Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

ENHANCED DERIVATIVE-FREE OPTIMIZATION USING ADAPTIVE CORRELATION-INDUCED FINITE DIFFERENCE ESTIMATORS

Guo Liang[1]

[1]Institute of Statistics and Big Data, Renmin University of China, Beijing, CHINA

ABSTRACT

Gradient-based methods are well-suited for derivative-free optimization (DFO), where finite-difference (FD) estimates are commonly used as gradient surrogates. Traditional stochastic approximation methods, such as Kiefer-Wolfowitz (KW) and simultaneous perturbation stochastic approximation (SPSA), typically utilize only two samples per iteration, resulting in imprecise gradient estimates and necessitating diminishing step sizes for convergence. In this paper, we combine a batch-based FD estimate and an adaptive sampling strategy, developing an algorithm designed to enhance DFO in terms of both gradient estimation efficiency and sample efficiency. Furthermore, we establish the consistency of our proposed algorithm and demonstrate that, despite using a batch of samples per iteration, it achieves the same sample complexity as the KW and SPSA methods. Additionally, we propose a novel stochastic line search technique to adaptively tune the step size in practice. Finally, comprehensive numerical experiments confirm the superior empirical performance of the proposed algorithm.

1 ALGORITHM

Gradient-based optimization algorithms are suitable for solving the unconstrained stochastic optimization problem $\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) = \mathbb{E}[F(\mathbf{x})]$, where gradient estimation and step size selection are two fundamental ingredients. In a noisy black-box setting, this paper integrates the adaptive sampling strategy (Bollapragada et al. 2024) and a novel gradient estimation method to determine the current descent direction, and proposes a stochastic line search to determine an appropriate step size in practical applications.

1.1 Correlation-Induced Central Finite Difference Estimation

In this section, we assume $\mathcal{X} \in \mathbb{R}$ and denote x_k by the point of interest. Our objective is to use n_k sample pairs (i.e., $2n_k$ function evaluations) to estimate the gradient. The procedure of our correlation-induced central FD (Cor-CFD) method is as follows:

- **Step 1.** Generates R perturbations $h_{k,1}, ..., h_{k,R}$ randomly from a pilot distribution $\mathcal{P}_0$.
- **Step 2.** Generate b_k ($n_k = b_k R$) sample pairs $(F_i(x_k + h_{k,r}), F_i(x_k - h_{k,r}))(i = 1, ..., b_k)$ for each perturbation $h_{k,r}(r = 1, ..., R)$ and construct CFD estimates $g_{b_k,h_{k,r}} := \sum_{i=1}^{b_k}(F_i(x_k + h_{k,r}) - F_i(x_k - h_{k,r}))/(2b_k)$ at each perturbation.
- **Step 3.** Use bootstrap and linear regression to estimate the optimal perturbation $\widehat{h}_k$ and transform $g_{b_k,h_{k,r}}$ to $g^{cor}_{b_k,r,\widehat{h}_k}, (r = 1, ..., R)$.
- **Step 4.** Average $g^{cor}_{b_k,r,\widehat{h}_k}$ and get the Cor-CFD estimate $g_k(x_k)$.

1.2 Adaptive Sampling

If n_k is too small, the error of $g_k(\mathbf{x}_k)$ is substantial; if n_k is too large, many samples may be wasted. The idea of adaptive sampling is to select an n_k that ensures the angle between $g_k(\mathbf{x}_k)$ and the true gradient is acute, which can, to some extent, be derived from $\mathbb{E}[|\boldsymbol{\epsilon}_k|^2|\mathcal{F}_k] \leq \theta^2|\nabla f(\mathbf{x}_k)|^2$ (*norm condition*).

Table 1: Optimality gap of different methods for the Rosenbrock function with noise.

Function Evaluations	2×10^3			2×10^4			2×10^5		
Method	Mean	SD	SR	Mean	SD	SR	Mean	SD	SR
AdaDFO	**2.85**	**2.69**	**100%**	**0.51**	**0.99**	**100%**	**0.32**	**0.34**	**100%**
NM	6.59	0.25	100%	6.58	0.22	100%	6.59	0.22	100%
STRONG	16.74	0.02	100%	4.99	0.55	100%	4.91	0.71	100%
SPSA	3.55	–	99.2%	1.08	–	98.4%	0.13	–	98.7%

Note that $\mathbb{E}[|\boldsymbol{\epsilon}_k|^2|\mathcal{F}_k]$ represents the sum of variances across all coordinates. We can employ sample variance to estimate the variance of each component of $g_k(\mathbf{x}_k)$ and subsequently $\mathbb{E}[|\boldsymbol{\epsilon}_k|^2|\mathcal{F}_k]$. For $|\nabla f(\mathbf{x}_k)|$, $|g_k(\mathbf{x}_k)|$ can be chosen as an appropriate surrogate. Specifically, the estimated version of the norm condition is $\sum_{i=1}^{d} \widehat{\sigma}_i^2/n_k \leq \theta^2|g_k(\mathbf{x}_k)|^2$, where $\widehat{\sigma}_i^2$ is the sample variance of $g_k(\mathbf{x}_k)$ at i-th coordinate in the k-th iteration, which is an estimated upper bound of the true variance.

1.3 Stochastic Line Search

- **Step 1.** Begin with an initial step size $a_k = \tilde{a}$ at k-th iteration and shrink $a_k \to l_2 a_k$ when
$$F(\mathbf{x}_k - a_k g_k(\mathbf{x}_k)) > F(\mathbf{x}_k) - l_1 a_k |g_k(\mathbf{x}_k)|^2 + 2\sigma_F \tag{1}$$
holds, where $l_1 < l_2 < 1$. Step 1 will stop until (1) does not hold.
- **Step 2.** While $a_k > \underline{a}$, shrink $a_k \to l_2 a_k$ if for any $N \leq N_0$,
$$\frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{x}_k - a_k g_k(\mathbf{x}_k)) > \frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{x}_k) - l_1 a_k |g_k(\mathbf{x}_k)|^2 - 2\frac{\sigma_F}{\sqrt{N}}. \tag{2}$$

In the end, we will obtain a step size either does not satisfy (1) and (2) or lies in $(l_2\underline{a}, \underline{a}]$ (small enough).

2 THEORETICAL RESULTS

- **Iteration Complexity.** Let $\mathbf{x}_0$ be the initial point and at k-th iteration, let $\max\{\mathbb{E}[|\boldsymbol{b}_k|^2|\mathcal{F}_k], \mathbb{E}[|\boldsymbol{\epsilon}_k|^2|\mathcal{F}_k]\} \leq \theta^2|\nabla f(\mathbf{x}_k)|^2$, where $0 < \theta < m/(2M)$ is a threshold. If $0 < a_k = a \leq 1/((2\theta^2 + 2\theta + 1)M)$ for any $k \geq 0$, then we have $\mathbb{E}\left[|\mathbf{x}_k - \mathbf{x}^*|^2\right] \leq (1 - (m - 2\theta M)a)^k\, |\mathbf{x}_0 - \mathbf{x}^*|^2$.
- **Sample Complexity.** Let $d = \mathcal{O}(1)$. Denote $\mathcal{S}(\epsilon)$ by the total stochastic function evaluations to get an ϵ-accurate solution. Under some mild conditions, we have $\mathbb{E}[\mathcal{S}(\epsilon)] \geq \mathcal{C}_1\epsilon^{-3/2} + \mathcal{C}_2$, where $\mathcal{C}_1$ and $\mathcal{C}_2$ are constants that depends on the threshold θ, step size a, problem dimension d, unknown function and the simulation error.

It follows the theoretical results that achieving an ϵ-accurate solution requires at least $\mathcal{O}\left(\epsilon^{-3/2}\right)$ function evaluations, matching the optimal performance of the KW algorithm.

3 NUMERICAL EXPERIMENTS

Table 1 demonstrates the effectiveness of AdaDFO under noise. Compared with NM and STRONG methods, the optimality gap of AdaDFO method is smaller. In addition, AdaDFO consistently achieves 100% success rate (SR) across all budgets compared with SPSA method.

REFERENCES

Bollapragada, R., C. Karamanli, and S. M. Wild. 2024. "Derivative-Free Optimization via Adaptive Sampling Strategies". *arXiv preprint arXiv:2404.11893*.