Single Observation Adaptive Search for Discrete and Continuous Stochastic Optimization

Seksan Kiatsupaibul[a,*], Robert L. Smith[b], Zelda B. Zabinsky[c]

[a] Department of Statistics, Social Innovation Research Unit, Greater Data Science Lab, Chulalongkorn University, Bangkok 10330, Thailand
[b] Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, Michigan 48109
[c] Department of Industrial and Systems Engineering, University of Washington, Seattle, Washington 98195

Abstract

Solving a stochastic optimization problem often involves performing repeated noisy function evaluations at points encountered during the algorithm. Recently, a continuous optimization framework for executing a single observation per search point was shown to exhibit a martingale property so that associated estimation errors are guaranteed to converge to zero. We generalize this martingale single observation approach to problems with mixed discrete-continuous variables. We establish mild regularity conditions for this class of algorithms to converge to a global optimum.

Keywords: Stochastic optimization, simulation optimization, adaptive search, martingale processes

1. Introduction

Optimization problems with an objective function that has little or no known structure (e.g., non-convex, non-differentiable) are challenging to solve. Adding to the challenge is the presence of noise when the objective function cannot be evaluated exactly but must be estimated. This is often accomplished by repeated calls to a computer program such as in a discrete-event simulation. This type of stochastic optimization problem is also referred to as simulation optimization [11]. Stochastic optimization has many diverse

*Corresponding author, seksan@cbs.chula.ac.th

applications areas including engineering, economics, computer science and the biological sciences.

We consider the stochastic optimization problem (P1),

$$\min_{x \in S} \quad f(x) \tag{1}$$

where

$$f(x) = \mathbb{E}\left[g(x, \xi)\right] \tag{2}$$

and the argument $x \in S \subset \mathbb{R}^d$ is a feasible point in a mixed discrete-continuous space S embedded in $\mathbb{R}^d$ that is equipped with the Euclidean metric and modeled by a finite union of compact convex sets and ξ is a random element over a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. In this setting, the class of optimization problems under consideration includes problems with all real-valued variables, mixed integer-continuous variables, and all integer variables.

Many approaches exist to solve problems of the form (P1). We consider a general setting where $f(x)$ cannot be evaluated directly, but must be estimated by observing the noisy performance function $g(x, \xi)$. When $g(x, \xi)$ is evaluated by running a discrete-event simulation, ξ represents a sequence of random numbers, typically initiated with a random seed. In this situation, the optimizer does not have direct access to ξ, and cannot make use of any structure of g given ξ. This is in contrast to sample average approximation [14, 15] or stochastic programming with recourse [6] where the optimizer may use known structure (e.g., linearity or convexity) of g under a specific scenario resulting from an observation of ξ.

A classic method to solve (P1) is stochastic approximation, which employs some form of a stochastic gradient to guide the search towards a stationary point where the gradient vanishes [7, 16, 22]. The ingenious element of stochastic approximation in the Robbins-Monro scheme is to use a single observation per design point and, by gradually reducing step sizes, indirectly average out *across iterations* the noise that is inherent in the stochastic gradients. However, this approach is not appropriate when the objective function is not differentiable, as occurs in a mixed integer-continuous space or an integer space. Also, a solution from stochastic approximation usually pertains to a local optimum, while we seek a global optimum.

Typically, adaptive random search algorithms are used for seeking a global optimum in a general domain (e.g., simulated annealing, genetic algorithms, model based algorithms, partitioning methods, meta-models, etc.) [1, 3, 19, 20, 24]. A common approach to handle noise in the objective function is to take several observations (repeated replications) at a feasible point x and estimate $f(x)$ with a sample mean. This leads to a question of how many replications to use and how to balance *exploration* with *estimation*. Determining the number of replications with a limited budget of evaluations has been explored in the optimal computation budget allocation (OCBA) approach [8]. Other methods [21] allow the number of repeated observations at a given sampled point to grow as the search progresses so that a law of large numbers can take effect and the estimate of the objective function converges to the true value.

Recently, a class of single observation search algorithms (SOSA) was introduced [13] that performs exactly one g function evaluation (e.g., simulation) per feasible point on problems in a continuous space. The idea in SOSA is to use nearby points, in a ball of shrinking radius, to estimate the true function value. SOSA can be coupled with a broad class of adaptive random search algorithms, and under mild assumptions, will converge with probability one to a global optimum. The convergence result relies on proving that the accumulated error of the search process follows a martingale process. However, the theory of SOSA in [13] is limited there to stochastic optimization in a continuous space.

In this paper, we generalize the single observation approach from a continuous space to a mixed discrete-continuous space. In [13], the feasible region is assumed to be a convex compact set; here we allow the feasible region S to be a finite union of convex compact sets. A bounded mixed-integer set with convex cross sections, and bounded integer lattices are included in this class. Two examples of feasible regions that are included in this paper are illustrated in Figure 1.

We first show in this paper that the function estimate $\hat{f}(x)$ using the single observation approach with a shrinking ball converges to the correct value $f(x)$ at each $x \in S$ with probability one (Theorem 1). This was not shown explicitly in [13] in a continuous convex space, and here we show it in a mixed discrete-continuous feasible region. In Theorem 2, we generalize the convergence of the optimal estimate to the global optimum from the continuous domain to a mixed discrete-continuous domain. The insight behind this result is that the radius of the ball goes to zero at a rate determined by the maximal dimension of the feasible region. This provides convergence results to a global optimum in a mixed discrete-continuous domain for a broad class of adaptive random search algorithms. When the domain is a finite set, such as a bounded integer lattice, we show that the single observation approach with a shrinking ball naturally achieves repeated observations in an efficient manner. In this finite case, we show that the function evaluations at the same point are independent and identically distributed, although there are dependencies in function evaluations across different points. Our approach thereby incorporates the traditional approach of multiple replications.

An extensive numerical experiment was reported in [17] documenting the effectiveness of the SOSA algorithm against various alternatives. In [17], SOSA showed promising results on 72 test problems, including 24 with all continuous variables, 24 with all integer variables, and 24 with mixed, half and half, integer and continuous variables. In this paper, we investigate the detailed behavior of SOSA when applied to a problem with mixed binary and continuous decision variables. We aim at complementing the results in [17] by illustrating the main features of the algorithm that deliver its effectiveness.

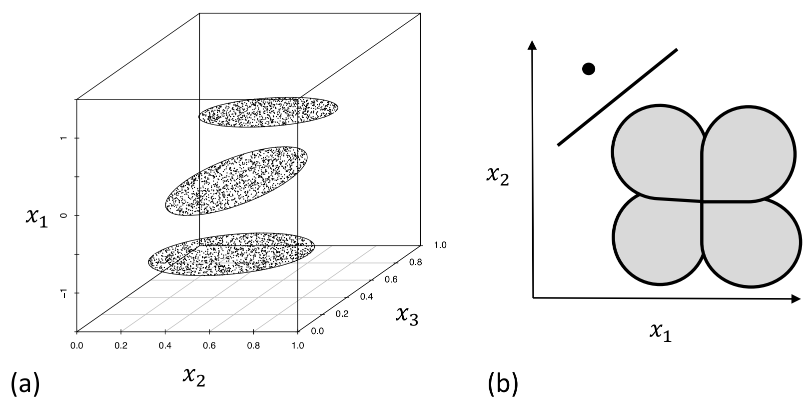

Figure 1: Two examples of feasible regions satisfying Assumption 1. The set S in panel (a) has one integer variable in the first dimension and two continuous variables in the second and the third dimensions. The feasible region is the union of three closed ellipses. The set $S \subset \mathbb{R}^2$ in panel (b) is the union of six convex sets of varying dimensions.

2. Martingale Search Algorithms

We now present SOSA, a class of single observation search algorithms, with slight modifications from [13], so that it can be applied to problems with a mixed discrete-continuous domain. During the course of the algorithm, points are sequentially sampled from the feasible set S, according to an adaptive sampling probability measure, denoted by q_n for iteration n. The measure q_n may be implicit when the method of generating the next point is based on a Markov chain Monte Carlo sampler, such as hit-and-run [18, 27] or another point generator. Thus a specific implementation of SOSA may be characterized by its method of generating sample points. As the algorithm proceeds, the objective function value at a point $x \in S$ is estimated by averaging the observed function evaluations of nearby points, in a ball of shrinking radii r_n.

Let $\mathcal{X}_n$ and $\mathcal{Y}_n$ be, respectively, the set of sample points obtained in the course of the algorithm and their corresponding observed function value up to iteration n. Let $B(x, r)$ denote a Euclidean ball centered at x with radius r. Note that we only make a *single observation* at each iteration.

Single Observation Search Algorithms (SOSA) cf. [13]

We are given:

- An initial probability q_1 for search on S and a family of adaptive search sampling distributions on S with conditional probability

$$q_n(\cdot \mid x_1, y_1, \ldots, x_{n-1}, y_{n-1}), n = 2, 3, \ldots,$$

  where x_n is the sample point at iteration n and y_n is its observed function value.

- A sequence of radii $r_n > 0$.

- A slowing sequence $i_n \leq n$.

Step 0: Sample x_1 from q_1, observe $y_1 = g(x_1, \xi_1)$ where ξ_1 is a random sample of ξ that is independent of x_1. Set $\mathcal{X}_1 = \{x_1\}$ and $\mathcal{Y}_1 = \{y_1\}$. Also, set $\hat{f}_1(x_1) = \hat{f}_1^* = y_1$ and $x^* = x_1$. Set $n = 2$.

Step 1: Given $x_1, y_1, \ldots, x_{n-1}, y_{n-1}$, sample the next point, x_n, from q_n. Independent of $x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n$, obtain ξ_n as a random sample of ξ, and observe

$$y_n = g(x_n, \xi_n).$$

Step 2: Update $\mathcal{X}_n = \mathcal{X}_{n-1} \cup \{x_n\}$ and $\mathcal{Y}_n = \mathcal{Y}_{n-1} \cup \{y_n\}$. For each $x \in \mathcal{X}_n$, estimate the objective function value as

$$\hat{f}_n(x) = \frac{\sum_{\{k \leq n:\ x_k \in B(x, r_k)\}} y_k}{|\{k \leq n:\ x_k \in B(x, r_k)\}|}$$

where $|A|$ for a set A denotes the number of elements in A. Note: a recursive sequential update of $\hat{f}_n(x)$ is given in [13].

Step 3: Estimate the optimal solution as

$$x_n^* \in \arg\min_{x \in \mathcal{X}_{i_n}} \hat{f}_n(x),$$

where $\mathcal{X}_{i_n}$ is the subset of $\mathcal{X}_n$ determined by the slowing sequence i_n. Estimate the optimal value as

$$\hat{f}_n^* = \hat{f}_n(x_n^*).$$

Step 4: If a stopping criterion is met, stop. Otherwise, update $n \leftarrow n + 1$ and go to Step 1.

3. Convergence Analysis

The convergence results of SOSA rely on proving that the accumulated error of the optimization search process follows a martingale process. We make the following assumptions regarding the stochastic optimization problem (P1).

Assumption 1: The feasible set $S \subset \mathbb{R}^d$ is a finite union of closed and bounded convex sets.

Figure 1 illustrates two feasible regions that satisfy Assumption 1. In Figure 1(a), the feasible set $S \subset \mathbb{R}^3$ is the union of three ellipses each a convex subset in $\mathbb{R}^2$, and in Figure 1(b), the feasible set $S \subset \mathbb{R}^2$ is the union of several convex sets with different dimensions including a point (zero-dimensional), a line segment (one-dimensional) and a flower-shaped figure (two-dimensional) of four convex pedals.

Assumption 2: Under the Euclidean metric on $\mathbb{R}^d$, the objective function

$f(x)$ is continuous on S.

Assumptions 1 and 2 ensure that an optimal solution exists, and we let f^* denote the optimal objective value and let $\mathcal{X}^* = \operatorname{Arg\,min}_{x \in S} f(x)$ denote the set of optimal points. Note that if $x \in S$ is an isolated point, any function f is continuous at x [23]. Therefore, in the case of a finite set of discrete points, such as an integer program, any objective function f automatically satisfies Assumption 2.

The convexity assumption, in Assumption 1, may be relaxed by extending each set to its convex hull, and then consistently, extending the objective function and sampling distribution to the convex hull while preserving all required properties (e.g., continuity of the expectation and boundedness of the noise).

As in [13], we denote the difference between the observed performance and the mean performance, called the *random observational error*, by

$$Z(x) = g(x, \xi) - f(x) \tag{3}$$

for $x \in S$.

Assumption 3: The random observational error $Z(x)$, as in (3), is uniformly bounded over $x \in S$, that is, there exists $0 < \alpha < \infty$ such that, for all $x \in S$, with probability one,

$$|g(x, \xi)] - f(x)| < \alpha.$$

The assumption of bounded noise, in Assumption 3, is commonly met in applications. In machine learning applications, it is usual to assume bounded noise [9]. When the objective function is the expected performance of a computer simulation, the noise is typically approximated by a truncated distribution which is bounded by construction.

However, in the case of a purely discrete feasible region, the boundedness condition on noise is not required for almost sure convergence. In the case of a mixed discrete-continuous feasible region, we can replace the assumption of uniformly bounded noise with bounded variance of the noise and prove convergence in probability, as discussed in [13].

Referring to SOSA and following the notation used in [13], let X_n and Y_n denote the sample point and its corresponding observed objective function evaluation at iteration n, for $n = 1, 2, \ldots$, that is,

$$Y_n = g(X_n, \xi_n), \tag{4}$$

where $\{\xi_n, n = 1, 2, \ldots\}$ are the random elements driving the value of Y_n.

Let $\mathcal{F}_0 = \sigma(X_1)$ be the initial sigma field generated by X_1, and let

$$\mathcal{F}_n = \sigma(X_1, \xi_1, \ldots, X_n, \xi_n, X_{n+1})$$

be the sigma field generated by $X_1, \xi_1, \ldots, X_n, \xi_n, X_{n+1}$. Note that X_n is $\mathcal{F}_{n-1}$ measurable. The process of (X_n, Y_n) is adapted to the filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$. From the algorithm construction, ξ_n's are not only independent and identically distributed (iid), but each ξ_n is also independent of $\mathcal{F}_{n-1}$.

Since ξ_n is independent of $\mathcal{F}_{n-1}$ and X_n is $\mathcal{F}_{n-1}$ measurable, the random observational error at iteration n, given by

$$Z_n = Y_n - f(X_n) = g(X_n, \xi_n) - f(X_n), \tag{5}$$

satisfies a crucial property; it is a *martingale difference*, i.e.,

$$\mathbb{E}\left[Z_n \mid \mathcal{F}_{n-1}\right] = 0 \quad \text{a.s.} \tag{6}$$

As a result, not only the sum of Z_n (the total error) is a martingale, but also any selective sum of Z_n also forms a martingale when the selection process is adapted to the filtration $\{\mathcal{F}_n\}_{n=0}^{\infty}$. This property is made formal in Lemma 1.

We are particularly interested in the sum of Z_k that contribute to the estimation of $f(x)$ on the n^{th} iteration. For any design point $x \in S$, define an indicator function $I_k(x)$ to identify if the point X_k sampled on the k^{th} iteration is in the ball $B(x, r_k)$,

$$I_k(x) = \begin{cases} 1 & \text{if } X_k \in B(x, r_k) \\ 0 & \text{if } X_k \notin B(x, r_k). \end{cases}$$

Observe that $I_k(x)$ is X_k measurable and, hence, $\mathcal{F}_{k-1}$ measurable.

For each $x \in S$, define the accumulated random error in estimating $f(x)$ on iteration n using the function evaluations from the points $X_k, k = 1, \ldots, n$ as the sum of random errors from sample points that fall into the balls around x,

$$M_n(x) = \sum_{k=1}^{n} I_k(x) Z_k. \tag{7}$$

Since $Z_k, k = 1, \ldots, n$ are martingale differences and $I_k(x)$ is $\mathcal{F}_{k-1}$ measurable, the sum of those Z_k included in $M_n(x)$ is also a martingale.

Lemma 1. *For each $x \in S$, the accumulated random error process $\{M_n(x), n = 1, 2, \ldots\}$ is a martingale.*

Proof. See the impossibility of systems in [10], page 213. □

Theorem 1 establishes that the function estimate at a given iteration point of the algorithm converges to the true value as long as the proportion of points in the balls remain positive with probability one. To specify this condition, let $L_n(x)$ denote the number of points X_k that are included in the estimate $\hat{f}_n(x)$,

$$L_n(x) = \sum_{k=1}^{n} I_k(x).$$

Theorem 1. *If Assumptions 1, 2 and 3 are satisfied, and if, for any feasible point $x \in S$, there exists $\kappa_x > 0$ such that*

$$\mathbb{P}\left(\liminf_{n\to\infty} \frac{L_n(x)}{n} > \kappa_x\right) = 1, \tag{8}$$

as $r_n \downarrow 0$, then the estimated objective function $\hat{f}_n(x)$ converges to the true objective function value $f(x)$ with probability one, i.e.,

$$\hat{f}_n(x) \to f(x) \text{ w.p.1.}$$

Proof. For $x \in S$, by the definition of $\hat{f}_n(x)$,

$$\hat{f}_n(x) = \frac{\sum_{k=1}^{n} I_k(x) Y_k}{L_n(x)}.$$

Since $Y_k = f(X_k) + Z_k$, we have

$$\hat{f}_n(x) = \frac{\sum_{k=1}^{n} I_k(x)(f(X_k) + Z_k)}{L_n(x)}$$

By adding and subtracting $f(x)$ and regrouping,

$$\begin{aligned} \hat{f}_n(x) &= f(x) + \left(\frac{\sum_{k=1}^{n} I_k(x) f(X_k)}{L_n(x)} - f(x)\right) + \frac{\sum_{k=1}^{n} I_k(x) Z_k}{L_n(x)} \\ &= f(x) + \left(\frac{\sum_{k=1}^{n} I_k(x) f(X_k)}{L_n(x)} - f(x)\right) + \frac{M_n(x)}{L_n(x)}. \end{aligned} \tag{9}$$

Consider a sample path from SOSA, denoted $x_1, x_2, \ldots, x_n, \ldots$, that satisfies the probability-one event in (8). For any $x \in S$, let $x_{i(n)}$ denote the subsequence of the sample path that falls in the balls around the point x, that is, $x_{i(n)} \in B(x, r_{i(n)})$ for $n = 1, 2, \ldots$. Also, let $l_n(x)$ be the number of points in the sample path that fall in the balls around x up to iteration n. Since $\liminf_{n\to\infty} (l_n(x)/n) > \kappa_x$, we have $l_n(x) \to \infty$ as $n \to \infty$. Since $r_n \downarrow 0$, we have $x_{i(n)} \to x$ as $n \to \infty$. By Assumption 2, $f(x_{i(n)}) \to f(x)$, since, as $r_n \downarrow 0$, the integer portion of the subsequence will stabilize, and the continuous portion will converge to its continuous counterpart.

By Cesáro's Lemma (see A30 in [5]),

$$\frac{\sum_{k=1}^{n} f(x_{i(k)})}{l_n(x)} - f(x) \to 0.$$

Therefore, the second term in (9) goes to zero with probability one, i.e.,

$$\frac{\sum_{k=1}^{n} I_k(x) f(X_k)}{L_n(x)} - f(x) \to 0 \quad \text{w.p.1.}$$

By Assumption 3 and the strong law of large numbers for a martingale (Theorem 3 page 243 in [10]), we have, for $x \in S$,

$$\frac{M_n(x)}{n} \to 0 \quad \text{w.p.1.}$$

This and condition (8) imply that the third term in (9) goes to zero with probability one, i.e.,

$$\frac{M_n(x)}{L_n(x)} = \frac{M_n(x)/n}{L_n(x)/n} \to 0 \quad \text{w.p.1.}$$

Therefore, $\hat{f}_n(x) \to f(x)$ w.p.1. □

Theorem 2 provides the main convergence result for SOSA. Theorem 2 states that, with some regularity conditions, the optimal value estimates generated by SOSA converge with probability one to the true global optimal value. A crucial condition for the convergence is to have the sample points generated by the algorithm cover the feasible region with sufficient density. To be precise, let $\tilde{L}(n)$ be a non-negative integer valued function of natural numbers. Define $D(n)$ as the event that there are at least $\tilde{L}(n)$ sample points in the balls around x, for each design vector x,

$$D(n) = \bigcap_{x \in S} \left\{ L_n(x) \geq \tilde{L}(n) \right\}.$$

Definition 1. *A function $h(n)$ is called $O(n^a)$ where $a \in \mathbb{R}$ if there is a $0 < \kappa_r < \infty$ such that for all $n \in \mathbb{N}$, $0 \leq h(n) \leq \kappa_r n^a$. A function $h(n)$ is called $\Omega(n^a)$ where $a \in \mathbb{R}$ if there is a $0 < \kappa_L < \infty$ such that for all $n \in \mathbb{N}, h(n) \geq \kappa_L n^a$.*

In Theorem 2, we use $\Omega(n^{-\beta})$ to specify a sequence of decreasing radii for the balls, and we use $\Omega(n^{\gamma})$ to bound below the number of points in each ball, for some $0 < \beta, \gamma < 1$.

The parameter d^* involved in the rate of decrease of radii is the maximum dimension of the convex sets comprising S, defined as follows. By Assumption 1, we can write S as a finite union of convex compact sets,

$$S = \cup_{i=1}^{K} S_i$$

where S_i is a convex compact set of dimension d_i for $i = 1, \ldots, K$. Note that $d_i = 0$ when S_i is a singleton. Let

$$d^* = \max\{1, d_1, \ldots, d_K\}.$$

For each iteration n, suppose that the sampling probability measure q_n for an implementation of SOSA can be written as a mixture of K probability distributions, i.e.,

$$q_n = \sum_{i=1}^{K} p_{n,i} q_{n,i} \tag{10}$$

where $\sum_{i=1}^{K} p_{n,i} = 1$, $p_{n,i} \geq 0$, and $q_{n,i}$ is a probability measure with S_i as its support, for $i = 1, \ldots, K$. We note that when S_i is a singleton, i.e., $d_i = 0$, for any $i = 1, \ldots, K$, the probability $q_{n,i}$ equals 1 on S_i, since S_i is its support. Properties of q_n are stated in the following assumption.

Assumption 4: For $n = 1, 2, \ldots$, q_n can be expressed as in (10), and for all $i = 1, \ldots, K$, $p_{n,i} > 0$, with $q_{n,i} = 1$ on S_i for $d_i = 0$, and for $d_i > 0$, the sampling probability $q_{n,i}$ is absolutely continuous with respect to d_i-dimensional

Lebesgue measure with density uniformly bounded away from zero on S.

Assumption 4 ensures that it is possible for an implementation of SOSA to accumulate enough sample points within the shrinking balls to satisfy the conditions of Theorem 2, that is, the balls shrink slowly at a rate that is dictated by d^*.

Theorem 2. *If Assumptions 1, 2, 3 and 4 are satisfied, and r_n is chosen to be $\Omega(n^{-\beta})$, where $\beta = (1-\gamma)/d^*$, $1/2 < \gamma < 1$ and $i_n \uparrow \infty$ such that $i_n \leq n^s$ where $0 < s < \gamma$, then a sequence of optimal value estimates $\hat{f}_n^*$ converges to the optimal value f^* with probability one, i.e.,*

$$\hat{f}_n^* \to f^* \text{ w.p.1.}$$

Proof. According to Theorem 3 in [13], the theorem is proved if there exists $\tilde{L}(n)$ that is $\Omega(n^\gamma)$ such that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(D(n)^c\right) < \infty, \tag{11}$$

where $D(n)^c$ is the complement of $D(n)$. Note that Theorem 3 in [13] did not require a convex feasible region. We next show that condition (11) is satisfied given the stated assumptions.

Let, for $i = 1, \dots, K$,

$$D_i(n) = \bigcap_{x \in S_i} \left\{ L_n(x) \geq \tilde{L}(n) \right\}$$

where $\tilde{L}(n)$ is of $\Omega(n^\gamma)$ and $S = \cup_{i=1}^K S_i$ from Assumption 1. Therefore,

$$D(n)^c = \bigcup_{i=1}^{K} D_i(n)^c,$$

and hence,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(D(n)^c\right) \leq \sum_{i=1}^{K} \sum_{n=1}^{\infty} \mathbb{P}\left(D_i(n)^c\right). \tag{12}$$

Consider S_i for a fixed $i \in \{1, \dots, K\}$. First, consider the case that $d_i = 0$, that is, S_i contains only one isolated point, say $\bar{x}$, and

$$\mathbb{P}\left(D_i(n)^c\right) = \mathbb{P}\left(L_n(\bar{x}) \geq \tilde{L}(n)\right).$$

By the coupling argument in the proof of Theorem 4 in [13],

$$\mathbb{P}\left(D_i(n)^c\right) \leq \sum_{k=0}^{\tilde{L}(n)-1} \binom{n}{k} \bar{p}^k (1-\bar{p})^{n-k},$$

where $\bar{p} > 0$ is a lower bound of $p_{n,i}$. Since $\tilde{L}(n)$ is $\Omega(n^\gamma)$ for γ, $1/2 < \gamma < 1$, applying the argument in the proof of Lemma 2 in [2], one can show that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(D_i(n)^c\right) < \infty, \tag{13}$$

when $d_i = 0$.

Now consider the case that $d_i > 0$. Note that a ball with radius r in $\mathbb{R}^d$ when projected on $\mathbb{R}^{d_i}$ is a ball with the same radius r. Since S_i is a closed and bounded convex set of d_i dimensions, Theorem 4 in [13], Assumption 4 and the condition on r_n imply that there exists $\tilde{L}(n)$ that is $\Omega(n^\gamma)$ such that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(D_i(n)^c\right) < \infty, \text{ where } d_i > 0. \tag{14}$$

By (12), (13) and (14), condition (11) holds, i.e.,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(D(n)^c\right) < \infty,$$

for some $\tilde{L}(n)$ that is $\Omega(n^\gamma)$ and, hence, Assumptions 1, 2, 3, and Theorem 3 in [13] imply the theorem. □

4. The Case of a Finite Feasible Set

We now consider the special case where S is a finite collection of isolated points. For example, S could be a bounded subset of the integer lattice $\mathbb{Z}^d$. While Theorem 2 holds for this special case, the finiteness of S allows us to remove the requirement of Assumption 3 and achieve the same results. Also in the finite case, SOSA does not require a slowing sequence in order to converge to the global optimal value. This is made formal in Theorem 3 and Corollary 1.

In the finite case, when n is large, the ball radius r_n becomes small enough that eventually there is only one feasible point x in the ball $B(x, r_n)$ for each $x \in S$. Thus a feasible point x will be revisited several times, and the noisy objective function values contributing to the objective function estimate at x, for n large, are from the same point. Thus the bias in the objective function estimate represented in the second term in (9) goes to zero as n increases. Furthermore, as Theorem 3 states, the noisy function evaluations at x are iid, which implies that the random error between the noisy function evaluations and the true value are also iid.

For a fixed x, define $N_1, N_2, \ldots$ as the stopping times representing the number of iterations until x is taken as a sample for the first time, the second time, and so on.

Theorem 3. *Consider any $x \in S$ and suppose that $X_n = x$ for iterations $n = N_1, N_2, \ldots$. The corresponding objective function evaluations $Y_{N_k} = g(X_{N_k}, \xi_{N_k})$ for $k = 1, 2, \ldots$ are independent and identically distributed random variables.*

Proof. Let $x \in S$ and suppose that $X_n = x$ for $n = N_1, N_2, \ldots$. We then have

$$Y_{N_k} = g(X_{N_k}, \xi_{N_k}) = g(x, \xi_{N_k}).$$

Hence, $Y_{N_k}, k = 1, 2, \ldots$ are iid if $\xi_{N_k}, k = 1, 2, \ldots$ are iid.

Let us adopt the inductive hypothesis that

$$\xi_{N_1}, \ldots, \xi_{N_k} \text{ are iid.}$$

For $A \in \mathcal{A}$ and $k \geq 1$,

$$\begin{aligned}
&\mathbb{P}\left(\xi_{N_{k+1}} \in A \mid \xi_{N_1} = \omega_1, \ldots, \xi_{N_k} = \omega_k\right) \\
&\quad = \sum_{n_1, \ldots, n_{k+1}} \mathbb{P}\left(\xi_{N_{k+1}} \in A \mid \xi_{N_1} = \omega_1, \ldots, \xi_{N_k} = \omega_k, N_1 = n_1, \ldots, N_{k+1} = n_{k+1}\right) \\
&\qquad\qquad \times \mathbb{P}\left(N_1 = n_1, \ldots, N_{k+1} = n_{k+1} \mid \xi_{N_1} = \omega_1, \ldots, \xi_{N_k} = \omega_k\right) \\
&\quad = \sum_{n_1, \ldots, n_{k+1}} \mathbb{P}\left(\xi_{n_{k+1}} \in A \mid \xi_{N_1} = \omega_1, \ldots, \xi_{N_k} = \omega_k, N_1 = n_1, \ldots, N_{k+1} = n_{k+1}\right) \\
&\qquad\qquad \times \mathbb{P}\left(N_1 = n_1, \ldots, N_{k+1} = n_{k+1} \mid \xi_{N_1} = \omega_1, \ldots, \xi_{N_k} = \omega_k\right) \qquad (15)
\end{aligned}$$

Since

$$\{N_1 = n_1, \ldots, N_{k+1} = n_{k+1}\} \in \mathcal{F}_{n_{k+1}-1},$$

and ξ_{N_k+1} is independent of $\mathcal{F}_{n_{k+1}-1}$, we have

$$\begin{aligned}&\mathbb{P}\left(\xi_{n_{k+1}} \in A \mid \xi_{N_1}=\omega_1,\ldots,\xi_{N_k}=\omega_k, N_1=n_1,\ldots,N_{k+1}=n_{k+1}\right)\\ &\qquad = \mathbb{P}\left(\xi_{n_{k+1}} \in A\right) = \mathbb{P}\left(\xi_1 \in A\right). \quad (16)\end{aligned}$$

Substituting (16) into (15), we obtain

$$\mathbb{P}\left(\xi_{N_{k+1}} \in A \mid \xi_{N_1}=\omega_1,\ldots,\xi_{N_k}=\omega_k\right) = \mathbb{P}\left(\xi_1 \in A\right), \quad (17)$$

and hence

$$\mathbb{P}\left(\xi_{N_{k+1}} \in A \mid \xi_{N_1},\ldots,\xi_{N_k}\right) = \mathbb{P}\left(\xi_1 \in A\right) \quad \text{w.p.1.} \quad (18)$$

But then

$$\begin{aligned}\mathbb{P}\left(\xi_{N_{k+1}} \in A\right) &= \mathbb{E}\left[\mathbb{P}\left(\xi_{N_{k+1}} \in A \mid \xi_{N_1},\ldots,\xi_{N_k}\right)\right]\\ &= \mathbb{E}\left[\mathbb{P}\left(\xi_1 \in A\right)\right]\\ &= \mathbb{P}\left(\xi_1 \in A\right). \quad (19)\end{aligned}$$

By (18) and (19), we have

$$\mathbb{P}\left(\xi_{N_{k+1}} \in A \mid \xi_{N_1},\ldots,\xi_{N_k}\right) = \mathbb{P}\left(\xi_{N_{k+1}} \in A\right) \quad \text{w.p.1.} \quad (20)$$

Now consider events $G_i = \{\xi_{N_i} \in A_i\}$, for $A_i \in \mathcal{A}, i = 1,\ldots,k,k+1$. Let I_{G_i} be the indicator function of G_i.

$$\begin{aligned}&\mathbb{P}\ \left(\xi_{N_i} \in A_i, i=1,\ldots,k,k+1\right)\\ &= \mathbb{E}\left[\Pi_{i=1}^{k+1} I_{G_i}\right]\\ &= \mathbb{E}\left[I_{G_{k+1}} \Pi_{i=1}^{k} I_{G_i}\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[I_{G_{k+1}} \Pi_{i=1}^{k} I_{G_i} \mid \xi_{N_1},\ldots,\xi_{N_k}\right]\right]\\ &= \mathbb{E}\left[\Pi_{i=1}^{k} I_{G_i} \mathbb{E}\left[I_{G_{k+1}} \mid \xi_{N_1},\ldots,\xi_{N_k}\right]\right]\end{aligned}$$

because $\Pi_{i=1}^{k} I_{G_i}$ is $\xi_{N_1},\ldots,\xi_{N_k}$ measurable,

$$\begin{aligned}&= \mathbb{E}\left[\Pi_{i=1}^{k} I_{G_i} \mathbb{P}\left(\xi_{N_{k+1}} \in A_{k+1} \mid \xi_{N_1},\ldots,\xi_{N_k}\right)\right]\\ &= \mathbb{E}\left[\Pi_{i=1}^{k} I_{G_i} \mathbb{P}\left(\xi_{N_{k+1}} \in A_{k+1}\right)\right]\end{aligned}$$

by Equation (20),

$$
\begin{aligned}
&= \mathbb{P}(\xi_{N_{k+1}} \in A_{k+1})\mathbb{E}[\Pi_{i=1}^k I_{G_i}] \\
&= \mathbb{P}(\xi_{N_{k+1}} \in A_{k+1})\Pi_{i=1}^k \mathbb{E}[I_{G_i}]
\end{aligned}
$$

by the induction hypothesis that $\xi_{N_1}, \ldots, \xi_{N_k}$ are iid,

$$
\begin{aligned}
&= \mathbb{P}(\xi_{N_{k+1}} \in A_{k+1})\Pi_{i=1}^k \mathbb{P}(\xi_{N_i} \in A_i) \\
&= \Pi_{i=1}^{k+1} \mathbb{P}(\xi_{N_i} \in A_i) \qquad (21)
\end{aligned}
$$

Since this is true for any arbitrary $A_i \in \mathcal{A}, i = 1, \ldots, k, k+1$, by the definition of independence, $\xi_{N_1}, \ldots, \xi_{N_k}, \xi_{N_{k+1}}$ are iid. This proves the theorem. $\square$.

In this finite case, SOSA does not require a slowing sequence in order to converge to the global optimal value. In other words, we can set $i_n = n$ and still guarantee the convergence property. This is shown in the following Corollary 1. Moreover, the strong law of large numbers of an iid sequence and the finiteness of S remove the requirement on the bounded random error in Assumption 3. The boundedness of the variance of the random error is not even required. The following Corollary 1 relaxes the assumptions in Theorem 2 for the finite case.

Corollary 1. *In the finite case, if Assumption 4 holds and we set $i_n = n$, then SOSA generates a sequence of optimal objective value estimates $\hat{f}_n^*$ that converges to the true optimal objective function f^* with probability one, i.e.,*

$$
\hat{f}_n^* \to f^* \ \textit{w.p.1.}
$$

Proof. Since the sampling distribution q_n is bounded away from zero on S for each iteration n (by Assumption 4), and S is finite, each x in S will be visited infinitely often by the search process as $n \to \infty$. Consider any $x \in S$ and iteration n, and suppose $X_i = x$ for iterations $i = N_1, N_2, \ldots, N_k$, where $N_k \leq n$. By Theorem 3, $Y_{N_1}, Y_{N_2}, \ldots, Y_{N_k}$ are iid. Without loss of generality, assume that $r_n < 1$ for all n. From the algorithm,

$$
\hat{f}_n(x) = \frac{\sum_{i=1}^k Y_{N_i}}{k}.
$$

Since Y_{N_i}'s are iid and x is visited infinitely often, the strong law of large numbers for an iid sequence implies

$$
\hat{f}_n(x) \to f(x) \quad \text{w.p.1.}
$$

Since each $x \in S$ is visited infinitely often, for some large N, the set of sample points visited on iteration n for $n > N$ is the entire set, that is, we have $\mathcal{X}_{i_n} = S$. Since S is finite,

$$\hat{f}_n^* = \arg\min_{x \in \mathcal{X}_{i_n}} \hat{f}_n(x) \to f^* \text{ w.p.1.}$$

$\square$

Theorem 3 is valid not only for the pure discrete case, but also for the mixed discrete-continuous case such as that in Figure 1(b) where the feasible region is composed of zero-dimensional points together with other continuous components. According to Theorem 3, the function evaluations corresponding to multiple visits to the same point are iid. Therefore, the function estimates of that point will converge to the true value. However, Corollary 1 is not valid for the mixed discrete-continuous case. Whenever the feasible region contains a continuous component, the slowing sequence and the boundedness of the random errors are required to guarantee the convergence of the optimal value estimates to the true optimal value as discussed in Theorem 2.

5. Numerical Examples

We performed an experiment by applying SOSA to the following test problem that has three continuous decision variables and four binary decision variables [26]. The objective is to illustrate some key features of SOSA that suggest its potential effectiveness over a multiple replication approach.

Test Problem: Yuan et al.[26]

$$\min \quad \mathbb{E}\left[f(x) + (1 + |f(x)|)U\right]$$

$$
\begin{aligned}
\text{s.t.} \qquad \sum_{i=1}^{6} x_i &\leq 5 \\
\sum_{i=1}^{3} x_i^2 + x_6^2 &\leq 5.5 \\
x_1 + x_4 &\leq 1.2 \\
x_2 + x_5 &\leq 1.8 \\
x_3 + x_6 &\leq 2.5 \\
x_1 + x_7 &\leq 1.2 \\
x_2^2 + x_5^2 &\leq 1.64 \\
x_3^2 + x_6^2 &\leq 4.25 \\
x_3^2 + x_5^3 &\leq 4.64 \\
x_i &\geq 0 \text{ for } i = 1, 2, 3 \\
x_i &\in \{0, 1\} \text{ for } i = 4, 5, 6, 7.
\end{aligned}
$$

where

$$
\begin{aligned}
f(x) = 0.5 \times [&(x_1 - 1)^2 + (x_2 - 2)^2 + (x_3 - 3)^2 + \\
&(x_4 - 1)^2 + (x_5 - 2)^2 + (x_6 - 1)^2 - \ln(x_7 + 1) - 4.5796]
\end{aligned}
$$

and $x \in \Re^3 \times \{0, 1\}^4$ and $U \sim \text{Uniform}[-0.1, 0.1]$. This problem contains a single global optimum at $x^* = (0.2, 0.8, 1.908, 1, 1, 0, 1)$ and $f(x^*) = 0$.

We employ the mixed integer-continuous hit-and-run sampler [4, 25] as the sample-point generator in the numerical experiment. The hit-and-run sampler implicitly determines the sampling sequence q_n.

The parameters to be determined for SOSA are s, the order of slowing sequence, γ, an upper bound on the slowing order, and β, the order of the shrinking ball. The choice of one parameter will affect the choice of the others according to Theorem 2. From numerical experiences, a small value of s will stall the algorithm. Therefore, in practice s should be set to a large value close to but less than 1. This choice of s will force β to be a small value close to but greater than zero. Here we set $s = 0.9$, $\gamma = 0.91$ and $\beta = 0.09/d^* = 0.09/3 = 0.03$.

We apply SOSA to solve the problem 100 times and take the averages of the performance measures at each iteration.

With the same experimental setting, we compare the performance of SOSA with that of a multiple replication approach formed by combining the same hit-and-run sampler with the adaptive search with resampling framework (ASR), given in [2]. The difference between SOSA and ASR is how they perform replications and estimate the objective function.

In brief, ASR maintains a schedule of increasing replications of the function evaluation for all feasible points sampled so far. It then adds more replications for promising points. ASR guarantees the same global convergence as SOSA when Assumptions 1 to 4 are satisfied, since the number of replications for ASR steadily increase. The main difference is that SOSA efficiently performs its function evaluation only once at each of *subsequent* sampled locations without reverting to perform multiple function evaluations at every *prior* sampled locations as ASR does. Since the estimation differs, the sequence of points sampled will differ, even though the same hit-and-run sample point generator is used. The numerical results are shown in Figure 2.

Figure 2 shows various performance metrics of the two algorithms at iteration $n = 1, \ldots, 12000$. Panels (a) and (b) show, respectively, the averages of the optimal value estimates $\hat{f}^*(n)$ and the averages of the (true) objective function of the optimal solution estimate $f(x_n^*)$. Panel (c) shows the averages of the number of samples whose function evaluations contribute to the optimal value estimates, l_n^*, for SOSA and ASR. For SOSA,

$$l_n^* = |\{k < n \ : \ x_k \in B(x_n^*, r_n)\}|.$$

For ASR,

$$l_n^* = |\{k < n \ : \ x_k = x_n^*\}|.$$

According to the theory, l_n^* of SOSA increases as the radius r_n decreases. In ASR, l_n^* also increases, but due to the scale in Figure 2, it is barely visible. It is interesting to see that, l_n^* of SOSA can increase, in this case, much faster than that of ASR.

Panel (d) shows the averages of the noises of the optimal solution estimates, $\hat{e}_n^*$, corresponding to the two algorithms. The noises are computed as the differences between the optimal value estimates and the function values of the optimal solution estimates. (They are the measurements in panel (a) subtracted by those in panel (b).) Since SOSA accumulates more points that contribute to the optimal value estimates than ASR does, as shown in panel (c), the averages of the noises from SOSA are smaller than those from ASR. Panel (d) implies that the optimal value estimate provided by SOSA is of higher quality compared with that from ASR.

In conclusion, with its adaptive sampling and averaging schemes, SOSA can not only efficiently produces the optimal solution estimates, but also, through boot strapping, provide high-quality estimates of the optimal value. A multiple replication approach such as ASR may assign many function evaluations repeating the samples with poor objective functions, resulting in unnecessarily slowing down the algorithm. The function evaluations gained by SOSA at the optimal solution estimate also enhance the accuracy of the optimal value estimate.

As demonstrated in the experiment, the strength of the proposed framework is its ability to accommodate *any* stochastic search for global optimization as long as it satisfies the assumptions as stated in Theorem 2. This provides an algorithm designer with the liberty to focus solely on deriving an efficient search algorithm that might exploit the problem structure at hand. For example, if one adopts the improving hit-and-run as the underlying search algorithm (as the one used in the experiment), one may employ the adaptive learning mechanism proposed by [12] to adapt the search direction to the feasible region. Adding such an adaptive learning mechanism into the search algorithm can improve the efficiency of the search algorithm and still preserves the asymptotic results guaranteed by the SOSA framework.

Acknowledgements

This work was supported in part by the National Science Foundation under Grant CMMI-1632793. We also thank Chulalongkorn University for supporting the co-authors' visits.

[1] Satyajith Amaran, Nikolaos V. Sahinidis, Bikram Sharda, and Scott J. Bury. Simulation optimization: A review of algorithms and applications. *CoRR*, abs/1706.08591, 2017. URL `http://arxiv.org/abs/1706.08591`.

[2] Sigrún Andradóttir and Andrei A. Prudius. Adaptive random search for continuous simulation optimization. *Naval Research Logistics*, 57(6): 583–604, 2010.

[3] Russell R. Barton and Martin Meckesheimer. Metamodel-based simulation optimization. In Shane G. Henderson and Barry L. Nelson, editors, *Handbook in Operations Research and Management Science*, pages 535–573. Elsevier, 2006.

[4] Stephen Baumert, Archis Ghate, Seksan Kiatsupaibul, Yanfang Shen, Robert L. Smith, and Zelda B. Zabinsky. Discrete hit-and-run for sampling points from arbitrary distributions over subsets of integer hyperrectangles. *Operations Research*, 57(3):727–739, 2009.

[5] Patrick Billingsley. *Probability and Measure*. Wiley, New York, 1995.

[6] John R. Birge and François Louveaux. *Introduction to Stochastic Programming*. Springer-Verlag, New York, NY, USA, 1997.

[7] Marie Chau and Michael C. Fu. An overview of stochastic approximation. In M. C. Fu, editor, *Handbook of Simulation Optimization*, pages 149–178. Springer, New York, 2015.

[8] Chun-Hung Chen and Loo-Hay Lee. *Stochastic Simulation Optimization: An Optimal Computing Budget Allocation*. World Scientific Publishing Co. Pte. Ltd., Singapore, 2011.

[9] Luc Devroye, László Györfi, Adam Krzyżak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, 22(3):1371–1385, 1994.

[10] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, New York, 1971.

[11] Michael Fu. *Handbook of Simulation Optimization*. International Series in Operations Research & Management Science. Springer, New York, 2015.

[12] David E. Kaufman and Robert L. Smith. Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, 46(1):84–95, 1998.

[13] Seksan Kiatsupaibul, Robert L. Smith, and Zelda B. Zabinsky. Single observation adaptive search for continuous simulation optimization. *Operations Research*, 66(6):1713–1727, 2018.

[14] Sujin Kim, Raghu Pasupathy, and Shane G. Henderson. A guide to sample average approximation. In M. C. Fu, editor, *Handbook of Simulation Optimization*, pages 207–243. Springer, New York, 2015.

[15] A. Kleywegt, A. Shapiro, and T. Homem-De-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal On Optimization*, 12(2):479–502, 2002.

[16] Harold Kushner and George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York, 2 edition, 2003.

[17] D. Linz, Z. B. Zabinsky, S. Kiatsupaibul, and R. L. Smith. A computational comparison of simulation optimization methods using single observations within a shrinking ball on noisy black-box functions with mixed integer and continuous domains. In W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, editors, *Proceedings of the 2017 Winter Simulation Conference*, pages 2045–2056, Las Vegas, NV, December 2017.

[18] H. O. Mete, Y. Shen, Z. B. Zabinsky, S. Kiatupaibul, and R. L. Smith. Pattern discrete and mixed hit-and-run for global optimization. *Journal of Global Optimization*, 50(4):597–627, 2011.

[19] Raghu Pasupathy and Soumyadip Ghosh. Simulation optimization: A concise overview and implementation guide. In H. Topaloglu, editor, *TutORial in Operations Research*, pages 122–150. INFORMS, 2013.

[20] Andrei A. Prudius and Sigrún Andradóttir. Averaging frameworks for simulation optimization with applications to simulated annealing. *Naval Research Logistics*, 59(6):411–429, 2012.

[21] L. M. Rios and N. V. Sahinidis. Derivative-free optimization: A review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293, 2013.

[22] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[23] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, NY, USA, 1976.

[24] Leyann Shi and Sigurdur Ólafsson. *Nested Partitions Method, Theory and Applications*. Springer Science+Business Media, LLC, 2009.

[25] Robert L. Smith. Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6): 1296–1308, 1984.

[26] X. Yuan, S. Zhang, L. Pibouleau, and S. D. A mixed-integer nonlinear-programming method for process design. *RAIRO - Recherche Opérationnelle-Operations Research*, 22:331–346, 1988.

[27] Zelda B. Zabinsky, Robert L. Smith, J. Fred McDonald, H. Edwin Romeijn, and David E. Kaufman. Improving hit-and-run for global optimization. *Journal of Global Optimization*, 3(2):171–192, 1993.

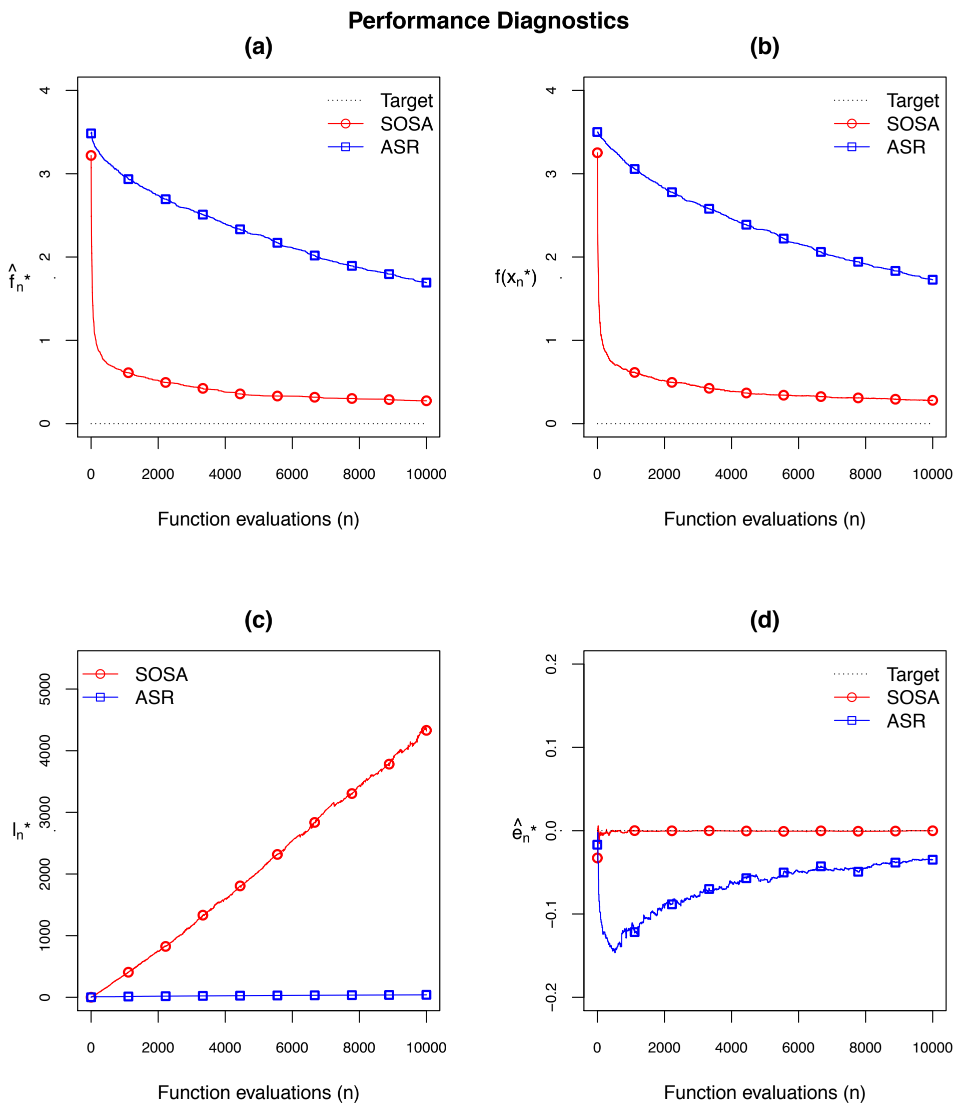

Figure 2: Performance diagnostics for Algorithm 1 with respect to the test problem. The mixed discrete-continuous hit-and-run is employed as the sample points generator. Panels (a) and (b) exhibit the optimal value estimate and the true objective function value at the optimal solution estimate (the best candidate), respectively. Panels (c) and (d) show the contributions to the best candidates and the average noises of the optimal solution estimate as functions of objective function evaluations, respectively. All the performance measures are compared against the alternative algorithm.