Optimal Budget Allocation for Sample Average Approximation

Johannes O. Royset and Roberto Szechtman
Operations Research Department, Naval Postgraduate School
Monterey, California, USA

June 1, 2011

Abstract. The sample average approximation approach to solving stochastic programs induces a sampling error, caused by replacing an expectation by a sample average, as well as an optimization error due to approximating the solution of the resulting sample average problem. We obtain an estimator of the optimal value of the original stochastic program after executing a finite number of iterations of an optimization algorithm applied to the sample average problem. We examine the convergence rate of the estimator as the computing budget tends to infinity, and characterize the allocation policies that maximize the convergence rate in the case of sublinear, linear, and superlinear convergence regime for the optimization algorithm.

1 Introduction

Sample average approximation (SAA) is a frequently used approach to solving stochastic programs where an expectation of a random function in the objective function is replaced by a sample average obtained by Monte Carlo sampling. The approach is appealing due to its simplicity and the fact that a large number of standard optimization algorithms are often available to optimize the resulting sample average problem. It is well known that under relatively mild assumptions global and local minimizers as well as stationary points of the sample average problem and the corresponding objective function values tend to the corresponding points and values of the stochastic program almost surely as the sample size increases to infinity. The asymptotic distribution of minimizers, minimum values, and related quantities for the sample average problem are also known under additional assumptions. We refer to Chapter 5 of [30] for a comprehensive presentation of results and [32, 29] for recent advances.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
01 JUN 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Optimal Budget Allocation for Sample Average Approximation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Operations Research Department ,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
in review

14. ABSTRACT
The sample average approximation approach to solving stochastic programs induces a sampling error, caused by replacing an expectation by a sample average, as well as an optimization error due to approximating the solution of the resulting sample average problem. We obtain an estimator of the optimal value of the original stochastic program after executing a finite number of iterations of an optimization algorithm applied to the sample average problem. We examine the convergence rate of the estimator as the computing budget tends to infinity, and characterize the allocation policies that maximize the convergence rate in the case of sublinear, linear, and superlinear convergence regime for the optimization algorithm.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**25**	

In view of the prevalence of uncertainty in planning problems, stochastic programs are formulated and solved by the SAA approach in a broad range of applications such as stochastic vehicle allocation and routing [21, 31, 20], electric power system planning [21, 20], telecommunication network design [21, 20], financial planning [28, 1, 32], inventory control [32], mixed logit estimation models [4], search theory [29], and engineering design [27, 29].

A main difficulty with the approach concerns the selection of an appropriate sample size. At one end, a large sample size provides small discrepancy in some sense between the stochastic program and the sample average problem, but results in a high computational cost as objective function and (sub)gradient evaluations in the sample average problem involve the averaging of a large number of quantities. At the other end, a small sample size is computationally inexpensive as the objective function and (sub)gradient evaluations in the sample average problem can be computed quickly, but yields poor accuracy as the sample average only coarsely approximates the expectation. It is usually difficult to select a sample size that balances accuracy and computational cost without extensive trial and error. This paper examines different policies for sample-size selection given a particular computing budget.

The issue of sample-size selection arises in most applications of the SAA approach. In this paper, however, we focus on stochastic programs where the corresponding sample average problems are solvable by a deterministic optimization algorithm with known rate of convergence such as in the case of subgradient, gradient, and Newtonion methods. This situation includes, for example, two-stage stochastic programs with continuous first-stage variables and a convex recourse function [15], conditional value-at-risk models [28, 32], and programs with convex smooth random functions. We do not deal with integer restrictions, which usually imply that the sample average problem is solvable in finite time, and random functions whose evaluation, or that of its subgradient, gradient, and Hessian (when needed), is difficult due to an unknown probability distribution or other complications. We also do not deal with chance constraints, i.e., situations where the feasible region is given in terms of random functions; see for example Chapter 4 of in [30]. We observe that there are several other approaches to solving stochastic programs (see for example [10, 14, 13, 17, 2, 3, 16, 24, 22]). However, this paper deals with the SAA approach exclusively.

There appears to be only a few studies dealing with the issue of determining a computationally efficient sample size within the SAA approach. Sections 5.3.1 and 5.3.2 of [30] provide estimates of the required sample size to guarantee that a set of near-optimal solutions of the sample average problem is contained in a set of near-optimal solutions of the stochastic program with a given confidence. While these results provide useful insights about the complexity of solving the stochastic program, the sample-size estimates are typically too conservative for practical use. The authors of [5] efficiently estimate the quality of a given

sequence of candidate solutions by Monte Carlo sampling using heuristically derived rules for selecting sample sizes, but do not deal with the sample size needed to generate the candidate solutions.

In the context of a variable SAA approach, where not only one, but a sequence of sample average problems are solved with increasing sample size, [26] constructs open-loop sample-size control policies using a discrete-time optimal control model. That study deals with linearly convergent optimization algorithms and cannot guarantee that the sample-size selections are optimal in some sense. However, the resulting sample-size control policies appear to lead to substantial computational savings over alternative selections.

The recent paper [25] also deals with a variable SAA approach. It defines classes of "optimal sample sizes" that best balance, in some asymptotic sense, the *sampling error* due to the difference between the stochastic program and the sample average problem with the *optimization error* caused by approximate solution of the sample average problems by an optimization algorithm. If the rate of convergence of the optimization algorithm is high, the optimization error will be small relative to that generated by an optimization algorithm with slower rate for a given computing budget. The paper [25] gives specific guidance how to select sample sizes tailored to optimization algorithm with sublinear, linear, and superlinear rate of convergence.

The simulation and simulation optimization literature (see [7] for a review) deals with how to optimally allocate effort across different task within the simulation given a specific computing budget. The allocation may be between exploration of different designs and estimation of objective function values at specific designs as in global optimization [9, 12], between estimation of different random variables nested by conditioning [19], or between estimation of different expected system performances in ranking and selection [8]. These studies typically define an optimal allocation as one that makes the estimator mean-squared error vanish at the fastest possible rate as the computing budget tends to infinity.

The present paper is related to these studies from the simulation and simulation optimization literature, and in particular the recent paper [25]. As in [25], we consider optimization algorithms with sublinear, linear, and superlinear rate of convergence for the solution of the sample average problem. However, we adopt more specific assumptions regarding these rates than in [25] and consider errors in objective function values instead of solutions, which allow us to avoid the potentially restrictive assumption about uniqueness of optimal solutions. Our assumptions are satisfied by standard optimization algorithms such as many subgradient, gradient, and Newtonian methods and allow us to develop refined results regarding the effect of various sample-size selection policies. For algorithms with a sublinear rate of convergence with optimization error of order n^{-p}, where n is the number of iterations, we examine the effect of the parameter $p > 0$. For linear algorithms with optimization

error of order θ^n, we study the influence of the rate of convergence coefficient $\theta \in (0,1)$. For superlinear algorithms with optimization error of order θ^{ψ^n}, the focus is on the power $\psi > 1$ and also the secondary effect due to $\theta > 0$. We determine the rate of convergence of the SAA approach as the computing budget tends to infinity, accounting for both sampling and optimization errors. Specifically, we view the value obtained after a finite number of iterations of an optimization algorithm as applied to a sample average problem with a finite sample size as an estimator of the optimal value of the original stochastic program, and examine the convergence rate of the estimator as the computing budget tends to infinity. To our knowledge, there has been no systematic study of this estimator, its convergence rate, and the influence of various sample-size selection policies on the rate. We determine optimal policies in a sense described below that lead to rates of convergence of order $c^{-\nu}$ for $0 < \nu < 1/2$, $(c/\log c)^{-1/2}$, and $(c/\log\log c)^{-1/2}$ as the computing budget c tends to infinity for sublinear, linear, and superlinear optimization algorithm, respectively. In the linear case, we also determine a policy with rate of convergence similar to the sublinear case that is robust to parameter misspecification.

The paper is organized as follows. The next section presents the stochastic program, the associated sample average problem, as well as underlying assumptions. Sections 3 to 5 consider the cases with sublinear, linear, and superlinear rate of convergence of the optimization algorithm, respectively. Section 6 presents numerical examples illustrating the sample-size selection policies.

2 Problem Statement and Assumptions

We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\Omega \subset \mathbb{R}^k$, a nonempty compact subset $X \subset \mathbb{R}^d$, and the function $f : X \to \mathbb{R}$ defined by

$$f(x) = \mathbb{E}[F(x,\omega)],$$

where $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$ and $F : \mathbb{R}^d \times \Omega \to \mathbb{R}$ is a random function. The following assumption, which ensures that $f(\cdot)$ is well-defined and finite valued as well as other properties, is used throughout the paper.

Assumption 1 *We assume that*

(i) *the expectation* $\mathbb{E}[F(x,\omega)^2] < \infty$ *for all* $x \in X$ *and*

(ii) *there exists a measurable function* $C : \Omega \to \mathbb{R}_+$ *such that* $\mathbb{E}[C(\omega)^2] < \infty$ *and*

$$|F(x,\omega) - F(x',\omega)| \le C(\omega)\|x - x'\|$$

for all $x, x' \in X$ *and almost every* $\omega \in \Omega$.

In view of Theorems 7.43 and 7.44 in [30], $f(\cdot)$ is well-defined, finite-valued, and Lipschitz continuous on X. We observe that weaker assumptions suffice for these properties to hold; see [30], pp. 368-369. However, in this paper we utilize a central limit theorem and therefore adopt these light-tail assumptions from the beginning for simplicity of presentation.

We consider the *stochastic program*

$$\mathbf{P}: \quad \min_{x \in X} f(x),$$

which from the continuity of $f(\cdot)$ and compactness of X has a finite optimal value denoted by f^*. We denote the set of optimal solutions of $\mathbf{P}$ by X^*.

In general, $f(x)$ cannot be computed exactly, and we approximate it using a sample average. We let $\overline{\Omega} = \Omega \times \Omega \times ...$ be the sample space corresponding to an infinite sequence of sample points and let $\overline{\mathbb{P}}$ be the probability distribution on $\overline{\Omega}$ generated by $\mathbb{P}$ under independent sampling. We denote subelements of $\overline{w} \in \overline{\Omega}$ by $\omega^j \in \Omega$, $j = 1, 2, ...$, i.e., $\overline{\omega} = (w^1, w^2, ...)$. Then, for $m \in \mathbb{N} = \{1, 2, 3, ...\}$, we define the *sample average* $f_m : \mathbb{R}^d \times \overline{\Omega} \to \mathbb{R}$ by

$$f_m(x, \overline{\omega}) = \sum_{j=1}^{m} F(x, \omega^j)/m.$$

Various sample sizes give rise to a family of approximations of $\mathbf{P}$. Let $\{\mathbf{P}_m(\overline{\omega})\}_{m \in \mathbb{N}}$ be this family, where, for any $m \in \mathbb{N}$, the *sample average problem* $\mathbf{P}_m(\overline{\omega})$ is defined by

$$\mathbf{P}_m(\overline{\omega}): \quad \min_{x \in X} f_m(x, \overline{\omega}).$$

Under Assumption 1 (and also under weaker assumptions), $f_m(\cdot, \overline{\omega})$ is Lipschitz continuous on X for almost every $\overline{\omega} \in \overline{\Omega}$. Hence, $\mathbf{P}_m(\overline{\omega})$ has a finite optimal value for almost every $\overline{\omega} \in \overline{\Omega}$, which we denote by $f_m^*(\overline{\omega})$.

The SAA approach consists of selecting a sample size m, generating a sample $\overline{\omega}$, and then approximately solving $\mathbf{P}_m(\overline{\omega})$ using an appropriate optimization algorithm. (In practice, this process may be repeated several times, possibly with variable sample size, to facilitate validation analysis of the obtained solutions and to reduce the overall computing time; see for example Section 5.6 in [30] and [29]. However, in this paper, we focus on a single replication.) A finite sample size induces a *sampling error* $f_m^*(\overline{\omega}) - f^*$, which typically is nonzero. However, as the sample size $m \to \infty$, the sampling error vanishes in some sense as the following proposition states, where $N(\mu, \sigma^2)$ stands for a normal random variable with mean μ and variance σ^2, $\sigma^2(x)$ for $Var[F(x, \omega)]$, and $\Rightarrow$ for weak convergence.

Proposition 1 *If Assumption 1 holds, then*

$$m^{1/2}(f_m^*(\overline{\omega}) - f^*) \Rightarrow \inf_{y \in X^*} N(0, \sigma^2(y)),$$

as $m \to \infty$.

Proof: The result follows directly from Theorem 5.7 in [30] as Assumption 1 implies the assumption of that theorem. □

Unless $\mathbf{P}_m(\overline{\omega})$ possesses a special structure such as in the case of a linear or quadratic program, it cannot be solved in finite computing time. Hence, the SAA approach is also associated with an optimization error. Given a deterministic optimization algorithm, let $A_m^n(x,\overline{\omega})$ be the solution obtained after $n \in \mathbb{N}$ iterations of that optimization algorithm, starting from $x \in X$, as applied to $\mathbf{P}_m(\overline{\omega})$. We assume that $A_m^n(x,\overline{\omega})$ is a random vector for any $n, m \in \mathbb{N}$ and $x \in X$, with $A_m^0(x,\overline{\omega}),\overline{\omega}) = x$. The *optimization error* is then $f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f_m^*(\overline{\omega})$. If the optimization algorithm converges to a globally optimal solution of $\mathbf{P}_m(\overline{\omega})$, then the optimization error vanishes as $n \to \infty$. However, the rate with which it vanishes depends on the rate of convergence of the optimization algorithm.

In this paper, we examine the trade-off between sampling and optimization errors in the SAA approach within a given *computing budget* c. A large sample size ensures a small sampling error, but, due to the computing budget, restricts the number of iterations of the optimization algorithm causing a potentially large optimization error. Similarly, a large number of iterations may result in a large sampling error. Naturally, therefore, the choice of sample size and number of iterations could depend on the computing budget and we sometimes write $m(c)$ and $n(c)$ to indicate this dependence. We refer to $\{(m(c), n(c)\}_{c\in\mathbb{N}}$, with $n(c), m(c) \in \mathbb{N}$ for all $c \in \mathbb{N}$, and $n(c), m(c) \to \infty$, as $c \to \infty$, as an *allocation policy*. An allocation policy specifies the number of iterations and sample size to adopt for a given computing budget c. We observe that the focus on unbounded sequences for both $n(c)$ and $m(c)$ is not restrictive as we are interested in situations where infinite number of iterations and sample size are required to ensure that both the optimization and sampling errors vanish.

The specifics of the trade-off between sampling and optimization errors depends on the computational effort needed to carry out n iterations of the optimization algorithm as a function of m. We adopt the following assumption.

Assumption 2 *For any $n, m \in \mathbb{N}$, $x \in X$, and $\overline{\omega} \in \overline{\Omega}$, the computational effort to obtain $A_m^n(x,\overline{\omega})$ is nm.*

Assumption 2 is reasonable in view of the fact that each function, (sub)gradient, and Hessian evaluation of the optimization algorithm when applied to $\mathbf{P}_m(\overline{\omega})$ requires the summation of m quantities. Hence, the effort per iteration would be proportional to m. This linear growth in m has also been observed empirically; see, e.g., p. 204 in [30]. Assuming that each iteration involves approximately the same number of operations, which is the case for single-point algorithm such as the subgradient, steepest descent, and Newton's methods, the computational effort to carry out n iterations would be proportional to nm. We observe that we could replace nm by γnm, where γ is a constant, in Assumption 2. However, this simply

amounts to a rescaling of the computing budget and has no influence on the subsequent analysis. An allocation policy $\{(n(c), m(c))\}_{c\in\mathbb{N}}$ that satisfies $n(c)m(c)/c \to 1$ as $c \to \infty$ is *asymptotically admissible*. Hence, an asymptotically admissible policy will, at least in the limit as c tends to infinity, satisfy the computing budget.

The two kinds of errors, due to sampling and optimization, contribute to the mean-squared error $\mathrm{MSE}(f_{m(c)}(A^{n(c)}_{m(c)}(x,\overline{\omega})) = E[(f_{m(c)}(A^{n(c)}_{m(c)}(x,\overline{\omega}),\overline{\omega}) - f^*)^2]$. In view of the discussion above, the MSE vanishes, in some sense, under mild assumptions as c tends to infinity. Of course, there is a large number of asymptotically admissible allocation policies, and ensuing convergence rates. In the next three sections we analyze the estimator convergence rate under the assumption of sublinear, linear, and superlinear rate of convergence of the optimization algorithm.

We use the following standard ordering notation in the remainder of the paper. A sequence of random variables $\{\xi_n\}_{n\in\mathbb{N}}$ is $O_p(1)$ if for all $\zeta > 0$ there exists a constant ϵ such that $P(|\xi_n| > \epsilon) < \zeta$ for all n sufficiently large. Similarly, the sequence is $O_p(0)$ if for $\epsilon > 0$ arbitrary $P(|\xi_n| > \epsilon) \to 0$, as $n \to \infty$. A deterministic sequence $\{\xi_n\}_{n\in\mathbb{N}}$ is $O(\alpha_n)$, for $\{\alpha_n\}_{n\in\mathbb{N}}$ a positive sequence, if $|\xi_n|/\alpha_n$ is bounded by a finite constant. We also write $\xi_n \sim \alpha_n$ if ξ_n/α_n tends to a finite constant as $n \to \infty$.

3 Sublinearly Convergent Optimization Algorithm

Suppose that the deterministic optimization algorithm used to solve $\mathbf{P}_m(\overline{\omega})$ converges sublinearly as stated in the next assumption.

Assumption 3 *There exists a $p > 0$ and a family of measurable functions $K_m : \overline{\Omega} \to \mathbb{R}_+$, $m \in \mathbb{N}$, such that $K_m(\overline{\omega}) \Rightarrow K \in [0,\infty)$, as $m \to \infty$, and*

$$f_m(A^n_m(x,\overline{\omega}),\overline{\omega}) - f^*_m(\overline{\omega}) \leq \frac{K_m(\overline{\omega})}{n^p}$$

for all $x \in X$, $n, m \in \mathbb{N}$, and almost every $\overline{\omega} \in \overline{\Omega}$.

Several standard algorithms satisfy Assumption 3 when $\mathbf{P}_m(\overline{\omega})$ is convex. For example, the subgradient method satisfies Assumption 3 with $p = 1/2$ and $K_m(\overline{\omega}) = D_X \sum_{j=1}^m C(\omega^j)/m$, where $C(\omega)$ is as in Assumption 1 and $D_X = \max_{x,x'\in X} \|x - x'\|$; see [23], pp. 142-143. When $F(\cdot,\omega)$ is Lipschitz continuously differentiable, Nesterov's optimal gradient method satisfies Assumption 3 with $p = 2$ and $K_m(\omega)$ proportional to the product of the average Lipschitz constant of $\nabla_x F(\cdot,\omega)$ on X and D^2_X; see p. 77 of [23].

In view of Assumption 3 and the optimality of $f^*_m(\overline{\omega})$ for $\mathbf{P}_m(\overline{\omega})$,

$$f^*_m(\overline{\omega}) - f^* \leq f_m(A^n_m(x,\overline{\omega}),\overline{\omega}) - f^* \leq f^*_m(\overline{\omega}) - f^* + K_m(\overline{\omega})/n^p, \tag{1}$$

for every $n, m \in \mathbb{N}$, $x \in X$, and almost every $\overline{\omega} \in \overline{\Omega}$. It follows that $\mathrm{MSE}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega})) \leq \max\{\mathrm{MSE}(f_m^*(\overline{\omega})), \mathrm{MSE}(f_m^*(\overline{\omega}) + K_m(\overline{\omega})/n^p)\}$. This suggests that a good allocation policy should balance the sampling error $f_m^*(\overline{\omega}) - f^*$, which contributes to both maximands and decays at rate $m^{-1/2}$, and the bias term $K_m(\overline{\omega})/n^p$ due to the optimization. More precisely, since under Assumption 2 increasing n and m are equally computationally costly, we would like to select an asymptotically admissible allocation policy $\{(n(c), m(c))\}_{c\in\mathbb{N}}$ such that $m(c)^{-1/2} \sim n(c)^{-p}$. This discussion is formalized in the next theorem, where to simplify the notation we write n and m instead of $n(c)$ and $m(c)$, respectively.

Theorem 1 *Suppose that Assumptions 1, 2, and 3 hold, $x \in X$, $\{(n(c), m(c))\}_{c\in\mathbb{N}}$ is an asymptotically admissible allocation policy, and $n(c)/c^{1/(2p+1)} \to a$, with $a \in (0,\infty)$, as $c \to \infty$. Then*

$$c^{p/(2p+1)}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*) = O_p(1).$$

Proof. By assumption, $m \to \infty$, as $c \to \infty$, and hence, in view of Proposition 1, $m^{1/2}(f_m^*(\overline{\omega}) - f^*) \Rightarrow \inf_{y\in X^*} N(0, \sigma^2(y))$, as $c \to \infty$. Let $r(c) = c^{p/(2p+1)}$. In view of Eq. (1),

$$r(c)(f_m^*(\overline{\omega}) - f^*) \leq r(c)(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*) \leq r(c)(f_m^*(\overline{\omega}) - f^* + K_m(\overline{\omega})/n^p) \quad (2)$$

for all $c \in \mathbb{N}$ and almost every $\overline{\omega} \in \overline{\Omega}$.

Since the policy is asymptotically admissible, $mn/c \to 1$, as $c \to \infty$. Moreover, $n/c^{1/(2p+1)} \to a \in (0,\infty)$, as $c \to \infty$, by assumption. We find that

$$\frac{r(c)}{m^{1/2}} = \left(\frac{c}{nm}\right)^{1/2} \left(\frac{n}{c^{1/(2p+1)}}\right)^{1/2} \to a^{1/2},$$

as $c \to \infty$. Similarly,

$$\frac{r(c)}{n^p} = \left(\frac{c^{1/(2p+1)}}{n}\right)^p \to a^{-p},$$

as $c \to \infty$.

Then, by a convergent together argument (p. 27 of [6]),

$$r(c)(f_m^*(\overline{\omega}) - f^* + K_m(\overline{\omega})/n^p) \Rightarrow a^{-p}K + a^{1/2} \inf_{y\in X^*} N(0, \sigma^2(y)), \quad (3)$$

as $c \to \infty$, where K is as in Assumption 3, and

$$r(c)(f_m^*(\overline{\omega}) - f^*) \Rightarrow a^{1/2} \inf_{y\in X^*} N(0, \sigma^2(y)), \quad (4)$$

as $c \to \infty$. From Eq. (2),

$$r(c)|f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*| \leq \max\{r(c)|f_m^*(\overline{\omega}) - f^*|, r(c)|f_m^*(\overline{\omega}) - f^* + K_m(\overline{\omega})/n^p|\}$$

for almost every $\overline{\omega} \in \overline{\Omega}$. Hence, in view of Eqs. (3) and (4), for any $\epsilon > 0$ there exists a constant $C > 0$ such that

$$\overline{\mathbb{P}}\left(r(c)|f_m^*(\overline{\omega}) - f^*| \geq C\right) < \frac{\epsilon}{2}$$

and

$$\overline{\mathbb{P}}\left(r(c)|f_m^*(\overline{\omega}) - f^* + K_m(\overline{\omega})/n^p| \geq C\right) < \frac{\epsilon}{2},$$

for all sufficiently large c. Therefore,

$$\begin{aligned}
& \overline{\mathbb{P}}\left(r(c)|f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*| \geq C\right) \\
\leq\ & \overline{\mathbb{P}}\left(\max\{r(c)|f_m^*(\overline{\omega}) - f^*|, r(c)|f_m^*(\overline{\omega}) - f^* + K_m(\overline{\omega})/n^p| \geq C\right) \\
\leq\ & \overline{\mathbb{P}}\left(r(c)|f_m^*(\overline{\omega}) - f^*| \geq C\right) + \overline{\mathbb{P}}\left(r(c)|f_m^*(\overline{\omega}) - f^* + K_m(\overline{\omega})/n^p| \geq C\right) \\
<\ & \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

for c sufficiently large. □

We see from Theorem 1 that for any finite $p > 0$, the convergence rate of $f_{m(c)}(A_{m(c)}^{n(c)}(x,\overline{\omega})$ is worse than the canonical rate $c^{-1/2}$ when only sampling is considered; see Proposition 1. Hence, $c^{1/2-p/(2p+1)} = c^{(1/2)/(2p+1)}$ is the "cost of optimization." Of course, if p is large, that cost is moderate. However, if $p = 1/2$ as for the subgradient method, then the cost of optimization is $c^{1/4}$ and the convergence rate is of order $c^{-1/4}$.

It follows from the proof of Theorem 1 that if n grows slower or faster than $c^{1/(2p+1)}$ then the convergence rate is slower than $c^{p/(2p+1)}$. Indeed, it is easy to see that $n \sim c^{1/(2p+1)+\epsilon}$, for $\epsilon > 0$, results in a convergence rate of order $c^{p/(2p+1)-\epsilon/2}$, while $n \sim c^{1/(2p+1)-\epsilon}$, for $0 < \epsilon < 1/(2p+1)$, leads to a convergence rate of order $c^{p/(2p+1)-p\epsilon}$. This lends support to our statement that $n \sim c^{1/(2p+1)}$ is the optimal allocation rule. Conveniently, however, the rate of convergence is robust to the choice of a.

4 Linearly Convergent Optimization Algorithm

Suppose that the deterministic optimization algorithm used to solve $\mathbf{P}_m(\overline{\omega})$ converges linearly with a rate of convergence coefficient independent of $\overline{\omega}$ and m as stated in the next assumption.

Assumption 4 *There exists a $\theta \in (0,1)$ such that*

$$f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f_m^*(\overline{\omega}) \leq \theta(f_m(A_m^{n-1}(x,\overline{\omega}),\overline{\omega}) - f_m^*(\overline{\omega}))$$

for all $x \in X$, $n, m \in \mathbb{N}$, and almost every $\overline{\omega} \in \overline{\Omega}$.

Assumption 4 is satisfied by many gradient methods such as the steepest descent method and projected gradient method when applied to $\mathbf{P}_m(\overline{\omega})$ under the assumption that $F(\cdot,\omega)$ is strongly convex and twice continuously differentiable for almost every $\overline{\omega} \in \overline{\Omega}$ and that X is convex. Moreover, the requirement in Assumption 4 that the rate of convergence coefficient θ holds "uniformly" for almost every $\overline{\omega} \in \overline{\Omega}$ follows when the largest and smallest eigenvalue of the Hessian of $F(\cdot,\omega)$ are bounded from above and below, respectively, with bounds independent of ω. The requirement of a twice continuously differentiable random function excludes at first sight two-stage stochastic programs with recourse [15], conditional Value-at-Risk minimization problems [28], inventory control problems [32], complex engineering design problems [27], and similar problems involving a nonsmooth random function. However, these nonsmooth functions can sometimes be approximated with high accuracy by smooth functions [1, 32, 29]. Hence, the results of this section as well as the next one, dealing with superlinearly convergent optimization algorithms, may also be applicable in such contexts.

In view of Assumption 4, we find that

$$f_m^*(\overline{\omega}) - f^* \leq f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^* \leq f_m^*(\overline{\omega}) - f^* + \theta^n(f_m(x,\overline{\omega}) - f_m^*(\overline{\omega})) \tag{5}$$

for every $n, m \in \mathbb{N}$ and almost every $\overline{\omega} \in \overline{\Omega}$. As in the sublinear case, a judicious approach to ensure that $\mathrm{MSE}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}))$ decays at the fastest possible rate is to equalize the sampling and optimization error decay rates. Since the first term decreases at a rate $m^{-1/2}$ and the second term at a rate θ^n, an allocation with $n(c) \sim \log c$ and $m(c) \sim c/\log c$ meets this criterion. The following theorem makes rigorous this argument.

Theorem 2 *Suppose that Assumptions 1, 2, and 4 hold, $x \in X$, $\{(n(c), m(c))\}_{c\in\mathbb{N}}$ is an asymptotically admissible allocation policy, and $n(c) - a\log c \to 0$, with $a > 0$, as $c \to \infty$.*

(i) If $a \geq (2\log(1/\theta))^{-1}$, then, as $c \to \infty$,

$$\left(\frac{c}{a\log c}\right)^{1/2}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*) \Rightarrow \inf_{y\in X^*} N(0,\sigma^2(y)).$$

(ii) If $0 < a < (2\log(1/\theta))^{-1}$, then

$$c^{a\log(\theta^{-1})}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*) = O_p(1).$$

Proof. Since the policy is asymptotically admissible, $mn/c \to 1$, and also $m \to \infty$, as $c \to \infty$. This fact and the Central Limit Theorem imply that $m^{1/2}(f_m(x,\overline{\omega}) - f(x)) \Rightarrow N(0,\sigma^2(x))$, as $c \to \infty$. Proposition 1 results in $m^{1/2}(f_m^*(\overline{\omega}) - f^*) \Rightarrow \inf_{y\in X^*} N(0,\sigma^2(y))$, as $c \to \infty$.

For any $r : \mathbb{N} \to \mathbb{R}_+$,

$$\begin{aligned} &r(c)(f_m^*(\overline{\omega}) + \theta^n(f_m(x,\overline{\omega}) - f_m^*(\overline{\omega})) - f^*) \\ &\qquad = r(c)\left[(f_m^*(\overline{\omega}) - f^*) + \theta^n(f_m(x,\overline{\omega}) - f(x) + f^* - f_m^*(\overline{\omega})) + \theta^n(f(x) - f^*)\right] \end{aligned} \tag{6}$$

for every $\overline{\omega} \in \overline{\Omega}$.

First we consider part (i) and let $r(c) = (c/(a \log c))^{1/2}$, with $a \geq (2 \log(1/\theta))^{-1}$. By the assumption on a,

$$r(c)\theta^n = r(c)e^{(n-a\log c)\log\theta}e^{a\log c\log\theta} = \left(\frac{1}{a\log c}\right)^{1/2} c^{1/2-a\log(\theta^{-1})}e^{(n-a\log c)\log\theta} \to 0,$$

as $c \to \infty$, so that $r(c)\theta^n/m^{1/2} \to 0$. Analogously,

$$\frac{r(c)}{m^{1/2}} = r(c)\left(\frac{c}{mn}\right)^{1/2}\left(\frac{n}{c}\right)^{1/2} = \left(\frac{c}{mn}\right)^{1/2}\left(\frac{n-a\log c}{a\log c}+1\right)^{1/2} \to 1,$$

as $c \to \infty$.

Then, by Eq. (6) and a converging together argument (p. 27 of [6]),

$$r(c)\left(f_m^*(\overline{\omega}) - f^* + \theta^n(f_m(x,\overline{\omega}) - f_m^*(\overline{\omega}))\right) \Rightarrow \inf_{y\in X^*} N(0,\sigma^2(y)),$$

and

$$r(c)\left(f_m^*(\overline{\omega}) - f^*\right) = r(c)\frac{m^{1/2}}{m^{1/2}}\left(f_m^*(\overline{\omega}) - f^*\right) \Rightarrow \inf_{y\in X^*} N(0,\sigma^2(y)),$$

as $c \to \infty$. In view of Eq. (5), the conclusion of part (i) follows.

Second, we consider part (ii) and let $r(c) = c^{a\log(\theta^{-1})}$, with $0 < a < (2\log(1/\theta))^{-1}$. Then,

$$r(c)\theta^n = r(c)e^{(n-a\log c)\log\theta}e^{a\log c\log\theta} = e^{(n-a\log c)\log\theta} \to 1,$$

as $c \to \infty$. Also,

$$\begin{aligned}\frac{r(c)}{m^{1/2}} &= r(c)\left(\frac{c}{nm}\right)^{1/2}\left(\frac{n}{c}\right)^{1/2} \\ &= c^{a\log(\theta^{-1})-1/2}\left(\frac{c}{nm}\right)^{1/2}(a\log c)^{1/2}\left(\frac{n-a\log c}{a\log c}+1\right)^{1/2} \to 0,\end{aligned}$$

as $c \to \infty$. Consequently,

$$\frac{r(c)\theta^n}{m^{1/2}} \to 0,$$

as $c \to \infty$. By Eq. (6) and a converging together argument,

$$r(c)\left(f_m^*(\overline{\omega}) + \theta^n(f_m(x,\overline{\omega}) - f_m^*(\overline{\omega})) - f^*\right) \Rightarrow f(x) - f^*, \tag{7}$$

and

$$r(c)\left(f_m^*(\overline{\omega}) - f^*\right) \Rightarrow 0, \tag{8}$$

as $c \to \infty$. From Eq. (5),

$$|f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*| \leq \max\{|f_m^*(\overline{\omega}) - f^*|, |f_m^*(\overline{\omega}) + \theta^n(f_m(x,\overline{\omega}) - f_m^*(\overline{\omega}))) - f^*|\}$$

for almost every $\overline{\omega} \in \overline{\Omega}$. From here on the argument follows the last part of the proof of Theorem 1, and is omitted.

□

In view of Theorem 2, we see that if $a \geq (2\log(1/\theta))^{-1}$, then $f_m(A_m^n(x,\overline{\omega}),\overline{\omega})$ tends to f^* at a rate $(c/\log c)^{-1/2}$, which is slower than the canonical rate $c^{-1/2}$ in the case of sampling only; see Proposition 1. Hence $(\log c)^{1/2}$ can be viewed as the cost of optimization. We note that the best choice of a is $(2\log(1/\theta))^{-1}$ and that the convergence rate worsens significantly when $a < (2\log(1/\theta))^{-1}$. Specifically, for $a = \delta(2\log(1/\theta))^{-1}$, with $\delta \in (0,1)$, the rate is $c^{-\delta/2}$, which is slower than $(c/\log)^{-1/2}$ for all c sufficiently large.

If X^* is a singleton and $a = (2\log(1/\theta))^{-1}$, then

$$\left(\frac{c}{\log c}\right)^{1/2}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*) \Rightarrow N(0,(-2\log\theta)^{-1}\sigma^2(x^*)),$$

as $c \to \infty$, which can be used to construct confidence interval for f^* if θ is known and $\sigma^2(x^*)$ can be estimated.

Often the rate of convergence coefficient, θ, of the optimization algorithm is theoretically known as in the case of the steepest descent and projected gradient methods with Armijo step size rule (see Section 6) and/or it can be accurately estimated from preliminary calculations using the optimization algorithm; see [26]. If the theoretical value of θ is excessively conservative relative to the actual progress made by the algorithm or preliminary calculations are impractical or unreliable, then it may be problematic to use the best allocation policy recommended by Theorem 2, i.e., selecting $\{(n(c),m(c)\}_{n\in\mathbb{N}}$ such that $n(c) - (2\log(1/\theta))^{-1}\log c \to 0$ and $n(c)m(c)/c \to 1$, as $c \to \infty$. A slight underestimation of θ would result in substantially slower rate as indicated by part (ii) of that theorem. In such a situation, it may be prudent to select a more conservative allocation policy that satisfies $n(c) \sim c^\nu$ for $0 < \nu < 1$, which guarantees the same convergence rate regardless of the value of θ; this is the same approach followed in a different context in [19]. The rate is worse than the optimal one of Theorem 2, but better than what can occur with a poor estimate of θ. This conservative asymptotic admissible allocation policy is discussed in the next proposition.

Proposition 2 *Suppose that Assumptions 1, 2, and 4 hold, $x \in X$, $\{(n(c),m(c))\}_{c\in\mathbb{N}}$ is an asymptotically admissible allocation policy, and $n(c)/c^\nu \to a > 0$, with $0 < \nu < 1$, as $c \to \infty$. Then,*

$$\frac{c^{\frac{1-\nu}{2}}}{a^{1/2}}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*) \Rightarrow \inf_{y\in X^*} N(0,\sigma^2(y)),$$

as $c \to \infty$.

Proof. Let $r(c) = c^{\frac{1-\nu}{2}}/a^{1/2}$. Then,

$$\frac{r(c)}{m^{1/2}} = r(c)\left(\frac{c}{mn}\right)^{1/2}\left(\frac{n}{c}\right)^{1/2} \to 1, \tag{9}$$

and

$$r(c)\theta^n = \frac{1}{a^{1/2}} e^{\frac{1-\nu}{2}\log c - \frac{n}{c^\nu}c^\nu \log\theta^{-1}} \to 0, \tag{10}$$

as $c \to \infty$. From these results it follows that

$$\frac{r(c)}{m^{1/2}}\theta^n \to 0, \tag{11}$$

as $c \to \infty$. Also, Proposition 1, Eq. (9), and a convergence together argument show that

$$r(c)(f_m^*(\overline{\omega}) - f^*) \Rightarrow \inf_{y \in X^*} N(0, \sigma^2(y)).$$

Proposition 1, the Central Limit Theorem, Eqs. (9)–(11), and a convergence together argument show that

$$r(c)\theta^n(f_m(x,\overline{\omega}) - f(x) + f^* - f_m^*(\overline{\omega})) \Rightarrow 0,$$

and

$$r(c)\theta^n(f(x) - f^*) \Rightarrow 0,$$

as $c \to \infty$. From this point on the proof resembles that of Theorem 2, and we omit the details. □

5 Superlinearly Convergent Optimization Algorithm

In this section, we assume that the optimization algorithm used to solve $\mathbf{P}_m(\overline{\omega})$ is superlinearly convergent as defined by the next assumption.

Assumption 5 *There exists a $\theta \in (0,\infty)$ and a $\psi \in (1,\infty)$ such that*

$$f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f_m^*(\overline{\omega}) \le \theta(f_m(A_m^{n-1}(x,\overline{\omega}),\overline{\omega}) - f_m^*(\overline{\omega}))^\psi$$

for all $x \in X$, $m,n \in \mathbb{N}$, and almost every $\overline{\omega} \in \overline{\Omega}$.

Assumption 5 holds for Newton's method with $\psi = 2$ when applied to $\mathbf{P}_m(\overline{\omega})$ with $F(\cdot,\omega)$ being strongly convex and twice Lipschitz continuously differentiable for almost every $\overline{\omega} \in \overline{\Omega}$, the starting point $x \in X$ is sufficiently close to the global minimizer of $\mathbf{P}_m(\overline{\omega})$, and if the Hessian of $F(\cdot,\omega)$ and its Lipschitz constant are bounded in some sense as ω ranges over Ω. Cases with $\psi \in (1,2)$ arise for example in Newtonian methods with infrequent Hessian updates.

For any $\overline{\omega} \in \overline{\Omega}$, it follows by induction from Assumption 5 that

$$f_m^*(\overline{\omega}) - f^* \leq f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^* \leq f_m^*(\overline{\omega}) - f^* + \theta^{-1/(\psi-1)}(\theta^{1/(\psi-1)}(f_m(x,\overline{\omega}) - f_m^*(\overline{\omega})))^{\psi^n}. \tag{12}$$

Similar to the discussion of the past sections, in order to guarantee that $\mathrm{MSE}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}))$ decreases at the fastest possible rate we equalize the sampling error rate (of order $m^{-1/2}$) with the optimization error decay rate (of order $(\theta^{1/(\psi-1)}(f_m(x,\overline{\omega}) - f_m^*(\overline{\omega})))^{\psi^n}$ as long as the element within parentheses is smaller than 1). Equalizing these rates suggests an allocation policy with $n(c) \sim \log\log c$. The formal result is stated next, where $\kappa(x) = \log(\theta^{-1/(\psi-1)}(f(x) - f^*)^{-1})$, which is positive for $\theta^{1/(\psi-1)}(f(x) - f^*) < 1$.

Theorem 3 *Suppose that Assumptions 1, 2, and 5 hold, $x \in X$, $\{(n(c), m(c))\}_{c\in\mathbb{N}}$ is an asymptotically admissible allocation policy, and that $n(c) - a\log\log c = O((\log c)^{-a\log\psi})$, with $a > 0$. Moreover, suppose that x is such that $\theta^{1/(\psi-1)}(f(x) - f^*) < 1$, where θ and ψ are as in Assumption 5.*

(i) If $a > 1/\log\psi$ or if $a = 1/\log\psi$ and $\kappa(x) \geq 1/2$ then, as $c \to \infty$,

$$\left(\frac{c}{a\log\log c}\right)^{1/2}(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*) \Rightarrow \inf_{y\in\mathcal{X}^*} N(0,\sigma^2(y)). \tag{13}$$

(ii) If $a < 1/\log\psi$ or if $a = 1/\log\psi$ and $\kappa(x) < 1/2$, then

$$\exp\left(\kappa(x)(\log c)^{a\log\psi}\right)(f_m(A_m^n(x,\overline{\omega}),\overline{\omega}) - f^*) = O_p(1). \tag{14}$$

Proof. Let $x \in X$ be such that $\theta^{1/(\psi-1)}(f(x) - f^*) < 1$. For any $r : \mathbb{N} \to \mathbb{R}_+$, we define

$$B_1(c) = \theta^{1/(\psi-1)} r(c)^{\psi^{-n}}(f_m(x,\overline{\omega}) - f(x)),$$
$$B_2(c) = \theta^{1/(\psi-1)} r(c)^{\psi^{-n}}(f^* - f_m^*(\overline{\omega})),$$
$$b_3(c) = \theta^{1/(\psi-1)} r(c)^{\psi^{-n}}(f(x) - f^*).$$

Then,

$$\begin{aligned} &r(c)\left[f_m^*(\overline{\omega}) - f^* + \theta^{-1/(\psi-1)}(\theta^{1/(\psi-1)}(f_m(x,\overline{\omega}) - f_m^*(\overline{\omega})))^{\psi^n}\right] \\ &\quad = r(c)(f_m^*(\overline{\omega}) - f^*) + \theta^{-1/(\psi-1)}\Big[B_1(c) + B_2(c) + b_3(c)\Big]^{\psi^n}. \end{aligned} \tag{15}$$

By the Mean Value Theorem, $\exp((a\log\log c - n)\log\psi) = 1 + ((a\log\log c - n)\log\psi)\exp(\xi_c)$, where ξ_c lies between $(a\log\log c - n)\log\psi$ and 0. By assumption, $\sup_c\{\xi_c\} < \infty$, so that $\exp((a\log\log c - n)\log\psi) = 1 + O((\log c)^{-a\log\psi})$. Therefore,

$$\begin{aligned} \log r(c)^{\psi^{-n}} &= \psi^{-n}\log r(c) \\ &= \exp((a\log\log c - n)\log\psi)\exp(-a\log\psi\log\log c)\log r(c) \\ &= (1 + O((\log c)^{-a\log\psi}))(\log c)^{-a\log\psi}\log r(c) \\ &= (\log c)^{-a\log\psi}\log r(c) + O((\log c)^{-2a\log\psi}\log r(c)). \end{aligned} \tag{16}$$

We first consider part (i), where $a > 1/\log\psi$ or $a = 1/\log\psi$ and $\kappa(x) \geq 1/2$. Let

$$r(c) = \left(\frac{c}{a\log\log c}\right)^{1/2}.$$

In view of Proposition 1 and the fact that

$$r(c)m^{-1/2} = \left(\frac{c}{nm}\left(\frac{n - a\log\log c}{a\log\log c} + 1\right)\right)^{1/2} \to 1 \tag{17}$$

as $c \to \infty$, we obtain by a converging together argument that

$$r(c)(f_m^*(\overline{\omega}) - f^*) \Rightarrow \inf_{y\in\mathcal{X}^*} N(0, \sigma^2(y)), \tag{18}$$

as $c \to \infty$. By Eqs. (12), (15), and (18) as well as a sandwich argument, the proof of part (i) will be complete once we show that $(B_1(c) + B_2(c) + b_3(c))^{\psi^n} \Rightarrow 0$, or, what is the same, that for an arbitrary $\epsilon > 0$, $\overline{\mathbb{P}}(|B_1(c) + B_2(c) + b_3(c)|^{\psi^n} > \epsilon) \to 0$ as $c \to \infty$. To wit,

$$\begin{aligned}
&\overline{\mathbb{P}}\left(|B_1(c) + B_2(c) + b_3(c)|^{\psi^n} > \epsilon\right) \\
&= \overline{\mathbb{P}}\left(\frac{B_1(c) + B_2(c) + b_3(c)}{\epsilon^{\psi^{-n}}} > 1\right) + \overline{\mathbb{P}}\left(\frac{B_1(c) + B_2(c) + b_3(c)}{\epsilon^{\psi^{-n}}} < -1\right).
\end{aligned} \tag{19}$$

From Eq. (16) it follows that

$$\begin{aligned}
&\log r(c)^{\psi^{-n}} \\
&= (\log c)^{-a\log\psi}\frac{1}{2}(\log c - \log(a\log\log c)) + O((\log c)^{1-2a\log\psi}) \\
&= \frac{1}{2}((\log c)^{1-a\log\psi} - (\log c)^{-a\log\psi}\log\log\log c) + O(1/\log c).
\end{aligned} \tag{20}$$

We observe that $\log r(c)^{\psi^{-n}} = O(1)$, so that $r(c)^{\psi^{-n}}/m^{1/2} \to 0$. Knowing that $m^{1/2}(f_m^*(\overline{\omega}) - f^*) \Rightarrow \inf_{y\in X^*} N(0, \sigma^2(y))$ and $m^{1/2}(f_m(x, \overline{\omega}) - f(x)) \Rightarrow N(0, \sigma^2(x))$, a convergence together argument results in $B_1(c) \Rightarrow 0$, and $B_2(c) \Rightarrow 0$.

Looking at the $b_3(c)$ term, we obtain using Eq. (20) that

$$\log b_3(c) = -\kappa(x) + \frac{1}{2}((\log c)^{1-a\log\psi} - (\log c)^{-a\log\psi}\log\log\log c) + O(1/\log c). \tag{21}$$

Therefore, if $a > 1/\log\psi$ then $\log b_3(c) \to -\kappa(x)$; i.e., $b_3(c) \to \theta^{1/(\psi-1)}(f(x) - f^*) < 1$. The fact that $\epsilon^{\psi^{-n}} \to 1$ as $c \to \infty$ and a convergence together argument then lead to

$$\frac{B_1(c) + B_2(c) + b_3(c)}{\epsilon^{\psi^{-n}}} \Rightarrow \theta^{1/(\psi-1)}(f(x) - f^*).$$

Since $0 \leq \theta^{1/(\psi-1)}(f(x) - f^*) < 1$, the latter implies that the r.h.s. of Eq. (19) converges to 0 as $c \to \infty$.

If $a = 1/\log\psi$ we get from (21) that

$$\log b_3(c) = -\kappa(x) + \frac{1}{2} - \frac{1}{2}\frac{\log\log\log c}{\log c} + O(1/\log c). \tag{22}$$

Hence, if $\kappa(x) > 1/2$ we get via a convergence together argument that the r.h.s. of Eq. (19) converges to 0. When $\kappa(x) = 1/2$, we need to treat the $\epsilon^{\psi^{-n}}$ term more carefully. Proceeding as in Eq. (16) we get that $\log \epsilon^{\psi^{-n}} = \log(\epsilon/\log c) + O((\log c)^{-2})$. Hence, Eqs. (16) and (22) yield

$$\log\left(\frac{b_3(c)}{\epsilon^{\psi^{-n}}}\right) = -\frac{1}{2}\frac{\log\log\log c}{\log c} + O(1/\log c),$$

meaning that $b_3(c)/\epsilon^{\psi^{-n}} < 1$ eventually, so the r.h.s. of Eq. (19) converges to 0 as $c \to \infty$.

Next we consider part (ii), where $a < 1/\log\psi$ or $a = 1/\log\psi$ and $\kappa(x) < 1/2$, and

$$r(c) = \exp\left(\kappa(x)(\log c)^{a\log\psi}\right).$$

Then

$$\begin{aligned}
& r(c)m^{-1/2} \\
&= \exp\left(\kappa(x)(\log c)^{a\log\psi}\right)\left(\frac{c}{nm}\right)^{1/2}\left(\frac{n}{a\log\log c}\right)^{1/2}\left(\frac{a\log\log c}{c}\right)^{1/2} \\
&= \exp\left(\kappa(x)(\log c)^{a\log\psi} - (\log c)/2\right)(a\log\log c)^{1/2}\left(\frac{c}{nm}\right)^{1/2}\left(\frac{n}{a\log\log c}\right)^{1/2} \\
&\to 0,
\end{aligned} \tag{23}$$

as $c \to \infty$. Thus, in view of Proposition 1 and a converging together argument we get that

$$r(c)(f_m^*(\overline{\omega}) - f^*) \Rightarrow 0, \tag{24}$$

as $c \to \infty$.

Also, since $\exp\left(\psi^{-n}\kappa(x)(\log c)^{a\log\psi}\right) \le \exp\left(\kappa(x)(\log c)^{a\log\psi}\right)$, Eq. (23) results in

$$\frac{r(c)^{\psi^{-n}}}{m^{1/2}} = \frac{\exp\left(\psi^{-n}\kappa(x)(\log c)^{a\log\psi}\right)}{m^{1/2}} \to 0,$$

as $c \to \infty$. The Central Limit Theorem and a convergent together argument result in $B_1(c) \Rightarrow 0$, as $c \to \infty$. Just the same, Proposition 1 and a converging together argument show that $B_2(c) \Rightarrow 0$, as $c \to \infty$.

Regarding the $b_3(c)$ term, we obtain from Eq. (16) that

$$\begin{aligned}
& \log b_3(c) \\
&= -\kappa(x) + (\log c)^{-a\log\psi}\log r(c) + O((\log c)^{-2a\log\psi}\log r(c)) \\
&= -\kappa(x) + (\log c)^{-a\log\psi}\kappa(x)(\log c)^{a\log\psi} + O((\log c)^{-a\log\psi}) \\
&= O((\log c)^{-a\log\psi}).
\end{aligned}$$

Also, an argument similar to the one leading to Eq. (16), we find that for any $\epsilon > 0$

$$\log \epsilon^{\psi^{-n}} = \log \epsilon (\log c)^{-a \log \psi} + O((\log c)^{-2a \log \psi}).$$

Hence, there is a finite ϵ such that for all c sufficiently large, $b_3(c) < \epsilon^{\psi^{-n}}$, meaning that $(B_1(c) + B_2(c) + b_3(c))^{\psi^n} = O_p(1)$. In conclusion, $r(c)(f_m(A_m^n(x, \overline{\omega}), \overline{\omega}) - f^*)$ is bounded below by an $O_p(0)$ term (cf., Eq. (24)), and above by an $O_p(1)$ term. The second statement of the theorem now follows using an argument similar to the one employed in the last part of the proof of Theorem 1. □

We observe from Theorem 3 that a should be selected as $1/\log \psi$ to obtain the most favorable coefficient in the rate expression, assuming that the initial solution $x \in X$ is sufficiently close to the optimal solution of $\mathbf{P}$. In this case, the convergence rate is essentially the canonical $c^{-1/2}$, only slightly reduced with a $\log \log c$ term. Hence, in the case of a superlinearly convergent optimization algorithm, the cost of optimization is essentially negligible. If a is selected smaller, the convergence rate may deteriorate, decaying at best at rate $c^{-\kappa(x)} > c^{-1/2}$, as in Theorem 2.

We see from the above result that a must be chosen larger when ψ approaches one to maintain the best rate of convergence. Consequently, n must also be chosen larger. Intuitively, as ψ tends to one, we expect that the above result tends to the one for the linear case. For example, suppose that ψ is a function of c. Specifically, let $\psi = \exp((\log \log c)/\log c)$, which tends to 1 as $c \to \infty$, and $\theta \in (0, 1)$. Then, if $a = \log c / \log \log c$, we obtain that $a \log \psi = 1$ and $\kappa(x) > 1/2$ for all c sufficiently large. Hence, we obtain that

$$\left(\frac{c}{a \log \log c} \right)^{1/2} = \left(\frac{c}{\log c} \right)^{1/2}, \tag{25}$$

which shows that the rate of Theorem 3, part (i), tends to the rate of Theorem 2 with $a = 1$, assuming that $\theta \leq \exp(-1/2)$. The rate is obtain in both cases using $n = \log c$.

6 Numerical Examples

We illustrate the above results using two problem instances. We solve the first problem instance, which arises in the optimization of an investment portfolio, using the sublinearly convergent subgradient method to illustrate the results of Section 3. The second problem instance is randomly generated and we solve it using the linearly convergent steepest descent method, to illustrate Section 4, as well as the quadratically convergent Newton's method, which relates to Section 5. We describe the problem instances and the corresponding numerical results in turn, with Subsections 6.1, 6.2, and 6.3 illustrating the sublinear, linear, and superlinear cases, respectively. We implement the problem instances and optimization

algorithms in Matlab Version 7.9 and run the calculations on a laptop computer with 2.26 GHz processor, 3.5 GB RAM, and Windows XP operating system.

6.1 Subgradient Method

The first problem instance is taken from [18] and considers $d-1$ financial instruments with random returns given by the $(d-1)$-dimensional random vector $\omega = \bar{R} + Qu$, where $\bar{R} = (\bar{R}_1, \bar{R}_2, ..., \bar{R}_{d-1})'$, with $\bar{R}_i$ being the expected return of instrument i, Q is an $(d-1)$-by-$(d-1)$ matrix, and u is a standard normal $(d-1)$-dimensional random vector. As in [18], we randomly generate $\bar{R}$ using an independent sample from a uniform distribution on $[0.9, 1.2]$ and Q using an independent sample from a uniform distribution on $[0, 0.1]$. We would like to distribute one unit of wealth across the $d-1$ instruments such that the Conditional Value-at-Risk of the portfolio return is minimized and the expected portfolio return is no smaller than 1.05. We let $x_i \in \mathbb{R}$ denote the amount of investment in instrument i, $i = 1, 2, ..., d-1$. This results in the random function (see [18, 28])

$$F(x, \omega) = x_d + \frac{1}{1-t} \max \left\{ -\sum_{i=1}^{d-1} \omega_i x_i - x_d, 0 \right\}, \tag{26}$$

where $x = (x_1, x_2, ..., x_d)'$, with $x_d \in \mathbb{R}$ being an auxiliary decision variable, and $t \in (0, 1)$ is a probability level. The feasible region

$$X = \left\{ x \in \mathbb{R}^d \;\middle|\; \sum_{i=1}^{d-1} x_i = 1, \sum_{i=1}^{d-1} \bar{R}_i x_i \geq 1.05, x_i \geq 0, i = 1, 2, ..., d-1 \right\}.$$

We use $d = 101$ and $t = 0.9$. The random function in Eq. (26) is not continuously differentiable everywhere for $\mathbb{P}$-almost every $\omega \in \Omega$. However, the function possesses a subdifferential and we consequently use the subgradient method with fixed step size $(n+1)^{-1/2}$, where n is the number of iterations. This step size is optimal in the sense of Nesterov; see [23], pp. 142-143. As stated in Section 3, the subgradient method satisfies Assumption 3 with $p = 1/2$ and $K_m(\overline{\omega}) = D_X \sum_{j=1}^{m} C(\omega^j)/m$, where $C(\omega)$ is as in Assumption 1 and $D_X = \max_{x,x' \in X} \|x - x'\|$. Of course, as pointed out in [18], this problem instance can be reformulated as a conic-quadratic programming problem and solved directly without the use of sampling. Hence, this is a convenient test instance as we are able to compute $f^* = -0.352604$ (rounded to six digits) using cvx [11]. We use initial solution $x = (0, 0, ..., 0, 1, 0, 0...., 0, -1)'$, where the 65-th component equals 1. In our data, the 65-th instrument has the largest expected return. Hence, the initial solution is the one with the largest expected portfolio return.

Figure 1 illustrates Theorem 2 and displays $\mathrm{MSE}(f_{m(c)}(A^{n(c)}_{m(c)}(x, \overline{\omega}), \overline{\omega}))$ for the subgradient method as a function of computing budget c on a logarithmic scale for the allocation

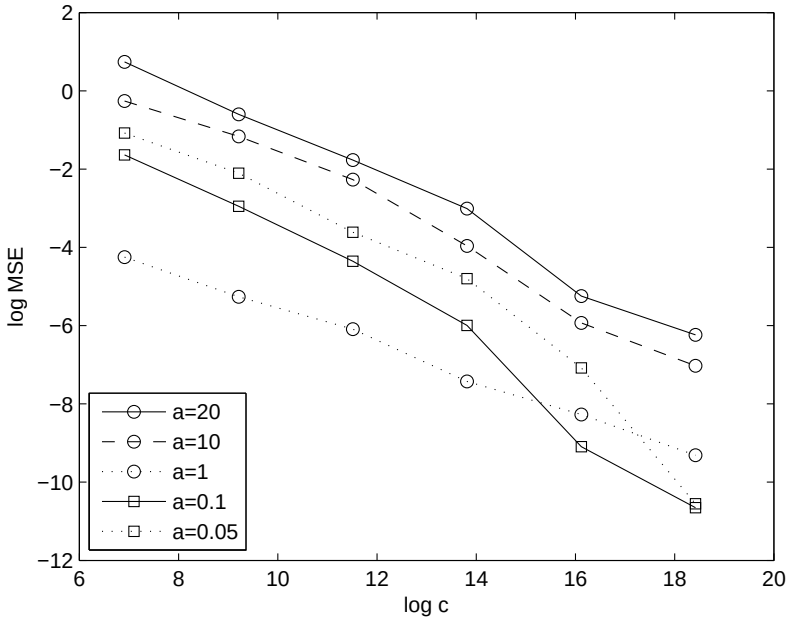

Figure 1: Estimates of MSE$((f_{m(c)}(A^{n(c)}_{m(c)}(x,\overline{\omega}),\overline{\omega}))$ for the subgradient method when applied to the first problem instance as a function of computing budget c on a logarithmic scale for the policy $n(c) = ac^{1/2}$, with $a = 20$, 10, 1, 0.1, and 0.05.

policy $n(c) = ac^{1/2}$, with $a = 20$, 10, 1, 0.1, and 0.05 as well as $c = 10^3$, 10^4, 10^5, 10^6, 10^7, and 10^8. Here and below, the MSE is estimated using 30 independent replications of $f_{m(c)}(A^{n(c)}_{m(c)}(x,\overline{\omega}),\overline{\omega})$. We see that the slopes of the lines in Figure 1 are approximately $-1/2$, which corresponds to a rate of convergence of order $c^{-1/2}$ for the MSE as predicted in Theorem 2 for $p = 1/2$. While the rate of convergence in Theorem 1 is independent of a, we do observe some sensitivity to a numerically. We find that the MSE initially decreases as a decreases until $a = 1$. As a becomes less than one, the picture is less clear and there appears to be little benefit from reducing a further. A examination of the proof of Theorem 2 shows that such a behavior can be expected as both the upper and lower bounds on the estimator depend on a; see (1), (3), and (4).

6.2 Steepest Descent Method

The second problem instance uses

$$F(x,\omega) = \sum_{i=1}^{20} a_i(x_i - b_i\omega_i)^2 \tag{27}$$

where b_i is randomly generated from a uniform distribution on $[-1,1]$ and a_i is randomly generated from a uniform distribution on $[1,2]$, $i = 1,2,...,20$. The vector $\omega = (\omega_1,\omega_2,...,\omega_{20})'$

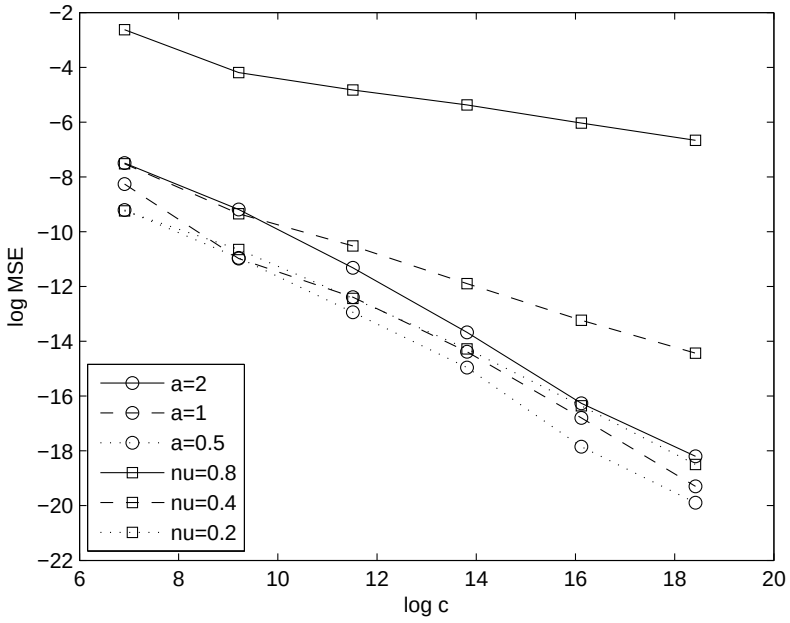

Figure 2: Estimates of MSE$((f_{m(c)}(A^{n(c)}_{m(c)}(x,\overline{\omega}),\overline{\omega}))$ for the steepest descent method when applied to the second problem instance as a function of computing budget c on a logarithmic scale for the policy $n(c) = a \log c$, with $a = 2$, 1, and 0.5 as well as the policy $n(c) = c^{\nu}$, with $\nu = 0.8$, 0.4, and 0.2.

consists of 20 independent and $[0, 1]$-uniformly distributed random variables. This problem instance is strongly convex with an unique global minimizer $x^* = (x_1^*, ..., x_{20}^*)'$, where $x_i^* = b_i/2$. The optimal value is $\sum_{i=1}^{20} a_i b_i^2/12 = 0.730706$ (rounded to six digits). We use initial solution $x = 0$.

The random function in this problem instance is continuously differentiable for all $\omega \in \Omega$ and we adopt the steepest descent method with Armijo step size rule as the optimization algorithm. This algorithm has at least a linear rate of convergence with rate of convergence coefficient $\theta = 1 - 4\lambda_{\min}\alpha(1-\alpha)\beta/\lambda_{\max}$, where $\alpha, \beta \in (0, 1)$ are Armijo step size parameters. We use $\alpha = 0.5$ and $\beta = 0.8$. Moreover, $\lambda_{\min}$ and $\lambda_{\max}$ are lower and upper bounds on the smallest and largest eigenvalues, respectively, of $\nabla^2 f(x)$ on $\mathbb{R}^d$. In this problem instance, we obtain that $\lambda_{\min} = 1.094895$ and $\lambda_{\max} = 1.991890$. Hence, the steepest descent method with Armijo step size rule satisfies Assumption 4 with $\theta = 0.56$.

Figure 2 illustrates Theorem 3 and displays MSE$((f_{m(c)}(A^{n(c)}_{m(c)}(x,\overline{\omega}),\overline{\omega}))$ for the steepest descent method when applied to the second problem instance as a function of computing budget c using the policy $n(c) = a \log c$, with $a = 2$, 1, and 0.5 (marked with circles) and the alternative policy $n(c) = c^{\nu}$, with $\nu = 0.8$, 0.4, and 0.2 (marked with boxes). As in the previous subsection, we consider $c = 10^3$, 10^4, 10^5, 10^6, 10^7, 10^8. The lines in Figure 2 marked with circles have slope quite close to -1, i.e., the MSE decays with rate of order c^{-1},

which is as predicted in Theorem 2. We see that the MSE decreases for smaller a. For the largest values of c examined, it appears that $a = 1$ has a slightly larger rate of decay than in the case of $a = 2$ and $a = 0.5$. These empirical results are in close correspondence with the asymptotic results of Theorem 2, which predict an improving rate of decay for decreasing a for $a \geq (2\log(1/\theta))^{-1}$, and a worsening of the rate for $a < (2\log(1/\theta))^{-1}$. In view of the above value of θ, $(2\log(1/\theta))^{-1} = 0.86$.

In the case of the alternative policy $n(c) = c^{\nu}$ (see Proposition 2), we see from the lines marked with squares in Figure 2 that the rate of decay of the MSE improves as ν decreases. However, the rate remains worse than the ones obtained using the policy $n(c) = a\log c$. These observations are consistent with the asymptotic results of Proposition 2 and Theorem 2: The policy $n(c) = c^{\nu}$ improves as ν tends to zero, but remains inferior to the policy $n(c) = a\log c$, with a sufficiently large. However, as this alternative policy is independent of θ, it may be easier to use in practice.

6.3 Newton's Method

We illustrates Theorem 3 by applying Newton's method with Armijo step size to the second problem instances defined in the previous subsection. On this problem instance, Newton's method has quadratic rate of convergence and satisfies Assumption 5, with $\psi = 2$ and some $\theta \in (0, \infty)$, when the initial solution $x \in \mathbb{R}^d$ is sufficiently close to the global minimizer of $\mathbf{P}_m(\overline{\omega})$. We use the same parameters in Armijo step size rule and initial solution as in the previous subsection.

Figure 3 presents similar results as in Figures 1 and 2 and considers the policy $n(c) = a\log\log c$, with $a = 3, 2, 1.4$, and 1. We again see that the slopes of the lines are close to -1, which is as predicted by Theorem 3. We see that the MSE decreases as a decreases from 3 to 1.4. However, the rate of decays are quite comparable. When $a = 1$, the MSE worsens. These empirical results are aligned with the asymptotic results of Theorem 3, which stipulate an improving rate of decay for decreasing a for $a > 1/\log\psi$. The quadratic convergence of Newton's method implies that $\psi = 2$ and consequently that the critical value $1/\log\psi$ is approximately 1.4. Moreover, Theorem 3 predicts a worsening rate of decay for $a < 1/\log\psi$ as observed empirically.

Comparing Figures 1, 2, and 3, we see that the MSE decreases and the rate of decay of the MSE increases as faster optimization algorithms are utilized. The improvement is most significant when moving from a sublinearly to a linearly convergent optimization algorithm. These results are reasonable as a faster optimization algorithm allows for fewer iterations and a larger sample size as compared to a slower optimization algorithm. The improvement is only slight when moving from a linearly to a superlinearly convergent optimization algorithm

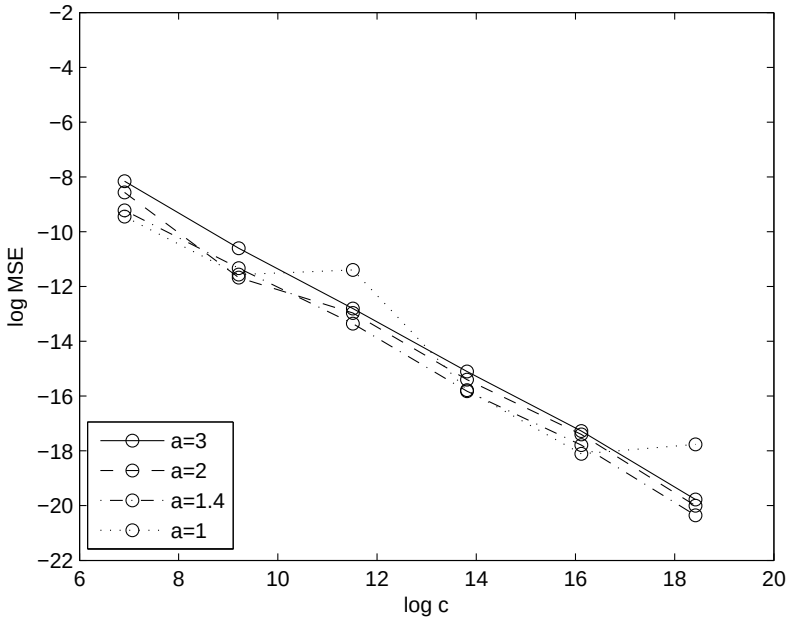

Figure 3: Estimates of MSE$((f_{m(c)}(A^{n(c)}_{m(c)}(x,\overline{\omega}),\overline{\omega}))$ for Newton's method when applied to the second problem instance as a function of computing budget c on a logarithmic scale for the policy $n(c) = a \log\log c$, with $a = 3$, 2, 1.4, and 1.

as it only allows a decrease in number of iterations from a value proportional to $\log c$ to a value proportional to $\log\log c$.

7 Conclusions

In this paper we characterize optimal computing budget allocation policies in the sample average approximation approach for solving stochastic programs. We find that in the case of a sublinearly convergent optimization algorithm for solving the sample average problem with rate of convergence of order n^{-p}, where n is the number of iterations and p is an algorithm specific parameter, the best achievable convergence rate is of order $c^{-p/(2p+1)}$. In the case of a linearly convergent optimization algorithm with rate of convergence of order θ^n for some parameter $\theta \in (0,1)$, the best overall convergence rate is of order $(c/\log c)^{-1/2}$. For a superlinearly convergent optimization algorithm with rate of convergence of order θ^{ψ^n}, where $\theta > 0$ and $\psi \in (0,\infty)$, the best convergence rate is of order $(c/\log\log c)^{-1/2}$. These rates are only obtained using particular policies for the selection of sample sizes and number of optimization iterations as identified in the paper. The policies depend on p in the sublinear case, on θ in the linear case, and on θ and ψ in the superlinear case. Other policies for sample size and number of iteration selection may result in substantially worse rates of decay as quantified in the paper. These results provide a detailed insight into the

challenging task of computing budget allocation within the sample average approximation approach and may spur further research into the development of efficient sampling-based algorithms for stochastic optimization.

Acknowledgement

This study is supported by the Air Force Office of Scientific Research Young Investigator and Optimization & Discrete Mathematics programs.

References

[1] S. Alexander, T.F. Coleman, and Y. Li. Minimizing CVaR and VaR for a portfolio of derivatives. *J. Banking & Finance*, 30:583–605, 2006.

[2] S. Andradottir. An overview of simulation optimization via random search. In *Elsevier Handbooks in Operations Research and Management Science: Simulation*, pages 617–631. Elsevier, 2006.

[3] R. R. Barton and M. Meckesheimer. Metamodel-based simulation optimization. In *Elsevier Handbooks in Operations Research and Management Science: Simulation*, pages 535–574. Elsevier, 2006.

[4] F. Bastin, C. Cirillo, and P.L Toint. An adaptive monte carlo algorithm for computing mixed logit estimators. *Computational Management Science*, 3(1):55–79, 2006.

[5] G. Bayraksan and D.P. Morton. A sequential sampling procedure for stochastic programming. *Operations Research*, to appear, 2011.

[6] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, New York, 1968.

[7] C.H. Chen, M. Fu, and L. Shi. Simulation and optimization. In *Tutorials in Operations Research*, pages 247–260, Hanover, Maryland, 2008. INFORMS.

[8] C.H. Chen, D. He, M. Fu, and L. H. Lee. Efficient simulation budget allocation for selecting an optimal subset. *INFORMS J. Computing*, 20(4):579–595, 2008.

[9] Y. L. Chia and P.W. Glynn. Optimal convergence rate for random search. In *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*. Available at www.informs-sim.org/2007informs-csworkshop/2007workshop.html, 2007.

[10] Y. Ermoliev. Stochastic quasigradient methods. In *Numerical Techniques for Stochastic Optimization*, Yu. Ermoliev and R.J-B. Wets (Eds.), New York, New York, 1988. Springer.

[11] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, December 2010.

[12] D. He, L. H. Lee, C.-H. Chen, M.C. Fu, and S. Wasserkrug. Simulation optimization using the cross-entropy method with optimal computing budget allocation. *ACM Transactions on Modeling and Computer Simulation*, 20(1):133–161, 2010.

[13] J. L. Higle and S. Sen. *Stochastic Decomposition: A Statistical Method for Large Scale Stochastic Linear Programming*. Springer, 1996.

[14] G. Infanger. *Planning Under Uncertainty: Solving Large-scale Stochastic Linear Programs*. Thomson Learning, 1994.

[15] P. Kall and J. Meyer. *Stochastic Linear Programming, Models, Theory, and Computation*. Springer, 2005.

[16] S. H. Kim and B. Nelson. Selecting the best system. In *Elsevier Handbooks in Operations Research and Management Science: Simulation*, pages 501–534. Elsevier, 2006.

[17] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.

[18] G. Lan. *Convex Optimization Under Inexact First-order Information*. Phd thesis, Georgia Institute of Technology, Atlanta, Georgia, 2009.

[19] S.-H. Lee and P.W. Glynn. Computing the distribution function of a conditional expectation via monte carlo: Discrete conditioning spaces. *ACM Transactions on Modeling and Computer Simulation*, 13(3):238–258, 2003.

[20] J. Linderoth, A. Shapiro, and S. Wright. The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research*, 142:215–241, 2006.

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

[22] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optimization*, 19(4):1574–1609, 2009.

[23] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer Academic, Boston, Massachusetts, 2004.

[24] S. Olafsson. Metaheuristics. In *Elsevier Handbooks in Operations Research and Management Science: Simulation*, pages 633–654. Elsevier, 2006.

[25] R. Pasupathy. On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. *Operations Research*, 58:889–901, 2010.

[26] E. Polak and J. O. Royset. Efficient sample sizes in stochastic nonlinear programming. *J. Computational and Applied Mathematics*, 217:301–310, 2008.

[27] R.T Rockafellar and J. O. Royset. On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety*, 95:499–510, 2010.

[28] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26:1443–1471, 2002.

[29] J. O. Royset. Optimality functions in stochastic programming. *Mathematical Programming*, to appear, 2011.

[30] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory.* Society of Industrial and Applied Mathematics, 2009.

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

[32] H. Xu and D. Zhang. Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. *Mathematical Programming*, 119:371–401, 2009.