Application of the stochastic smoothing method for solving problems with a zero-order oracle

Molozhavenko Alexander
MIPT

Introduction

Many approaches to solving optimization problems with zeroth-order oracle are constrained by non-smooth object functions. In this work we are focused on smooth problems with only zero-order information accessible. Such tasks are still viable, since sometimes we can not use autograd for obtaining first order information ([1]). The approach is based on "stohastic smoothing" ([2]) using stohastic gradient for gradient approximation. This article provides iteration and zeroth-order oracle call number bounds.

The Problem

We consider *optimization problem* in the setting of *zeroth-order* oracle with noise that is bounded by a small $\Delta > 0$:

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) \tag{1}$$

defining $Q_\gamma = Q + B_2^d(\gamma)$, where $B_2^d(\gamma)$ is a zero-centered Euclidean ball with radius γ, we make assumptions:

- Q is a convex set and the function f is convex on the set Q_γ
- The function f is Lipschitz-continuous with constant $M(= M_2, p = 2)$, *i.e.*

$$|f(y) - f(x)| \leq M\|y - x\|_p \; \forall x, y \in Q_\gamma, p \in [1, 2]$$

- The functions f gradient is Lipschitz-continuous with constant L, *i.e.*

$$\|\nabla f(y) - \nabla f(x)\|_q \leq L\|y - x\|_p \; \forall x, y \in Q_\gamma, 1/q + 1/p = 1$$

The Approach

The approach will be *stochastic smoothing*. Using

$$\nabla f_\gamma(x, e) = d\frac{f(x + \gamma e) - f(x - \gamma e)}{2\gamma}e, \tag{2}$$

where e is uniformly distributed random vector in $S_2^d(1)$ (zero-centered sphere), we apply $\mathbb{A}(L, \sigma^2)$ *i.e.*batched algorithm, that solves (1), with L - Lipschitz constant of gradient of f and by using stochastic first-order oracle that depends on a random variable η returns sothastic gradient at point x :

$$\mathbb{E}_\eta[\|\nabla_x f(x, \eta) - \nabla f(x)\|_q^2] \leq \sigma^2$$

For $\mathbb{A}(L, \sigma^2)$ we also make assumptions:

- To reach ε-suboptimality in expectation, this algorithm requires $N(L, \varepsilon)$ iterations and $T(L, \sigma^2, \varepsilon)$ stochastic first-order oracle calls
- Allows batch parallelization with the average batch size $B(L, \sigma^2, \varepsilon) = \frac{T(L,\sigma^2,\varepsilon)}{N(L,\varepsilon)}$

Theorem: properties of $\nabla f_\gamma(x, e)$

For all $x \in Q$

- Unbiased:

$$\mathbb{E}_e\{\nabla f_\gamma(x, e)\} = \nabla f(x), \gamma \to 0; \tag{3}$$

-

$$\forall \gamma, \mathbb{E}_e\{\nabla f_\gamma(x, e)\} = \nabla f(x) + \mathbb{E}_e\{e\}(O(\gamma) + \frac{d\Delta}{\gamma}) \tag{4}$$

- Bounded variance:

$$\begin{aligned} &\mathbb{E}_e\{\|\nabla f_\gamma(x, e)\|_q^2\} = \kappa(p, d)(dM_2^2 + \frac{d^2\Delta^2}{\gamma^2}), \\ &1/p + 1/q = 1, \\ &\kappa(p, d) = O(\sqrt{\mathbb{E}_e\|e\|_q^4}) = \begin{cases} O(1), p = 2; \\ O((\ln d)/d), p = 1. \end{cases} \end{aligned} \tag{5}$$

Theorem proof

Full proof can be found in a link

In a nutshell, using Taylor expansion for $\nabla f_\gamma(x, e)$:

$$\vec{g} := d\mathbb{E}_e\{\frac{\delta(x + \gamma e) + \langle \nabla f(x), \gamma e\rangle - (\delta(x - \gamma e) - \langle \nabla f(x), \gamma e\rangle) + O(\|\gamma e\|_2^2)}{2\gamma}e\}$$

and bounding oracle noise $|\delta(x \pm \gamma e)| \leq \Delta$ we obtain the theorem's statement as well as bounds for oracles noise:

For ε sub-optimality the noise should not be greater then $\frac{\varepsilon^2}{R^2}$

Obtained bounds

In case $\mathbb{A}(L, \sigma^2)$ implemented with (2) we obtain zeroth-order method for our smooth problem. Solving (1) with ε-accuracy will require:

$$\begin{aligned} &N(L, \varepsilon) \text{ successive iterations} \\ &2T(L, 2\kappa(p, d)dM_2^2, \varepsilon) \text{ zeroth-order oracle calls.} \end{aligned} \tag{6}$$

Gradient Descent Case

If as A the gradient descent method is used ($p = 2$), then we obtain, that:

$$\mathbb{E}\{f(\overline{x}^N)\} - f(x_*) \leq \sqrt{\frac{d}{N}} \cdot M_2 R, \gamma \to 0 \implies N \approx d\frac{M_2^2 R^2}{\varepsilon}$$

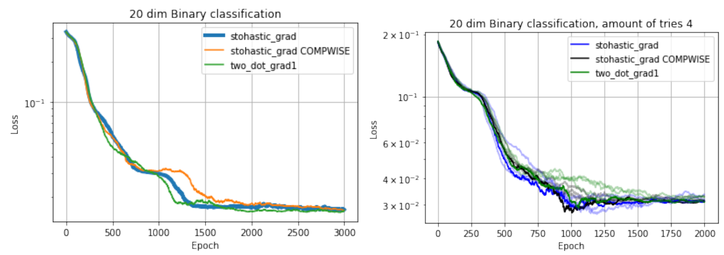
Which is close to optimal estimations for set of convex Lipschitz continuous task with zero order oracle $N \approx \frac{M_2^2 R^2}{\varepsilon}$

Numerical example

Consider a numerical example with $f(\vec{x}, \vec{w}) = \sum_{i=1}^{m}(y_i - \frac{1}{1+\exp z})^2$ with $y_i \in \{0, 1\}, \forall i \in \overline{1, \dots, m}$ and $z = w_0 + \sum_{j=1}^{n} w_j x_j$. Since the task is *stochastic optimization problem* we will solve it via batched gradient method $\mathbb{A}$ SGD with forward coordinate finite differences for gradient, and with our approach for gradient with different generation of random vector.

Results

Binary Classification with different methods. Code for 2 dimensional experiment (link) for 20 dimensional experiment (link) advanced wandb graphs (link)



Conclusion

During our research we have investigated a new method based on stochastic gradient. We have theoretically proved its' properties that allow using it and compared it with different methods for gradient approximation in example of 20 dimensional binary classification task. Further work could be conducted with research of different methods for finding uniformly distributed vector on hyper-sphere.

Acknowledgements

This material is based upon work supported by Alexander Gasnikov.

References

[1] Lev Bogolubsky, Pavel Dvurechensky, Alexander Gasnikov, Gleb Gusev, Yurii Nesterov, Andrey Raigorodskii, Aleksey Tikhonov, and Maxim Zhukovskii. Learning supervised pagerank with gradient-based and gradient-free optimization methods, 2016.

[2] Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takáč, Pavel Dvurechensky, and Bin Gu. The power of first-order smooth optimization for black-box non-smooth problems, 2022.