leq 0.057$ with probability 0.95.

3.3. Performance Bounds. If additional replications are not likely to lead to a better solution, a decision should be made to either increase the sample size or to stop. To assist in making this decision, as well as for other performance evaluation purposes, it is useful to evaluate the quality of a solution $\hat{x} \in \mathcal{S}$, not only relative to other candidate solutions, but also relative to the optimal value v^*. To do this, we would like to compute the optimality gap $g(\hat{x}) - v^*$. Unfortunately, the very reason for the approach described in this paper is that both terms of the optimality gap are hard to compute. As before,

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

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

An estimator of v^* is given by

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

where $\hat{v}_N^m$ denotes the optimal objective value of the m-th 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 2.3 indicates that this bias tends to be bigger for problems with larger sets $\mathcal{S}^*$ 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

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

we have that on average the above estimator overestimates the optimality gap $g(\hat{x}) - v^*$. Norkin, Pflug, and Ruszczyński [9] and Mak, Morton, and Wood [6] showed 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

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

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 $S^2_{N'}(\hat{x})/N' + S^2_M/M$.

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

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

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

$$\hat{g}_N^m(\hat{x}) \;\equiv\; \frac{1}{N}\sum_{n=1}^{N} G(\hat{x}, W^{mn}).$$

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

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

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 the sample sizes N, N', and M. For many applications, one would expect positive correlation between $\hat{g}^m_{N}(\hat{x})$ and $\hat{v}^m_{N}$. The additional computational effort to compute $\hat{g}^m_{N}(\hat{x})$ for $m = 1, \ldots, M$, should also be taken into account when evaluating any such variance reduction. Either way, the Central Limit Theorem can be applied to the optimality gap estimators $\hat{g}_{N'}(\hat{x}) - \bar{v}^M_{N}$ and $\bar{g}^M_{N}(\hat{x}) - \bar{v}^M_{N}$, 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 $z_\alpha \equiv \Phi^{-1}(1-\alpha)$, 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

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

or

$$\bar{g}^M_{N}(\hat{x}) - \bar{v}^M_{N} + z_\alpha \frac{\bar{S}_{M}}{\sqrt{M}}$$

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

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

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 Bechhofer, Santner, and Goldsman [1]. In the numerical tests described in Section 4, a combined procedure was used, as described in Nelson, Swann, Goldsman, and Song [8]. During the first stage of the combined procedure a subset $\mathcal{S}''$ of the candidate solutions $\mathcal{S}' \equiv \left\{\hat{x}^1_{N}, \ldots, \hat{x}^M_{N}\right\}$ are chosen (called screening) for further evaluation, based on their sample average values $\hat{g}_{N'}(\hat{x}^m_{N})$. During the second stage, sample sizes $N'' \geq N'$ are determined for more detailed evaluation, based on the sample variances $S^2_{N'}(\hat{x}^m_{N})$. Then $N'' - N'$ additional observations are generated, and then 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 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 successive SAA replications without improvement, using the Bayesian guideline if needed), 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 2.2.
    - 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 Compute $\hat{g}_{N'}(\hat{x}_N^m)$ and compare with $\hat{g}_{N'}(\hat{x}_N^{m'})$, the value of the best solution $\hat{x}_N^{m'}$ found so far, $m' < m$. Let $\hat{x}$ denote the solution among $\hat{x}_N^{m'}$ and $\hat{x}_N^m$ with the best value of $\hat{g}_{N'}(x)$.
3. Estimate the optimality gap $g(\hat{x}) - v^*$, and the variance of the gap estimator.
4. If the optimality gap is too large, increase the sample sizes N and/or N', and return to step 2. Otherwise, choose the best solution $\hat{x}$ among all candidate solutions $\hat{x}_N^m$, $m = 1, \ldots, M$, using a screening and selection procedure. Stop.

4. Numerical Tests. In this section we describe an application of the SAA method to some optimization problems. 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 well-conditioning measure $\alpha(\varepsilon)$, on the performance of the method.

It is insightful to note that the number of decision variables and the well-conditioning measure $\alpha(\varepsilon)$ are related. To illustrate this relationship, consider a discrete optimization problem with feasible set $\mathcal{S}$ given by the vertices of the unit hypercube in $\mathbb{R}^k$, i.e., $\mathcal{S} = \{0,1\}^k$. Suppose the origin is the unique optimal solution of the true problem, i.e., $\mathcal{S}^* = \{0\}$. Let us restrict attention to linear objective functions. Thus the optimization problem is

$$\min_{x \in \{0,1\}^k} \sum_{i=1}^{k} c_i x_i$$

where $c_i > 0$ for all $i \in \{1, \ldots, k\}$. It is easy to see that $\alpha(0) = \min_{i \in \{1,\ldots,k\}} c_i$. Thus, by choosing c_i arbitrarily small for some i, the well-conditioning measure $\alpha(0)$ can be made arbitrarily poor. It is more interesting to investigate how good the well-conditioning measure $\alpha(0)$ can be. Thus we want to choose c to

$$\max_{c \in \mathbb{R}^k_+} \left\{ \alpha(0) = \min_{i \in \{1,\ldots,k\}} c_i \right\}$$

To make the result independent of scale, we restrict c to satisfy $\|c\| = 1$. For example, using $\|\cdot\|_p$, the constraint is $\sum_{i=1}^k c_i^p = 1$, where $1 \leq p < \infty$. It is easily seen that the best choice of c is to make all the components equal, i.e., to take $c_i = 1/k^{1/p}$ for all $i \in \{1, \dots, k\}$. Hence the best value of the well-conditioning measure $\alpha(0)$ is $1/k^{1/p}$. This indicates that the well-conditioning measure tends to be poorer if the number k of decision variables is larger.

4.1. Resource Allocation Problem. First 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 $W = (W_1, \dots, W_k)$. 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 $[x]^+ \equiv \max\{x, 0\}$. 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 to fit the items in. 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. Another 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 $\{p_1, \dots, p_k\}$ 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 $x_i = 1$ denotes 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).

Still 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

$$\min_{x\in\mathcal{X}} \; I\!\!E[C(x)]$$

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

Objective functions with terms such as $I\!\!E\left[\sum_{i=1}^{k} W_i x_i - q\right]^{+}$ are also interesting for the following reason. For many stochastic optimization problems good solutions can be obtained by replacing the random variables by their means and then solving the resulting deterministic optimization problem, called the expected value problem (Birge and Louveaux [2]). 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 is also 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.

4.2. Numerical Results. It was shown in section 2 that the well-conditioning measure

$$\alpha(\varepsilon) \;\equiv\; \min_{x\in\mathcal{S}\setminus\mathcal{S}^{\varepsilon}} g(x) - v^* - \varepsilon$$

is an important factor affecting the convergence rate of the SAA method for stochastic discrete optimization. A disadvantage of this measure is that it is too hard to compute beforehand for the types of optimization problems that the SAA method is intended for. Nevertheless, it is insightful to investigate the effect of $\alpha(\varepsilon)$ on the performance of the SAA method, and in the examples that follow we attempt to demonstrate some of this effect.

Another factor which may have an important effect on the performance of the SAA method, is the number of decision variables. As discussed earlier, the well-conditioning measure $\alpha(\varepsilon)$ tends to become poorer with an increase in the number of decision variables. Therefore, 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. Each set has one instance (called instances 10D and 20D, respectively) that was chosen "deterministically" to be hard, in the sense that its well-conditioning measure $\alpha(\varepsilon)$ is poor, and five instances (called instances 10R1 through 10R5 and 20R1 through 20R5, respectively) that were generated randomly.

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, we have then that the random variable $Y(x) \equiv \sum_{i=1}^k W_i x_i - q$ 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 $Y(x) \sim N(\mu(x), \sigma(x)^2)$, that

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

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Thus, we obtain

$$g(x) \;=\; \sum_{i=1}^k r_i x_i - c\left[\mu(x)\Phi(\mu(x)/\sigma(x)) + \frac{\sigma(x)}{\sqrt{2\pi}} \exp\left(-\frac{\mu(x)^2}{2\sigma(x)^2}\right)\right]. \tag{4.2}$$

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 (Cohn and Barnhart [3]), and for evaluation of solutions produced by the SAA Algorithm. Good numerical approximations are available for computing $\Phi(\cdot)$, such as Algorithm AS66 (Hill [5]).

Figure 4.1 and Figure 4.2 show the distributions of the relative objective values $g(x)/v^*$ of the best 5% of the solutions of the first set of instances, and the best 0.01% of the solutions of the second set of instances, respectively. The objective values $g(x)$ were computed using (4.2), and v^* and the best solutions were identified by enumeration. The harder instances (10D and 20D) have many more solutions with objective values close to optimal than the random instances.

Figure 4.3 and Figure 4.4 show the well-conditioning measure $\alpha(\varepsilon)$ as a function of ε, for instances 10D, 10R1, 10R5, 20D, 20R1, and 20R5. It can be seen that the well-conditioning measure $\alpha(\varepsilon)$ is much worse for the harder instances (10D and 20D) than for the randomly generated instances.

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 well-conditioning measure $\alpha(\varepsilon)$. Figure 4.5 to Figure 4.10 show the estimated probability that a SAA optimal solution $\hat{x}_N$ has objective value $g(\hat{x}_N)$ within relative tolerance d of the optimal value v^*, i.e., $P[v^* - g(\hat{x}_N) \leq d\, v^*]$, as a function of the sample size N, for different values of d. The experiment was conducted by generating

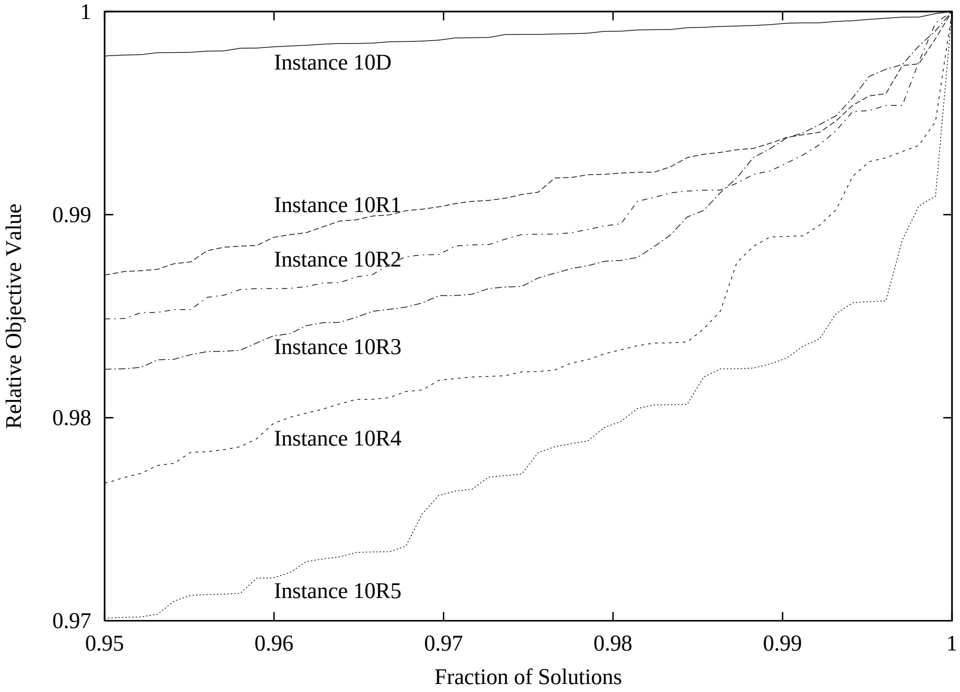

Fig. 4.1. *Distribution of relative objective values $g(x)/v^*$ of best 5% of solutions of instances with 10 decision variables.*

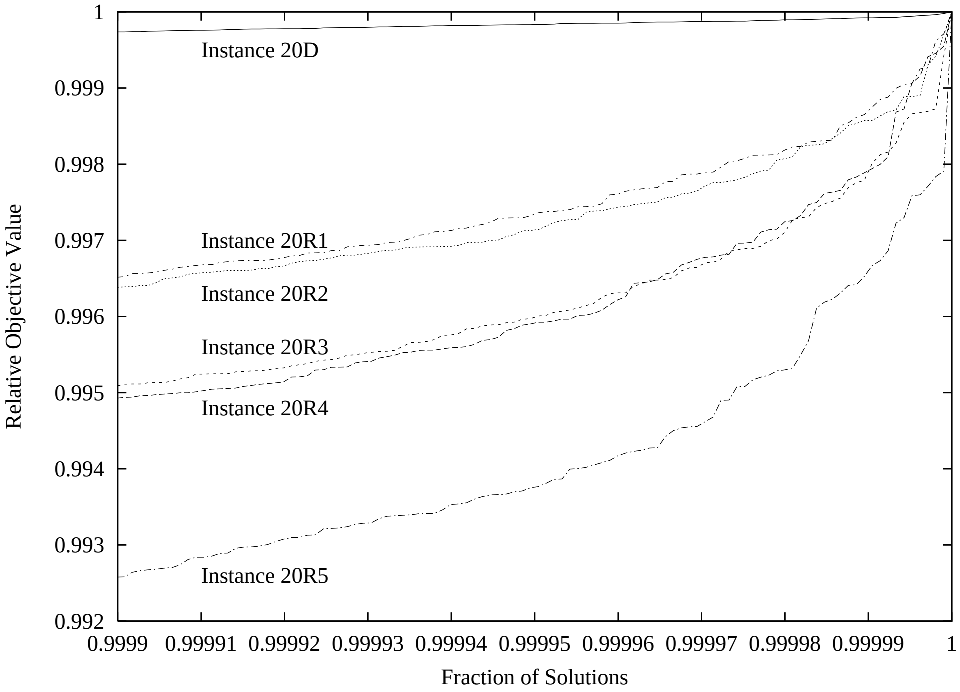

Fig. 4.2. *Distribution of relative objective values $g(x)/v^*$ of best 0.01% of solutions of instances with 20 decision variables.*

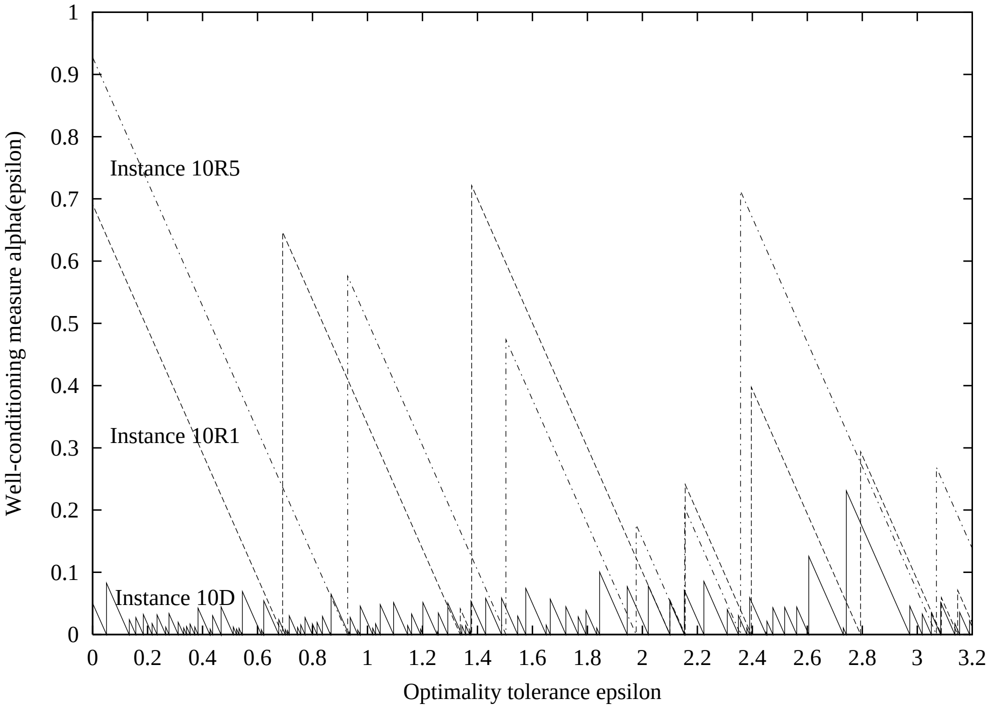

FIG. 4.3. *Well-conditioning measure $\alpha(\varepsilon)$ as a function of ε, for instances 10D, 10R1, and 10R5.*

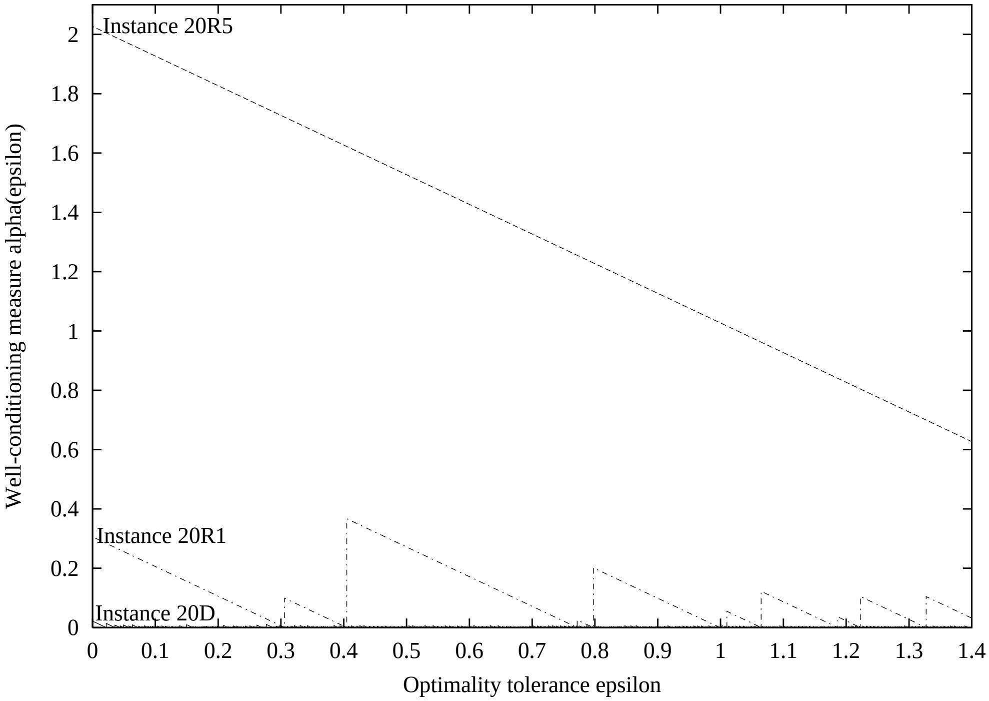

FIG. 4.4. *well-conditioning measure $\alpha(\varepsilon)$ as a function of ε, for instances 20D, 20R1, and 20R5.*

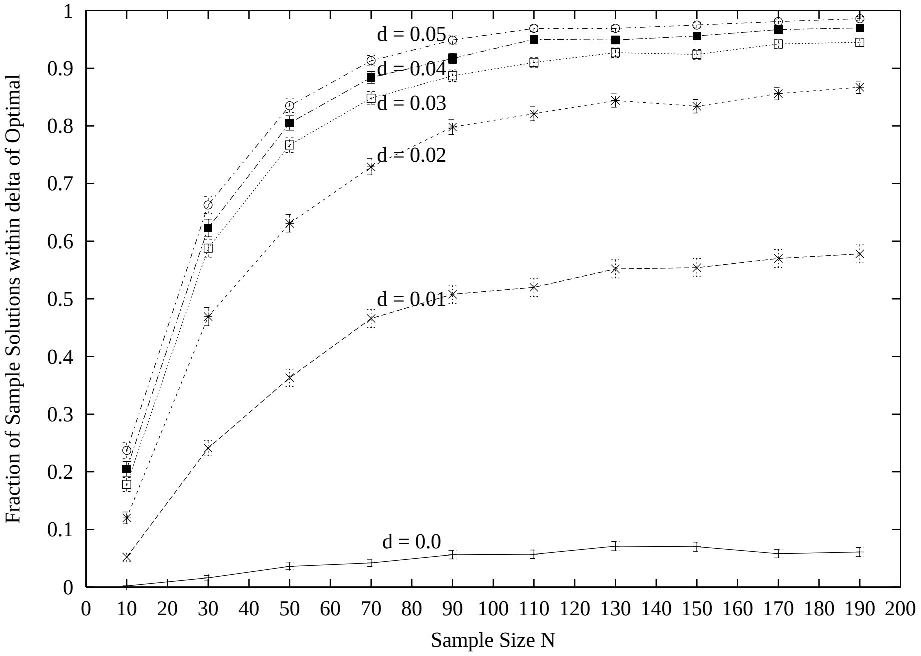

FIG. 4.5. *Probability of SAA optimal solution $\hat{x}_N$ having objective value $g(\hat{x}_N)$ within relative tolerance d of the optimal value v^*, $\hat{P}[v^* - g(\hat{x}_N) \le d\,v^*]$, as a function of sample size N, for different values of d, for Instance 10D.*

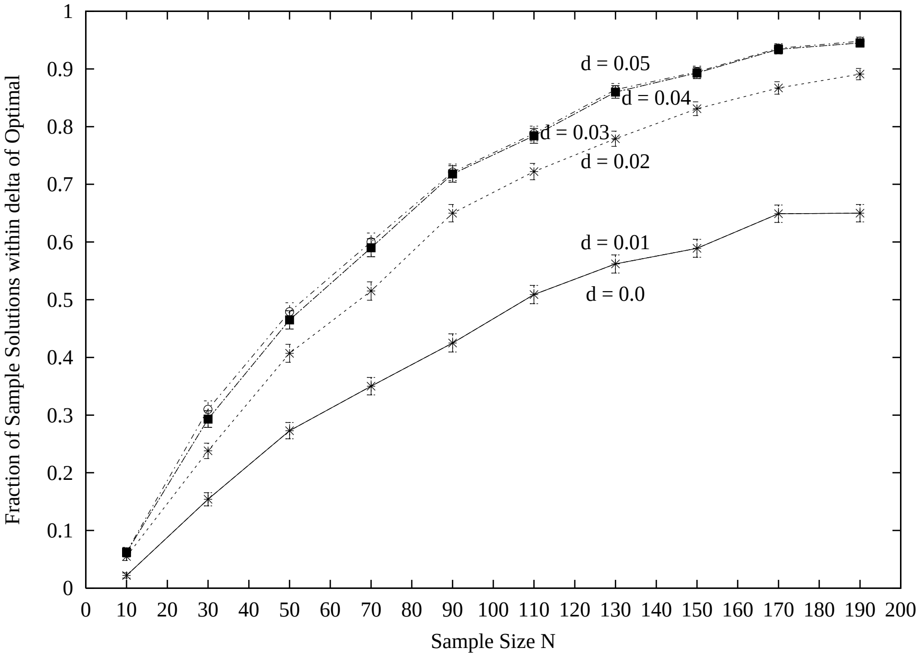

FIG. 4.6. *Probability of SAA optimal solution $\hat{x}_N$ having objective value $g(\hat{x}_N)$ within relative tolerance d of the optimal value v^*, $\hat{P}[v^* - g(\hat{x}_N) \le d\,v^*]$, as a function of sample size N, for different values of d, for Instance 10R1.*

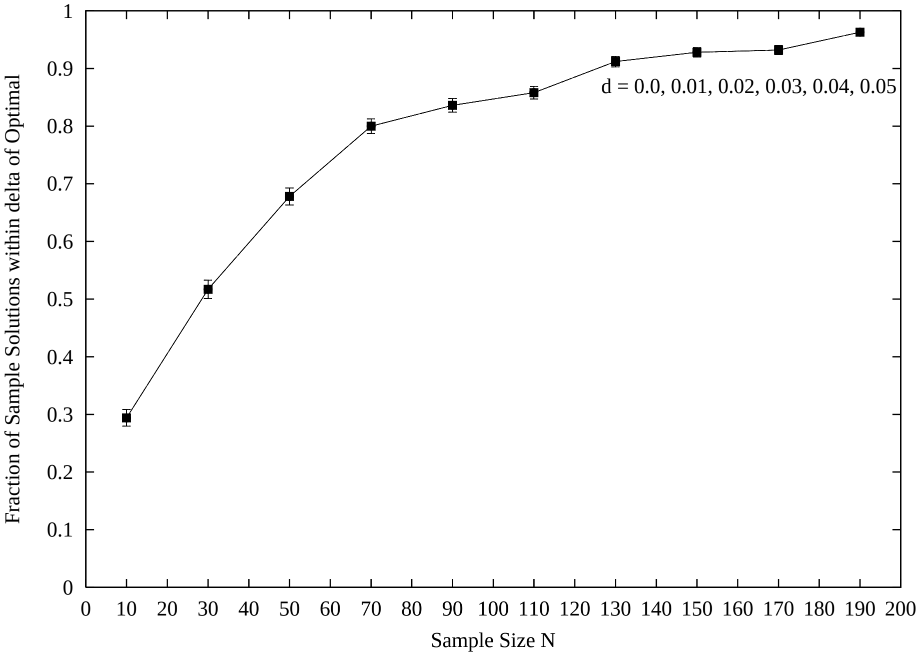

FIG. 4.7. *Probability of SAA optimal solution $\hat{x}_N$ having objective value $g(\hat{x}_N)$ within relative tolerance d of the optimal value v^*, $\hat{P}[v^* - g(\hat{x}_N) \leq d\,v^*]$, as a function of sample size N, for different values of d, for Instance 10R5.*

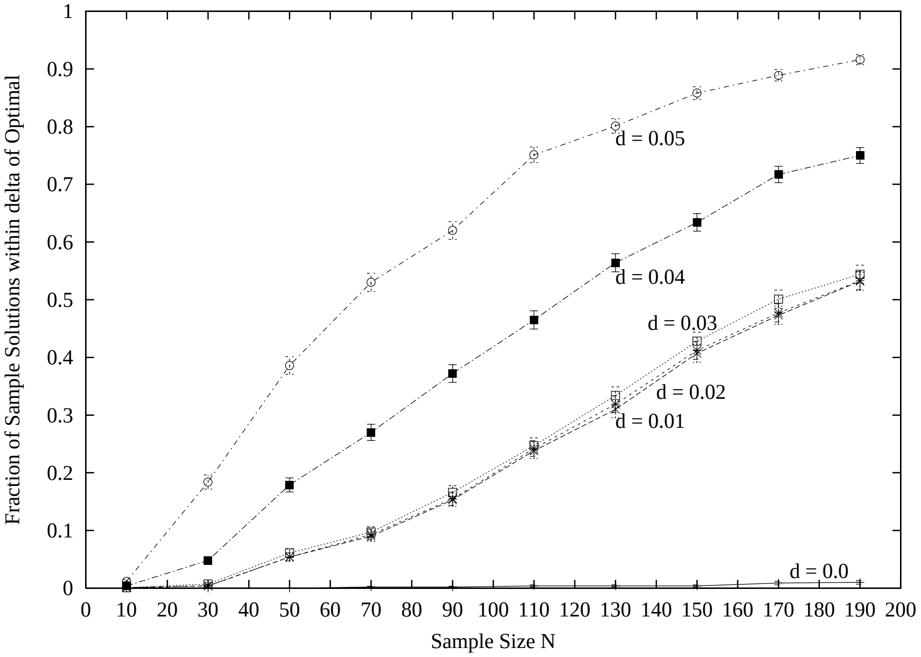

FIG. 4.8. *Probability of SAA optimal solution $\hat{x}_N$ having objective value $g(\hat{x}_N)$ within relative tolerance d of the optimal value v^*, $\hat{P}[v^* - g(\hat{x}_N) \leq d\,v^*]$, as a function of sample size N, for different values of d, for Instance 20D.*

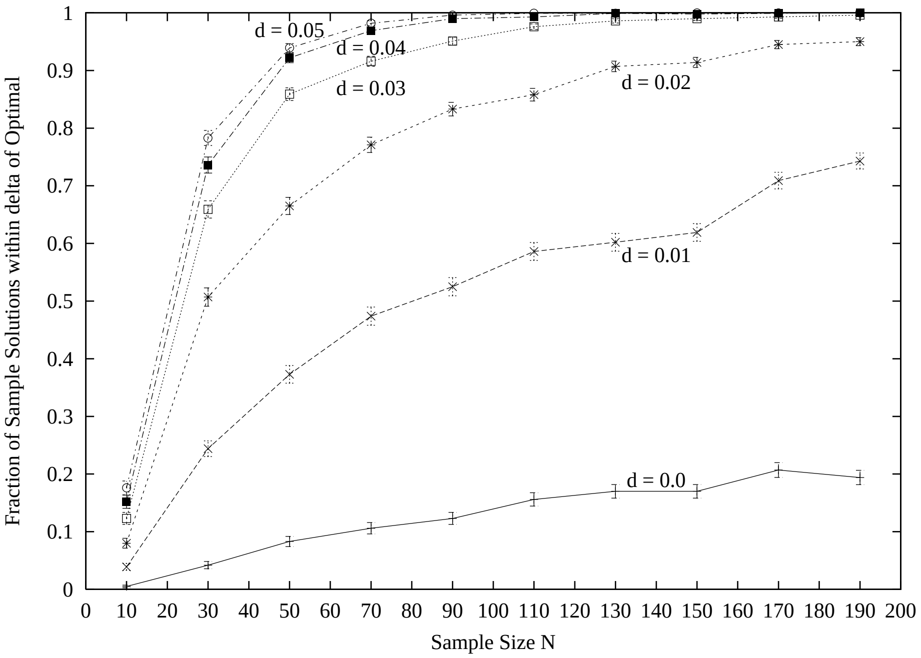

FIG. 4.9. *Probability of SAA optimal solution $\hat{x}_N$ having objective value $g(\hat{x}_N)$ within relative tolerance d of the optimal value v^*, $\hat{P}[v^* - g(\hat{x}_N) \leq d\,v^*]$, as a function of sample size N, for different values of d, for Instance 20R1.*

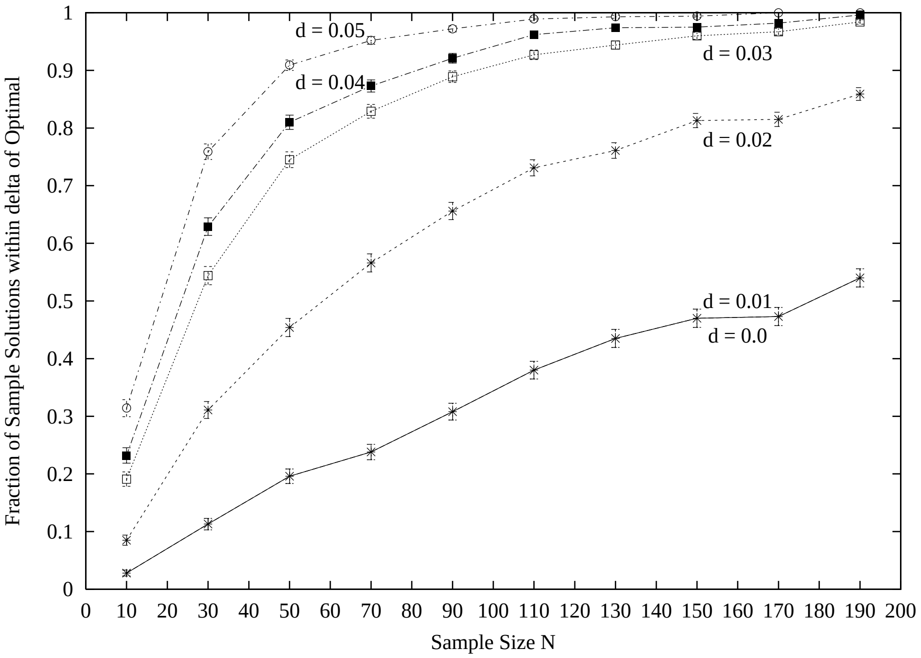

FIG. 4.10. *Probability of SAA optimal solution $\hat{x}_N$ having objective value $g(\hat{x}_N)$ within relative tolerance d of the optimal value v^*, $\hat{P}[v^* - g(\hat{x}_N) \leq d\,v^*]$, as a function of sample size N, for different values of d, for Instance 20R5.*

$M = 1000$ independent SAA replications for each sample size N, computing SAA optimal solutions $\hat{x}_N^m$, $m = 1, \ldots, M$, and their objective values $g(\hat{x}_N^m)$ using (4.2), 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 errorbars 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 a SAA replication generates an optimal solution ($d = 0$) increases much slower with increase in the sample size N for the harder instances (10D and 20D) with poor well-conditioning measures $\alpha(\varepsilon)$ than for the randomly generated instances with good well-conditioning measures. However, the probability that a SAA replication generates a reasonably good solution (eg., $d = 0.05$) increases quite fast with increase in the sample size N for both the harder instances and for the randomly generated instances. Also, it seems that the probability that a SAA replication generates an optimal solution ($d = 0$) increases somewhat slower with increase in the sample size N for the instances with more decision variables than for the instances with fewer decision variables, for both the harder instances as well as 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 well-conditioning measure $\alpha(\varepsilon)$. In this experiment the maximum number of successive SAA replications without improvement with the same sample size N was chosen as $M' = 50$. Besides that, 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' successive SAA replications without improvement had occurred. At that stage the optimality gap estimator (3.3) was computed. If the optimality gap estimator was greater than a specified tolerance, then the sample size N was increased and the procedure was repeated. Otherwise, if the optimality gap estimator 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.

Figure 4.11 to Figure 4.16 show the objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ produced during the course of the algorithm. The figures show several noticeable effects. First, for all the instances 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 estimator 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 for the harder instances. Third, for the instances with fewer decision variables 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 for the

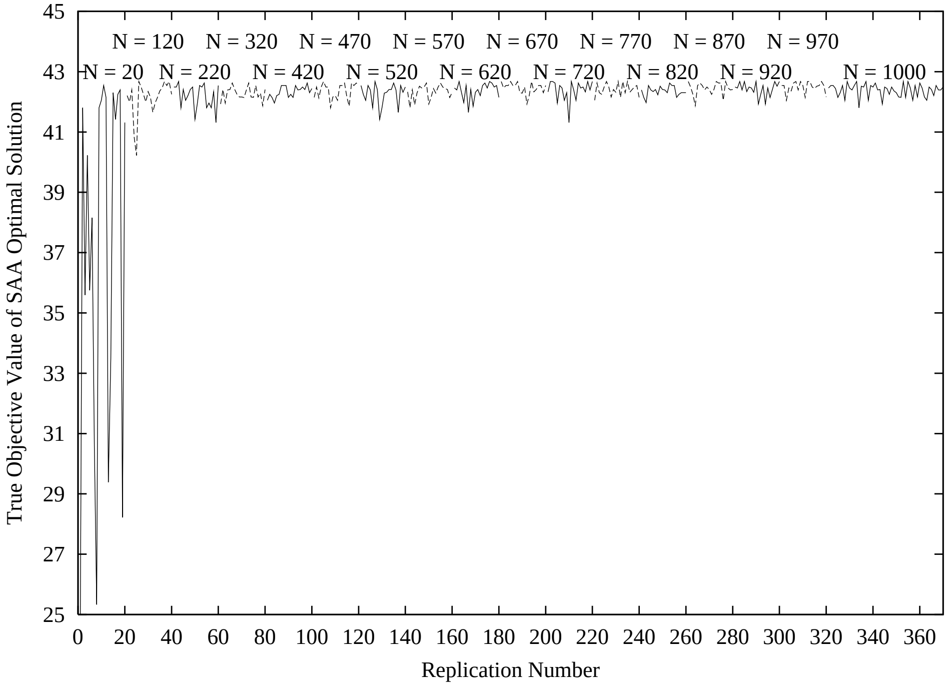

FIG. 4.11. *Improvement of objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ as the sample size N increases, for Instance 10D.*

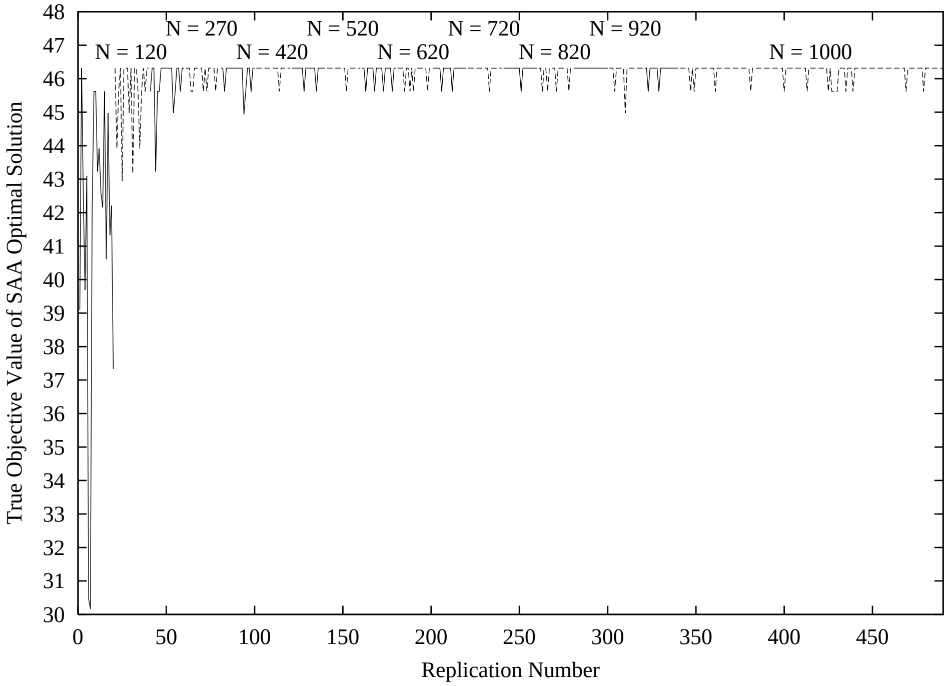

FIG. 4.12. *Improvement of objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ as the sample size N increases, for Instance 10R1.*

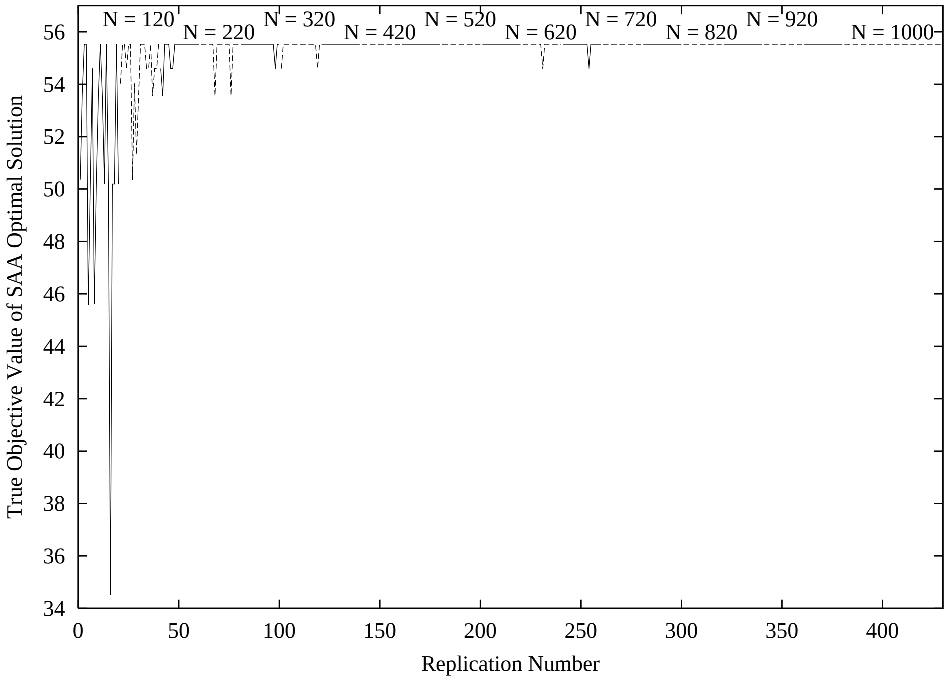

FIG. 4.13. *Improvement of objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ as the sample size N increases, for Instance 10R5.*

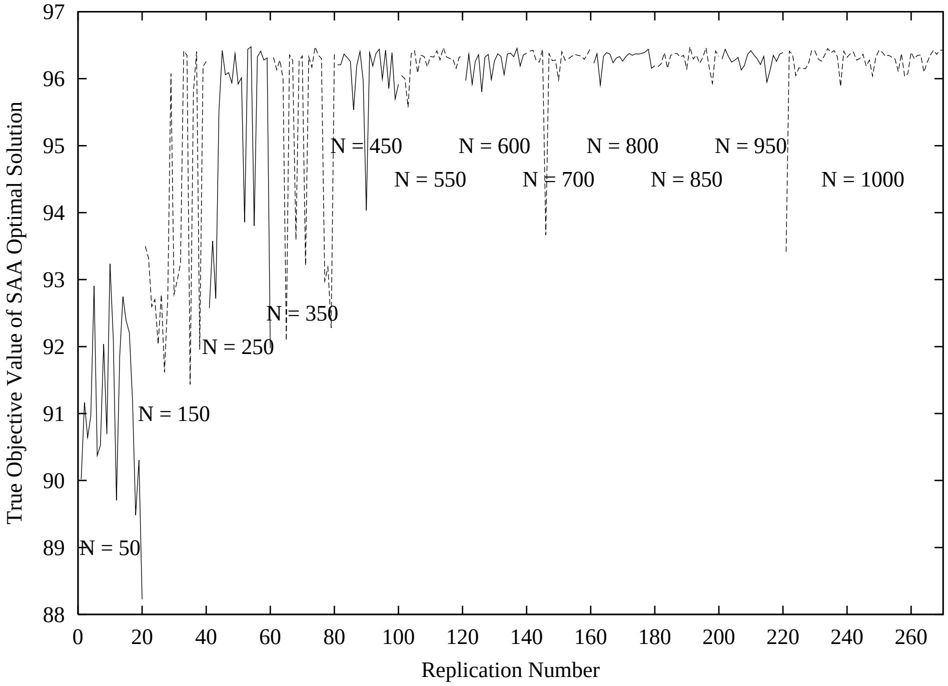

FIG. 4.14. *Improvement of objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ as the sample size N increases, for Instance 20D.*

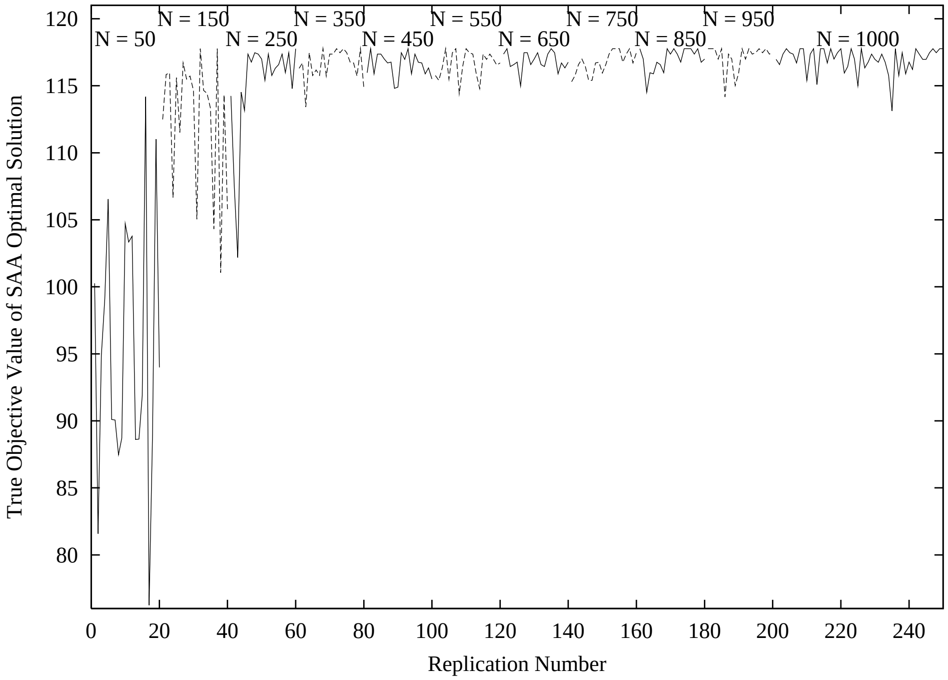

FIG. 4.15. *Improvement of objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ as the sample size N increases, for Instance 20R1.*

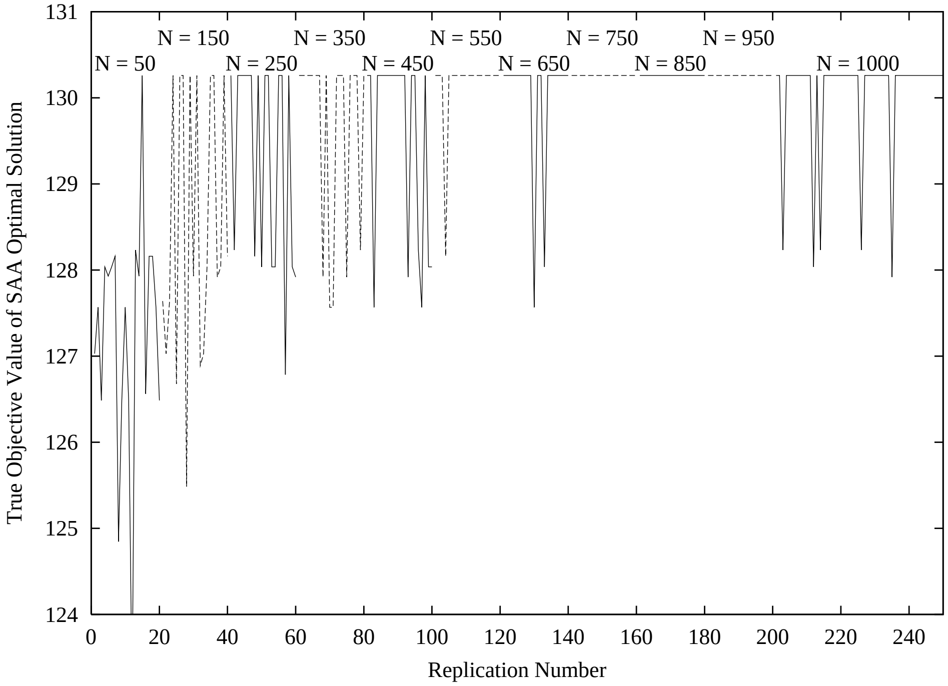

FIG. 4.16. *Improvement of objective values $g(\hat{x}_N^m)$ of SAA optimal solutions $\hat{x}_N^m$ as the sample size N increases, for Instance 20R5.*

instances with more decision variables. 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 $N = 120$; for instance 10R1, the optimal solution was first produced after $m = 2$ replications with sample size $N = 20$; and for instance 10R5, the optimal solution was first produced after $m = 3$ 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); for instance 20R1, the optimal solution was first produced after $m = 12$ replications with sample size $N = 150$; and for instance 20R5, the optimal solution was first produced after $m = 15$ replications with sample size $N = 50$.

As mentioned above, in the second numerical experiment it was noticed that often the optimality gap estimator is large, even if an optimal solution has been found, i.e., $v^* - g(\hat{x}) = 0$, (which is also a common problem in deterministic discrete optimization). Consider the components of the optimality gap estimator 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 $z_\alpha(S^2_{N'}(\hat{x})/N' + S^2_M/M)^{1/2}$ 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 $\bar{v}^M_N - v^*$, and the fact that $\mathbb{E}[\bar{v}^M_N] - v^* \geq 0$, as identified in (3.1). It was also mentioned that the bias decreases as the sample size N increases. However, the second numerical experiment indicates 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 investigates the effect of the number of decision variables and the well-conditioning measure $\alpha(\varepsilon)$ on the bias in the optimality gap estimator. Figure 4.17 and Figure 4.18 show how the relative bias $\bar{v}^M_N/v^*$ of the optimality gap estimator changes as the sample size N increases, for different instances. The most noticeable effect is that the bias decreases much slower for the harder instances than for the randomly generated instances as the sample size N increases. This is in accordance with the asymptotic result (2.8) of Proposition 2.3. Also, the bias seems to decrease slower for the instances with more decision variables than for the instances with fewer decision variables.

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 well-conditioning of the problem, which in turn tends to become poorer with an increase in the number of decision variables. It was also 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 many fold 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.

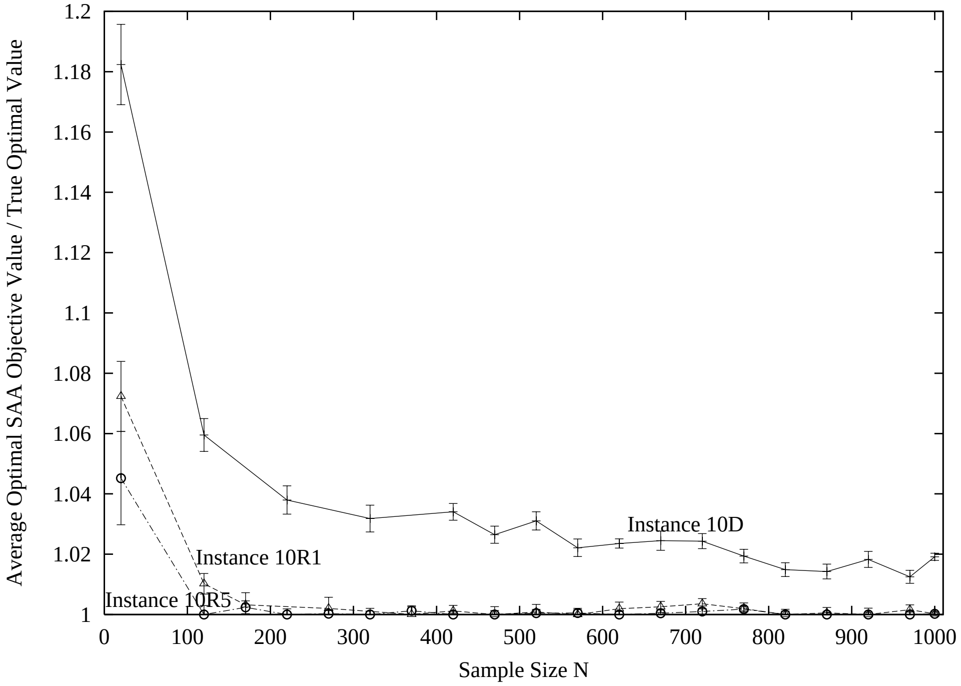

FIG. 4.17. *Relative bias $\bar{v}_N^M/v^*$ of the optimality gap estimator as a function of the sample size N, for instances 10D, 10R1, and 10R5, with 10 decision variables.*

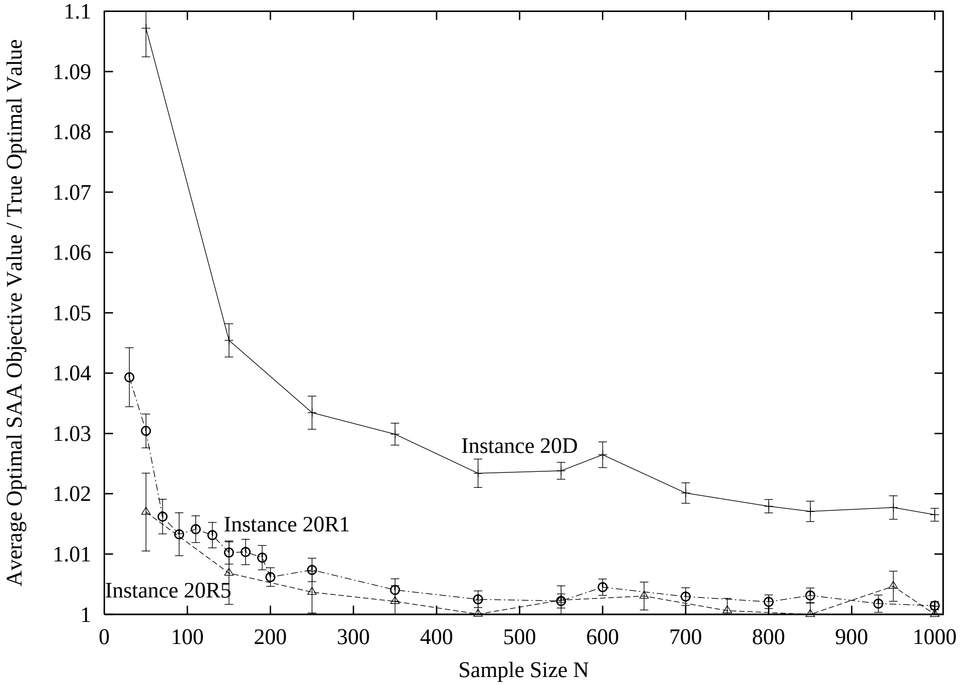

FIG. 4.18. *Relative bias $\bar{v}_N^M/v^*$ of the optimality gap estimator as a function of the sample size N, for instances 20D, 20R1, and 20R5, with 20 decision variables.*

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

[2] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*, Springer Series in Operations Research, Springer-Verlag, New York, 1997.

[3] 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, San Juan, Puerto Rico, June 1998, TRISTAN III.

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

[5] I. D. Hill, *Algorithm as 66: The normal integral*, Applied Statistics, 22 (1973), pp. 424–427.

[6] 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), pp. 47–56.

[7] 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, Netherlands, 1998, ch. 5, pp. 149–168.

[8] 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.* preprint, 1999.

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

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

[11] A. Shapiro, *Asymptotic analysis of stochastic programs*, Annals of Operations Research, 30 (1991), pp. 169–186.

[12] A. Shapiro and T. Homem-de-Mello, *On rate of convergence of Monte Carlo approximations of stochastic programs.* Preprint, available at: Stochastic Programming E-Print Series, `http://dochost.rz.hu-berlin.de/speps/`, 1999.