Gradient-free methods in convex optimization

Vitaliy Eroshin

Advisers: A. V. Gasnikov, A. V. Lobanov

Moscow Institute of Physics and Technology

March 2024

The "Black-Box" Optimization Problem

Consider a common stochastic convex optimization problem:

$$\min_x \left[f(x) := \mathbb{E}_{\xi \sim \mathcal{D}} \left[f(x, \xi) \right] \right]$$

We study the case where we only have access to value $f(x)$ and nothing else.

Such problem can be classified as a "*black-box*".

Motivation

This class of optimization problem recently has received significant attention in the setting of reinforcement learning, federated learning, distributed learning and overparameterized models.

In addition, the "black-box" problem arises when computation of gradient is too expensive or not available.

Zero-Order Oracle

We introduce a concept of *Zero-Order (gradient-free) Oracle* .

Such oracle cannot provide a value of gradient, but it is able to return function value.

In addition, we study a noisy oracle $\widetilde{f}$:

$$\widetilde{f} = f(x, \xi) + \delta$$

δ is a noise. Two popular concepts of noise:

- determined $\delta(x) < \Delta$
- stochastic $\mathbb{E}\left[\delta^2\right] \leq \Delta^2$

Gradient Approximation

Convex optimization with gradient computation is well studied.

Idea: approximate gradient with only zero-order oracle calls

Kernel-based Approximation

Approximation was introduced in [1]. It requires f to have increased smoothness and has the following form:

$$g(x, e) = d\frac{\widetilde{f}(x + hre) - \widetilde{f}(x - hre)}{2h}K(r)e$$

- $h > 0$ - smoothing parameter
- $e \in S_2^d(1)$ - uniformly distributed
- $r \in [0, 1]$ - uniformly distributed
- $K : [-1, 1] \to \mathbb{R}$ - kernel function

Such approximation can be used as gradient oracle, so we can apply it to known gradient methods.

Strong Growth Condition

Strong Growth Condition:
There exists constants ρ, such that

$$\mathbb{E}\|\nabla f(x,\xi)\|^2 \leq \rho\|\nabla f(x)\|^2$$

More general condition of SG:
There exists constants ρ, σ^2, such that

$$\mathbb{E}\|\nabla f(x,\xi)\|^2 \leq \rho\|\nabla f(x)\|^2 + \sigma^2$$

Biased Gradient Oracle

Biased Gradient Oracle:
A map $\mathbf{g} : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^d$ for a bias $\mathbf{b}$: $\mathbb{R} \to \mathbb{R}$ such that:

$$\mathbf{g}(x, \xi) = \nabla f(x, \xi) + \mathbf{b}(x)$$

Bounded bias:
There exists constant $\delta \geq 0$ such that:

$$\|\mathbf{b}(x)\| = \|\mathbb{E}\left[\mathbf{g}(x, \xi)\right] - \nabla f(x)\| \leq \delta$$

Bounded noise:
There exists constants ρ, σ^2, such that

$$\mathbb{E}\|g(x, \xi)\|^2 \leq \rho\|\nabla f(x)\|^2 + \sigma^2$$

Remember generalized SGC?

Accelerated SGD by Nesterov

Update rules of AccSGD algorithm:

$$
\begin{aligned}
x_{k+1} &= y_k - \eta \mathbf{g}(y_k, \xi_k) \\
y_k &= \alpha_k z_k + (1 - \alpha_k) x_k \\
z_{k+1} &= \zeta_k z_k + (1 - \zeta_k) y_k - \gamma_k \eta \mathbf{g}(y_k, \xi_k)
\end{aligned}
$$

Where

x_k, y_k, z_k - sequences, updated in each iteration

α_k, ζ_k - tunable parameters

Convergence of AccSGD

Theorem 1. ([2])
Let the function f is L-smooth, and the unbiased gradient oracle $g(x,\xi) = \nabla f(x,\xi)$ satisfies SGC, then the accelerated Stochastic Gradient Descent by Nesterov with chosen parameters:

$$\gamma_k = \frac{\rho^{-1}+\sqrt{\rho^{-2}+4\gamma_{k-1}^2}}{2}; \quad a_{k+1} = \gamma_k\sqrt{\eta\rho}; \quad \alpha_k = \frac{\gamma_k\eta}{\gamma_k\eta+a_k^2}; \quad \eta = \frac{1}{\rho L}$$

has the following rate of convergence:

$$\mathbb{E}\left[f(X_n)\right] - f^* \lesssim \frac{\rho^2 L R^2}{N^2} + \frac{N\sigma^2}{L\rho^2}$$

Convergence of AccSGD with biased gradient

Theorem 2. ([4], Theorem 3.1)
Let the function f is L-smooth, and the gradient oracle $g(x,\xi) = \nabla f(x,\xi)$ has bounded bias and noise, then the accelerated SGD with batching by Nesterov with $\rho_B = \max\{1, \frac{\rho}{B}\}$ and chosen parameters:

$$\gamma_k = \frac{\rho_B^{-1} + \sqrt{\rho_B^{-2} + 4\gamma_{k-1}^2}}{2}; \quad a_{k+1} = \gamma_k \sqrt{\eta \rho_B}; \quad \alpha_k = \frac{\gamma_k \eta}{\gamma_k \eta + a_k^2}; \quad \eta = \frac{1}{\rho_B L}$$

has the following rate of convergence:

$$\mathbb{E}\left[f(X_n)\right] - f^* \lesssim \frac{\rho_B^2 L R^2}{N^2} + \frac{N\sigma^2}{LB\rho_B^2} + \delta\widetilde{R} + \frac{N}{L}\delta^2$$

Zero-Order Accelerated Stochastic Gradient Descent

Recently was introduced ZO-AccSGD algorithm [4], which is obtained by using gradient approximation in accelerated by Nesterov Stochastic Gradient Descent for biased gradient [3].

ZO-AccSGD improves the convergence result (iteration complexity) of previous state-of-the-art algorithms for our problem formulation.

Plans & Contribution

- Convergence results for ZO-AccSGD were obtained in concept of stochastic oracle.

  We plan to generalize [4] and get convergence results of ZO-AccSGD in concept of oracle with determined bounded noise.

- Further work towards research of gradient-free methods.

References I

[1] B. T. Polyak and A. B. Tsybakov. "Optimal orders of accuracy for search algorithms of stochastic optimization". In: *Problemy Peredachi Informatsii* 26.2 (1990), pp. 45–53. ISSN: 0555-2923.

[2] Sharan Vaswani, Francis Bach, and Mark Schmidt. "Fast and Faster Convergence of SGD for Over-Parameterized Models and an Accelerated Perceptron". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 16–18 Apr 2019, pp. 1195–1204. URL: https://proceedings.mlr.press/v89/vaswani19a.html.

References II

[3] Ahmad Ajalloeian and Sebastian U. Stich. "Analysis of SGD with Biased Gradient Estimators". In: *CoRR* abs/2008.00051 (2020). arXiv: 2008.00051. URL: https://arxiv.org/abs/2008.00051.

[4] Aleksandr Lobanov, Nail Bashirov, and Alexander Gasnikov. *The Black-Box Optimization Problem: Zero-Order Accelerated Stochastic Method via Kernel Approximation*. 2023. arXiv: 2310.02371 [math.OC].