Single Observation Adaptive Search for Continuous Simulation Optimization

Seksan Kiatsupaibul Robert L.Smith Zelda B. Zabinskyc

Presented by Kexin Qin, Fudan University

Contents

Contents

Overview

This paper proposes a framework for adaptive search algorithms that perform exactly one simulation per design point, which are called single observation search algorithms (SOSA).

There are two main points of this framework:

- Estimating a point with an average of observed values from previously visited nearby points within a shrinking ball

Overview

This paper proposes a framework for adaptive search algorithms that perform exactly one simulation per design point, which are called single observation search algorithms (SOSA).

There are two main points of this framework:

- Estimating a point with an average of observed values from previously visited nearby points within a shrinking ball
- Convergence to a global optimum for this class of SOSA algorithms under some mild regularity conditions

Contents

Problem Formulation

The stochastic optimization problem we consider is

$$\min_{x \in S} f(x),$$

where $x \in S \subset \Re^d$ and

$$f(x) = \mathbb{E}[g(x, U)].$$

$f(x)$ cannot be evaluated exactly, the performance at a design point $x \in S \subset \Re^d$ is given by $g : S \times \Omega \to \Re$, where U is a random element over a probability space denoted $(\Omega, \mathscr{A}, \mathbb{P})$.

Problem Formulation

Assume that f is continuous and S is compact so that a minimum exists. Let $\mathscr{X}^* = \arg\min_{x \in S} f(x)$ denote the set of optimal solutions, f^* be the optimal value.

We estimate $f(x)$ by observing the output, $g(x, u)$, where u is a realization of the random variable U. The difference between the observed performance and mean performance, denoted

$$Z(x) = g(x, U) - f(x),$$

represents the random observational error.

Dependencies Among the Errors

- When the random observational errors are i.i.d across all iterations of the algorithm, according to strong law of large numbers, the error goes to 0 as iterations go to infinity.

Dependencies Among the Errors

- When the random observational errors are i.i.d across all iterations of the algorithm, according to strong law of large numbers, the error goes to 0 as iterations go to infinity.
- However, for an adaptive algorithm, the random errors are in general, neither identically distributed nor independent. An adaptive algorithm favors "better" design points and the optimal value estimates tend to be negatively biased.
  - The accumulated error of the process after a point has been evaluated forms a martingale

- When the random observational errors are i.i.d across all iterations of the algorithm, according to strong law of large numbers, the error goes to 0 as iterations go to infinity.

Dependencies Among the Errors

- When the random observational errors are i.i.d across all iterations of the algorithm, according to strong law of large numbers, the error goes to 0 as iterations go to infinity.
- However, for an adaptive algorithm, the random errors are in general, neither identically distributed nor independent. An adaptive algorithm favors “better” design points and the optimal value estimates tend to be negatively biased.

Dependencies Among the Errors

Examples

Suppose S is the union of two non-overlapping balls L and R. Moreover, suppose that the objective function values $f(x)$, for $x \in L$, are better (less) than those in R.
Step 1. sample from ball L
Step 2. sample from the other ball R
Step 3. The third point will be sampled from the ball with the smaller observed value.

Given point 3, point 2 is dependent on point 1:
$\Rightarrow$ Suppose the third point is in R. In this case, a negative error at the first point indicates that the error at the second point must also be negative.

Dependencies Among the Errors

- Errors looking backward from the current iteration point are dependent (e.g., looking at the first and second points, having sampled the third).

Dependencies Among the Errors

- Errors looking backward from the current iteration point are dependent (e.g., looking at the first and second points, having sampled the third).
- Errors looking forward when conditioning on the identity of the current iteration point (e.g., looking at the fourth point, having sampled the third) are independent of past errors.
    - The accumulated error of the process after a point has been evaluated forms a martingale

Assumptions

Assumption 1

The feasible set $S \subset \Re^d$ is a closed and bounded convex set with nonempty interior.

Assumption 2

The objective function $f(x)$ is continuous on S.

Assumptions

Assumption 3

The random error $(g(x, U) - f(x))$ is uniformly bounded over $x \in S$; that is, there exists $0 < \alpha < \infty$ such that, for all $x \in S$, with probability one,

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

Assumption 3 does not include distributions having infinite support, such as normal or gamma distributions.

Assumptions

Assumption 3

The random error $(g(x, U) - f(x))$ is uniformly bounded over $x \in S$; that is, there exists $0 < \alpha < \infty$ such that, for all $x \in S$, with probability one,

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

Assumption 3 does not include distributions having infinite support, such as normal or gamma distributions.

Assumption 3′

The random error $(g(x, U) - f(x))$ has bounded variance over $x \in S$.

$\Rightarrow$ Assumption 3 leads to a stronger convergence result (convergence with probability one) than Assumption 3' (convergence in probability).

Contents

Notations

$B(x, r)$	the ball centered at x with radius r
$\mathscr{X}_n$	the set of sample points obtained up to iteration n
$\mathscr{Y}_n$	the set of funtion evaluations up to iteration n.
$\hat{f}_n(x_i)$	the objective function estimate of x_i
$l_n(x_i)$	the number of points that fall into the balls centered at x_i
q_n	sampling density
r_n	a sequence of radii.

Single Observation Search Algorithms (SOSA)

We are given:

- A continuous initial sampling density for search on $S : q_1(x)$, and a family of continuous adaptive search sampling distributions on S with density:

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

  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 sequence $i_n < n$.

Single Observation Search Algorithms (SOSA)

Step 0: Sample x_1 from q_1, observe $y_1 = g\left(x_1, u_1\right)$, where u_1 is a sample value from distribution U and independent of x_1. Set $\mathscr{X}_1 = \{x_1\}$ and $\mathscr{Y}_1 = \{y_1\}$. Also, set $\hat{f}_1\left(x_1\right) = \hat{f}_1^*\left(x_1\right) = y_1$, $l_1\left(x_1\right) = 1$, and $x_1^* = x_1$. 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 and evaluate the objective function value $y_n = g\left(x_n, u_n\right)$.

Step 2: Update $\mathscr{X}_n = \mathscr{X}_{n-1} \cup \{x_n\}$ and $\mathscr{Y}_n = \mathscr{Y}_{n-1} \cup \{y_n\}$. For each $x \in \mathscr{X}_n$, update the contribution and the estimate of the objective function value $\hat{f}_n(x)$. Estimate the optimal value as $\hat{f}_n^*$ and optimal solution x_n^*.

Step 3: If a stopping criterion is met, stop. Otherwise, update $n \leftarrow n + 1$ and go to Step 1 .

Step 2: Update $\mathscr{X}_n = \mathscr{X}_{n-1} \cup \{x_n\}$ and $\mathscr{Y}_n = \mathscr{Y}_{n-1} \cup \{y_n\}$. For each $x \in \mathscr{X}_n$, update $l_n(x)$ and $\hat{f}_n(x)$ as:

$$l_n(x) = |\{k \leq n : x_k \in B(x, r_k)\}| = \begin{cases} l_{n-1}(x) & \text{if } x_n \notin B(x, r_n) \\ l_{n-1}(x) + 1 & \text{if } x_n \in B(x, r_n), \end{cases}$$

$$\begin{aligned} \hat{f}_n(x) &= \frac{\sum_{\{k \leq n : x_k \in B(x, r_k)\}} y_k}{|\{k \leq n : x_k \in B(x, r_k)\}|} \\ &= \begin{cases} \hat{f}_{n-1}(x), & \text{if } x_n \notin B(x, r_n), \\ \Big((l_n(x) - 1)\, \hat{f}_{n-1}(x) + y_n \Big) / l_n(x), & \text{if } x_n \in B(x, r_n), \end{cases} \end{aligned}$$

Single Observation Search Algorithms (SOSA)

Step 2: Estimate the optimal value as:

$$\hat{f}_n^* = \min_{x \in \mathscr{X}_{in}} \hat{f}_n(x)$$

and estimate the optimal solution as

$$x_n^* \in \left\{ x \in \mathscr{X}_{i_n} : \hat{f}_n(x) = \hat{f}_n^* \right\},$$

where $\mathscr{X}_{i_n}$ is the subset of $\mathscr{X}_n$ that only includes points through i_n.

Trick: The algorithm takes the estimate of the optimal value from a subset of the function estimates up to i_n. The idea is that the shrinking balls around points used to estimate a global optimum shrink slowly enough to allow for the number of points in those balls to grow to infinity.

Contents

Martingale Property of Random Error

Recall that $Z(x) = g(x, U) - f(x)$, and because:

$$\mathbb{E}[Z(x)] = \mathbb{E}[g(x, U)] - f(x) = f(x) - f(x) = 0, \quad x \in S,$$

Z(x) is a random error with zero expectation. Let X_n and Y_n denote the sample point and its corresponding evaluation at iteration n. Then

$$Y_n = g(X_n, U_n)$$

where $\{U_n, n = 1, 2, \ldots\}$ are random elements, i.i.d..
We can construct a filtration, starting with $\mathscr{F}_0 = \sigma(X_1)$, the σ-field generated by X_1, and then, $\mathscr{F}_n = \sigma(X_1, U_1, \ldots, X_n, U_n, X_{n+1})$, the σ-field generated by $X_1, U_1, \ldots, X_n, U_n, X_{n+1}$.
X_n, Y_n is $\mathscr{F}_n$ measurable. The process of (X_n, Y_n) is then adapted to the filtration $\{\mathscr{F}_n\}_{n=0}^{\infty}$.

Martingale Property of Random Error

Denote the random error at iteration n by Z_n, where

$$Z_n = Y_n - f(X_n)$$

Z(x) is a random error with zero expectation. Let X_n and Y_n denote the sample point and its corresponding evaluation at iteration n. Then

$$Y_n = g(X_n, U_n)$$

Because X_n is $\mathscr{F}_{n-1}$ measurable and U_n is independent of $\mathscr{F}_{n-1}$,

$$\begin{aligned} \mathbb{E}[Y_n \mid \mathscr{F}_{n-1}] = \mathbb{E}[g(X_n, U_n) \mid \mathscr{F}_{n-1}] = \mathbb{E}[g(X_n, U_n) \mid X_n] \\ = \mathbb{E}[g(X_n, U) \mid X_n] = f(X_n) \end{aligned}$$

Then, we derive the martingale property of random error:

$$\begin{aligned} &\mathbb{E}[Z_n \mid \mathscr{F}_{n-1}] = \mathbb{E}[Y_n - f(X_n) \mid \mathscr{F}_{n-1}] = \mathbb{E}[Y_n \mid \mathscr{F}_{n-1}] - f(X_n) = 0, \\ &\mathbb{E}[Z_n] = \mathbb{E}[\mathbb{E}[Z_n \mid \mathscr{F}_{n-1}]] = \mathbb{E}[0] = 0. \end{aligned}$$

Accumulated Error

At iteration n, and for a sample point X_i, $i \leq n$, let $M_n(X_i)$ be the accumulated error in estimating $f(X_i)$ using evaluations from the points X_k, $k = 1, \ldots, n$ that fall into balls around X_i. Define an indicator function of points in balls around X_i,

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

for $k = 1, \ldots, n$. Using the indicator function, we have

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

Note that $\{M_n(X_i), n = 1, 2, \ldots\}$ for $i > 1$ is not a martingale, owing to the dependencies on early sample points in the sequence.

Accumulated Error

Decompose the accumulated error $M_n(X_i)$ into two parts (the error from the sample points that preceded X_i and the error from the sample points that were sampled after X_i):

$$M_n(X_i) = \sum_{k=0}^{i-1} I_k(X_i) Z_k + M_n^i(X_i)$$

$$M_n^i(X_i) = \sum_{k=i}^{n} I_k(X_i) Z_k, \quad n = i, i+1, \ldots$$

Theorem 1

Theorem 1

For any i, $i = 1, 2, \ldots,$ $\{M_n^i(X_i), n = i, i+1, \ldots\}$ is a martingale with respect to the filtration $\{\mathscr{F}_n, n = i, i+1, \ldots\}$.

Theorem 1

For any i, $i = 1, 2, \ldots,$ $\left\{M_n^i\left(X_i\right), n = i, i+1, \ldots\right\}$ is a martingale with respect to the filtration $\left\{\mathscr{F}_n, n = i, i+1, \ldots\right\}$.

Proof:
Define $\tilde{M}_n^i = \sum_{k=i}^n Z_k$ as the accumulated error from all points sampled on the iterations from iteration i through iteration n.
First show that $\left\{\tilde{M}_n^i, n = i, i+1, \ldots\right\}$ is a martingale with respect to the filtration $\{\mathscr{F}_n, n = i, i+1, \ldots,\}$.
This is equivalent to showing that:

$$\mathbb{E}[|\tilde{M}_n^i|] < \infty, \text{ and } \mathbb{E}[\tilde{M}_n^i \mid \mathscr{F}_{n-1}] = \tilde{M}_{n-1}^i \text{ for all } n \geq i.$$

Theorem 1

For any i, $i = 1, 2, \ldots,$ $\{M_n^i(X_i), n = i, i+1, \ldots\}$ is a martingale with respect to the filtration $\{\mathscr{F}_n, n = i, i+1, \ldots\}$.

Proof:
By Assumption 3, $\mathbb{E}[|Z_n|] < \alpha < \infty$. Then, $\mathbb{E}[|\tilde{M}_n^i|] \leq (n-i+1)\alpha < \infty$ In addition:

$$\begin{aligned}\mathbb{E}[\tilde{M}_n^i \mid \mathscr{F}_{n-1}] &= \mathbb{E}[Z_n + \tilde{M}_{n-1}^i \mid \mathscr{F}_{n-1}] \\ &= \mathbb{E}[Z_n \mid \mathscr{F}_{n-1}] + \mathbb{E}[\tilde{M}_{n-1}^i \mid \mathscr{F}_{n-1}] \\ &= \tilde{M}_{n-1}^i\end{aligned}$$

Now, for $n = i, i+1, \ldots$

$$M_n^i(X_i) = \sum_{k=i}^{n} I_k(X_i) Z_k = M_{n-1}^i(X_i) + I_n(X_i)(\tilde{M}_n^i - \tilde{M}_{n-1}^i)$$

By *the impossibility of systems* (Feller 1971), $\{M_n^i(X_i), n = i, i+1, \ldots\}$ is a martingale.

Estimate of Function value

For a fixed i, let $L_n(X_i)$ be the number of sample points that fall into the balls $B(X_i, r_k)$ around X_i where $k = 1, \ldots, n$ and $n \geq i$; that is,

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

The estimate of the function value at point X_i can be expressed as

$$\hat{f}_n(X_i) = \frac{\sum_{k=1}^{n} I_k(X_i) Y_k}{L_n(X_i)} = \frac{\sum_{k=1}^{n} I_k(X_i) f(X_k)}{L_n(X_i)} + \frac{M_n(X_i)}{L_n(X_i)'}$$

where the first term includes the systematic bias and the second term is the accumulated error.

Estimate of Function value

The estimate of the optimal value f^* is

$$\hat{f}_n^* = \min_{i=1,\dots,i_n} \left\{ \hat{f}_n(X_i) \right\} = \min_{i=1,\dots,i_n} \left\{ \frac{\sum_{k=1}^{n} I_k(X_i) f(X_k)}{L_n(X_i)} + \frac{M_n(X_i)}{L_n(X_i)} \right\}$$

The size of the subset i_n is a control parameter required to ensure the convergence of the optimal value estimate $\hat{f}^*$ to the true optimal value f^* by slowing the search for an optimum.

Assumption 4

Given a function of natural numbers $\tilde{L}(n)$, define $D(n)$ to be the event that each x has at least $\tilde{L}(n)$ sample points in the balls around x; that is,

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

The key idea is that the number of points in the balls around x grows at least as fast as $\tilde{L}(n)$ even though the radii of the balls are shrinking.

$\Rightarrow$ the balls cannot shrink too quickly, they must maintain a threshold of sample points.

Assumption 4

Assumption 4

Assume there exists $1/2 < \gamma < 1$ and a function $\tilde{L}(n)$ that is $\Omega(n^{\gamma})$ such that

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

where $D(n)^c$ is the complement of event $D(n)$ and γ is called an order of local sample density.

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

Assumption 4

Assumption 4

Assume there exists $1/2 < \gamma < 1$ and a function $\tilde{L}(n)$ that is $\Omega\left(n^{\gamma}\right)$ such that

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

where $D(n)^c$ is the complement of event $D(n)$ and γ is called an order of local sample density.

* A function $h(n)$ is called $\Omega\left(n^{p}\right)$, where $p \in \mathbb{R}$ if there is a $0 < \kappa_L < \infty$ such that for all $n \in \mathbb{N}, h(n) \geq \kappa_L n^{p}$.
Assumption 4 ensures that there are on the order of n^{γ} evaluations used in the estimate of every point. If the search sampling density $q_n, n = 1, 2, \ldots$ is uniformly bounded away from zero on S and r_n is of $\Omega\left(n^{-(1-\gamma)/d}\right)$, then Assumption 4 is satisfied.

Convergence Analysis

Expand the estimate of the function value in as

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

- the first term: the correct value
- the second term: the bias due to nearby points
- the third term: the non-martingale accumulated error
- the fourth term: the martingale accumulated error

Convergence Analysis

- the first term: the correct value
- **the second term:** the bias due to nearby points.
  employ Cesa'ro's Lemma with the shrinking ball mechanism to show that the bias term is washed away by averaging.

Cesáro's Lemma

If $x, x_1, x_2, \ldots$ are real numbers such that $x_n \to x$ as $n \to \infty$, then

$$\frac{\sum_{k=1}^{n} x_k}{n} \to x$$

Convergence Analysis

- the first term: the correct value
- the second term: the bias due to nearby points
- **the third term:** the non-martingale accumulated error
  the slowing sequence, i_n, is employed to slow the growth of this term, causing this non-martingale random error to diminish to zero when divided by the number of points in the associated balls.

Convergence Analysis

- **the fourth term:** the martingale accumulated error
  the slowing sequence together with the martingale property through the Azuma–Hoeffding inequality causes the martingale random error to disappear.

Azuma-Hoeffding Inequality

Let $M_1, \ldots, M_n$ be a martingale with mean $\mu = \mathbb{E}[M_n]$. Let $M_0 = \mu$ and suppose that, for $k \geq 1$,

$$|M_k - M_{k-1}| \leq \alpha_k,$$

where $\alpha_k > 0, k = 1, 2, \ldots$. Then, for all $n \geq 0$ and any $\epsilon > 0$,

$$\mathbb{P}(|M_n - M_0|) \geq \epsilon) \leq 2 \exp\left(-\frac{\epsilon^2}{2\sum_{k=1}^{n} \alpha_k^2}\right)$$

Theorem 2

Define $A(n, \varepsilon)$ as the event that, when we consider only the early portion of the sequence up to i_n, at least one objective function estimate is incorrect by more than the target error ε allowed for $\varepsilon > 0$; that is,

$$A(n, \varepsilon) = \bigcup_{i=1}^{i_n} \left\{ \left| \hat{f}_n(X_i) - f(X_i) \right| \geq \varepsilon \right\}.$$

Theorem 2

If Assumptions 1, 2, 3, and 4 are satisfied, and if $i_n \uparrow \infty$ such that $i_n \leq n^s$, where $0 < s < \gamma$, then, for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(A(n, \varepsilon)) < \infty.$$

Proof of Theorem 2

By Assumption 4 , there exist $1/2 < \gamma < 1$ and κ_L such that $\tilde{L}(n) \geq \kappa_L n^\gamma$ and

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

Observe that

$$A(n, \varepsilon) = [A(n, \varepsilon) \cap D(n)] \cup [A(n, \varepsilon) \cap D(n)^c] .$$

$$\begin{aligned}\mathbb{P}(A(n, \varepsilon)) &\leq \mathbb{P}(A(n, \varepsilon) \cap D(n)) + \mathbb{P}\left(A(n, \varepsilon) \cap D(n)^c\right) \\ &\leq \mathbb{P}(A(n, \varepsilon) \cap D(n)) + \mathbb{P}(D(n))^c)\end{aligned}$$

By Assumption 4, $\sum_{n=1}^{\infty} \mathbb{P}\left(D(n)^c\right) < \infty$. Therefore, it suffices to show that:

$$\sum_{n=1}^{\infty} \mathbb{P}(A(n, \varepsilon) \cap D(n)) < \infty$$

Proof of Theorem 2

Let

$$A(n,\varepsilon) = \bigcup_{i=1}^{i_n} E(n,i)$$

where $E(n,i) = \left\{ \left| \hat{f}_n(X_i) - f(X_i) \right| \geq \varepsilon \right\}$. The total error can be decomposed into three components, $E_1(n,i), E_2(n,i)$ and $E_3(n,i)$:

$$E(n,i) \subseteq E_1(n,i) \cup E_2(n,i) \cup E_3(n,i)$$

where:

$$E_1(n,i) = \left\{ \left| \frac{\sum_{k=1}^{n} I_k(X_i) f(X_k)}{L_n(X_i)} - f(X_i) \right| \geq \frac{\varepsilon}{2} \right\}$$

$$E_2(n,i) = \left\{ \left| \frac{\sum_{k=0}^{i-1} I_k(X_i) Z_k}{L_n(X_i)} \right| \geq \frac{\varepsilon}{4} \right\} \quad \text{and} \quad E_3(n,i) = \left\{ \left| \frac{M_n^i(X_i)}{L_n(X_i)} \right| \geq \frac{\varepsilon}{4} \right\}$$

Proof of Theorem 2

and, hence,

$$\mathbb{P}(A(n,\varepsilon)\cap D(n)) \leq \sum^{i_n} \mathbb{P}\left(E_1(n,i)\cap D(n)\right) + \sum^{i_n} \mathbb{P}\left(E_2(n,i)\cap D(n)\right) + \sum^{i_n} \mathbb{P}\left(E_3(n,i)\cap D(n)\right)$$

Proof of Theorem 2

Consider first $\sum_{i=1}^{i_n} \mathbb{P}(E_1(n,i) \cap D(n))$.
By Assumptions 1 and 2, f is uniformly continuous on S. Since the radii of balls $r_n \downarrow 0$, if $x_k \in B(x, r_k), k = 1, \ldots, n$, we have $f(x_n) \to f(x)$. By Cesáro's Lemma, there exist K such that, for all $x \in S$ and $x_k \in B(x, r_k), k = 1, \ldots, m$, we have that $m > K$ implies

$$\left| \frac{\sum_{k=1}^{m} f(x_k)}{m} - f(x) \right| < \varepsilon/2$$

Since $\tilde{L}(n) \geq \kappa_L n^\gamma$, if $n > (K/\kappa_L)^{1/\gamma}$, then $\tilde{L}(n) > K$. Let $n > (K/\kappa_L)^{1/\gamma}$ and fix $i \leq n$. Suppose $D(n)$ occurs. Consider when

$$X_1 = x_1, \ldots, X_n = x_n$$

and $x_{i_k} \in B(x_i, r_{i_k})$ for $k = 1, \ldots, m$. Since $D(n)$ occurs, by Assumption 4, we have $L_n(x_i) = m \geq \tilde{L}(n) > K$.

Proof of Theorem 2

By Azuma–Hoeffding inequality,

$$\left|\frac{\sum_{k=1}^{n} I_k\left(X_i\right) f\left(X_k\right)}{L_n\left(X_i\right)} - f\left(X_i\right)\right| = \left|\frac{\sum_{k=1}^{m} f\left(x_{i_k}\right)}{m} - f\left(x_i\right)\right| < \varepsilon/2$$

Therefore, $E_1(n,i)$ does not occur. Hence, for $n > \left(K/\kappa_L\right)^{1/\gamma}$ and $i \leq n$, we have $E_1(n,i) \cap D(n) = \emptyset$ and, hence,

$$\mathbb{P}\left(E_1(n,i) \cap D(n)\right) = 0$$

Since this is true for all $i \leq n$, we also have

$$\sum_{i=1}^{i_n} \mathbb{P}\left(E_1(n,i) \cap D(n)\right) = 0$$

That means $\sum_{i=1}^{i_n} \mathbb{P}\left(E_1(n,i) \cap D(n)\right) > 0$ for only finitely many n. Hence,

$$\sum_{n=1}^{\infty} \sum_{i=1}^{i_n} \mathbb{P}\left(E_1(n,i) \cap D(n)\right) < \infty$$

Proof of Theorem 2

Now we show that $\sum_{n=1}^{\infty}\sum_{i=1}^{i_n}\mathbb{P}\left(E_2(n,i)\cap D(n)\right)<\infty$. Fix $i_n \le n^s$.

$$E_2(n,i)\cap D(n)=\left\{\left|\frac{\sum_{k=0}^{i-1} I_k\left(X_i\right)Z_k}{L_n\left(X_i\right)}\right|\ge\frac{\varepsilon}{4}\right\}\cap D(n)$$
$$\subseteq\left\{\left|\frac{i\alpha}{L_n\left(X_i\right)}\right|\ge\frac{\varepsilon}{4}\right\}\cap D(n)$$

by the bounded variance assumption in Assumption 3 , and since
$\tilde{L}(n)\ge\kappa_L n^{\gamma}$

$$\subseteq\left\{\left|\frac{n^s\alpha}{\kappa_L n^{\gamma}}\right|\ge\frac{\varepsilon}{4}\right\}$$

Since $i_n\le n^s$ and $0<s<\gamma,\mathbb{P}\left(E_2(n,i)\cap D(n)\right)=0$ for all $i=1,\ldots,i_n$, when n is large enough. Hence,

$$\sum_{n=1}^{\infty}\sum_{i=1}^{i_n}\mathbb{P}\left(E_2(n,i)\cap D(n)\right)<\infty$$

Proof of Theorem 2

Now show that $\sum_{i=1}^{i_n} \mathbb{P}\left(E_3(n,i) \cap D(n)\right) \to 0$ as $n \to \infty$.

$$E_3(n,i) \cap D(n) = \left\{ \left| \frac{M_n^i\left(X_i\right)}{L_n\left(X_i\right)} \right| \geq \frac{\varepsilon}{4} \right\} \cap D(n)$$
$$= \left\{ \left| M_n^i\left(X_i\right) \right| \geq L_n\left(X_i\right) \frac{\varepsilon}{4} \right\} \cap D(n)$$
$$\subseteq \left\{ \left| M_n^i\left(X_i\right) \right| \geq \tilde{L}(n) \frac{\varepsilon}{4} \right\}$$

Therefore, for each $i = 1, \ldots, i_n$,

$$\mathbb{P}\left(E_3(n,i) \cap D(n)\right) \leq \mathbb{P}\left(\left| M_n^i\left(X_i\right) \right| \geq \tilde{L}(n) \frac{\varepsilon}{4} \right)$$
$$\leq 2 \exp\left(-\frac{\tilde{L}(n)^2 \varepsilon^2}{32(n-i+1)\alpha^2} \right)$$

by Azuma–Hoeffding inequality and Theorem 1

$$\leq 2 \exp\left(-\frac{\kappa_L^2 n^{2\gamma} \varepsilon^2}{32(n-i+1)\alpha^2} \right)$$

Proof of Theorem 2

since $\tilde{L}(n) \geq \kappa_L n^\gamma$

$$\leq 2 \exp\left(-\frac{\kappa_L^2 n^{2\gamma} \varepsilon^2}{32 n \alpha^2}\right)$$

$$\leq 2 \exp\left(-\frac{\kappa_L^2 n^{(2\gamma-1)} \varepsilon^2}{32 \alpha^2}\right)$$

Combining the probabilities for all $i = 1, \ldots, i_n \leq n^s$, we have the following.

$$\sum_{i=1}^{i_n} \mathbb{P}\left(E_3(n,i) \cap D(n)\right) \leq 2n^s \exp\left(-\frac{\kappa_L^2 n^{(2\gamma-1)} \varepsilon^2}{32 \alpha^2}\right)$$

The right hand side of the last inequality has finite infinite sum because $1/2 < \gamma < 1$. Therefore,

$$\sum_{n=1}^{\infty} \sum_{i=1}^{i_n} \mathbb{P}\left(E_3(n,i) \cap D(n)\right) < \infty$$

Theorem 2

Define $A(n, \varepsilon)$ as the event that, when we consider only the early portion of the sequence up to i_n, at least one objective function estimate is incorrect by more than the target error ε allowed for $\varepsilon > 0$; that is,

$$A(n, \varepsilon) = \bigcup_{i=1}^{i_n} \left\{ \left| \hat{f}_n(X_i) - f(X_i) \right| \geq \varepsilon \right\}.$$

Theorem 2

If Assumptions $1, 2, 3$, and 4 are satisfied, and if $i_n \uparrow \infty$ such that $i_n \leq n^s$, where $0 < s < \gamma$, then, for all $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}(A(n, \varepsilon)) < \infty.$$

Convergence Analysis

Once the estimated errors of sample points are controlled as described in Theorem 2, the optimal value estimates converge to the true value.

Theorem 3

If Assumptions $1, 2, 3$, and 4 are satisfied, and if $i_n \uparrow \infty$ such that $i_n \leq n^s$, where $0 < s < \gamma$, then $\hat{f}_n^* \to f^*$ with probability one.

If Assumption 3 is relaxed to Assumption $3'$, we have a weaker convergence in probability result.

Corollary 1

If Assumptions $1, 2, 3'$, and 4 are satisfied, and if $i_n \uparrow \infty$ such that $i_n \leq n^s$, where $0 < s < \gamma$, then, for all $\varepsilon > 0$,(i.e., $\hat{f}_n^* \to f^*$ in probability.)

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\hat{f}_n^* - f^*\right| \geq \varepsilon\right) = 0,$$

Proof of Theorem 3

By Assumptions 1,2, for a fixed ε, there exists δ such that, for $x, y \in S$, $\|x-y\|<\delta$ implies $|f(x)-f(y)|<\varepsilon/2$. Select an optimal solution $x^* \in \mathcal{X}^*$. Whenever $\|x-x^*\|<\delta$, we have $|f(x)-f(x^*)|=|f(x)-f^*|<\varepsilon/2$. Define

$$F(n, \varepsilon/2)=\bigcap_{i=1}^{i_n}\{\|X_i-x^*\| \geq \delta\}$$

Therefore, $\bigcap_{i=1}^{i_n}\{|f(X_i)-f^*| \geq \varepsilon/2\} \subseteq F(n, \varepsilon/2)$ and

$$\begin{aligned}
\left\{\left|\hat{f}_n^*-f^*\right| \geq \varepsilon\right\} & \subseteq A(n, \varepsilon/2) \cup F(n, \varepsilon/2) \\
& \subseteq A(n, \varepsilon/2) \cup[F(n, \varepsilon/2) \cap D(n)] \cup D(n)^c
\end{aligned}$$

Hence,

$$\sum_{n=1}^{\infty} \mathbb{P}(|\hat{f}_n^*-f^*| \geq) \leq \sum_{n=1}^{\infty} \mathbb{P}(A(n, \varepsilon/2))+\sum_{n=1}^{\infty} \mathbb{P}(F(n, \varepsilon/2) \cap D(n))+\sum_{n=1}^{\infty} D(n)^c.$$

Proof of Theorem 3

Theorem 2 states that the first term on the right hand side is finite.
Assumption 4 implies that the last term on the right hand side is finite.
Now consider the second term.
Since $r_n \downarrow 0, i_n \uparrow \infty$ and $\tilde{L}(n) \uparrow \infty$, when n large enough $F(n, \varepsilon/2) \cap D(n) = \emptyset$. Therefore, $\mathbb{P}(F(n, \varepsilon/2) \cap D(n)) > 0$ for finitely many n, which implies

$$\sum_{n=1}^{\infty} \mathbb{P}(F(n, \varepsilon/2) \cap D(n)) < \infty.$$

Hence,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\hat{f}_n^* - f^*\right| \geq \varepsilon\right) < \infty$$

Therefore, $\hat{f}_n^* \to f^*$ with probability 1.

Contents

Experiments Settings

Apply four algorithms to two problems. Four algorithms:
1. SOSA with IHR sampler (IHR-SO);
2. SOSA with AP sampler (AP-SO);
3. ASR with IHR sampler (IHR-ASR);
4. ASR with AP sampler (AP-ASR).

Experiments Settings

Problem 1 (Shifted Sinusoidal Problem).

$$\begin{array}{ll} \min & \mathbb{E}[f(x) + (1 + |f(x)|)U] \\ \text{s.t.} & 0 \le x_i \le \pi, \quad i = 1, \dots, 10, \end{array}$$

where

$$f(x) = -\left[2.5\Pi_{i=1}^{10} \sin\left(x_i - \pi/6\right) + \Pi_{i=1}^{10} \sin\left(5\left(x_i - \pi/6\right)\right)\right] + 3.5,$$

$x \in \Re^{10}$, and $U \sim$ Uniform $[-0.1, 0.1]$. According to Ali et al. (2005), this problem contains $4, 882, 813$ local optima with a single global optimum at $x^* = (4\pi/6, \dots, 4\pi/6)$ and $f\left(x^*\right) = 0$.

Experiments Settings

Problem 2 (Rosenbrock Problem).

$$
\begin{aligned}
\text{min} \quad & \mathbb{E}[f(x) + (1 + |f(x)|)U] \\
\text{s.t.} \quad & -10 \le x_i \le 10, \quad i = 1, \ldots, 10,
\end{aligned}
$$

where $f(x) = 10^{-6} \times \sum_{i=1}^{d-1} \left((1 - x_i)^2 + 100 \left(x_{i+1} - x_i^2 \right)^2 \right)$, $x \in \mathfrak{R}^{10}$, and $U \sim$ Uniform $[-0.1, 0.1]$. The global minimum is at $(1, \ldots, 1)$ and $f^* = 0$.

Experiments Results

Table 1. Statistics of the Optimal Value Estimates $\hat{f}_n^*$ of the Four Algorithms at Termination

Problem	Algorithm	Mean	Mean squared error	Best	Percentile 25	Percentile 50	Percentile 75	Worst
Problem 1	IHR-SO	0.3181	0.5103	0.0631	0.0915	0.1073	0.1433	2.6569
	AP-SO	0.5354	0.7961	0.1109	0.1211	0.1278	0.9007	2.5117
	IHR-ASR	1.6085	3.8991	0.0751	0.2948	2.5052	2.5897	3.1575
	AP-ASR	0.8905	1.6840	0.0436	0.0922	0.1780	1.8247	2.7173
Problem 2	IHR-SO	−0.0402	0.0017	−0.0524	−0.0466	−0.0407	−0.0352	−0.0247
	AP-SO	−0.0079	0.0002	−0.0231	−0.0145	−0.0104	−0.0025	0.0615
	IHR-ASR	−0.0065	0.0002	−0.0274	−0.0144	−0.0070	−0.0010	0.0243
	AP-ASR	0.0499	0.0035	−0.0020	0.0254	0.0471	0.0682	0.1593

Notes. The experiments for Problem 1 terminate with n = 12,000. The experiments for Problem 2 terminate with n = 4,000.

Experiments Results

Figure 1. (Color online) Performance Diagnostics for IHR-SO, AP-SO, IHR-ASR, and AP-ASR with Respect to Problem 1

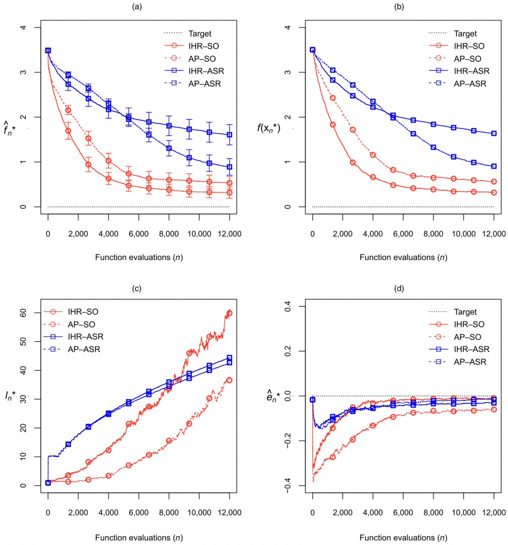

Figure 2. (Color online) Performance Diagnostics for IHR-SO, AP-SO, IHR-ASR, and AP-ASR with Respect to Problem 2