Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

ON ADAPTIVE SAMPLING RULES FOR STOCHASTIC RECURSIONS

Fatemeh S. Hashemi

The Grado Dept of Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061, USA

Soumyadip Ghosh

T.J. Watson IBM Research Center
Yorktown Heights, NY 10598, USA

Raghu Pasupathy

Department of Statistics
Purdue University
West Lafayette, IN 47905, USA

ABSTRACT

We consider the problem of finding a zero of an unknown function, or optimizing an unknown function, with only a stochastic simulation that outputs noise-corrupted observations. A convenient paradigm to solve such problems takes a deterministic recursion, e.g., Newton-type or trust-region, and replaces function values and derivatives appearing in the recursion with their sampled counterparts. While such a paradigm is convenient, there is as yet no clear guidance on how much simulation effort should be expended as the resulting recursion evolves through the search space. In this paper, we take the first steps towards answering this question. We propose using a fully sequential Monte Carlo sampling method to adaptively decide how much to sample at each point visited by the stochastic recursion. The termination criterion for such sampling is based on a certain relative width confidence interval constructed to ensure that the resulting iterates are consistent, and efficient in a rigorous (Monte Carlo canonical) sense. The methods presented here are adaptive in the sense that they "learn" to sample according to the algorithm trajectory. In this sense, our methods should be seen as refining recent methods in a similar context that use a predetermined sequence of sample sizes.

1 INTRODUCTION AND MOTIVATION

We study adaptive sampling within stochastic recursions involving quantities estimated using a stochastic simulation. The prototypical example setting is Simulation Optimization (SO) (Henderson and Nelson 2006, Pasupathy and Ghosh 2013), where an optimization problem $\min_{\theta\in\Theta} g(\theta)$, $g:\mathbb{R}^d\to\mathbb{R}$, is to be solved using only a stochastic simulation capable of providing estimates of the objective function and constraints at a requested point. Another closely related example setting is the Stochastic Root Finding Problem (SRFP) (Pasupathy and Kim 2011, Pasupathy 2010, Pasupathy and Schmeiser 2009), where the zero of a vector function $h(\theta)$ is sought over the feasible domain Θ, with only simulation-based estimates of the function involved. (A first-order optimality condition for an SO problem is that the gradient of the function $\nabla_\theta g(\theta) = f(\theta)$ matches zero, which is an SRFP.) SO and SRFPs have recently generated great attention because they allow the function involved in the problem to have an implicit representation through a stochastic simulation, thereby allowing the embedding of virtually any level of complexity. Such flexibility has resulted in widespread adoption. Examples include logistics (Homem-de-Mello, Shapiro, and Spearman 1999, Atlason, Epelman, and Henderson 2008), healthcare (Alagoz, Schaefer, and Roberts 2009, Deng and Ferris 2006), and traffic (Osorio and Bierlaire 2013, Lu and Li 2009).

A recently popular solution paradigm for solving SO and SRFPs simply mimics what an algorithm might do within a deterministic context, after estimating any needed function and derivative values using the available stochastic simulation. An example serves to illustrate such a technique best. Consider the Newton-type recursion

$$\theta_{k+1} = \theta_k - H_g^{-1}(\theta_k)\nabla g(\theta_k) \tag{1}$$

for finding a first-order critical point of the real-valued function $g : \mathbb{R}^d \to \mathbb{R}$, recalling that $H_g(\cdot)$ in (1) is an approximation of the Hessian of g. If only "noisy" simulation-based estimates of g are available, then a reasonable adaptation of (1) might be to use the recursion

$$\theta_{k+1} = \theta_k - \tilde{H}_g^{-1}(m_k, \theta_k)\tilde{\nabla} g(m_k, \theta_k) \tag{2}$$

where $\tilde{\nabla} g(m_k, \theta_k)$ and $\tilde{H}_g^{-1}(m_k, \theta_k)$ are the simulation-based approximations of the gradient and Hessian of g respectively, obtained with simulation effort m_k. The simulation effort m_k is general and might represent the number of replications in the case of terminating simulations or the simulation run length in the case of non-terminating simulations (Law 2007).

Effectively implementing a recursion of the sort (2) relies crucially on the choice of the sample size sequence $\{m_k\}$. To understand the trade-offs involved in such choice, note that the error in the iterates $\{\theta_k\}$ generated by (2) has two sources. The first, *structural error*, arises due to the mechanics of the employed recursion. Structural error is not specific to the simulation context and arises in any recursive approximation setting. The second, *sampling error*, arises as a result of the inherent stochasticity of the simulation output, and is hence directly related to the choice of $\{m_k\}$. Guaranteeing that the recursion (2) produces iterates that converge to the correct solution stipulates adequate sampling, that is, the sequence $\{m_k\}$ should be so large that the observations do not lead the iterates astray per chance. At the same time, choosing the sample size m_k too large would be inefficient since this would mean that the resulting sampling error will be small compared to the structural error. Thus it seems likely that consistency and efficiency should dictate an "optimal regime" for the sample sizes $\{m_k\}$.

Recent work by (Pasupathy, Glynn, Ghosh, and Hashemi 2014) explores the notion of optimal sample size regimes for stochastic recursions such as (2), by introducing and analyzing the broader context of Sampling-Controlled Stochastic Recursions (SCSR) for finding the zero of an unknown function $h : \mathbb{R}^d \to \mathbb{R}^d$:

$$\theta_{k+1} = \theta_k - \frac{1}{\beta}\tilde{h}_{m_k}(\theta_k) \tag{SCSR}$$

where $\tilde{h}_m$ is the estimator of h, and β is some constant chosen based on some prior informtaion on curvature of the function h. (Pasupathy, Glynn, Ghosh, and Hashemi 2014) characterize the rates at which the sample sizes $\{m_k\}$ should increase in (SCSR) in order to guarantee the consistency and efficiency of the resulting iterates. In particular, they demonstrate that the speed of the underlying recursive function h, its estimator $\tilde{h}_{m_k}$, and the optimal regime of sample sizes are intimately linked, with faster recursions allowing for a wider range of sample sizes while remaining efficient. For instance, (Pasupathy, Glynn, Ghosh, and Hashemi 2014) show that when linearly converging recursions are employed, certain geometrically increasing sample size sequences are efficient; likewise, when superlinearly converging recursions are employed, all geometrically increasing sample size sequences, and certain super-exponential sample-size sequences are efficient.

Remark 1 The celebrated Stochastic Approximation (Kushner and Yin 2003) method is a competitor to what we propose here but falls within the purview of SCSR. A restriction within SA is that all sample sizes $m_k = m$, that is, a fixed sample size is used across iterations. The literature on SA is enormous, and its success somewhat mixed (Pasupathy and Ghosh 2013) due to the presence of certain algorithmic sequences that need to be chosen by the user. Theoretical prescriptions for the choice of this sequence in order to preserve the correct direction through the search space and the fast convergence towards the true

solution, is throroughly discussed in the literature (Hashemi and Pasupathy 2012, Spall 2000, Spall 2012). Meanwhile there are continuing efforts (Broadie, Cicek, and Zeevi 2010, Broadie, Cicek, and Zeevi 2009) to make SA implementable by devising schemes that automatically choose these algorithmic sequences.

The sampling regimes characterized in (Pasupathy, Glynn, Ghosh, and Hashemi 2014) constitute an important step towards making stochastic recursions implementable, since they serve to provide some guidance on sampling. Such guidance is only broad, however, and still leaves a lot of room for choice. For instance, when using linearly converging recursions, (Pasupathy, Glynn, Ghosh, and Hashemi 2014) demonstrate that choosing $\{m_k\}$ such that $m_k/m_{k-1} = \gamma$, for all $k \geq 2$ produces iterates that are consistent and efficient (in a certain rigorous sense) as long as $\gamma \in (0, c)$, where $c > 1$ is a constant. This directive is useful but, depending on the specific problem, different choices of γ, or even varying γ across iterations, may be needed to produce robust algorithm performance. In general, good algorithm performance in finite time entails inferring and reacting to specific problem structure, perhaps by using the trajectory of algorithm iterates and their corresponding function estimates. Regardless of the problem structure, the analysis in (Pasupathy, Glynn, Ghosh, and Hashemi 2014) is asymptotic, leaving an enormous range of possible choices of $\{m_k\}$ that still guarantee efficiency.

Can sample sizes $\{m_k\}$ be chosen adaptively, by reacting to function information that is obtained as the iterates evolve through the search space? Moreover, can such adaption happen in way that also ensures consistency and efficiency in the rigorous sense of (Pasupathy, Glynn, Ghosh, and Hashemi 2014)? There have been some recent proposals in the literature towards answering this question. For instance, Byrd et al. (2012) propose the following two-stage sampling procedure to determine the sample size at any iteration k:

$$M_k = \frac{\alpha \hat{\sigma}^2_{M_{k-1}}(\theta_k)}{\|\tilde{h}_{M_{k-1}}(\theta_k)\|^2}, \tag{3}$$

where the estimate $\tilde{h}_{M_{k-1}}$ and its variance $\hat{\sigma}^2_{M_{k-1}}$ are constructed from the M_{k-1} samples gathered in the earlier iteration. The expression in (3) can be interpreted as the result of balancing the squared bias and the variance of the function estimator; it can also be interpreted as the minimum sample size required to declare with some certainty that the function value $h(\theta_k)$ at the current iterate θ_k has been estimated to a sufficient level of accuracy to rule out θ_k being the solution. Byrd et al. (2012) show that under the sampling rule (3), the resulting iterates converge to a zero of h and the samples M_k grow geometrically. The proof for this convergence requires a strong condition (Eq 4.20), in part because of the two-stage nature of the procedure, and it is unclear how such a condition can be checked a priori.

A competing fully *sequential* rule proposed by Pasupathy and Schmeiser (2010) has the following form:

$$M_k = \inf_m : \ a_m \frac{\hat{\sigma}_m(\theta_k)}{\sqrt{m}} < \alpha \|\tilde{h}_m(\theta_k)\|, \quad \alpha > 0, \tag{4}$$

(A simpler version of (4) was proposed in Anscombe (1953) within the context of estimating a confidence interval on the mean.) Pasupathy and Schmeiser (2010) conjecture that the use of the fully sequential stopping rule (4) in (SCSR) results in convergent and asymptotically efficient iterates.

1.1 Contributions

We investigate the use of adaptive sampling within stochastic recursions (SCSR) for solving SO and SRFPs. The adaptive sampling schemes we introduce are a fully sequential version of (3), and are constructed to balance the estimated variance and squared bias of the (recursive) function estimates at each visited point. There is emerging evidence that schemes similar to what we propose work well in practice and come closer to the goal of achieving robust finite-time performance with no user-intervention. However, the analysis of such fully adaptive schemes turns out to be challenging, and there appears to be no clear analysis of the consistency and efficiency of the resulting iterates to date. In this paper, we present two results that

take us closer to the construction of provably consistent and efficient adaptive sampling schemes within stochastic recursions.

(1) We first analyze a simple adaptive sampling rule similar to (4) obtained by replacing the $\tilde{h}_m(\theta_k)$ on the right-hand side with a geometrically decreasing deterministic sequence γ^{-k}, for some $\gamma \in (1, \bar{\gamma})$, where $\bar{\gamma}$ is defined based on some prior curvature information . We show that under such a rule, the iterates converge efficiently. This result is a slight generalization of the geometric sample growth rates that are shown to be efficient in SCSRs ((Pasupathy, Glynn, Ghosh, and Hashemi 2014)), allowing the sample sizes m_k to also react to local estimation conditions ($\sigma_m(\theta_k)$).

(2) We next analyze a version of the adaptive sampling rule (4) that replaces $\tilde{h}_m(\theta_k)$ on the right-hand side with the actual function value $h(\theta_k)$. Our proposed scheme adapts to the local conditions of the iterations of SCSR by determining the amount of sampling needed solely based on the relative accuracy of the function estimate at the current iterate. We start with known results for sequential estimation methods, first described by (Chow and Robbins 1965) for iid populations with unknown variance, and extend their analysis to show that SCSR augmented with the proposed sequential sampling rule is asymptotically efficient.

2 NOTATION AND BASICS

We will adopt the following notation through the section. (i) For a sequence of random variables $\{X_n\}$, we say $X_n \xrightarrow{\text{p}} X$ if $\{X_n\}$ converges to X in probability; similarly, we say $X_n \xrightarrow{\text{d}} X$ to mean that $\{X_n\}$ converges to X in distribution, and finally $X_n \overset{\text{wp1}}{\to} X$ to mean that $\{X_n\}$ converges to X with probability one. we say $a_n = o(1)$ if $\lim_{n\to\infty} a_n = 0$; and $a_n = O(1)$ if $\{a_n\}$ is bounded, i.e., $\exists c \in (0,\infty)$ with $|a_n| < c$ for large enough n. We say that $a_n = \Theta(1)$ if $a_n = O(1)$ but a_n is not $o(1)$. (v) For a sequence of real numbers $\{a_n\}$, we say $a_n = o_p(1)$ if $a_n \to 0$ as $n \to \infty$; and $a_n = O_p(1)$ if $\{a_n\}$ is stochastically bounded, that is, for given $\varepsilon > 0$ there exists $c(\varepsilon) \in (0,\infty)$ with $\Pr\{|a_n| < c(\varepsilon)\} > 1-\varepsilon$ for large enough n. We say that $a_n = \Theta_p(1)$ if $a_n = O_p(1)$ but a_n is not $o_p(1)$.

Also, the following notion will help our exposition.

Definition 1 *(Growth rate of a sequence.)* A sequence $\{m_k\}$ is said to exhibit Geometric(c) growth if $m_{k+1} = c\,m_k, k = 1,2,\ldots$ for some $c \in (1,\infty)$,

We place the following standing conditions on the function $h(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$ of interest to the SRFP problem of determining a $\theta^* \in \Theta$ that satisfies $h(\theta^*) = \bar{0}$.

Assumption 2.1 The function $h(\theta)$ satisfies the following:

A1 There exists a unique $\theta^* \in \Theta$ such that $h(\theta^*) = \bar{0}$,
A2 for all $\theta \in \Theta$, $(\theta - \theta^*)^T h(\theta) \geq l_0 \|\theta - \theta^*\|^2$,
A3 h is locally Lipschitz continuous at θ^*, that is, there exists $l_1 > 0$, such that for all $\theta \in \Theta$, $\|h(\theta)\| \leq l_1 \|\theta - \theta^*\|$.

Analogously, the function $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ of interest to the SO problem $\{\min_\theta f(\theta)\}$ is assumed to satisfy the following.

Assumption 2.2 (*Strong Convexity*) The function $f(\theta)$ is twice continuously differentiable and there exist constants $0 < \lambda < \beta$, such that

$$\lambda \|u\|^2 \leq u^T \nabla h(\theta) u \leq \beta \|u\|^2, \text{ for all } \theta \text{ and } u. \tag{5}$$

In the context of SRFPs, the function h of Assumption 2.1 relates to the function f of Assumption 2.2 as $\nabla f(\theta) = h(\theta)$, and the conditions can be verified to yield the same SCSR error structure.

The stochastic recursion (SCSR) requires an estimate of the function $h(\theta)$. Assume that i.i.d. observations $Y_1, Y_2, \ldots$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ are available such that their (unknown) mean $\mathbb{E}[Y_i(\theta)] = h(\theta)$,

and the (unknown) non-singular $(d \times d)$-covariance matrix Σ, with $\text{tr}(\Sigma) \leq \sigma^2$ for a finite positive constant σ. The estimator for $h(\theta)$ from m copies of these observations is

$$\tilde{h}_m(\theta) = m^{-1} \sum_{i=1}^{m} Y_i(\theta),$$

The following linear transformation of the sample covariance matrix will play an important role in the sequel:

$$\hat{\sigma}_m^2 = \text{tr}(\frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \tilde{h}_m)(Y_i - \tilde{h}_m)^T).$$

Finally, we call a stochastic recursion asymptotically efficient if the following holds.

Definition 2 *(Asymptotically Efficient)* Denote $\Gamma_k := \sum_{i=1}^{k} M_i$ as the total samples used up till the kth iteration. If there exists a sequence $\{\nu_k\}$ such that $\Gamma_k = O_p(\nu_k)$, then a stochastic recursion with iterations defined by (SCSR) converges asymptotically efficiently if

$$\mathbb{E}[h(\theta_k)] \;=\; \mathbb{E}[h(\theta_k) - h(\theta^*)] \;=\; O(\nu_k^{-1}). \tag{6}$$

Pasupathy et al. (2014) show in Theorem 6.1 that this rate is the fastest that any sampling-controlled stochastic recursion (SCSR) under the assumed conditions can achieve. We anticipate that this result will be true for stochastic recursions with dynamic random sample sizes M_k; this paper assumes this in the definition of asymptotic efficiency.

3 MAIN RESULTS

In this section we investigate the behavior of sampling-controlled stochastic recursions

$$\theta_{k+1} = \theta_k - \frac{1}{\beta} \tilde{h}_{m_k}(\theta_k) \tag{SCSR}$$

when augmented with sequential rules for choosing the sample size m_k. Under each of the two rules we consider, the sample size is, conditional on the current iterate θ_k, a random variable. We shall use the notation M_k to emphasize this distinction. The random variable M_k is a stopping time adapted to the sequence $\{Y_i\}$ in both methods, and is determined as the lowest sample size that matches the confidence interval of the estimate $\tilde{h}_{M_k}$ to a target value. A measure of the squared half-width of the confidence interval is $\sigma^2_{M_k}/M_k$.

The first stopping rule matches the confidence interval widths to a pre-specified sequence of target values γ_k, where $\gamma_k \to 0$ and $k \to \infty$, but $\sum_k \gamma_k < \infty$. Theorem 1 shows that the sequence of iterates $\{\theta_k\}$ converges to θ^* a.s. under these mild conditions on γ_k. Furthermore, if γ_k were to grow geometrically, then the recursion is asymptotically work-efficient for all geometric growth factors up to a finite upper bound. The main tool used for analysis of the proposed sequential sampling rules embedded in stochastic recursions, is to study asymptotic theories for randomly stopped random sequences, traces back to (Anscombe 1953).

Theorem 1 Let $\{\gamma_k\}_{k\geq 1}$ be a fixed positive sequence for which we have $\sum_{i=1}^{\infty} \gamma_i < \infty$. Let the function $h(\cdot)$ satisfy Assumption 2.1,

(i) Let for $\alpha > 0$, β in (SCSR) satisfy $\frac{(1+\alpha)l_1^2}{2l_0} < \beta < \infty$. Denote $\{M_k\}_{k\geq 1}$ as a sequence of random variables in the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Considering the stochastic recursion

$$\theta_{k+1} = \theta_k - \frac{1}{\beta} \tilde{h}_{M_k}(\theta_k),\ k = 1, 2, \ldots, \tag{7}$$

denote the history of the method up to time k by

$$\mathscr{F}_k = \{\theta_0, (M_1, S_{M_1}), (M_2, S_{M_2}), \ldots, (M_{k-1}, S_{M_{k-1}})\},$$

where $S_j = (Y_1, Y_2, \ldots, Y_j)$. Assume either of the following holds.

(R1) Y_1 is normally distributed, and $M(\theta_k) = \inf\{m > 3 : \frac{\hat{\sigma}^2_m(\theta_k)}{m} < \alpha\gamma_k | \mathscr{F}_k\}$, $k = 1, 2, \ldots$;

(R2) $\mathbb{E}[Y_1^6] < \infty$, and for $0 < \alpha < 1$, $M(\theta_k) = \inf\{m > \max(2, [\alpha\gamma_k]^{-1/2} + 1) : \frac{\hat{\sigma}^2_m(\theta)}{m} < \alpha\gamma_k | \mathscr{F}_k\}$.

Then (7) satisfies $\theta_{k+1} \to \theta^*$ a.s..

(ii) Further, letting $\gamma_k^{-1} = [\gamma]^k$, $1 < \gamma \leq (1 - \frac{2l_0}{\beta} + \frac{(1+\alpha)l_1^2}{\beta^2})^{-1}$, the algorithm is asymptotically efficient.

The following Lemma is used in proving Theorem 1.

Lemma 3.1 Let X_is , $i = 1, \ldots$ be iid observations from $N(\mu, \Sigma)$, whose m-sample mean is denoted by $\mathscr{Z}_m = 1/m \sum_{i=1}^m X_i$. Consider the following sequential procedure:

$$M_c = \inf\{m \geq 1 : \frac{\hat{\sigma}^2_m}{m} < c\}, \tag{8}$$

where σ^2_m is the trace of the sample covariance matrix, c is a positive constant that is allowed to approach zero. Letting $\sigma^2 = \mathrm{tr}(\Sigma)$, we have

(i) M_c is a stopping time with respect to $\{X_i\}_{1 \leq i \leq m}$, and $\lim_{c \to 0} (\frac{c}{\sigma})^2 M_c = 1$ a.s.;

(ii) for the stopped process $\mathscr{Z}_{M_c}$ we have $\mathrm{Var}(\mathscr{Z}_{M_c}) = \mathbb{E}[\sigma^2 M_c^{-1}]$.

Proof. First we prove that M_c is a well-defined stopping time with respect to $\{X_i\}_{1 \leq i \leq m}$. Consider the stochastic process $X = \{X_m : m \in \mathbb{N}_\}$ defined on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. We note that M as defined in (8) is a random time for the stochastic process $X = \{X_m : m \geq 1\}$, as M_c is a discrete random variable on the same probability space as X. For $m \in \mathbb{N}$, let $\mathscr{F}_m = \sigma\{X_s, s \in \mathbb{N}, s \leq m\}$, be the σ-algebra of events up to time m. The random time M_c, is a stopping time since

$$\{M_c > m\} \in \mathscr{F}_m,$$

for each $m \in \mathbb{N}$. In other words, $\{M_c = m\}$, is $\mathscr{F}_m$ measurable, which means that the event $\{M_c = m\}$ is completely determined by the total information up to time m, $\{X_1, X_2, \ldots, X_m\}$, and is not dependent on the future $X_{m+1}, X_{m+2}, \ldots$.

Then part (i) follows by Lemma 2 in (Chow and Robbins 1965). For the second part, we note that the probability distribution of M_c is defined for any $m \geq 1$ by

$$P(M_c = m) = P\{\hat{\sigma}^2_k < ck \text{ for } k = m \text{ but not for any } k < m\}. \tag{9}$$

Since X_is are normally distributed, we know from Basu (1955) that $\hat{\sigma}^2_m$ is statistically independent of $\mathscr{Z}_m$. Therefore

$$P(M = m | \mathscr{Z}_M) = P(M = m). \tag{10}$$

Hence the event $\{M = m\}$ is independent of $\mathscr{Z}_M$, and so

$$\begin{aligned} \mathrm{Var}(\mathscr{Z}_M) &= \mathbb{E}[\mathrm{Var}(\mathscr{Z}_M | M = m)] + \mathrm{Var}(\mathbb{E}[\mathscr{Z}_M | M = m]) \\ &= \mathbb{E}[\sigma^2 M^{-1}]. \end{aligned}$$

□

Proof of Theorem 1(i). First we find a finite time upper bound on the squared error. To this end, by (7), letting $Z_k = \theta_k - \theta^*$, we have for all k,

$$\begin{aligned} Z_{k+1}^2 &= Z_k^2 - \frac{2}{\beta} Z_k^T \tilde{h}_{M_k}(\theta_k) + \frac{1}{\beta^2}\|\tilde{h}_{M_k}(\theta_k)\|^2, \\ &= Z_k^2 - \frac{2}{\beta} Z_k^T (\tilde{h}_{M_k}(\theta_k) - h(\theta_k)) - \frac{2}{\beta} Z_k^T h(\theta_k) + \frac{1}{\beta^2}\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|^2 \\ &\quad + \frac{1}{\beta^2}\|h(\theta_k)\|^2 + \frac{2}{\beta^2} h(\theta_k)^T (\tilde{h}_{M_k}(\theta_k) - h(\Theta_k)). \end{aligned}$$

By Assumption 2.1(A2), we have

$$\begin{aligned} \mathbb{E}_\Omega[Z_{k+1}^2|\mathscr{F}_k] &= Z_k^2 - \frac{2}{\beta} Z_k^T h(\theta_k) + \frac{1}{\beta^2}\|h(\theta_k)\|^2 + \frac{1}{\beta^2}\mathbb{E}_\Omega[\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|^2|\mathscr{F}_k], \\ &\leq Z_k^2 - \frac{2l_0}{\beta} Z_k^2 + \frac{l_1^2}{\beta^2} Z_k^2 + \frac{1}{\beta^2}\mathbb{E}_\Omega[\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|^2|\mathscr{F}_k], \\ &= (1 - \frac{2l_0}{\beta} + \frac{l_1^2}{\beta^2}) Z_k^2 + \frac{1}{\beta^2}\mathbb{E}_\Omega[\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|^2|\mathscr{F}_k]. \end{aligned} \tag{11}$$

Letting $a := 1 - \frac{2l_0}{\beta} + \frac{l_1^2}{\beta^2}$, since $\beta > \frac{l_1^2}{2l_0}$, $0 < a < 1$.

Under (R1), by Lemma 3.1 part (ii) we have

$$\mathbb{E}_\Omega[(\tilde{h}_{M_k}(\theta_k) - h(\theta_k))^2|\mathscr{F}_k] = \mathrm{tr}(\Sigma)\mathbb{E}_\Omega[M_k^{-1}|\mathscr{F}_k].$$

Hence by (11) and Theorem 3_* of (Starr 1966), when $\gamma_k \to 0$,

$$\begin{aligned} \mathbb{E}_\Omega[Z_{k+1}^2|\mathscr{F}_k] &\leq a Z_k^2 + \frac{\sigma^2}{\beta^2}\mathbb{E}_\Omega[M_k^{-1}|\mathscr{F}_k], \\ &\leq a Z_k^2 + \frac{\alpha}{\beta^2}\gamma_k. \end{aligned} \tag{12}$$

Since $\sum_{i=1}^\infty \gamma_k < \infty$, by Lemma 2 in (Yousefian, Nedić, and Shanbhag 2012), we have that $\theta_k \to \theta^*$a.s..

Under (R2), by (11) and the relation (2.4) in (Mukhopadhyay and Datta 1996), (12) holds, and the claim follows accordingly.

Proof of Theorem 1(ii)

By Assumptions 2.1(A2) and (A3), we have

$$l_0^2 z_k^2 \leq \|h(\theta_k)\|^2 \leq l_1^2 z_k^2. \tag{13}$$

Considering $\Gamma_k = O_p(\sum_{i=1}^k [a + \alpha(\frac{l_1}{\beta})^2]^{-i})$, asymptotic efficiency of (7) follows by (12), corollary 4.3 of (Byrd, Chin, Nocedal, and Wu 2012) and Definition 2. □

Theorem 1 closely matches a result for SCSRs under the same conditions; Theorem 6.5 in (Pasupathy, Glynn, Ghosh, and Hashemi 2014) shows that pre-determined sample sizes m_k that grow geometrically with the same growth rate restrictions as Theorem 1(ii) are asymptotically efficient. Note that the sequential sampling rule introduced in Theorem 1 results in larger samples than the lower bound of the geometrically growing γ_k^{-1}, and is sensitive to the quality of the estimator $\tilde{h}_m$ at the current estimate. This sequential stopping rule is easy to implement given the chosen sequence $\{\gamma_k^{-1}\}$ since the update of the variance estimator $\tilde{\sigma}_m$ is a constant-computational-effort operation. However, we do not have a truly hands-off method yet since the user needs to still pick the geometric growth factor γ carefully to ensure efficiency.

However note that the random sample size $M_k \sim O_p(\gamma_k^{-1})$, and the sequences $\{\gamma_k^{-1}\}$ that were judged efficient grow exactly as the inverse of $E[h(\theta_k)]$ by Theorem 1(ii). This motivates the next sampling rule 14, which will sample till the sampling error at the current iterate is just smaller than the optimality gap of the iterate. Will SCSR augmented with such sampling rule be efficient ?"

Accordingly, we introduce another sequential procedure that replaces the sequence $\{\gamma_k\}$ with purely local information:

$$M_k = \inf\{m > \max\{3, \gamma_k\} : \ \frac{\hat{\sigma}_m^2(\theta_k)}{m} < \alpha \|h(\theta_k)\|^2\}, \tag{14}$$

for $\alpha > 0$, Theorem 2 analyses the asymptotic behavior of the corresponding SCSR method.

Theorem 2 The function $h(\cdot)$ satisfies Assumption 2.1. Let $\{\gamma_k\}_{k\geq 1}$ be a fixed positive sequence for which we have $\sum_{i=1}^{\infty} 1/\gamma_i < \infty$. Denote $\{M_k\}_{k\geq 1}$ as a sequence of random variables in probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Considering the following stochastic recursion,

$$\theta_{k+1} = \theta_k - \frac{1}{\beta}\tilde{h}_{M_k}(\theta_k), \ k = 1, 2, \ldots, \tag{15}$$

denote the history of the method up to time k by

$$\mathscr{F}_k = \{\theta_0, (M_1, S_{M_1}), (M_2, S_{M_2}), \ldots, (M_{k-1}, S_{M_{k-1}})\},$$

where $S_j = (Y_1, Y_2, \ldots, Y_j)$. Given $\mathscr{F}_k$. Assume either of the following conditions hold.

(C1) (Parametric Setup) Y_1 is normally distributed, and for $0 < \alpha < 1$, $M_k := M(\theta_k) = \inf\{m > \max\{3, \gamma_k\} : \frac{\hat{\sigma}_m^2(\theta_k)}{m} < \alpha \|h(\theta_k)\|^2 | \mathscr{F}_k\}$;

(C2) (Nonparametric Setup) Let $\mathbb{E}[Y_1^8] < \infty$, and for $0 < \alpha < 1/4$, $\zeta > 0$, $M_k := M(\theta_k) = \inf\{m > \max\{3, \gamma_k\} : \frac{\hat{\sigma}_m(\theta_k)}{m} + m^{-(1+\alpha)} < \zeta \|h(\theta_k)\|^2 | \mathscr{F}_k\}$.

Then the SCSR iterates (15) are (a) almost surely convergent to the true solution θ^*, and (b) asymptotically efficient.

Proof. First we note that under either $(C1)$, or $(C2)$, by Lemma 3.1, M_k is a stopping time with respect to $\mathscr{F}_k$.

Under $(C1)$, a.s. convergence follows by slight changes in Theorem 1. For efficiency of (15), by (12) we get

$$\mathbb{E}_\Omega[Z_{k+1}^2 | \mathscr{F}_k] \ \leq \ aZ_k^2, \tag{16}$$

where $a := 1 - \frac{2l_0}{\beta} + \frac{(1+\alpha)l_1^2}{\beta^2}$, and by $\beta > \frac{(1+\alpha)l_1^2}{2l_0}$, $0 < a < 1$. Hence letting $b_k := \mathbb{E}[Z_k^2]$, $q_k := \frac{1}{\beta^2}\mathbb{E}_\Omega[\|\tilde{H}_{M_k}(\theta_k) - h(\theta_k)\|^2]$ and $d_k := b_1(1 - \frac{2l_0}{\beta})^k + \sum_{i=2}^{k-1}(1 - \frac{2l_0}{\beta})^{k-i}q_i + q_k$, and

$$\eta = \max(b_{k_0}a^{1-k_0}, \max_{1\leq k\leq k_0}\{a^{-k}d_k\}),$$

for all $k \geq 1$, we have $b_k \leq \eta a^k$.

Since for all k, $b_k \leq a\eta$, Z_k^2 is uniformly integrable, and so is $\|h(\theta_k)\|^2$, by (13). Hence $\mathbb{E}\|h(\theta_k)\|^2$ approaches to zero with the same geometric rate as b_k.

Moreover since for all k, $\Pr(M_k < \infty) = 1$, we have

$$\begin{aligned} \mathbb{E}[\|h(\theta_k)\|^2 M_k] &= \mathbb{E}[\mathbb{E}[\|h(\theta_k)\|^2 M_k | \mathscr{F}_k]], \\ &= \mathbb{E}\|h(\theta)\|^2 \mathbb{E}[M_k | \mathscr{F}_k]. \end{aligned} \tag{17}$$

Thus, when for a given ε, $\|Z_k\| \leq \varepsilon$, $\mathbb{E}\|h(\theta_k)\|^2 M_k \to \sigma^2/\alpha$ a.s. and we have

$$\frac{\log M_k}{k} = \frac{\log M_k \mathbb{E}\|h(\theta_k)\|^2}{k} - \frac{\log \mathbb{E}\|h(\theta_k)\|^2}{k} \approx \frac{\log 1/b_k}{k} \to 1/a,$$

which proves that as $\theta_k \to \theta^*$, M_k is geometrically growing with constant $1/a$.

Accordingly, asymptotic efficiency of the method follows by considering $\Gamma_k = O_p(\sum_{i=1}^{k}[a]^{-i}) = O_p([a]^{-k})$, and $\nu_k = [a]^k$ in Definition 2.

Under $(C2)$, first we prove that (15) is a.s. convergent. Letting $c_k := \alpha\|f(\theta_k)\|^2$, for all iterations we have

$$P(M_k = \infty) = \lim_{m\to\infty} P\{M_k > m\} \leq P\{\sigma_m^2(\theta_k)/m > c_k \text{ for all } m \geq \gamma_k\} = 0 \tag{18}$$

as σ_m^2 is convergent a.s. as $m \to \infty$. Therefore $P\{M_k < \infty\} = 1$, and

$$\sqrt{M_k}(\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|) \leq \sup_m \sqrt{m}(\|\tilde{h}_m(\theta_k) - h(\theta_k)\|).$$

Besides since $\mathbb{E}[\|Y_1\|^2] < \infty$, $\mathbb{E}[\sup_m m\|\tilde{h}_m(\theta_k) - h(\theta_k)\|^2] < \infty$; together with

$$\mathbb{E}_\Omega[\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|^2|\mathscr{F}_k] \leq \frac{1}{\gamma_k}\mathbb{E}[(\sqrt{M_k}(\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|))^2|\mathscr{F}_k],$$

by (11), $\sum_k \gamma_k^{-1} < \infty$ and Lemma 2 in (Yousefian, Nedić, and Shanbhag 2012), we conclude (15) is a.s. convergent.

In order to prove part (ii) of the theorem for condition $(C2)$, we first note that by (Ghosh and Mukhopadhyay 1979), when $\theta_k \to \theta^*$, $\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|^2\|h(\theta_k)\|^{-2}$ is uniformly integrable and we have

$$\mathbb{E}\|\tilde{h}_{M_k}(\theta_k) - h(\theta_k)\|/\|h(\theta_k)\| \to \zeta.$$

Therefore by (A3) and (11), there exists k_0, such that for all $k > k_0$,

$$\begin{aligned}
\mathbb{E}_\Omega[Z_{k+1}^2|\mathscr{F}_k] &\leq (1 - \frac{2l_0}{\beta} + \frac{l_1^2}{\beta^2})Z_k^2 + \frac{1}{\beta^2}\mathbb{E}_\Omega[\|\tilde{H}_{M_k}(\theta_k) - h(\theta_k)\|^2|\mathscr{F}_k] \\
&= (1 - \frac{2l_0}{\beta} + \frac{l_1^2}{\beta^2})Z_k^2 + \frac{1}{\beta^2}\|h(\theta_k)\|^2\mathbb{E}_\Omega[\|\tilde{H}_{M_k}(\theta_k) - h(\theta_k)\|^2/\|h(\theta_k)\|^2||\mathscr{F}_k] \\
&\leq (1 - \frac{2l_0}{\beta} + \frac{l_1^2}{\beta^2})Z_k^2 + \frac{\zeta l_1^2}{\beta^2}Z_k^2.
\end{aligned} \tag{19}$$

Accordingly efficiency of (15) follows by the same approach as in part $(C1)$. □

The following corollary is an immediate consequence of Theorem 2 and the equivalence of Assumption 2.1 for $h(\cdot)$ and Assumption 2.2 for $f(\cdot)$ where $h(\theta) = \nabla_\theta f(\theta)$.

Corollary 3 The function $h(\cdot)$ satisfies the conditions in Assumption 2.2. Let $\{\gamma_k\}_{k\geq 1}$ be a fixed positive sequence for which we have $\sum_{i=1}^{\infty} 1/\gamma_i < \infty$. Define $\{M_k\}_{k\geq 1}$ as a sequence of random variables in probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Let $\mathbb{E}[Y_1^8] < \infty$, and for $0 < \alpha < 1/4$, $\zeta > 0$, $M_k := M(\theta_k) = \inf\{m > \max\{3, \gamma_k\} : \frac{\hat{\sigma}_m(\theta_k)}{m} + m^{-(1+\alpha)} < \zeta\|h(\theta_k)\|^2|\mathscr{F}_k\}$. Then the SA iterates

$$\theta_{k+1} = \theta_k - 1/\beta\tilde{h}_{M_k}(\theta_k),\ k = 1, 2, \ldots, \tag{20}$$

are (a) almost surely convergent to the true solution θ^*, and (b) asymptotically efficient.

The version of the sequential rule used in Theorem 2 and Corollary3 has the true function $h(\theta_k)$ value on the right-hand side of (14), thereby leaving no critical parameters for choice by the user. Thus, this version of the stochastic recursion is truly parameter-free and fully adaptive, in that the sample size needed in each iteration is determined solely by local functional and estimation properties. Such a rule is, however, not implementable because the function $h(\theta_k)$ is not known. To reach a completely adaptive version that can be implemented easily, one can modify (14) to:

$$M_k = \inf\{m > \max\{3, \gamma_k\} : \frac{\sigma_m^2(\theta_k)}{m} < \alpha \|\tilde{h}_{M_{k-1}}(\theta_k)\|^2\}, \tag{21}$$

This rule is easy to implement as a sequential rule if the estimator $\tilde{h}_{M_{k-1}}$ and the variance function σ_m^2 can be updated using constant-effort computations. The results in Theorem 2 indicate that one can expect similar a.s. and efficient convergence properties. Unlike the rule (14) which compares the *absolute* confidence interval to a fixed target $h(\theta_k)$, the rule (21) compares the *relative* confidence interval $\sigma_{M_k}^2/(M_k\|\tilde{h}_{M_k}\|^2)$ to a target. Thus, the convergence under (21) is not a straightforward consequence of the proof method of Theorem 2. The convergence properties under the rule (21) is of active research interest to the authors.

4 CONCLUDING REMARKS

Our goal is to develop an adaptive sampling rule for use in parameter-free stochastic recursions to produce iterates that perform well in finite time while enjoying provable asymptotic consistency and efficiency under mild restrictions. The main idea underlying our adaptive sampling proposal is to continue sampling at a point until there is enough probabilistic evidence that the subsequent iterate θ_{k+1} is of a higher quality (in terms of objective function value) than the current iterate θ_k. The corresponding sample size M_k will then be used in estimating the function h and its derivatives at the incumbent point. The sample size determining rules M_k are designed to provide the stochastic recursion the flexibility to adapt to the problem structure and exhibit good performance in both finite and infinite time. This is as opposed traditional algorithms like SAA (Pasupathy and Ghosh 2013) and SA where the sample size growth follows a deterministic rule (e.g. geometric) as the algorithm searches through potential solutions in the search space.

While adaptive sampling schemes such as those we have presented have been shown to be effective in a number of recent studies, their analysis when used within (SCSR) has posed challenges. The results we provide in this paper take the first steps towards the analyzing the consistency and efficiency of a broad swathe of adaptive sampling strategies within (SCSR).

REFERENCES

Alagoz, O., A. J. Schaefer, and M. S. Roberts. 2009. "Optimization in Organ Allocation". In *Handbook of Optimization in Medicine*, edited by P. Pardalos and E. Romeijn. Kluwer Academic Publishers.

Anscombe, F. J. 1953. "Sequential estimation". *Journal of the Royal Statistical Society. Series B (Methodological)*:1–29.

Atlason, J., M. A. Epelman, and S. G. Henderson. 2008. "Optimizing call center staffing using simulation and analytic center cutting plane methods". *Management Science* 54 (2): 295–309.

Basu, D. 1955. "On statistics independent of a complete sufficient statistic". *Sankhyā: The Indian Journal of Statistics*:377–380.

Broadie, M., D. M. Cicek, and A. Zeevi. 2009. "An adaptive multidimensional version of the Kiefer-Wolfowitz Stochastic Approximation Algorithm". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 601–612. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Broadie, M., D. M. Cicek, and A. Zeevi. 2010. "General Bounds and Finite-Time Improvement for the Kiefer-Wolfowitz Stochastic Approximation Algorithm". *Operations Research* 59:1211–1224. To appear.

Byrd, R. H., G. M. Chin, J. Nocedal, and Y. Wu. 2012. "Sample size selection in optimization methods for machine learning". *Mathematical programming* 134 (1): 127–155.

Chow, Y. S., and H. Robbins. 1965. "On the asymptotic theory of fixed-width sequential confidence intervals for the mean". *The Annals of Mathematical Statistics*:457–462.

Deng, G., and M. C. Ferris. 2006. "Adaptation of the UOBQYA algorithm for noisy functions". In *Proceedings of the 2006 Winter Simulation Conference*, edited by L. Perrone, F. Wieland, J. Liu, B. Lawson, D. Nicol, and R. Fujimoto, 312–319: Institute of Electrical and Electronics Engineers: Piscataway, New Jersey.

Ghosh, M., and N. Mukhopadhyay. 1979. "Sequential point estimation of the mean when the distribution is unspecified". *Communications in Statistics-Theory and Methods* 8 (7): 637–652.

Hashemi, F., and R. Pasupathy. 2012. "Averaging and derivative estimation within stochastic approximation algorithms". In *Proceedings of the 2012 Winter Simulation Conference*, edited by R. P. O. R. C. Laroque, J. Himmelspach and A. M. Uhrmacher, 1–9. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Henderson, S. G., and B. L. Nelson. (Eds.) 2006. Volume 13 of *Handbooks in Operations Research and Management Science: Simulation*. Elsevier.

Homem-de-Mello, T., A. Shapiro, and M. L. Spearman. 1999. "Finding optimal release times using simulation based optimization". *Management Science* 45:86–102.

Kushner, H. J., and G. G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. New York, NY.: Springer-Verlag.

Law, A. M. 2007. *Simulation Modeling and Analysis*. New York, NY.: McGraw-Hill.

Lu, D., and W. V. Li. 2009. "A note on multivariate gaussian estimates.". *Journal of Mathematical Analysis and Applications* 354:704–707.

Mukhopadhyay, N., and S. Datta. 1996. "On sequential fixed-width confidence intervals for the mean and second-order expansions of the associated coverage probabilities". *Annals of the Institute of Statistical Mathematics* 48 (3): 497–507.

Osorio, C., and M. Bierlaire. 2013. "A Simulation-Based Optimization Framework for Urban Transportation Problems". *Operations Research* 61 (6): 1333–1345.

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

Pasupathy, R., and S. Ghosh. 2013. "Simulation Optimization: A Concise Overview and Implementation Guide". Tutorials in Operations Research,, 122–150. INFORMS, http://dx.doi.org/10.1287/educ.2013.0118.

Pasupathy, R., P. Glynn, S. Ghosh, and F. Hashemi. 2014. "How Much to Sample in Simulation-Based Stochastic Recursions?". http://filebox.vt.edu/users/pasupath/pasupath.htm.

Pasupathy, R., and S. Kim. 2011. "The stochastic root-finding problem: overview, solutions, and open questions". *ACM TOMACS* 21 (3): 19.

Pasupathy, R., and B. W. Schmeiser. 2009. "Retrospective-approximation algorithms for multidimensional stochastic root-finding problems". *ACM TOMACS* 19 (2).

Pasupathy, R., and B. W. Schmeiser. 2010. "DARTS — Dynamic Adaptive Random Target Shooting". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1255–1262. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Spall, J. C. 2000. "Adaptive stochastic approximation by the simultaneous perturbation method". *Automatic Control, IEEE Transactions on* 45 (10): 1839–1853.

Spall, J. C. 2012. "Stochastic optimization". In *Handbook of computational statistics*, 173–201. Springer.

Starr, N. 1966. "On the asymptotic efficiency of a sequential procedure for estimating the mean". *The Annals of Mathematical Statistics*:1173–1185.

Yousefian, F., A. Nedić, and U. V. Shanbhag. 2012. "On stochastic gradient and subgradient methods with adaptive steplength sequences". *Automatica* 48 (1): 56–67.

AUTHOR BIOGRAPHIES

FATEMEH S. HASHEMI is a PhD candidate in the Grado Department of Industrial and Systems Engineering at Virginia Tech under supervision of Dr. Raghu Pasupathy. Currently, she is a research Co-op at IBM Thomas J. Watson Research Center. Her research interests are methods for sampling within stochastic recursions, and applications to simulation optimization and stochastic root finding. Fatemeh's research finds application in large scale machine learning, stochastic adaptive controls, system design for dynamical systems and artificial neural networking. She is the recipient of INFORMS-DGWGOR finalist award, ACM-SIGSIM Student Travel Award, Best Score Award in Research Symposium and Research Travel Grant Award in Virginia Tech. Her email address is fatemeh@vt.edu and her website is at https://filebox.vt.edu/users/fatemeh/.

SOUMYADIP GHOSH is a Research Staff Member in the Business Analytics and Mathematical Sciences Department at the IBM T.J. Watson Research Center. His current research interests lie in simulation based optimization techniques for stochastic optimization problems, with a focus on applications in Energy and Power systems and supply chain management. His email is ghoshs@us.ibm.com and his web page is at https://researcher.ibm.com/researcher/view.php?person=us-ghoshs.

RAGHU PASUPATHY is an associate professor in the Department of Statistics at Purdue University. His research interests lie broadly in Monte Carlo methods with a specific focus on simulation optimization. He is a member of INFORMS, IIE, and ASA, and serves as an associate editor for *Operations Research* and *INFORMS Journal on Computing*. He is the Area Editor for the Simulation Desk at IIE Transactions. His email address is pasupath@purdue.edu and his web page is https://filebox.vt.edu/users/pasupath/pasupath.htm.