Gradient-free methods in convex optimization

Vitaliy Eroshin

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

Moscow Institute of Physics and Technology

May 2024

Contents

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 an access to value $f(x)$ (probably with some noise) and nothing else.

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

Motivation

"Black-Box" problems usually arises when computation of gradient is unavailable, or it is too expensive.

Nowadays, such problems often appear in various settings of reinforcement learning [7], federated learning [6], distributed learning and overparameterization

Methods

Usually, "black-box" problems are being solved using gradient-free (zero-order) methods.

There are two main approaches:

- Evolutionary algorithms [1]
- Exploit first-order methods with gradient approximation

Zero-Order algorithms

There is a number of algorithms, acquired by using gradient approximation for various settings.

- ZO-RSMD [3] - gradient-free version of Robust Stochastic Mirror Descent
- ZO-Clip-SMD - gradient clipping
- ZO-MB-SGD [5] - Mini-Batch SGD
- etc...

Zero-Order Accelerated SGD

Recently was introduced ZO-AccSGD algorithm [4], which is obtained by using gradient approximation in accelerated by Nesterov Stochastic Gradient Descent for biased gradient [2].

Original work proposes convergence and noise estimations in the concept of oracle with stochastic noise

Contribution

Obtained convergence rate of ZO-AccSGD and estimations for allowed noise in concept of oracle with deterministic noise.

Theorem

Convergence of ZO-AccSGD for oracle with deterministic noise

Let the following assumptions be satisfied:

1. f has higher order smoothness
2. $g(x, e)$ (kernel approximation) has bounded bias and noise

Then ZO-AccSGD with $\rho_B = \max\{1, \frac{4d\kappa}{B}\}$ and with chosen algorithm parameters:

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

converges to desidred ε accuracy, $\mathbb{E}[f(x_N)] - f^* \leq \varepsilon$

Case	N	Δ
$B \in [1, 4d\kappa]$, $h \lesssim \varepsilon^{3/4}$	$O\left(\sqrt{\frac{d^2 L R^2}{B^2 \varepsilon}}\right)$	$\min\{\frac{\varepsilon^{3/2}}{d^{3/2}}, \frac{\varepsilon^{7/4}}{d}\}$
$B > 4d\kappa$, $h \lesssim \varepsilon^{1/(\beta-1)}$	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$	$\min\{\frac{\varepsilon^{\frac{3\beta+1}{4(\beta-1)}} B^{1/2}}{d}, \frac{\varepsilon^{1+\frac{1}{\beta-1}}}{d}, \frac{\varepsilon^{\frac{3\beta+1}{4(\beta-1)}}}{d^{3/2}}\}$

Proof idea

Firstly we need to bound a bias of gradient approximation

$$
\begin{aligned}
&\|\mathbb{E}[g(x_k,\xi,e)] - \nabla f(x_k)\| \leq \\
&\quad \leq \left\|\mathbb{E}\left[\frac{f(x+hre) - f(x-hre) + \delta_1 - \delta_2}{2h} deK(r)\right] - \nabla f(x)\right\| \leq \\
&\quad\quad \leq dLh^{b-1}\kappa_\beta \mathbb{E}\left[\|e^\beta\|\right] + \frac{d}{h}\Delta \leq Lh^{b-1}\kappa_B + \frac{d\Delta}{h}
\end{aligned} \tag{1}
$$

And bounding second moment of gradient approximation

$$
\mathbb{E}\left[\|g(x_k,\xi,e)\|^2\right] \leq 4d\kappa\|\nabla f(x_k)\|^2 + 4d\kappa L^2 h^2 + \frac{\kappa d^2\Delta^2}{h^2} \tag{2}
$$

Proof idea

(1) and (2) equations give us constants

$$\rho = 4d\kappa \qquad \sigma^2 = 4d\kappa L^2 h^2 + \frac{\kappa d^2 \Delta^2}{h^2} \qquad \delta = L h^{b-1}\kappa_B + \frac{d\Delta}{h}$$

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:

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

Proof idea

After substitution, we can carefully estimate interesting parameters N and δ. That gives:

Case	N	Δ
$B \in [1, 4d\kappa],\ h \lesssim \varepsilon^{3/4}$	$O\left(\sqrt{\frac{d^2 L R^2}{B^2 \varepsilon}}\right)$	$\min\{\frac{\varepsilon^{3/2}}{d^{3/2}}, \frac{\varepsilon^{7/4}}{d}\}$
$B > 4d\kappa,\ h \lesssim \varepsilon^{1/(\beta-1)}$	$O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$	$\min\{\frac{\varepsilon^{\frac{3\beta+1}{4(\beta-1)}} B^{1/2}}{d}, \frac{\varepsilon^{1+\frac{1}{\beta-1}}}{d}, \frac{\varepsilon^{\frac{3\beta+1}{4(\beta-1)}}}{d^{3/2}}\}$

Overview

Interest in gradient-free algorithms grows in recent years, so it is important to study bounds of what such algorithms capable of.

This work generalizes bounds and estimations for ZO-AccSGD, which can be used in the corresponding applications.

References I

[1] A. Auger and N. Hansen. "A restart CMA evolution strategy with increasing population size". In: *2005 IEEE Congress on Evolutionary Computation*. Vol. 2. 2005, 1769–1776 Vol. 2. DOI: 10.1109/CEC.2005.1554902.

[2] 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.

[3] Nikita Kornilov et al. *Gradient-Free Methods for Non-Smooth Convex Stochastic Optimization with Heavy-Tailed Noise on Convex Compact*. 2023. arXiv: 2304.02442 [math.OC].

[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].

References II

[5] Aleksandr Lobanov, Alexander Gasnikov, and Fedor Stonyakin. *Highly Smoothness Zero-Order Methods for Solving Optimization Problems under PL Condition*. 2023. arXiv: 2305.15828 [math.OC].

[6] Wang Lu et al. *ZooPFL: Exploring Black-box Foundation Models for Personalized Federated Learning*. 2023. arXiv: 2310.05143 [cs.AI].

[7] Lei Song et al. *Reinforced In-Context Black-Box Optimization*. 2024. arXiv: 2402.17423 [cs.LG].