Survey of modern results in gradient-free convex optimization

Alexander Gasnikov, Aleksandr Lobanov

Innopolis University, Innopolis, Russia
Moscow Institute of Physics and Technology, Dolgoprudny, Russia

gasnikov@yandex.ru, lobbsasha@mail.ru

October 11, 2024

Gratient-free algorithms

We consider a convex optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Question

When should gradient-free algorithms be used?

Gratient-free algorithms

We consider a convex optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Question

When should gradient-free algorithms be used?

Approaches for creating randomized gradient-free methods

- **Non-smooth case**
  - Smoothing scheme with l_1 randomization
  - Smoothing scheme with l_2 randomization
- **Smooth case**
  - l_1 randomization
  - l_2 randomization
- **Case with increased smoothness**
  - Kernel-based approximation

Gratient-free algorithms

We consider a convex optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Question

When should gradient-free algorithms be used?

Optimality criteria

1. Number of oracle calls: T
2. Number of consecutive method iterations: N
3. Maximum allowable noise level Δ

Motivation to find the maximum noise level

Figure: Resource saving

Figure: Robustness to attacks

Figure: Confidentiality

1. Resource Saving. The more accurately the objective function value is calculated, the more expensive this process to be performed.
2. Robustness to Attacks. Improving the maximum noise level makes the algorithm more robust to adversarial attacks.
3. Confidentiality. Some companies, due to secrecy, can't hand over all the information.

Table of Contents

Assumptions on the Objective Function

Problem Formulation

We study a standard convex optimization problem:

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

where $f : \mathbb{R}^d \to \mathbb{R}$ convex function that we want to minimize on the convex set Q.

Assumption. (Higher order smoothness)

Let l denote maximal integer number strictly less than β. Let $\mathcal{F}_\beta(L)$ denote the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ which are differentiable l times and $\forall x, z \in Q$ the Hölder-type condition:

$$\left| f(z) - \sum_{0 \leq |n| \leq l} \frac{1}{n!} D^n f(x)(z-x)^n \right| \leq L_\beta \|z - x\|^\beta,$$

where $L_\beta > 0$, the sum is over multi-index $n = (n_1, ..., n_d) \in \mathbb{N}^d$, we used the notation $n! = n_1! \cdots n_d!$, $|n| = n_1 + \cdots + n_d$, and $\forall v = (v_1, ..., v_d) \in \mathbb{R}^d$ we defined $D^n f(x) v^n = \frac{\partial^{|n|} f(x)}{\partial^{n_1} x_1 \cdots \partial^{n_d} x_d} v_1^{n_1} \cdots v_d^{n_d}$.

Assumptions on the Zero-Order Oracle

Zero-order oracle

We assume that the oracle $\tilde{f}$ can only return the function value $f(x)$ at the requested point x with some stochastic noise ξ:

$$\tilde{f} = f(x) + \xi.$$

Assumption. (Stochastic noise)

We assume that the following holds

- $\xi_1 \neq \xi_2$ such that $\mathbb{E}[\xi_1^2] \leq \Delta^2$ and $\mathbb{E}[\xi_2^2] \leq \Delta^2$, $\Delta \geq 0$ is level noise;
- the random variables ξ_1 and ξ_2 are independent from $\mathbf{e}$ and r.

Gradient approximation with one point feedback

$$\mathbf{g}(x, \mathbf{e}) = d\frac{f(x + hr\mathbf{e}) + \xi_1 - f(x - hr\mathbf{e}) - \xi_2}{2h} K(r)\mathbf{e}.$$

Assumptions on the Gradient Approximation

Definition. (Kernel Approximation)

$$\mathbf{g}(x, \mathbf{e}) = d\frac{f(x + hr\mathbf{e}) + \xi_1 - f(x - hr\mathbf{e}) - \xi_2}{2h}K(r)\mathbf{e}.$$

where $h > 0$ is a smoothing parameter, $\mathbf{e} \in S_2^d(1)$ is a vector uniformly distributed on the Euclidean unit sphere, r is a random value uniformly distributed on the interval $r \in [0, 1]$.

Assumptions on the Gradient Approximation

Definition. (Kernel Approximation)

$$\mathbf{g}(x, \mathbf{e}) = d\frac{f(x + hr\mathbf{e}) + \xi_1 - f(x - hr\mathbf{e}) - \xi_2}{2h} K(r)\mathbf{e}.$$

where $h > 0$ is a smoothing parameter, $\mathbf{e} \in S_2^d(1)$ is a vector uniformly distributed on the Euclidean unit sphere, r is a random value uniformly distributed on the interval $r \in [0, 1]$.

Assumption. (Kernel function)

Let $K : [-1, 1] \to \mathbb{R}$ is a kernel function that satisfies

$$\mathbb{E}[K(u)] = 0,\ \mathbb{E}[uK(u)] = 1,$$
$$\mathbb{E}[u^j K(u)] = 0,\ j = 2, ..., l,\ \mathbb{E}[|u|^\beta |K(u)|] < \infty.$$

Assumptions on the Gradient Approximation

Definition. (Kernel Approximation)

$$\mathbf{g}(x, \mathbf{e}) = d\frac{f(x + hr\mathbf{e}) + \xi_1 - f(x - hr\mathbf{e}) - \xi_2}{2h}K(r)\mathbf{e}.$$

where $h > 0$ is a smoothing parameter, $\mathbf{e} \in S_2^d(1)$ is a vector uniformly distributed on the Euclidean unit sphere, r is a random value uniformly distributed on the interval $r \in [0, 1]$.

Assumption. (Kernel function)

Let $K : [-1, 1] \to \mathbb{R}$ is a kernel function that satisfies

$$\mathbb{E}[K(u)] = 0,\ \mathbb{E}[uK(u)] = 1,$$
$$\mathbb{E}[u^j K(u)] = 0,\ j = 2, ..., l,\ \mathbb{E}[|u|^\beta |K(u)|] < \infty.$$

Example of Kernel function

A weighted sum of Legendre polynoms is an example of such kernels.

Background

References	Iteration Complexity	Maximum Noise Level
Bach, Perchet (2016) [1]	$\mathcal{O}\left(\frac{d^{2+\frac{2}{\beta-1}}\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗
Novitskii, Gasnikov (2020) [2]	$\tilde{\mathcal{O}}\left(\frac{d^{1+\frac{1}{\beta-1}}\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗
Akhavan, Chzhen, Pontil, Tsybakov (2023) [3]	$\tilde{\mathcal{O}}\left(\frac{d^2\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗

Background

References	Iteration Complexity	Maximum Noise Level
Bach, Perchet (2016) [1]	$\mathcal{O}\left(\frac{d^{2+\frac{2}{\beta-1}}\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗
Novitskii, Gasnikov (2020) [2]	$\tilde{\mathcal{O}}\left(\frac{d^{1+\frac{1}{\beta-1}}\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗
Akhavan, Chzhen, Pontil, Tsybakov (2023) [3]	$\tilde{\mathcal{O}}\left(\frac{d^2\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗

Is estimation on iteration complexity unimprovable?

Background

References	Iteration Complexity	Maximum Noise Level
Bach, Perchet (2016) [1]	$\mathcal{O}\left(\frac{d^{2+\frac{2}{\beta-1}}\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗
Novitskii, Gasnikov (2020) [2]	$\tilde{\mathcal{O}}\left(\frac{d^{1+\frac{1}{\beta-1}}\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗
Akhavan, Chzhen, Pontil, Tsybakov (2023) [3]	$\tilde{\mathcal{O}}\left(\frac{d^2\Delta^2}{\varepsilon^{2+\frac{2}{\beta-1}}}\right)$	✗

Is estimation on iteration complexity unimprovable?

What is the maximum noise level that can be taken?

Problem Statement and Main Assumptions

Problem Statement

We reformulate the initial optimization problem as follows:

$$f^* = \min_{x \in Q \subseteq \mathbb{R}^d} \{f(x) := \mathbb{E}[f(x, \xi)]\}.$$

Assumption (Convexity)

Function f is convex if it holds

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in Q.$$

Assumption (L-smooth)

Function f is L-smooth if it holds

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \quad \forall x, y \in Q.$$

Gradient Oracle and Assumptions

Definition (Biased Gradient Oracle)

A map $\mathbf{g} : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}^d$ s.t.

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

for a bias $\mathbf{b} : \mathbb{R}^d \to \mathbb{R}^d$ and unbiased stochastic gradient $\mathbb{E}\left[\nabla f(x, \xi)\right] = \nabla f(x)$.

Assumption (Bounded bias)

There exists constant $\delta \geq 0$ s.t. $\forall x \in \mathbb{R}^d$

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

Assumption (Bounded noise)

There exists constants $\rho, \sigma^2 \geq 0$ such that the more general condition of strong growth is satisfied $\forall x \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\mathbf{g}(x, \xi)\|^2\right] \leq \rho \|\nabla f(x)\|^2 + \sigma^2.$$

Generalization of convergence results to the biased oracle

Convergence of the accelerated algorithms [4]

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

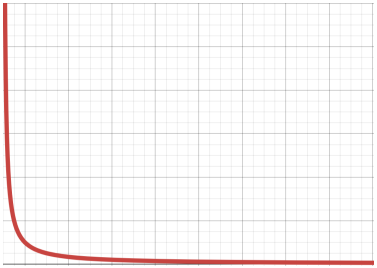


Figure: Case without bias

Generalization of convergence results to the biased oracle

Convergence of the accelerated algorithms (with biased gradient oracle) [Our result]

$$\mathbb{E}\left[f(x_N)\right] - f^* \lesssim \frac{\rho_B^2 L R^2}{N^2} + \frac{N\sigma^2}{\rho_B^2 LB} + \underbrace{\delta\tilde{R}}_{③} + \underbrace{\frac{N}{L}\delta^2}_{④}.$$

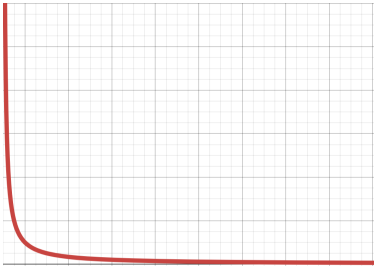


Figure: Case without bias

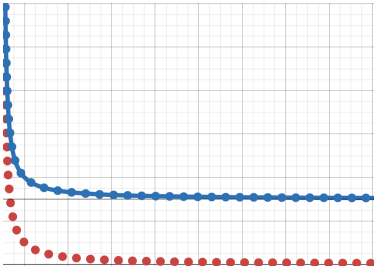


Figure: Case with bias

Zero-Order Accelerated Stochastic Gradient Descent

Algorithm 1 Zero-Order Accelerated Stochastic Gradient Descent

Input: iteration number N, batch size B, Kernel $K : [-1, 1] \to \mathbb{R}$, step size η, smoothing parameter h, $x_0 = y_0 = z_0 \in \mathbb{R}^d$, $\alpha_0 = \gamma_0 = 0$.

for $k = 0$ **to** $N - 1$ **do**

1. Sample vectors $\mathbf{e}_1, \mathbf{e}_2 ..., \mathbf{e}_B$ uniformly distributed on the unit sphere $S_2^d(1)$ and scalars $r_1, r_2, ..., r_B$ uniformly distributed on the interval [-1, 1] independently
2. Define $\mathbf{g}(x_k, \mathbf{e}_i) = d\frac{\tilde{f}(x_k + hr_i\mathbf{e}_i) - \tilde{f}(x_k - hr_i\mathbf{e}_i)}{2h} K(r_i)\mathbf{e}_i$ via (6)
3. Calculate $\mathbf{g}_k = \frac{1}{B}\sum_{i=1}^{B} \mathbf{g}(x_k, \mathbf{e}_i)$
4. $x_{k+1} \leftarrow y_k - \eta \mathbf{g}_k$
5. $z_{k+1} \leftarrow z_k - \gamma_k \eta \mathbf{g}_k$
6. $y_{k+1} \leftarrow \alpha_{k+1} z_{k+1} + (1 - \alpha_{k+1}) x_{k+1}$

end for

Return: x_N

Finding the bias of the gradient approximation

The bias of the gradient approximation [3]

$$\|\mathbf{b}(x)\| = \|\mathbb{E}\left[\mathbf{g}(x_k, \mathbf{e})\right] - \nabla f(x_k)\| \lesssim \kappa_\beta L h^{\beta-1}.$$

Second moment of gradient approximation [3]

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

Convergence of the gradient-free algorithm

$$\mathbb{E}\left[f(x_N)\right] - f^* \lesssim \underbrace{\frac{\rho_B^2 L R^2}{N^2}}_{①} + \underbrace{\frac{N d\kappa L^2 h^2}{\rho_B^2 L B}}_{②} + \underbrace{\frac{N\kappa d^2 \Delta^2}{h^2 \rho_B^2 L B}}_{③} + \underbrace{\tilde{R}\kappa_\beta L_\beta h^{\beta-1}}_{④} + \underbrace{\frac{N\kappa_\beta^2 L_\beta^2 h^{2(\beta-1)}}{L}}_{⑤}.$$

Main Theorem

Theorem 4.1 (Convergence results) *Let the function f satisfy Assumption 2.2 and the gradient approximation $\mathbf{g}(x, \mathbf{e})$ of (7) satisfies Assumptions 2.3 and 2.4, then Zero-Order Accelerated Stochastic Gradient Descent (see Algorithm 1) with $\rho_B = \max\{1, \frac{4d\kappa}{B}\}$, and with the 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 the desired ε accuracy, $\mathbb{E}\left[f(x_N)\right] - f^* \leq \varepsilon$

- *in the case $B \in [1, 4d\kappa]$, $h \lesssim \varepsilon^{3/4}$ and $\beta \geq \frac{7}{3}$ after*

$$N = \mathcal{O}\left(\sqrt{\frac{d^2 L R^2}{B^2 \varepsilon}}\right); \qquad T = N \cdot B = \mathcal{O}\left(\sqrt{\frac{d^2 L R^2}{\varepsilon}}\right)$$

number of iterations and gradient-free oracle calls, respectively, at

$$\Delta \lesssim \frac{\varepsilon^{3/2}}{\sqrt{d}} \quad \text{maximum noise level;}$$

- *in the case $B > 4d\kappa$ and $h \lesssim \varepsilon^{1/(\beta-1)}$ after*

$$N = \mathcal{O}\left(\sqrt{\frac{L R^2}{\varepsilon}}\right); \quad T = N \cdot B = \max\left\{\mathcal{O}\left(\sqrt{\frac{d^2 L R^2}{\varepsilon}}\right), \mathcal{O}\left(\frac{d^2 \Delta^2}{\varepsilon^{2 + \frac{2}{\beta - 1}}}\right)\right\}$$

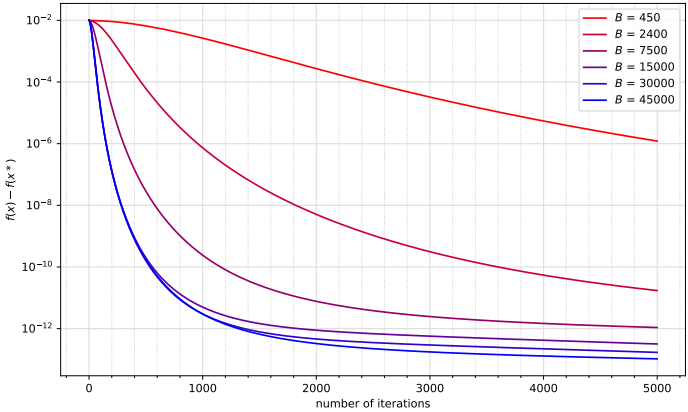
number of iterations and gradient-free oracle calls, respectively, at

$$\Delta \lesssim \frac{\varepsilon^{\frac{3\beta+1}{4(\beta-1)}}}{d} B^{1/2} \quad \text{maximum noise level;}$$

Experiments

Function for the minimization problem

$$f(w) := -y \log \left[\frac{1}{1 + \exp\left(-w^\mathsf{T} X\right)} \right] + (1 - y) \log \left[1 - \frac{1}{1 + \exp\left(-w^\mathsf{T} X\right)} \right].$$



Useful links

Where were the materials sourced from?

- **Smooth optimization problem ($\beta \geq 2$)**
  - *The "Black-Box" Optimization Problem: Zero-Order Accelerated Stochastic Method via Kernel Approximation*

Thank you for your attention!

Figure: Contact Alexander Gasnikov

Figure: Contact Aleksandr Lobanov

Reference I

[1] Francis Bach and Vianney Perchet. "Highly-smooth zero-th order online optimization". In: *Conference on Learning Theory*. PMLR. 2016, pp. 257–283.

[2] Vasilii Novitskii and Alexander Gasnikov. "Improved exploiting higher order smoothness in derivative-free optimization and continuous bandit". In: *arXiv preprint arXiv:2101.03821* (2021).

[3] Arya Akhavan et al. "Gradient-free optimization of highly smooth functions: improved analysis and a new algorithm". In: *arXiv preprint arXiv:2306.02159* (2023).

[4] Sharan Vaswani, Francis Bach, and Mark Schmidt. "Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron". In: *The 22nd international conference on artificial intelligence and statistics*. PMLR. 2019, pp. 1195–1204.