Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

ROOT FINDING VIA DARTS — DYNAMIC ADAPTIVE RANDOM TARGET SHOOTING

Raghu Pasupathy

Industrial and Systems Engineering
Virginia Tech
Blacksburg, VA 24061, USA

Bruce W. Schmeiser

School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, USA

ABSTRACT

Consider multi-dimensional root finding when the equations are available only implicitly via a Monte Carlo simulation oracle that for any solution returns a vector of point estimates. We develop DARTS, a stochastic-approximation algorithm that makes quasi-Newton moves to a new solution whenever the current sample size is large compared to the estimated quality of the current solution and estimated sampling error. We show that DARTS converges in a certain precise sense, and discuss reasons to expect substantial computational efficiencies over traditional stochastic approximation variations.

1 INTRODUCTION

In this paper we revisit Stochastic Approximation (SA) for solving the Stochastic Root-Finding Problem (SRFP) — that of identifying a solution x to the vector equation $g(x) = 0$ ($g : \mathbb{R}^q \to \mathbb{R}^q$ is a vector-valued function) when a stochastic simulation capable of "generating" a consistent estimator of g is all that is available. The reader might recognize SRFPs as the stochastic analogue of the problem of solving a nonlinear system of equations — something that has been investigated in tremendous detail ever since Sir Isaac Newton, in the mid-seventeenth century, first introduced a method to successively approximate the roots of polynomials. SRFPs, by contrast, first gained attention in 1951 through a seminal paper by Robbins and Monro (1951) introducing the now famous SA recursion. The relevance of SRFPs as a class of problems in their own right is now fairly well-established. See, for example, (Pasupathy 2010, Pasupathy and Schmeiser 2009, Pasupathy and Kim 2010) for elaborate accounts on the contexts in which SRFPs occur, specific motivating examples, and their relation to simulation-optimization problems.

Our main aim in this paper is much less the introduction of yet another algorithm for solving SRFPs, than using SRFPs as a context to set the stage for a novel sampling-based variant of SA. Our hope is that, much like the original paper by Robbins and Monro (1951), this paper forms the first step toward using sampling-based SA variants within simulation-optimization problems. As we elaborate in subsequent sections, our motivation for improving SA is essentially the same as Broadie, Cicek, and Zeevi (2009a, 2009b) — to come up with a simple SA recursion that does not rely on the user for tuning algorithmic parameters to ensure good performance. While Broadie, Cicek, and Zeevi (2009a, 2009b) and numerous other authors before them have attempted this through rules that dynamically tune algorithm parameters as SA recurses through the solution space, our strategy is fundamentally different. As we shall see, the algorithm we propose attempts to dispense with all parameters within SA through the judicious use of sampling.

1.1 Problem Statement

We now present a formal problem statement for SRFPs.
Given: A simulation that generates, for any $x \in D \subset \mathbb{R}^q$, an estimator $G_m(x)$ of the function $g : D \to \mathbb{R}^q$ such that $G_m(x) \xrightarrow{d} g(x)$ ("$\xrightarrow{d}$" denotes convergence in distribution) as $m \to \infty$, for all $x \in D$.
Find: A root $x^* \in D$ of g, i.e., find x^* such that $g(x^*) = \gamma$, assuming one exists.

As stated, the SRFP makes no assumptions about the nature of $G_m(x)$ except that $G_m(x) \xrightarrow{d} g(x)$ as $m \to \infty$, where m is the "sample size," i.e., some measure of simulation effort that is usually well-defined depending on the context. In the context of terminating simulations (Law 2007, pp. 490,491), m usually refers to the number of times the simulation is called in computing the estimator $G_m(x)$. In non-terminating simulations, m usually refers to the "length of time" the simulation is executed when computing the estimator $G_m(x)$. That $G_m(x) \xrightarrow{d} g(x)$ as $m \to \infty$ is a standing assumption. For the purposes of this paper, the feasible set D is assumed to be known, i.e., any constraint functions involved in the specification of D are observed without error.

1.2 Notation and Terminology

The following is a list of key notation and definitions adopted in the paper: (i) π^* and x^* denote the set of true solutions to the SRFP and a solution in the set π^*, respectively; (ii) If $x = (x_1, \ldots, x_q)$ is a $q \times 1$ vector, then the L_2 norm of x is defined by $\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_q^2}$; (iii) The sequence of random vectors $\{X_k\}$ is said to converge to a random vector X almost surely, written as $X_k \to X$ wp1, if $\Pr\{\lim_{k\to\infty} X_k = X\} = 1$; (iv) The sequence of random vectors $\{X_k\}$ is said to converge to a random vector X in probability, written as $X_k \xrightarrow{p} X$, if $\forall \varepsilon > 0, \lim_{k\to\infty} \Pr\{\|X_k - X\| > \varepsilon\} = 0$; (v) The sequence of random vectors $\{X_k\}$ is said to converge to a random vector X in distribution, written as $X_k \xrightarrow{d} X$, if $\lim_{k\to\infty} \Pr\{X_k \le t\} = \Pr\{X \le t\}$ for all continuity points t of $\Pr\{X \le t\}$; (vi) For a sequence of real numbers $\{a_k\}$, we say $a_k = o(1)$ if $\lim_{n\to\infty} a_k = 0$; and $a_k = O(1)$ if $\{a_k\}$ is bounded, i.e., $\exists c > 0$ with $|a_k| < c, \forall k$; (vii) For a sequence of random variables $\{X_k\}$, we say $X_k = O_p(1)$, if $\{X_k\}$ is stochastically bounded (or bounded in probability), if $\forall \varepsilon > 0$ $\exists M > 0$ such that $\Pr\{|X_k| > M\} < \varepsilon, \forall k$; (viii) $\text{dist}(x, A) = \inf\{\|x - y\| : y \in A\}$ denotes the Euclidean distance between a point $x \in \mathbb{R}^q$ and a set $A \subset \mathbb{R}^q$; (ix) The words "root" and "zero" of a function $f : D \subset \mathbb{R}^q \to \mathbb{R}^q$ are used interchangeably to refer to $x \in D$ satisfying $f(x) = 0$; (x) The indicator function $I_A(x) = 1$ if $x \in A$ and 0 otherwise; (xi) The region $B_A(\varepsilon) = \{x : \text{dist}(x, A) \le \varepsilon\}$.

2 STOCHASTIC APPROXIMATION

We now provide a very broad overview of SA algorithms. SA has a long history and a tremendous amount has been written (see, for example, Kushner and Clark (1978), Kushner and Yin (2003)) on the subject over the last four decades. What we present here is thus by no means a comprehensive account. Instead, we focus on just the basics of SA with a view toward easier exposition of DARTS in the subsequent section.

Classical Stochastic Approximation (CSA) is the original stochastic root-finding algorithm developed by Robbins and Monro (1951). Much of the literature on solving SRFPs is based on CSA. The algorithm as was originally proposed has the simple iterative structure

$$X_{k+1} = X_k - a_k(\overline{Y}_k - \gamma), k = 0, 1, \ldots, \tag{1}$$

where X_0 is the initial guess, $\overline{Y}_k = \sum_{i=1}^m Y_i(X_k)/m$, $\{Y_1(x), \ldots, Y_m(x)\}$ is a random sample from the distribution of $Y(x)$, and $\{a_k\}_{k=0}^\infty$ is a predetermined sequence of positive constants satisfying $\sum_{i=0}^\infty a_k = \infty$ and $\sum_{i=0}^\infty a_k^2 < \infty$. Owing to its simple structure, CSA converges in mean square under fairly general conditions.

The last four decades have seen an enormous number of variations on the basic iteration in (1). Most of these variations have focused on ways to accelerate convergence of the original iteration through better step-size and/or direction choices within the SA iteration. For example, a notable variant of CSA is the Accelerated Stochastic Approximation (ASA) algorithm proposed by Kesten (1958) where an adaptive random step-size that decreases only when the previous two iterates have "bracketed" the target value γ is introduced. ASA is a slightly improved version of CSA but still suffers from the existence of a predetermined sequence of parameters $\{a_k\}$, and the non-utilization of slope information. Following Kesten's work in 1958, various other authors (e.g., Fabian (1968), Venter (1967), Wasan (1969)) have either extended CSA, or expounded on the specific properties of the CSA iterates. More recently, Andradóttir has proposed the Scaled Stochastic Approximation (Andradóttir 1996) and the Projected Stochastic Approximation (Andradóttir 1991) algorithms as modifications to CSA. Scaled Stochastic Approximation, for example, uses the iteration

$$X_{k+1} = X_k - a_k \left(\frac{\overline{Y}_k^{(1)} - \gamma}{\max\{\varepsilon, |\overline{Y}_k^{(2)} - \gamma|\}} + \frac{\overline{Y}_k^{(2)} - \gamma}{\max\{\varepsilon, |\overline{Y}_k^{(1)} - \gamma|\}} \right), k = 0, 1, \ldots,$$

where ε is a predetermined constant, and given X_k, $\overline{Y}_k^{(1)}$ and $\overline{Y}_k^{(2)}$ are two independent estimators of $g(X_k)$ at a specified sample size. These two algorithms, like most other earlier modifications to CSA focus either on relaxing convergence conditions or on randomizing the step-size sequence while achieving optimal asymptotic efficiency. All these variants of CSA have a simple structure, extend naturally to multiple dimensions, and usually require user-tuning of parameters.

Arguably the most popular current method for solving SRFPs is Spall's Simultaneous Perturbation Stochastic Approximation (SPSA) method. The first-order and second-order versions of this method are discussed in (Spall 2000, Spall 2003). The second-order method, called Adaptive Simultaneous Perturbation (ASP), is the true stochastic analogue of the modified Newton's method for finding the zeros of a deterministic non-linear function. The iteration for ASP is usually expressed as

$$X_{n+1} = \Pi_{\mathscr{D}}(X_k - a_k U(X_k)^{-1} G_m(X_k)), \tag{2}$$

where $\Pi_{\mathscr{D}}[x]$ is the point in the set $\mathscr{D}$ that is closest to x, and $U(X_k)$ is the Jacobian estimate of the function g at the point X_k estimated by sampling at most one or two extra points in the vicinity of X_k. One of the important contributions of ASP is the efficient incorporation of gradient estimates for direction finding into the SA recursion. Spall (2000, 2003) shows that even with such parsimonious Jacobian estimation, there is no loss in asymptotic efficiency. There have been numerous important papers focusing on techniques to accelerate convergence of the iteration in (2). Three notable examples are the idea of averaging iterates (Polyak and Juditsky 1992), the idea of having multiple time scales within the basic SA iteration (Bhatnagar and Borkar 1997, Bhatnagar and Borkar 1998, Bhatnagar, Fu, and Marcus 2001), and most recently the idea of scaling and shifting for dynamic tuning of parameters within Stochastic Approximation (SA) (Broadie, Cicek, and Zeevi 2009a, Broadie, Cicek, and Zeevi 2009b).

3 DARTS - AN ADAPTIVE SAMPLING VARIATION OF SA FOR ROOT FINDING

Despite the tremendous amount of work on SA variations over the last four decades, the crucial question of automatically choosing the parameters $\{a_k\}$ in SA algorithms has remained somewhat elusive. For instance, Spall mentions in (Spall 2003, pp. 197) that while ASP performs well when the initial solution is "sufficiently close" to the true root, the iteration generally requires careful choice of parameters for efficient performance. Similar sentiments have been expressed in (Yakowitz, L'Ecuyer, and Vazquez-Abad 2000), and demonstrated with plenty of evidence in (Pasupathy and Schmeiser 2009, Pasupathy and Kim 2010).

The main issue within virtually all SA variations is that the sequence $\{a_k\}$ is never chosen "just right" for a given problem. The parameters are either "too small" resulting in SA "stalling," or they are chosen "too big" resulting in SA exhibiting "wild fluctuations." See (Broadie, Cicek, and Zeevi 2009b, Broadie, Cicek, and Zeevi 2009a) for a more elaborate description. Such unsatisfactory finite-time behavior is in spite of assurances of optimal performance in the asymptotic sense. This is somewhat unsurprising because optimal asymptotic performance only stipulates that the sequence $\{a_k\}$ be chosen to satisfy certain asymptotic properties, thereby leaving an infinite number of possible choices for the user. DARTS (Dynamic Adaptive Random Target Shooting), the stochastic root-finding algorithm that we propose in this paper, adopts a fundamentally different strategy to circumvent this issue of choosing parameters within SA variations. *In broad terms, DARTS is an SA variation where the sequence $\{a_k\}$ is almost fully dispensed with, through a judicious use of sampling as the algorithm evolves through the search space.*

For exposition, let us indulge in a "thought experiment" where the estimator $G_m(\cdot)$ provided as part of the problem has zero variance, i.e., each call to the simulation at a candidate solution x returns the exact value of the function $g(x)$. If this information is available ahead of time, a user will simply set $m = 1$, i.e., call the simulation exactly once on each visit to a candidate solution $x \in \mathscr{D}$. More importantly, the user will dispense with the parameter sequence $\{a_k\}$ introduced within the SA variation. The resulting recursion will coincide with one of the numerous available Newton-based recursions (Ortega and Rheinboldt 1970, Kelly 2006, Kelly 1995) that are available for solving deterministic root-finding problems.

This last observation tells us that the role of the parameter sequence $\{a_k\}$ in SA is purely to protect the SA recursion from ill-effects of the randomness inherent in the simulation estimator $G_m(\cdot)$. Specifically, due to the variance inherent to the estimator $G_m(\cdot)$, the recursion only has a rough sense of the quality of an incumbent solution. In fact, even if the recursion accidentally had a correct zero $x^* \in \pi^*$ of the function $g(x)$, it simply would not know this fact without doing an infinite amount of sampling. The luxury of having an accurate measure of quality, e.g., $\|g(x) - \gamma\|$, is thus lost in the stochastic context. This is precisely why SA recursions have to introduce artificial constructs such as the parameter sequence $\{a_k\}$ to ensure that the recursion takes shorter and shorter steps, i.e., converges, in the limit. (Of course, due to the possibility of

converging "too quickly" to the wrong zero, not every sequence $\{a_k\}$ that converges to zero will be fit for the purpose.) Such artificial constructs are not needed in the deterministic context because the known quantity $\|g(x)-\gamma\|$ provides an indication, after appropriate scaling, as to whether the recursion should be taking large or small steps.

DARTS aims to dispense with the parameter sequence $\{a_k\}$ in the SA iteration through a simple but judicious sampling sampling framework. The main idea is to augment the basic Newton search inherent in SA iterations with differential sampling that is commensurate with the inferred quality of an incumbent solution. Specifically, and as we will explain in further detail, the expected sample size used at an incumbent solution x increases in inverse proportion to $\|g(x)-\gamma\|$. (We note that all of the SA variations discussed in Section 2 explicitly or implicitly assume that a fixed sample size m is used to construct the estimator G_m as the SA algorithm evolves through the search space.) We argue that the sampling scheme incorporated within DARTS is an intuitive and automatically-implemented efficiency mechanism, that facilitates rapid convergence to a zero of the function g.

3.1 Algorithm Listing

We now provide a formal listing of the algorithm DARTS. For convenience, the algorithm listing is presented assuming that the estimator G_m is a sample mean. This restrictions is purely for expository purposes, and is easily relaxed.

Algorithm DARTS:

Given: target γ; a simulation that returns $G_m(x)=\sum_{i=1}^m Y_i(x)$ for given $x\in\mathscr{D}$ and given sample size m.
Find: a root $x^*\in\mathscr{D}$ satisfying $g(x^*)=\gamma$.
Algorithmic Parameters: initial guess X_0; a positive constant c satisfying $0<c<1$.

0. Set $k=0$.
1. Simulate at X_k with sample size

$$M(X_k)=\inf\{m:\|G_m(x)-\gamma\|>c\hat{\sigma}_m(x)/\sqrt{m}\},$$

where $\hat{\sigma}_m(x)=\sqrt{(m-1)^{-1}\sum_{i=1}^m (Y_i(x)-G_m(x))^T(Y_i(x)-G_m(x))}$, and $Y_i, i=1,2,\ldots,m$ are column vectors.
2. Set

$$X_{k+1}=\Pi_{\mathscr{D}}(X_k-U(X_k)^{-1}G_{M(X_k)}(X_k)),$$

where $U(X_k)$ is an appropriately chosen Jacobian estimator of $g(X_k)$.
3. Set $k=k+1$ and go to Step 1.

Let us first note that the search step in DARTS (Step 2) does not have the usual sequence of parameters $\{a_k\}$ found in SA variations. Next, as can be seen from the algorithm listing, DARTS does nothing different from modern SA iterations in terms of the search procedure — the search step consists of a Newton step followed by a projection, if necessary. The key deviation arises in Step 1 where the sample size $M(X_k)$ to be used at the incumbent solution X_k is determined dynamically. DARTS continues to sample at X_k until the estimated deviation from the target $\|G_{m_k}(X_k)-\gamma\|$ exceeds a small fraction of an appropriate summary measure of sampling variability. The justification for the sampling strategy in Step 2 is based on the idea that the iteration should not progress until DARTS is reasonably certain that the deviation of $G_{m_k}(X_k)$ from the target γ is much less due to the sampling variability of $G_{m_k}(X_k)$ than due to the bias of $G_{m_k}(X_k)$ with respect γ.

Why should we expect DARTS to converge in any sense? As we shall see, the sampling framework within DARTS is designed to make the iterates spend most of their time around a zero. This is because, in the proximity of a zero, $\|g(X_k)-\gamma\|$ tends to be small by definition, and sampling continues in Step 2 while the sampling variability in $G_m(X_k)$ exceeds $\|G_m(X_k)-\gamma\|$. By the same reasoning, DARTS spends little time far away from the set of zeros where the standard for a move is much less stringent due to the large magnitude of $\|g(X_k)-\gamma\|$. In fact, our numerical experience demonstrates exactly that — the iterates in DARTS spend an inordinate amount of time around a zero, interspersed with brief forays far away from the set of zeros. The forays are a clear result of randomness that is inherent in the estimator G_m, and are decidedly brief because the sampling scheme dictates that the sample size should be small when the inferred solution quality is poor.

(As we demonstrate in Section 3.2, the sampling rate at solution x in Step 2 of DARTS should increase fast enough with distance $\|g(x)-\gamma\|$ in order to guarantee convergence.)

The simple sampling strategy in Step 2 also seems to naturally promote efficiency. Unlike in traditional SA variations, where the operative sample size does not vary across incumbent solutions, sampling within DARTS is performed selectively and (only) to the extent of reliably identifying a better incumbent solution through the search step. Accordingly, sample sizes tend to be low at locations far away from a zero, since identifying better candidates is relatively easy; locations close to a zero, by similar reasoning, are usually associated with a high sample size. The latter point implies that the trajectory of DARTS, in the vicinity of a zero, in general looks very similar to the (hypothetical) trajectory of a deterministic Newton recursion when executed on the function $g(x)$.

3.2 Algorithm Analysis

In this section, we provide a brief analysis of the asymptotic behavior of DARTS. Specifically, we demonstrate under mild conditions that the iterates $\{X_k\}$ in DARTS converge to the set of zeros in probability, as the total amount of expended effort tends to infinity. All results are stated without proofs, and we have no corresponding rate results on $\{X_k\}$ at the moment.

Recall that X_k is the incumbent solution at the end of k iterations obtained after expending an amount of effort $\sum_{i=1}^{k} M(X_i)$, where $M(X_i)$ is the sample-size at iterate X_i. We now note that the double-sequence $\{X_k, M(X_k)\}$ is not a semi-Markov process (Çinlar 1975, Ross 1995). This is because, depending on the Jacobian estimate used within DARTS, $\{X_k, M(X_k)\}$ most probably satisfies

$$\Pr\{X_{k+1}\in\mathscr{A}, M_{k+1}\le m | X_0, X_1, \ldots, X_k; M_0, M_1, \ldots, M_k\} = \Pr\{X_{k+1}\in\mathscr{A}, M_{k+1}\le m | X_k; M_k\};$$

and never

$$\Pr\{X_{k+1}\in\mathscr{A}, M_{k+1}\le m | X_0, X_1, \ldots, X_k; M_0, M_1, \ldots, M_k\} = \Pr\{X_{k+1}\in\mathscr{A}, M_{k+1}\le m | X_k\}$$

for measurable sets $\mathscr{A}$ and $m\in(0,\infty)$. In other words, while the distribution of $(X_{k+1}, M(X_{k+1}))$ (conditional on the entire history) most probably depends only on the most recent state (X_k, M_k), it depends on both the iterate X_k and on how much sampling was done at the iterate X_k. This disqualifies $\{X_k, M_k\}$ from being a semi-Markov process, thereby making some of our analysis a little more nuanced.

Before proceeding further, we assume without proof that there exists a well-defined random variable X such that $X_k \xrightarrow{d} X$ and that $\Pr\{X\in\pi^*\} = \mu(\pi^*) > \delta$ for some $\delta > 0$. That the sequence $\{X_k\}$ achieves such steady state is fairly straightforward to show. Under mild conditions, most notably on the nature of the estimator G_m, the Jacobian estimator used within the search step of DARTS, and the underlying function g, the method of proof closely follows that established within (Meyn and Tweedie 2009, Chapter 10) for showing the existence of a steady state distribution for a discrete time Markov chain on a continuous state space.

We are now ready to describe key properties of the sequence of sample sizes $\{M_k\}$ used within DARTS.

Proposition 1. *Let* $G_m(x) = m^{-1}\sum Y_i(x)$ *and* $\hat{\sigma}_m(x) = \sqrt{(m-1)^{-1}\sum_{i=1}^{m}(Y_i(x)-G_m(x))^T(Y_i(x)-G_m(x))}$, *where* $Y_i(x), i=1,2,\ldots$ *are iid random (column) vectors having a finite variance matrix. Denote*

$$M(x) = \inf\{m : \|G_m(x)-\gamma\| > c\hat{\sigma}_m(x)/\sqrt{m}\}, \tag{3}$$

where c is some positive constant. Then the following hold.

- *(i)* $E[sup_{x\notin B_{\pi^*}(\varepsilon)} M(x)] < \infty$.
- *(ii)* $M(x)$ *is* $O_p(\|g(x)-\gamma\|^{-2})$, $M(x)$ *is not* $o_p(\|g(x)-\gamma\|^{-2})$, *and* $E[M(x)] = O(\|g(x)-\gamma\|^{-2})$ *as* $dist(x,\pi^*)\to 0$.

The first assertion of Proposition 1 states that the expected sample size at any solution lying outside the region $B_{\pi^*}(\varepsilon)$ is uniformly bounded. This is not surprising considering that the random variables $Y_i, i=1,2,\ldots,m$ making up the estimator G_m have finite variance. The proof of assertion (i) is very similar to the famous result by Chow and Robbins (1965) for the context of sequential stopping when constructing fixed-width confidence intervals. Assertion (ii) in Proposition 1 establishes the exact rate at which the sample size tends to infinity as x tends to the set of zeros of g. This result is easily understood upon noting (loosely) that wp1, $G_m(x)$ tends to $g(x)$ and $\hat{\sigma}_m^2$ tends to the trace $\sigma^2(x)$ of the matrix $\mathrm{Var}(Y_1)$ as $m\to\infty$. From (3), this means that $M(x)$ "looks like" $c\sigma^2/\|g(x)-\gamma\|^2$ for large m. Such intuition is again based on arguments in (Chow and Robbins 1965).

For ease of exposition of Proposition 2, let us introduce the random variable Z_t related to the iterate X_k as

$$Z_t = X_{J(t)}, \text{ where } J(t) = \min\{i : \sum_{j=1}^{i} M(X_j) \geq t\}. \tag{4}$$

In words, Z_t is the incumbent solution in DARTS after expending an amount of sampling effort t. Proposition 2 is a statement on the asymptotic behavior of Z_t.

Proposition 2. *Let* $G_m(x) = m^{-1}\sum Y_i(x)$ *and* $\hat{\sigma}_m(x) = \sqrt{(m-1)^{-1}\sum_{i=1}^{m}(Y_i(x)-G_m(x))^T(Y_i(x)-G_m(x))}$, *where* $Y_i(x), i = 1,2,\ldots$ *are iid random (column) vectors having a finite variance matrix. Also, for* $x^* \in \pi^*$, *let* $g(x) = o(\sqrt{x-x^*})$ *as dist*$(x,\pi^*) \to 0$. *Then, denoting* $\mu_k(\cdot)$ *as the probability measure associated with* X_k, *for any* $\varepsilon > 0$,

(i) $\limsup_{k\to\infty} \int_{x\in\mathscr{D}} I\{X_k \notin B_{\pi^*}(\varepsilon)\}M(x)\mu_{X_k}(dx) < \infty$;
(ii) $\liminf_{k\to\infty} \int_{x\in\mathscr{D}} I\{X_k \in B_{\pi^*}(\varepsilon)\}M(x)\mu_{X_k}(dx) = \infty$;
(iii) $\lim_{t\to\infty} Pr\{Z_t \notin B_{\pi^*}(\varepsilon)\} = 0$.

The condition $g(x) = o(\sqrt{\|x-x^*\|})$ as $\text{dist}(x,\pi^*) \to 0$ is a structural condition that stipulates a "minimum decrease" on the function g around a zero. Assertion (i) of Proposition 2 states that the asymptotic expected effort spent outside the region $B_{\pi^*}(\varepsilon)$ is finite. A proof follows somewhat simply from assertion (i) in Proposition 1. Likewise, assertion (ii) of Proposition 2 states that the asymptotic expected effort spent inside the region $B_{\pi^*}(\varepsilon)$ is infinite. A proof for this assertion follows from the assumed nature of the limiting distribution of X_k and assertion (ii) of Proposition 1. Specifically, we have assumed that the limiting distribution of X_k assigns positive probability mass to the region $B_{\pi^*}(\varepsilon)$. Furthermore, we know from assertion (ii) of Proposition 1, and the structural condition imposed on the function g, that the integrand in (ii) probabilistically increases to infinity (as $\text{dist}(x,\pi^*) \to 0$) at a rate faster than the rate at which the function x^{-1} tends to infinity around $x = 0$. These two facts together imply assertion (ii). Assertion (iii) of Proposition 2 is crucial from an algorithmic standpoint. It implies that the incumbent solution converges to the set of solutions π^* in probability. A proof follows in a straightforward fashion from assertions (i) and (ii) .

Thus far, we have called Z_t as the "incumbent solution," i.e., the solution that would be returned to the user if the algorithm was stopped after expending an amount of effort t. An alternative solution that seems equally attractive for reporting to the user is the location $X^*(t)$ where the largest sample size was observed thus far, i.e.,

$$X^*(t) = X_{J^*(t)} \text{ where } J^*(t) = arg\,max\{M(X_j) : j \leq J(t)\} \text{ and } J(t) = \min\{i : \sum_{j=1}^{i} M(X_j) \geq t\}.$$

Proposition 3 asserts that such an incumbent solution converges in probability as well.

Proposition 3. *Let the conditions of Propositions 1 and 2 hold. Then for any* $\varepsilon > 0$,

$$\lim_{t\to\infty} Pr\{X^*(t) \notin B_{\pi^*}(\varepsilon)\} = 0.$$

We end this section by noting that the structural condition imposed on the function g in the vicinity of its zeros (assumed in Proposition 2) can be relaxed if an additional parameter $\beta \geq 1$ is introduced in the expression for the sample size chosen at a solution x. Formally, instead of the definition in (3), if we define

$$M(x) = \inf\{m : \|G_m(x) - \gamma\|^{\beta} > c\hat{\sigma}_m(x)/\sqrt{m}\},$$

then the required structural condition on g can be weakened to $g(x) = o(\|x-x^*\|^{1/2\beta})$ as $\text{dist}(x,\pi^*) \to 0$. This has the effect of further increasing sampling in the vicinity of an incumbent solution whose quality is inferred to be high. While this is elegant from a theoretical standpoint, the choice of the parameter β generally poses practical problems.

4 CONCLUDING REMARKS AND ONGOING RESEARCH

The problem of automatically choosing parameters within SA algorithms has long proved elusive. This is essentially because the conditions that stipulate optimal asymptotic performance of SA still leave a large

number of potential parameter sequences for possible choice, all of which do not produce good finite-time performance. DARTS is a new variation of SA that attempts to circumvent this problem by completely dispensing with algorithm parameters through judicious sampling within the search space. The sampling framework is such that the sampling effort spent at an incumbent solution is commensurate with the inferred quality of the solution. This strategy plays the dual role of ensuring convergence and efficiency, while naturally providing clues (through the sampling trail) into where the high quality solutions lie.

While DARTS has been presented for the context of SRFPs in this paper, we believe that the sampling ideas inherent in DARTS apply more generally, particularly in simulation optimization problems. Our ongoing research attempts to construct an analogous sampling-based SA algorithm for continuous local simulation-optimization problems.

REFERENCES

Andradóttir, S. 1991. A projected stochastic approximation algorithm. In *Proceedings of the 1991 Winter Simulation Conference*, ed. B. L. Nelson, D. W. Kelton, and G. M. Clark, 854–957: Institute of Electrical and Electronics Engineers: Piscataway, New Jersey.

Andradóttir, S. 1996. A scaled stochastic approximation algorithm. *Management Science* 42:475–498.

Bhatnagar, S., and V. S. Borkar. 1997. Multiscale stochastic approximation for parametric optimization of hidden markov models. *Probability in the Engineering and Informational Sciences* 11:509–522.

Bhatnagar, S., and V. S. Borkar. 1998. A two time scale stochastic approximation scheme for simulation based parametric optimization. *Probability in the Engineering and Informational Sciences* 12:519–531.

Bhatnagar, S., M. C. Fu, and S. I. Marcus. 2001. Two-timescale algorithms for simulation optimization of hidden markov models. *IIE Transactions* 33:245–258.

Broadie, M., D. M. Cicek, and A. Zeevi. 2009a. An adaptive multidimensional version of the Kiefer-Wolfowitz stochastic approximation algorithm. In *Proceedings of the 2009 Winter Simulation Conference*, ed. M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 601–612: Institute of Electrical and Electronics Engineers: Piscataway, New Jersey.

Broadie, M., D. M. Cicek, and A. Zeevi. 2009b. General bounds and finite-time improvement for the kiefer-wolfowitz stochastic approximation algorithm. manuscript.

Çinlar, E. 1975. *Introduction to stochastic processes*. New Jersey: Prentice-Hall.

Chow, Y. S., and H. E. Robbins. 1965. On the asymptotic theory of fixed-width confidence intervals for the mean. *Annals of Mathematical Statistics* 36:457–462.

Fabian, V. 1968. On asymptotic normality in stochastic approximation. *Annals of Mathematical Statistics* 39:1327–1332.

Kelly, C. T. 1995. *Iterative methods for linear and nonlinear equations*. Philadelphia, PA.: SIAM.

Kelly, C. T. 2006. *Solving nonlinear equations with Newton's method*. Philadelphia, PA.: SIAM.

Kesten, H. 1958. Accelerated stochastic approximation. *Annals of Mathematical Statistics* 21:41–59.

Kushner, H., and D. Clark. 1978. *Stochastic approximation methods for constrained and unconstrained systems*. New York, NY.: Springer-Verlag.

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.

Meyn, S., and R. L. Tweedie. 2009. *Markov chains and stochastic stability*. Cambridge, UK: Cambridge University Press.

Ortega, J. M., and W. C. Rheinboldt. 1970. *Iterative solution of nonlinear equations in several variables*. New York, NY.: Academic Press.

Pasupathy, R. 2010. On choosing parameters in retrospective-approximation algorithms for stochastic root finding and simulation optimization. To Appear in *Operations Research*.

Pasupathy, R., and S. Kim. 2010. The stochastic root-finding problem: overview, solutions, and open questions. *ACM Transactions on Modeling and Computer Simulation*. Under revision.

Pasupathy, R., and B. W. Schmeiser. 2009. Retrospective-approximation algorithms for multidimensional stochastic root-finding problems. *ACM Transactions on Modeling and Computer Simulation* 19 (2): 5:1–5:36.

Polyak, B. T., and A. B. Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* 30 (4): 838–855.

Robbins, H., and S. Monro. 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 22:400–407.

Ross, S. 1995. *Stochastic processes*. New York, NY.: Wiley.

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

Spall, J. C. 2003. *Introduction to stochastic search and optimization*. Hoboken, NJ.: John Wiley & Sons, Inc.

Venter, H. J. 1967. An extension of the Robbins-Monro procedure. *Annals of Mathematical Statistics* 38:181–190.

Wasan, M. T. 1969. *Stochastic approximation*. Cambridge, UK: Cambridge University Press.

Yakowitz, S., P. L'Ecuyer, and F. Vazquez-Abad. 2000. Global stochastic optimization with low-dispersion point sets. *Operations Research* 48:939–950.

AUTHOR BIOGRAPHIES

RAGHU PASUPATHY is an assistant professor in the Industrial and Systems Engineering Department at Virginia Tech. His research interests lie broadly in Monte Carlo methods with a specific focus on simulation optimization and stochastic root finding. He is a member of INFORMS, IIE, and ASA, and serves as an Associate Editor for ACM TOMACS and INFORMS Journal on Computing. His e-mail address is ⟨pasupath@vt.edu⟩ and his web page is ⟨https://filebox.vt.edu/users/pasupath/pasupath.htm⟩.

BRUCE SCHMEISER is professor of Industrial Engineering at Purdue University. His research interests center on developing methods for better simulation experiments. He is a member of INFORMS, is a Fellow of IIE, and has been active within the Winter Simulation Conference for many years. His e-mail address is ⟨bruce@purdue.edu⟩ and his web page is ⟨gilbreth.ecn.purdue.edu/~bruce/⟩.