Surrogate-Based Promising Area Search for Lipschitz Continuous Simulation Optimization

Qi Fan, Jiaqiao Hu

Presented by Qingkai Zhang

School of Management, Fudan University

2021.12.20

Outline

Outline

Idea behind the algorithm

This paper draw upon ideas from the following highly successful techniques:

- the shrinking ball method
- COMPASS
- surrogate models/metamodels/response surface methods (RSMs)

Idea behind the algorithm

- They propose an algorithm called surrogate-based promising area search (SPAS) for solving Lipschitz continuous simulation optimization problems.
- SPAS proceeds iteratively by constructing and optimizing a sequence of surrogate models,
- which are approximations of the objective function on promising subsets of the solution space

Idea behind the algorithm

- Each iteration of the algorithm consists of the following steps:
  1. Generate a set of candidate solutions by randomly sampling from the promising region constructed in the previous iteration
  2. Use the shrinking ball technique to estimate the performance of the sampled solutions.
  3. Use all candidate solutions generated thus far to build a surrogate model of the objective function.
  4. Optimize the surrogate model and construct a new promising region that contains the optimal solution to the model.

Idea behind the algorithm

- Each iteration of the algorithm consists of the following steps:
  1. Generate a set of candidate solutions by randomly sampling from the promising region constructed in the previous iteration
  2. Use the shrinking ball technique to estimate the performance of the sampled solutions.
  3. Use all candidate solutions generated thus far to build a surrogate model of the objective function.
  4. Optimize the surrogate model and construct a new promising region that contains the optimal solution to the model.

Idea behind the algorithm

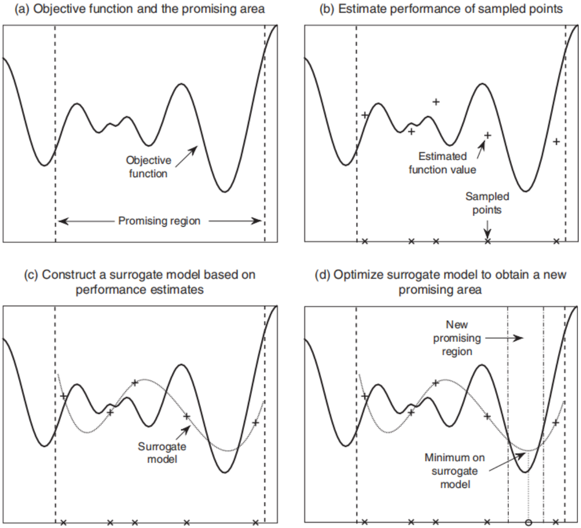
- Each iteration of the algorithm consists of the following steps:
  1. Generate a set of candidate solutions by randomly sampling from the promising region constructed in the previous iteration
  2. Use the shrinking ball technique to estimate the performance of the sampled solutions.
  3. Use all candidate solutions generated thus far to build a surrogate model of the objective function.
  4. Optimize the surrogate model and construct a new promising region that contains the optimal solution to the model.

Idea behind the algorithm

- Each iteration of the algorithm consists of the following steps:
  1. Generate a set of candidate solutions by randomly sampling from the promising region constructed in the previous iteration
  2. Use the shrinking ball technique to estimate the performance of the sampled solutions.
  3. Use all candidate solutions generated thus far to build a surrogate model of the objective function.
  4. Optimize the surrogate model and construct a new promising region that contains the optimal solution to the model.

Idea behind the algorithm

Figure 1. A Schematic Description of SPAS



Idea behind the algorithm

- Intuitively, the shrinking ball method reduces the simulation noise at a sampled solution by averaging observations at solutions that are close to it, avoiding the need to allocate multiple simulation replications to the same point.
- The use of a promising area helps to concentrate the computational effort on subsets of the solution space.
- Additionally, the surrogate model is able to successively predict the response surface of the objective function by using past sampling information.

Idea behind the algorithm

- Note that since the sampling of new solutions is performed within the promising region (as opposed to the entire solution domain),
- the use of the surrogate model in our approach is not intended to provide a global fit of the underlying response surface,
- but rather aims to accurately predict the objective function values at unsampled points within the current search area.
- This facilitates the discovery of better solutions by intensifying the search in the new promising area surrounding the best point predicted by the model.

Locally convergent

- Under some appropriate conditions, they show that the sequence of surrogate model optimizers converges with probability one to the set of local optimal solutions to the original problem.

Relationship with trust region framework

- The algorithm shares some similarities with a class of algorithms developed under the so-called trust region framework (e.g., Deng and Ferris 2009, Chang et al. 2013)
- where the common idea is to use a low-order (linear or quadratic) surrogate model to approximate the true response surface over a predefined trust region and then adaptively adjust the size of the region based on the approximation quality of the model.

Relationship with trust region framework

- However, unlike this approach, which does not use gradient information,
- the analysis of trust-region-based methods typically relies on the twice differentiability of the objective function,
- and some of these algorithms (e.g., Chang et al. 2013) also require the use of gradient and Hessian estimates in constructing local models and determining solution quality.
- In addition, for highly nonlinear problems, because of the limited approximation capability of low-order models, the trust region radius in these algorithms may become very small, which limits the size of the region to be explored.

Relationship with trust region framework

- In contrast, SPAS adopts an interpolation-based fitting strategy and allows for the use of more sophisticated yet practical surrogate models.
- Such models have been shown efficient in approximating high-dimensional nonlinear functions,
- and when used in conjunction with promising region search, may quickly identify areas of the search space with high-quality solutions at no extra simulation effort.

Outline

Targeted problem

- To solve the following general simulation optimization problem

$$\min_{x \in \mathbb{X}} \{H(x) = \mathbb{E}[h(x, \phi)]\},$$

where the solution space $\mathbb{X}$ is a full-dimensional convex, compact subset of $\mathbb{R}^d$ with nonempty interior.

- We assume that the expectation cannot be computed analytically and instead needs to be estimated.

Mathematical notation

- N_k is the number of candidate solutions sampled at the k th iteration of the algorithm.
- Λ_k is the set of sampled solutions at the k th iteration.
- V_k is the collection of all candidate solutions sampled up to the k th iteration,
- $\{r_k\}_{k\geq 1}$ is a sequence of deterministic positive real numbers,
- $B(x, r) = \{y \in \mathbb{X} : d(x, y) < r\}$ is an open ball of center x and radius r.
- For two given points x and y in $\mathbb{X}$, we use $d(x, y)$ to denote the Euclidean distance between them,
- whereas for a set $A \subseteq \mathbb{X}$, the distance between a point x and the set A is defined and denoted by $d(x, A) = \inf_{y\in A} d(x, y)$.
- Finally, let S_k and $P_k \subseteq \mathbb{X}$ be the respective surrogate model and promising area constructed at the k th iteration of SPAS.

Algorithm Description

Step 0

- Set the iteration counter $k = 0$, $V_0 = \varnothing$, and $P_0 = \mathbb{X}$.
- Specify a small positive constant $\delta > 0$, a sequence of numbers $\{\alpha_k\}_{k \geq 1}$ satisfying $\alpha_k \in [0, 1)$, $\forall k$, and a shrinking ball strategy $\{r_k\}_{k \geq 1}$.

Algorithm Description

step 1

- Let $k = k + 1$.
- Uniformly and independently sample a set of N_k candidate solutions $\Lambda_k = \left\{x_1^k, x_2^k, \ , \ldots, x_{N_k}^k\right\}$ from the current promising area P_{k-1}.
- Let $V_k = V_{k-1} \cup \Lambda_k$.
- Obtain the sample performance at each point in Λ_k and use the shrinking ball method to construct performance estimates $\tilde{H}_k(x)$ for all $x \in V_k$ as follows:

$$\tilde{H}_k(x) = \alpha_k \frac{\sum_{y \in B(x,r_k) \cap V_k} h(y)}{|B(x, r_k) \cap V_k|} + (1 - \alpha_k) \frac{\sum_{y \in B(x,r_k) \cap \Lambda_k} h(y)}{|B(x, r_k) \cap \Lambda_k|}$$

- where $|A|$ represents the cardinality of a set A.

Algorithm Description

Step 2

- Build a surrogate model $S_k(x)$ that interpolates the objective function estimates $\tilde{H}_k(x)$ at all sampled points $x \in V_k$.

Step 3

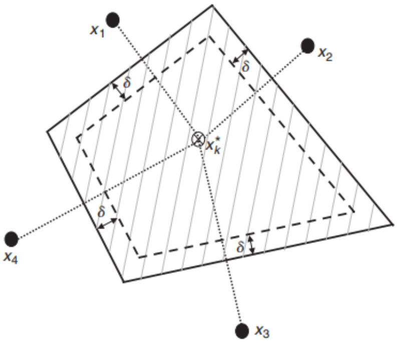
- Optimize the surrogate model $S_k(x)$ on P_{k-1} to obtain a minimizer x_k^*,that is, $x_k^* \in \arg\min_{x \in P_{k-1}} S_k(x)$.
- Construct a new promising area P_k based on x_k^* as follows:

$$P_k = \left\{ y \in \mathbb{X} : d\left(y, x_k^*\right) \leq d\left(y, x + 2\left(x - x_k^*\right) \frac{\delta}{d\left(x_{k'}^*, x\right)} \right), \forall x \in V_k \right\}.$$

- Reiterate from Step 1 until a stopping condition is satisfied.

Some remarks on promising area

Figure 2. Graphical Illustration of the Promising Area in Two Dimensions, Where x_k^* Is a Surrogate Model Minimizer and $x_1, \ldots, x_4$ Are Four Sampled Solutions



Notes. The region circumscribed by the dashed line represents the set of points that are closer in distance to x_k^* than to the sampled solutions. The promising area (shaded region) is formed by expanding the boundaries of the set by δ unit(s) in the directions of the vectors $\overrightarrow{x_k^* x_i}$, $i = 1, \ldots, 4$.

Some remarks on promising area

- In Step 3, we construct a new promising subset P_k, which is defined as the set of points in X whose distances to x_k^* are less than 2δ plus their distances to the set of sampled solutions.
- The use of the constant $\delta > 0$ ensures P_k to have a nonempty interior and prevents it from degenerating into a single point when the set of sampled points becomes dense in the neighborhood of x_k^*.

Some remarks on promising area

- This is conducted in a way that is very similar to the COMPASS approach.
- The major difference is that now the construction is based on the best point predicted by the surrogate model rather than the one with the current best estimated performance.

Some remarks on promising area

- The intuition is that when the solution space is continuous, the current best sampled solution may be far from being optimal;
- consequently, using the point to directly construct the promising area may result in the search of new solutions being conducted in a region that is very distant from the set of true (local) optimizers, leading to slow convergence or inferior local solutions.

Some remarks on promising area

- The surrogate model, on the other hand, retains the previous simulation information in predicting the simulation responses at unsampled solutions.
- Thus, if the model can correctly capture the behavior of the true response surface, then its optimizer would be a more reliable estimate of the true (local) optimal solution than the best sampled solution itself.

Remarks on convex combination technique in step 2

- Another subtle issue worth mentioning is that since the construction of promising areas is adaptive, the performance estimates obtained at successive iterations of the algorithm are generally not independent.
- For example, the sample performance $h(x)$ at a point x generated in the k th iteration will affect the shape and size of the promising region P_k obtained at Step 3.
- This will in turn determine the chance/likelihood of the points to be produced in the next iteration.

Remarks on convex combination technique in step 2

- Thus, for a given sampled solution $x \in V_k$, the observations at points that were generated preceding it are correlated,
- and a straightforward estimation of its true performance $H(x)$ by averaging past observations (such as the shrinking ball method) will result in an extra bias.
- They address this issue by taking the estimator $\tilde{H}_k(x)$ as the convex combination of the average of the observations collected at all points in $B(x, r_k) \cap V_k$ and the average of the observations at points in $B(x, r_k) \cap \Lambda_k$.

Remarks on convex combination technique in step 2

- The second average in the combination only depends on points sampled at the current iteration Λ_k and does not suffer from the correlation bias, whereas the first term relies on past sampling information and is hence biased.
- This bias effect is discounted by putting a weight parameter $\alpha_k \in [0, 1)$ that diminishes as more points are generated.

Remarks on convex combination technique in step 2

- Intuitively speaking, since there are only a few points generated in the early iterations, setting the initial values of α_k large (close to 1) helps to effectively use the performance estimates collected at previously sampled points to reduce the variance of the estimator.
- On the other hand, as sampling gets more focused on the current promising area, the variance of the second term becomes smaller while the correlation bias accumulated in the first term can be removed by letting α_k decrease to zero.

Relationship with other approaches

- It is interesting to observe that in SPAS,
- if there is no surrogate model and the promising area is taken to be the entire feasible region in all iterations,
- then the algorithm is identical to the deterministic shrinking ball method discussed in Andradóttir and Prudius (2010).

Relationship with other approaches

- On the other hand, if the solution space is (discrete) integer ordered, then since each ball $B(x, r_k)$ will only contain x itself (when r_k becomes small enough), the shrinking ball strategy reduces to the usual sample average approximation.
- Thus, the algorithm (without the surrogate model) reduces to the COMPASS algorithm.
- In this respect, SPAS can essentially be seen as the extension of COMPASS to continuous simulation optimization.

Outline

Some definations

- We define

$$\mathscr{F}_k = \sigma\left\{\Lambda_1, \left\{h(x), x \in \Lambda_1\right\}, \ldots, \Lambda_k, \left\{h(x), x \in \Lambda_k\right\}\right\}, k = 1, 2, \ldots$$

  as the sequence of increasing σ-fields generated by the set of all sampled solutions and their corresponding sample performance measures obtained up to iteration k.
- In the rest of the paper, F_k denotes the uniform sampling measure (conditional on $\mathscr{F}_{k-1}$) used at the k th iteration,
- a sequence a_k is said to be $\Omega\left(k^n\right)$ if $\exists c > 0$ and $k_0 > 0$, s.t. $\forall k \geq k_0, a_k \geq ck^n$
- and to be $\Theta\left(k^n\right)$ if $\exists c_1, c_2 > 0$ and $k_0 > 0$, s.t. $\forall k \geq k_0, c_1 k^n \leq a_k \leq c_2 k^n$.

Assumptions

Assumption 1 (A1)

The objective function $H(x)$ is Lipschitz continuous on $\mathbb{X}$ with Lipschitz constant L_1.

Assumption 2 (A2)

Conditional on $\mathscr{F}_{k-1}$ and given Λ_k, the simulation noises $h(x) - H(x)$ at all $x \in \Lambda_k$ are independent with mean zero. In addition, $h(x) - H(x)$ is uniformly bounded on $\mathbb{X}$, that is, there exists $0 < \mathscr{B} < \infty$ such that $|h(x) - H(x)| < \mathscr{B}$ for all $x \in \mathbb{X}$ w.p.1.

Assumptions

Assumption 3 (A3)

The surrogate model $S_k(x)$ satisfies $S_k(x) = \tilde{H}_k(x), \forall x \in V_k$. Moreover, all S_k 's are Lipschitz continuous on $\mathbb{X}$ with their Lipschitz constants uniformly bounded by L_2 for all k w.p.1.

Assumption 4 (A4)

$N_k = \Theta\left(k^t\right), r_k = \Omega\left(k^{-p/d}\right)$ with $\lim_{k\to 0} r_k = 0$, where t and pare two positive constants satisfying $p < t$. The weight parameter α_k satisfies $\alpha_k \in [0, 1) \forall k$ and $\lim_{k\to\infty} \alpha_k = 0$.

Convergence analysis

- Let $\mathscr{M}$ be the set of all local minimizers.
- Our main result is to show that the sequence of the surrogate model minimizers $\{x_k^*\}_{k\geq 1}$ will converge to $\mathscr{M}$ with probability one.
- Our analysis proceeds in several steps. First, we prove the following result, which implies that the collection of sampled solutions will eventually become dense in P_k.

Lemma 1

For any $\epsilon > 0$ and $x_{k-1} \in P_{k-1}$, define the event $A_k(x_{k-1}, \epsilon) = \{\exists y \in \Lambda_k, d(x_{k-1}, y) < \epsilon\}$. If Assumption A4 holds, then $\sum_{k=1}^{\infty} P\left(\bar{A}_k(x_{k-1}, \epsilon) \mid \mathscr{F}_{k-1}\right) < \infty$ w.p.1.

Convergence analysis

- The next result shows that for any point x in the promising area P_{k-1}, its true objective function value $H(x)$ can be closely approximated by the surrogate model $S_k(x)$ as the number of iterations gets large.

Lemma 2

If Assumptions $\mathrm{A1} - \mathrm{A4}$ hold, then for any $\epsilon > 0$ and $x_{k-1} \in P_{k-1}$, we have $\sum_{k=1}^{\infty} P\left(|S_k\left(x_{k-1}\right) - H\left(x_{k-1}\right)| > \epsilon \,|\mathscr{F}_{k-1}\right) < \infty$ w.p.1.

Convergence analysis

- The next result is a strengthened version of Lemma 2 , which shows that the objective function $H(x)$ can be closely approximated by the surrogate model $S_k(x)$ uniformly for all points x in the promising area P_{k-1}.

Proposition 1

If Assumptions $\mathrm{A1} - \mathrm{A4}$ hold, then for any $\epsilon > 0$,
$P\left(\max_{x \in P_{k-1}} |S_k(x) - H(x)| > \epsilon \ i.o.\right) = 0.$

Convergence analysis

- Since $x_k^* \in \operatorname{argmin}_{x \in P_{k-1}} S_k(x)$ and $S_k(x)$ is close to $H(x)$ uniformly over P_{k-1},
- it is reasonable to expect that $H(x_k^*)$ should also be close to the minimum of the function $H(x)$ over the promising area P_{k-1}. This intuition is formalized next.

Lemma 3

If Assumptions $\mathrm{A1} - \mathrm{A4}$ hold, then for any $\epsilon > 0$,
$P\left(\left|H(x_k^*) - \min_{x \in P_{k-1}} H(x)\right| > \epsilon \text{ i.o. }\right) = 0.$

Convergence analysis

- In addition, since $S_k(x)$ is uniformly close to the true objective function $H(x)$ on P_{k-1},
- Proposition 1 also suggests that the distance between x_k^* and the set of minimizers of H on the promising area P_{k-1} will approach zero as k tends to infinity.
- This leads to Proposition 2 .

Proposition 2

If Assumptions A1-A4 hold, then

$$P\left(\lim_{k\to\infty} d\left(x_k^*, \operatorname*{argmin}_{x\in P_{k-1}} H(x)\right) = 0\right) = 1.$$

Convergence analysis

- Finally, we arrive at the following convergence result for the SPAS algorithm.

Theorem 1

If Assumptions $\mathrm{A1} - \mathrm{A4}$ hold, then

$$P\left(\lim_{k\to\infty} d(x_k^*, \mathscr{M}) = 0\right) = 1.$$

Outline

Implementation of the algorithm

- The surrogate model is constructed using the radial basis function (RBF) approximation method, which has been successfully used as a curve fitting tool in surrogate-based optimization.
- The specific approximator considered here is a linear combination of RBFs of the following form: $S_k(x) = \sum_{i=1}^{|V_k|} w_i \psi\left(\|x - x_i\|\right)$, where $\psi(r) = r^3$, x_i are the sampled solutions, and w_i are the weights of the basis functions, which can be computed by solving a system of linear equations.
- Note that since ψ is chosen to be a polynomial of degree 3 , the derivative of $S_k(x)$ admits an explicit expression and its minimization over the promising area at Step 3 of the algorithm can be conveniently carried out using a straightforward gradient descent method.

Implementation of the algorithm

- The surrogate model is constructed using the radial basis function (RBF) approximation method, which has been successfully used as a curve fitting tool in surrogate-based optimization.
- The specific approximator considered here is a linear combination of RBFs of the following form: $S_k(x) = \sum_{i=1}^{|V_k|} w_i \psi\left(\|x - x_i\|\right)$, where $\psi(r) = r^3, x_i$ are the sampled solutions, and w_i are the weights of the basis functions, which can be computed by solving a system of linear equations.
- Note that since ψ is chosen to be a polynomial of degree 3 , the derivative of $S_k(x)$ admits an explicit expression and its minimization over the promising area at Step 3 of the algorithm can be conveniently carried out using a straightforward gradient descent method.

Implementation of the algorithm

- The surrogate model is constructed using the radial basis function (RBF) approximation method, which has been successfully used as a curve fitting tool in surrogate-based optimization.
- The specific approximator considered here is a linear combination of RBFs of the following form: $S_k(x) = \sum_{i=1}^{|V_k|} w_i \psi\left(\|x - x_i\|\right)$, where $\psi(r) = r^3, x_i$ are the sampled solutions, and w_i are the weights of the basis functions, which can be computed by solving a system of linear equations.
- Note that since ψ is chosen to be a polynomial of degree 3 , the derivative of $S_k(x)$ admits an explicit expression and its minimization over the promising area at Step 3 of the algorithm can be conveniently carried out using a straightforward gradient descent method.

Deterministic Functions with Added Noise

- Tests were performed on 10 deterministic functions with added noise.
- These functions are well known and have been widely used in the literature to investigate the performance of various optimization algorithms.

Deterministic Functions with Added Noise

- In particular, problems h_1, h_2, and h_3 are unimodal, each with a unique local (global) minimizer.
- Functions h_4 and h_5 are low-dimensional problems with a few local minima,
- while the last five functions $h_6 - h_{10}$ are highly multimodal with the number of local minima grows exponentially with the problem dimension.

Deterministic Functions with Added Noise

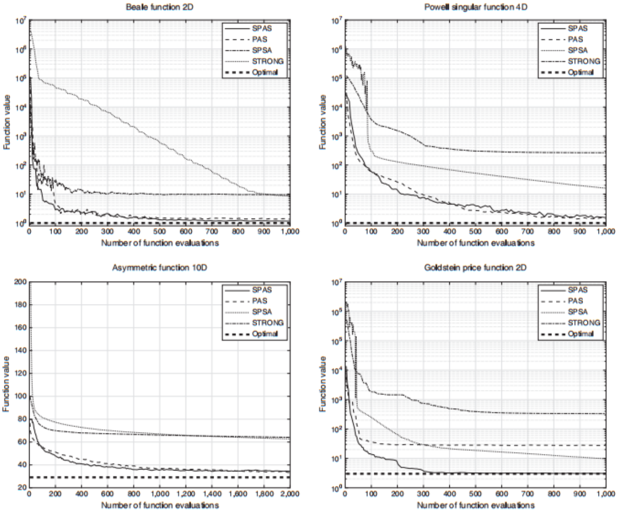
- In each case, the added noise is assumed to follow a zeromean truncated normal distribution $\mathscr{T}\mathcal{N}\left(0,\sigma^2\right)$, which is the normal distribution $\mathcal{N}\left(0,\sigma^2\right)$ truncated over the region $[-3\sigma, 3\sigma]$.

Deterministic Functions with Added Noise

- For comparison purposes, they have also applied the simultaneous perturbation stochastic approximation (SPSA) algorithm and the STRONG method on the 10 testing problems.
- To further illustrate the benefit of using surrogate models in the proposed algorithm, they have also included a simplified version of SPAS, called PAS.
- PAS has the same structure as SPAS but without the surrogate model approximation step, and the promising region is constructed at each iteration based on the current best sampled solution.
- Thus, PAS is essentially a version of COMPASS applied to continuous simulation optimization.

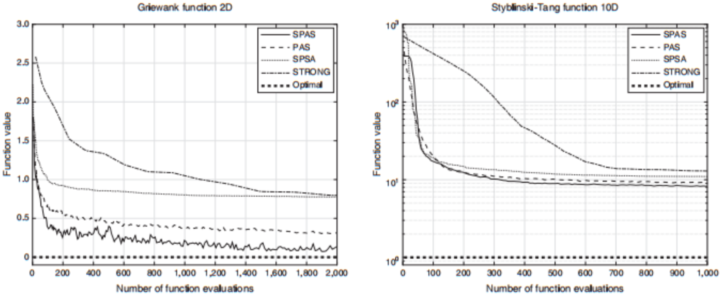
Deterministic Functions with Added Noise

Figure 3. Averaged Performance of Comparison Algorithms on Test Functions h_1–h_6



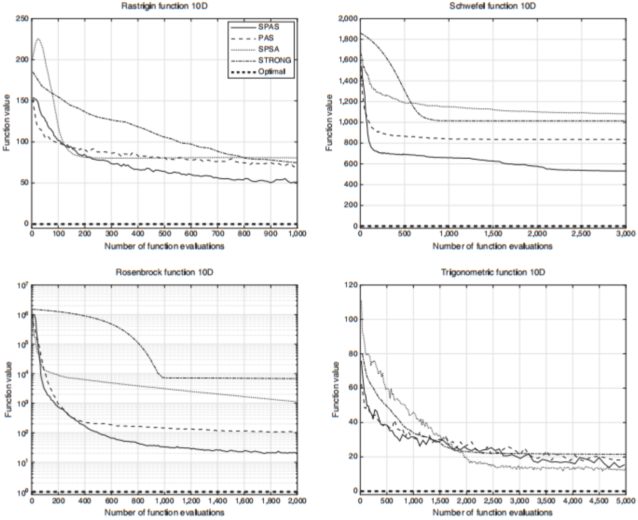
Deterministic Functions with Added Noise

Figure 3. (Continued)



Deterministic Functions with Added Noise

Figure 4. Averaged Performance of Comparison Algorithms on Test Functions h_7–h_{10}



An Inventory Control Example

- They also consider a discrete-time (s, S) inventory control problem with independent and identically distributed exponentially distributed demands.
- The inventory level is reviewed at the beginning of each time period.
- When the inventory position falls below the level s, an order is placed to increase the inventory position to S.
- The objective is to find the optimal threshold values, s^* and S^*, in order to minimize the long-run average cost per period.

An Inventory Control Example

Table 3. Four Test Cases of the Inventory Problem

Case	$E[\mathcal{D}_t]$	p	K	$J(s^*, S^*)$
1	20	1	10	40.00
2	20	10	100	102.68
3	200	10	100	740.95
4	200	100	1,000	1,470.30

Table 4. Performance of Different Algorithms on the Four Test Cases (Standard Errors in Parentheses)

Case	N_{rep}	SPAS	PAS	SPSA	Strong
1	200	40.47 (0.16)	98.12 (22.90)	422.29 (36.09)	636.50 (61.03)
2	200	103.94 (0.24)	147.47 (18.85)	426.78 (36.14)	585.47 (66.83)
3	1,000	750.83 (2.12)	756.51 (4.32)	841.29 (10.25)	941.59 (29.52)
4	1,000	1,499.01 (5.84)	1,498.68 (5.74)	1,664.13 (63.44)	2,476.70 (322.37)

An Inventory Control Example

Table 5. Computational Run Times (in Seconds) of Different Algorithms (Standard Errors in Parentheses)

Case	N_{rep}	SPAS	PAS	SPSA	Strong
1	200	6.24 (0.11)	1.98 (0.03)	0.84 (1.8e−3)	0.99 (0.01)
2	200	6.17 (0.09)	1.89 (0.03)	0.86 (3.7e−3)	1.02 (0.03)
3	1,000	48.47 (0.28)	23.37 (0.19)	4.13 (3.6e−3)	5.16 (0.36)
4	1,000	47.69 (0.36)	24.64 (0.16)	4.49 (1.5e−2)	6.02 (0.48)

Conclusions

- In this paper, by integrating ideas from the shrinking ball method, surrogate model approximation, and promising region search, we have proposed a novel approach, called SPAS, for solving Lipschitz continuous simulation optimization problems.
- Under appropriate conditions, we have shown that the algorithm converges almost surely to the set of local optimal solutions.
- The performance of SPAS has been illustrated on a set of 10 benchmark testing problems and an inventory control example.
- Empirical results on these examples indicate that the algorithm is promising and may significantly outperform some existing methods that exploit gradient information.

Extension

- Surrogate model can be neural network or Kriging.
- N_k may only need to be a constant.
- Correlation bias issue may can also be tackled with martingale approach.

Thanks for listening !