Gradient-Free Methods with Inexact Oracle for Convex-Concave Stochastic Saddle-Point Problem

Aleksandr Beznosikov[1,2(✉)], Abdurakhmon Sadiev[1(✉)], and Alexander Gasnikov[1,2,3,4(✉)]

[1] Moscow Institute of Physics and Technology, Dolgoprudny, Russia
beznosikov.an@phystech.edu, sadiev1998@mail.ru, gasnikov@yandex.ru
[2] Sirius University of Science and Technology, Krasnoyarsk, Russia
[3] Institute for Information Transmission Problems RAS, Moscow, Russia
[4] Caucasus Mathematical Center, Adyghe State University, Maykop, Russia

Abstract. In the paper, we generalize the approach Gasnikov et al. 2017, which allows to solve (stochastic) convex optimization problems with an inexact gradient-free oracle, to the convex-concave saddle-point problem. The proposed approach works, at least, like the best existing approaches. But for a special set-up (simplex type constraints and closeness of Lipschitz constants in 1 and 2 norms) our approach reduces $n/\log n$ times the required number of oracle calls (function calculations). Our method uses a stochastic approximation of the gradient via finite differences. In this case, the function must be specified not only on the optimization set itself, but in a certain neighbourhood of it. In the second part of the paper, we analyze the case when such an assumption cannot be made, we propose a general approach on how to modernize the method to solve this problem, and also we apply this approach to particular cases ofsomeclassical sets.

Keywords: Zeroth-order optimization · Saddle-point problem · Stochastic optimization

1 Introduction

In the last decade in the ML community, a big interest cause different applications of Generative Adversarial Networks (GANs) [10], which reduce the ML problem to the saddle-point problem, and the application of gradient-free methods for Reinforcement Learning problems [17]. Neural networks become rather

The research of A. Beznosikov was partially supported by RFBR, project number 19-31-51001. The research of A. Gasnikov was partially supported by RFBR, project number 18-29-03071 mk and was partially supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) no 075-00337-20-03.

Y. Kochetov et al. (Eds.): MOTOR 2020, CCIS 1275, pp. 105–119, 2020.
https://doi.org/10.1007/978-3-030-58657-7_11

popular in Reinforcement Learning [13]. Thus, there is an interest in gradient-free methods for saddle-point problems

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\varphi(x,y). \tag{1}$$

One of the natural approach for this class of problems is to construct a stochastic approximation of a gradient via finite differences. In this case, it is natural to expect that the complexity of the problem (1) in terms of the number of function calculations is $\sim n$ times large in comparison with the complexity in terms of number of gradient calculations, where $n = \dim\mathcal{X} + \dim\mathcal{Y}$. Is it possible to obtain better result? In this paper, we show that this factor can be reduced in some situation to a much smaller factor $\log n$.

We use the technique, developed in [8,9] for stochastic gradient-free non-smooth convex optimization problems (gradient-free version of mirror descent [2]) to propose a stochastic gradient-free version of saddle-point variant of mirror descent [2] for non-smooth convex-concave saddle-point problems.

The concept of using such an oracle with finite differences is not new (see [5,16]). For such an oracle, it is necessary that the function is defined in some neighbourhood of the initial set of optimization, since when we calculate the finite difference, we make some small step from the point, and this step can lead us outside the set. As far as we know, in all previous works, the authors proceed from the fact that such an assumption is fulfilled or does not mention it at all. We raise the question of what we can do when the function is defined only on the given set due to some properties of the problem.

1.1 Our Contributions

In this paper, we present a new method called zeroth-order Saddle-Point Algorithm (`zoSPA`) for solving a convex-concave saddle-point problem (1). Our algorithm uses a zeroth-order biased oracle with stochastic and bounded deterministic noise. We show that if the noise $\sim \varepsilon$ (accuracy of the solution), then the number of iterations necessary to obtain $\varepsilon-$solution on set with diameter $\Omega \subset \mathbb{R}^n$ is $\mathcal{O}\left(\frac{M^2\Omega^2}{\varepsilon^2}n\right)$ or $\mathcal{O}\left(\frac{M^2\Omega^2}{\varepsilon^2}\log n\right)$ (depends on the optimization set, for example, for a simplex, the second option with $\log n$ holds), where M^2 is a bound of the second moment of the gradient together with stochastic noise (see below, (3)).

In the second part of the paper, we analyze the structure of an admissible set. We give a general approach on how to work in the case when we are forbidden to go beyond the initial optimization set. Briefly, it is to consider the "reduced" set and work on it.

Next, we show how our algorithm works in practice for various saddle-point problems and compare it with full-gradient mirror descent.

One can find the proofs together and additional numerical experiments in the full version of this paper available on arXiv [4].

2 Notation and Definitions

We use $\langle x, y\rangle \overset{\text{def}}{=} \sum_{i=1}^{n} x_i y_i$ to define inner product of $x, y \in \mathbb{R}^n$ where x_i is the i-th component of x in the standard basis in $\mathbb{R}^n$. Hence we get the definition of ℓ_2-norm in $\mathbb{R}^n$ in the following way $\|x\|_2 \overset{\text{def}}{=} \sqrt{\langle x, x\rangle}$. We define ℓ_p-norms as $\|x\|_p \overset{\text{def}}{=} \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$ for $p \in (1, \infty)$ and for $p = \infty$ we use $\|x\|_\infty \overset{\text{def}}{=} \max_{1\le i\le n} |x_i|$. The dual norm $\|\cdot\|_q$ for the norm $\|\cdot\|_p$ is defined in the following way: $\|y\|_q \overset{\text{def}}{=} \max\left\{\langle x, y\rangle \mid \|x\|_p \le 1\right\}$. Operator $\mathbb{E}[\cdot]$ is full mathematical expectation and operator $\mathbb{E}_\xi[\cdot]$ express conditional mathematical expectation.

Definition 1 (M-Lipschitz continuity). *Function $f(x)$ is M-Lipschitz continuous in $X \subseteq \mathbb{R}^n$ with $M > 0$ w.r.t. norm $\|\cdot\|$ when*

$$|f(x) - f(y)| \le M\|x - y\|, \quad \forall\ x, y \in X.$$

Definition 2 (μ-strong convexity). *Function $f(x)$ is μ-strongly convex w.r.t. norm $\|\cdot\|$ on $X \subseteq \mathbb{R}^n$ when it is continuously differentiable and there is a constant $\mu > 0$ such that the following inequality holds:*

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2, \quad \forall\ x, y \in X.$$

Definition 3 (Prox-function). *Function $d(z) : \mathcal{Z} \to \mathbb{R}$ is called prox-function if $d(z)$ is 1-strongly convex w.r.t. $\|\cdot\|$-norm and differentiable on $\mathcal{Z}$ function.*

Definition 4 (Bregman divergence). *Let $d(z) : \mathcal{Z} \to \mathbb{R}$ is prox-function. For any two points $z, w \in \mathcal{Z}$ we define Bregman divergence $V_z(w)$ associated with $d(z)$ as follows:*

$$V_z(w) = d(z) - d(w) - \langle \nabla d(w), z - w\rangle.$$

We denote the Bregman-diameter $\Omega_\mathcal{Z}$ of $\mathcal{Z}$ w.r.t. $V_{z_1}(z_2)$ as $\Omega_\mathcal{Z} \overset{\text{def}}{=} \max\{\sqrt{2V_{z_1}(z_2)} \mid z_1, z_2 \in \mathcal{Z}\}$.

Definition 5 (Prox-operator). *Let $V_z(w)$ Bregman divergence. For all $x \in \mathcal{Z}$ define prox-operator of ξ:*

$$\text{prox}_x(\xi) = \arg\min_{y\in\mathcal{Z}} \left(V_x(y) + \langle \xi, y\rangle\right).$$

3 Main Result

3.1 Non-smooth Saddle-Point Problem

We consider the saddle-point problem (1), where $\varphi(\cdot, y)$ is convex function defined on compact convex set $\mathcal{X} \subset \mathbb{R}^{n_x}$, $\varphi(x, \cdot)$ is concave function defined on compact convex set $\mathcal{Y} \subset \mathbb{R}^{n_y}$.

We call an inexact stochastic zeroth-order oracle $\widetilde{\varphi}(x, y, \xi)$ at each iteration. Our model corresponds to the case when the oracle gives an inexact noisy function value. We have stochastic unbiased noise, depending on the random variable ξ and biased deterministic noise. One can write it the following way:

$$\tilde{\varphi}(x, y, \xi) = \varphi(x, y, \xi) + \delta(x, y),$$
$$\mathbb{E}_\xi[\widetilde{\varphi}(x, y, \xi)] = \widetilde{\varphi}(x, y), \quad \mathbb{E}_\xi[\varphi(x, y, \xi)] = \varphi(x, y), \tag{2}$$

where random variable ξ is responsible for unbiased stochastic noise and $\delta(x, y)$ – for deterministic noise.

We assume that exists such positive constant M that for all $x, y \in \mathcal{X} \times \mathcal{Y}$ we have

$$\|\nabla\varphi(x, y, \xi)\|_2 \leq M(\xi), \quad \mathbb{E}[M^2(\xi)] = M^2. \tag{3}$$

By $\nabla\varphi(x, y, \xi)$ we mean a block vector consisting of two vectors $\nabla_x\varphi(x, y, \xi)$ and $\nabla_y\varphi(x, y, \xi)$. One can prove that $\varphi(x, y, \xi)$ is $M(\xi)$-Lipschitz w.r.t. norm $\|\cdot\|_2$ and that $\|\nabla\varphi(x, y)\|_2 \leq M$.

Also the following assumptions are satisfied:

$$|\widetilde{\varphi}(x, y, \xi) - \varphi(x, y, \xi)| = |\delta(x, y)| \leq \Delta. \tag{4}$$

For convenience, we denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and then $z \in \mathcal{Z}$ means $z \overset{\text{def}}{=} (x, y)$, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$. When we use $\varphi(z)$, we mean $\varphi(z) = \varphi(x, y)$, and $\varphi(z, \xi) = \varphi(x, y, \xi)$.

For $\mathbf{e} \in \mathcal{RS}_2^n(1)$ (a random vector uniformly distributed on the Euclidean unit sphere) and some constant τ let $\tilde{\varphi}(z + \tau\mathbf{e}, \xi) \overset{\text{def}}{=} \tilde{\varphi}(x + \tau\mathbf{e}_x, y + \tau\mathbf{e}_y, \xi)$, where $\mathbf{e}_x$ is the first part of $\mathbf{e}$ size of dimension $n_x \overset{\text{def}}{=} \dim(x)$, and $\mathbf{e}_y$ is the second part of dimension $n_y \overset{\text{def}}{=} \dim(y)$. And $n \overset{\text{def}}{=} n_x + n_y$. Then define estimation of the gradient through the difference of functions:

$$g(z, \xi, \mathbf{e}) = \frac{n\left(\tilde{\varphi}(z + \tau\mathbf{e}, \xi) - \tilde{\varphi}(z - \tau\mathbf{e}, \xi)\right)}{2\tau} \begin{pmatrix} \mathbf{e}_x \\ -\mathbf{e}_y \end{pmatrix}. \tag{5}$$

$g(z, \xi, \mathbf{e})$ is a block vector consisting of two vectors.

Next we define an important object for further theoretical discussion – a smoothed version of the function $\widetilde{\varphi}$ (see [15,16]).

Definition 6. *Function $\hat{\varphi}(x, y) = \hat{\varphi}(z)$ defines on set $\mathcal{X} \times \mathcal{Y}$ satisfies:*

$$\hat{\varphi}(z) = \mathbb{E}_\mathbf{e}\left[\varphi(z + \tau\mathbf{e})\right].$$

Note that we introduce a smoothed version of the function only for proof; in the algorithm, we use only the zero-order oracle (5). Now we are ready to present our algorithm:

Algorithm 1. Zeroth-Order Saddle-Point Algorithm (`zoSPA`)

Input: Iteration limit N.
Let $z_1 = \operatorname*{argmin}_{z\in\mathcal{Z}} d(z)$.
for $k = 1, 2, \ldots, N$ **do**
  Sample $\mathbf{e}_k$, ξ_k independently.
  Initialize γ_k.
  $z_{k+1} = \text{prox}_{z_k}(\gamma_k g(z_k, \xi_k, \mathbf{e}_k))$.
end for
Output: $\bar{z}_N$,

where

$$\bar{z}_N = \frac{1}{\Gamma_N}\left(\sum_{k=1}^{N}\gamma_k z_k\right), \quad \Gamma_N = \sum_{k=1}^{N}\gamma_k. \tag{6}$$

In Algorithm 1, we use the step γ_k. In fact, we can take this step as a constant, independent of the iteration number k (see Theorem 1).

Note that we work only with norms $\|\cdot\|_p$, where p is from 1 to 2 (q is from 2 to ∞). In the rest of the paper, including the main theorems, we assume that p is from 1 to 2.

Lemma 1 (see Lemma 2 from [3]). *For $g(z,\xi,\mathbf{e})$ defined in (5) the following inequalitie holds:*

$$\mathbb{E}\left[\|g(z,\xi,\mathbf{e})\|_q^2\right] \leq 2\left(cnM^2 + \frac{n^2\Delta^2}{\tau^2}\right)a_q^2,$$

where c is some positive constant (independent of n) and a_q^2 is determined by $\sqrt{\mathbb{E}[\|e\|_q^4]} \leq a_q^2$ and the following statement is true

$$a_q^2 = \min\{2q-1, 32\log n - 8\}n^{\frac{2}{q}-1}, \quad \forall n \geq 3. \tag{7}$$

Note that in the case with $p=2,\ q=2$ we have $a_q = 1$, this follows not from (7), but from the simplest estimate. And from (7) we get that with $p=1,\ q=\infty$ – $a_q = \mathcal{O}(^{\log n}/_{n})$ (see also Lemma 4 from [16]).

Lemma 2 (see Lemma 8 from [16]). *Let $\mathbf{e}$ be from $\mathcal{RS}_2^n(1)$. Then function $\hat{\varphi}(z,\xi)$ is convex-concave and satisfies :*

$$\sup_{z\in\mathcal{Z}}|\hat{\varphi}(z) - \varphi(z)| \leq \tau M + \Delta.$$

Lemma 3 (see Lemma 10 from [16] and Lemma 2 from [3]). *It holds that*

$$\tilde{\nabla}\hat{\varphi}(z) = \mathbb{E}_{\mathbf{e}}\left[\frac{n\left(\varphi(z+\tau\mathbf{e}) - \varphi(z-\tau\mathbf{e})\right)}{2\tau}\begin{pmatrix}\mathbf{e}_x\\ -\mathbf{e}_y\end{pmatrix}\right],$$

$$\|\mathbb{E}_{\mathbf{e}}[g(z,\mathbf{e})] - \tilde{\nabla}\hat{\varphi}(z)\|_q \leq \frac{\Delta n a_q}{\tau},$$

where

$$g(z, \mathbf{e}) = \mathbb{E}_\xi \left[g(z, \xi, \mathbf{e})\right] \\ = \frac{n\left(\tilde{\varphi}(z + \tau\mathbf{e}) - \tilde{\varphi}(z - \tau\mathbf{e})\right)}{2\tau} \begin{pmatrix} \mathbf{e}_x \\ -\mathbf{e}_y \end{pmatrix}.$$

Hereinafter, by $\tilde{\nabla}\hat{\varphi}(z)$ *we mean a block vector consisting of two vectors* $\nabla_x \hat{\varphi}(x, y)$ *and* $-\nabla_y \hat{\varphi}(x, y)$.

Lemma 4 (see Lemma 5.3.2 from [2]). *Define* $\Delta_k \overset{def}{=} g(z_k, \xi_k, \mathbf{e}_k) - \tilde{\nabla}\hat{\varphi}(z_k)$. *Let* $D(u) \overset{def}{=} \sum_{k=1}^N \gamma_k \langle \Delta_k, u - z_k \rangle$. *Then we have*

$$\mathbb{E}\left[\max_{u \in \mathcal{Z}} D(u)\right] \le \Omega^2 + \frac{\Delta \Omega n a_q}{\tau} \sum_{k=1}^N \gamma_k + M_{all}^2 \sum_{k=1}^N \gamma_k^2,$$

where $M_{all}^2 \overset{def}{=} 2\left(cnM^2 + \frac{n^2\Delta^2}{\tau^2}\right) a_q^2$ *is from Lemma 1.*

Theorem 1. *Let problem (1) with function* $\varphi(x, y)$ *be solved using Algorithm 1 with the oracle* $g(z_k, \xi_k, \mathbf{e}_k)$ *from (5). Assume, that the function* $\varphi(x, y)$ *and its inexact modification* $\widetilde{\varphi}(x, y)$ *satisfy the conditions (2), (3), (4). Denote by* N *the number of iterations. Let step in Algorithm 1* $\gamma_k = \frac{\Omega}{M_{all}\sqrt{N}}$. *Then the rate of convergence is given by the following expression*

$$\mathbb{E}\left[\varepsilon_{sad}(\bar{z}_N)\right] \le \frac{3M_{all}\Omega}{\sqrt{N}} + \frac{\Delta\Omega n a_q}{\tau} + 2\tau M,$$

where $\bar{z}_N$ *is defined in (6),* Ω *is a diameter of* $\mathcal{Z}$, $M_{all}^2 = 2\left(cnM^2 + \frac{n^2\Delta^2}{\tau^2}\right) a_q^2$ *and*

$$\varepsilon_{sad}(\bar{z}_N) = \max_{y' \in \mathcal{Y}} \varphi(\bar{x}_N, y') - \min_{x' \in \mathcal{X}} \varphi(x', \bar{y}_N),$$

$\bar{x}_N$, $\bar{y}_N$ *are defined the same way as* $\bar{z}_N$ *in (6).*

Next we analyze the results.

Corollary 1. *Under the assumptions of the Theorem 1 let* ε *be accuracy of the solution of the problem (1) obtained using Algorithm 1. Assume that*

$$\tau = \Theta\left(\frac{\varepsilon}{M}\right), \quad \Delta = \mathcal{O}\left(\frac{\varepsilon^2}{M\Omega n a_q}\right), \tag{8}$$

then the number of iterations to find ε*-solution*

$$N = \mathcal{O}\left(\frac{\Omega^2 M^2 n^{2/q}}{\varepsilon^2} C^2(n, q)\right),$$

where $C(n, q) \overset{def}{=} \min\{2q - 1, 32\log n - 8\}$.

Consider separately cases with $p = 1$ and $p = 2$.

Note that in the case with $p = 2$, we have that the number of iterations increases n times compared with [2], and in the case with $p = 1$ – just $\log^2 n$ times (Table 1).

Table 1. Summary of convergence estimation for non-smooth case: $p = 2$ and $p = 1$.

p, $(1 \leqslant p \leqslant 2)$	q, $(2 \leqslant q \leqslant \infty)$	N, Number of iterations
$p = 2$	$q = 2$	$\mathcal{O}\left(\frac{\Omega^2 M^2}{\varepsilon^2} n\right)$
$p = 1$	$q = \infty$	$\mathcal{O}\left(\frac{\Omega^2 M^2}{\varepsilon^2} \log^2(n)\right)$

3.2 Admissible Set Analysis

As stated above, in works (see [5,16]), where zeroth-order approximation (5) is used instead of the "honest" gradient, it is important that the function is specified not only on an admissible set, but in a certain neighborhood of it. This is due to the fact that for any point x belonging to the set, the point $x + \tau\mathbf{e}$ can be outside it.

But in some cases we cannot make such an assumption. The function and values of x can have a real physical interpretation. For example, in the case of a probabilistic simplex, the values of x are the distribution of resources or actions. The sum of the probabilities cannot be negative or greater than 1. Moreover, due to implementation or other reasons, we can deal with an oracle that is clearly defined on an admissible set and nowhere else.

In this part of the paper, we outline an approach how to solve the problem raised above and how the quality of the solution changes from this.

Our approach can be briefly described as follows:

- Compress our original set X by $(1 - \alpha)$ times and consider a "reduced" version X^α. Note that the parameter α should not be too small, otherwise the parameter τ must be taken very small. But it's also impossible to take large α, because we compress our set too much and can get a solution far from optimal. This means that the accuracy of the solution ε bounds α: $\alpha \le h(\varepsilon)$, in turn, α bounds τ: $\tau \le g(\alpha)$.
- Generate a random direction $\mathbf{e}$ so that for any $x \in X^\alpha$ follows $x + \tau\mathbf{e} \in X$.
- Solve the problem on "reduced" set with $\varepsilon/2$-accuracy. The α parameter must be selected so that we find ε-solution of the original problem.

In practice, this can be implemented as follows: 1) do as described in the previous paragraph, or 2) work on the original set X, but if $x_k + \tau\mathbf{e}$ is outside X, then project x_k onto the set X^α. We provide a theoretical analysis only for the method that always works on X^α.

Next, we analyze cases of different sets. General analysis scheme:

- Present a way to "reduce" the original set.
- Suggest a random direction $\mathbf{e}$ generation strategy.
- Estimate the minimum distance between X^α and X in ℓ_2-norm. This is the border of τ, since $\|\mathbf{e}\|_2$.
- Evaluate the α parameter so that the $\varepsilon/2$-solution of the "reduced" problem does not differ by more than $\varepsilon/2$ from the ε-solution of the original problem.

The first case of set is a **probability simplex**:

$$\triangle_n = \left\{ \sum_{i=1}^n x_i = 1, \quad x_i \geq 0, \quad i \in 1 \dots n \right\}.$$

Consider the hyperplane

$$\mathcal{H} = \left\{ \sum_{i=1}^n x_i = 1 \right\},$$

in which the simplex lies. Note that if we take the directions $\mathbf{e}$ that lies in $\mathcal{H}$, then for any x lying on this hyperplane, $x + \tau \mathbf{e}$ will also lie on it. Therefore, we generate the direction $\mathbf{e}$ randomly on the hyperplane. Note that $\mathcal{H}$ is a subspace of $\mathbb{R}^n$ with size dim$\mathcal{H} = n - 1$. One can check that the set of vectors from $\mathbf{R}^n$

$$\mathbf{v} = \begin{pmatrix} \mathbf{v}_1 = {}^1\!/\!\sqrt{2}(1, -1, 0, 0, \dots 0), \\ \mathbf{v}_2 = {}^1\!/\!\sqrt{6}(1, 1, -2, 0, \dots 0), \\ \mathbf{v}_3 = {}^1\!/\!\sqrt{12}(1, 1, 1, -3, \dots 0), \\ \dots \\ \mathbf{v}_k = {}^1\!/\!\sqrt{k+k^2}(1, \dots 1, -k, \dots, 0), \\ \dots \\ \mathbf{v}_{n-1} = {}^1\!/\!\sqrt{n-1+(n-1)^2}(1, \dots, 1, -n+1) \end{pmatrix},$$

is an orthonormal basis of $\mathcal{H}$. Then generating the vectors $\tilde{\mathbf{e}}$ uniformly on the euclidean sphere $\mathcal{RS}_2^{n-1}(1)$ and computing $\mathbf{e}$ by the following formula:

$$\mathbf{e} = \tilde{\mathbf{e}}_1 \mathbf{v}_1 + \tilde{\mathbf{e}}_2 \mathbf{v}_2 + \dots + \tilde{\mathbf{e}}_k \mathbf{v}_k + \dots \tilde{\mathbf{e}}_{n-1} \mathbf{v}_{n-1}, \tag{9}$$

we have what is required. With such a vector $\mathbf{e}$, we always remain on the hyperplane, but we can go beyond the simplex. This happens if and only if for some i, $x_i + \tau \mathbf{e}_i < 0$. To avoid this, we consider a "reduced" simplex for some positive constant α:

$$\triangle_n^\alpha = \left\{ \sum_{i=1}^n x_i = 1, \quad x_i \geq \alpha, \quad i \in 1 \dots n \right\}.$$

One can see that for any $x \in \triangle_n^\alpha$, for any $\mathbf{e}$ from (9) and $\tau < \alpha$ follows that $x + \tau \mathbf{e} \in \triangle_n$, because $|\mathbf{e}_i| \leq 1$ and then $x_i + \tau \mathbf{e}_i \geq \alpha - \tau \geq 0$.

The last question to be discussed is the accuracy of the solution that we obtain on a "reduced" set. Consider the following lemma (this lemma does not apply to the problem (1), for it we prove later):

Lemma 5. *Suppose the function $f(x)$ is M-Lipschitz w.r.t. norm $\|\cdot\|_2$. Consider the problem of minimizing $f(x)$ not on original set X, but on the "reduced" set X_α. Let we find x_k solution with ${}^\varepsilon\!/\!2$-accuracy on $f(x)$. Then we found $({}^\varepsilon\!/\!2 + rM)$-solution of original problem, where*

$$r = \max_{x \in X} \left\| x - \operatorname*{argmin}_{\hat{x} \in X^\alpha} \|x - \hat{x}\|_2 \right\|_2.$$

It is not necessary to search for the closest point to each x and find r. It's enough to find one that is "pretty" close and find some upper bound of r. Then it remains to find a rule, which each point x from X associated with some point $\hat{x}$ from X_α and estimate the maximum distance $\max_X \|\hat{x}-x\|_2$. For any simplex point, consider the following rule:

$$\hat{x}_i = \frac{(x_i + 2\alpha)}{(1+2\alpha n)}, \qquad i = 1, \dots n.$$

One can easy to see, that for $\alpha \leq {}^1\!/_{2n}$:

$$\sum_{i=1}^n \hat{x}_i = 1, \qquad \hat{x}_i \leq \alpha, \qquad i = 1, \dots n.$$

It means that $\hat{x} \in X_\alpha$. The distance $\|\hat{x} - x\|_2$:

$$\|\hat{x} - x\|_2 = \sqrt{\sum_{i=1}^n (\hat{x}_i - x_i)^2} = \frac{2\alpha n}{1+2\alpha n}\sqrt{\sum_{i=1}^n \left(\frac{1}{n} - x_i\right)^2}.$$

$\sqrt{\sum\limits_{i=1}^n \left(\frac{1}{n} - x_i\right)^2}$ is a distance to the center of the simplex. It can be bounded by the radius of the circumscribed sphere $R = \sqrt{\frac{n-1}{n}} \leq 1$. Then

$$\|\hat{x} - x\|_2 \leq \frac{2\alpha n}{1+2\alpha n} \leq 2\alpha n. \tag{10}$$

(10) together with Lemma 5 gives that $f(x_k) - f(x^*) \leq \frac{\varepsilon}{2} + 2\alpha n M$. Then by taking $\alpha = {}^\varepsilon\!/_{4nM}$ (or less), we find ε-solution of the original problem. And it takes $\tau \leq \alpha = {}^\varepsilon\!/_{4nM}$.

The second case is a **positive orthant**:

$$\perp_n = \{x_i \geq 0, \quad i \in 1 \dots n\}.$$

We propose to consider a "reduced" set of the following form:

$$\perp_n^\alpha = \{y_i \geq \alpha, \quad i \in 1 \dots n\}.$$

One can note that for all i the minimum of the expression $y_i + \tau \mathbf{e}_i$ is equal to $\alpha - \tau$, because $\mathbf{e}_i \geq -1$ and $y_i \geq \alpha$. Therefore, it is necessary that $\alpha - \tau \geq 0$. It means that for any $\mathbf{e} \in \mathcal{RS}_2^n(1)$, for the vector $y + \tau\mathbf{e}$ the following expression is valid:

$$y_i + \tau \mathbf{e}_i \geq 0, \quad i \in 1 \dots n.$$

The projection onto $\perp_n^\alpha$ is carried out as well as onto $\perp_n$: if $x_i < \alpha$ then $x_i \to \alpha$.

Then let find r in Lemma 5 for orthant. Let for any $x \in \perp_n$ define $\hat{x}$ in the following way:

$$\hat{x}_i = \begin{cases} \alpha, & x_i < \alpha, \\ x_i, & x_i \geq \alpha, \end{cases} \qquad i = 1, \dots n.$$

One can see that $\hat{x}_i \in \perp_n^\alpha$ and

$$\|\hat{x} - x\|_2 = \sqrt{\sum_{i=1}^n (\hat{x}_i - x_i)^2} \leq \sqrt{\sum_{i=1}^n \alpha^2} = \alpha\sqrt{n}.$$

By Lemma 5 we have that $f(x_k) - f(x^*) \leq \frac{\varepsilon}{2} + \alpha\sqrt{n}M$. Then by taking $\alpha = {}^{\varepsilon}/_{2\sqrt{n}M}$ (or less), we find ε-solution of the original problem. And it takes $\tau \leq \alpha = {}^{\varepsilon}/_{2\sqrt{n}M}$.

The above reasoning can easily be generalized to an arbitrary orthant:

$$\tilde{\perp}_n = \{b_i x_i \geq 0, \quad b_i = \pm 1, \quad i \in 1 \dots n\}.$$

The third case is a **ball in p-norm** for $p \in [1; 2]$:

$$\mathcal{B}_p^n(a, R) = \{\|x - a\|_p \leq R\},$$

where a is a center of ball, R – its radii. We propose reducing a ball and solving the problem on the "reduced" ball $\mathcal{B}_p^n(a, R(1-\alpha))$. We need the following lemma:

Lemma 6. *Consider two concentric spheres in p norm, where $p \in [1; 2]$, $\alpha \in (0; 1)$:*

$$\mathcal{S}_p^n(a, R) = \{\|x - a\|_p = R\}, \quad \mathcal{S}_p^n(a, R(1 - \alpha)) = \{\|y - a\|_p = R(1 - \alpha)\}.$$

Then the minimum distance between these spheres in the second norm

$$m = \frac{\alpha R}{n^{1/p - 1/2}}.$$

Using the lemma, one can see that for any $x \in \mathcal{B}_n^\alpha(a, R(1-\alpha))$, $\tau \leq {}^{\alpha R}/_{n^{1/p-1/2}}$ and for any $\mathbf{e} \in \mathcal{RS}_2^n(1)$, $x + \tau\mathbf{e} \in \mathcal{B}_n(a, R)$.

Then let find r in Lemma 5 for ball. Let for any x define $\hat{x}$ in the following way:

$$\hat{x}_i = a + (1 - \alpha)(x_i - a), \qquad i = 1, \dots n.$$

One can see that $\hat{x}_i$ is in the "reduced" ball and

$$\|\hat{x} - x\|_2 = \sqrt{\sum_{i=1}^n (\hat{x}_i - x_i)^2} = \sqrt{\sum_{i=1}^n (\alpha(x_i - a))^2} = \alpha\sqrt{\sum_{i=1}^n (x_i - a)^2} \leq \alpha \sum_{i=1}^n |x_i - a|.$$

By Holder inequality:

$$\|\hat{x} - x\|_2 \leq \alpha \sum_{i=1}^n |x_i - a| \leq \alpha n^{\frac{1}{q}} \left(\sum_{i=1}^n |x_i - a|^p\right)^{\frac{1}{p}} = \alpha n^{\frac{1}{q}} R.$$

By Lemma 5 we have that $f(x_k) - f(x^*) \le \frac{\varepsilon}{2} + \alpha n^{1/q} RM$. Then by taking $\alpha = \varepsilon/2n^{1/q}RM$ (or less), we find ε-solution of the original problem. And it takes $\tau \le \alpha R / n^{1/p - 1/2} = \varepsilon / 2M\sqrt{n}$.

The fourth case is a **product of sets** $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We define the "reduced" set Z^α as

$$Z^\alpha = X^\alpha \times Y^\alpha,$$

We need to find how the parameter α and τ depend on the parameters α_x, τ_x and α_y, τ_y for the corresponding sets X and Y, i.e. we have bounds: $\alpha_x \le h_x(\varepsilon)$, $\alpha_y \le h_y(\varepsilon)$ and $\tau_x \le g_x(\alpha_x) \le g_x(h_x(\varepsilon))$, $\tau_y \le g_y(\alpha_y) \le g_y(h_y(\varepsilon))$. Obviously, the functions g, h are monotonically increasing for positive arguments. This follows from the physical meaning of τ and α.

Further we are ready to present an analogue of Lemma 5, only for the saddle-point problem.

Lemma 7. *Suppose the function $\varphi(x, y)$ inthe saddle-point problem is M-Lipschitz. Let we find $(\tilde{x}, \tilde{y})$ solution on X^α and Y^α with $\varepsilon/2$-accuracy. Then we found $(\varepsilon/2 + (r_x + r_y)M)$-solution of the original problem, where r_x and r_y we define in the following way:*

$$r_x = \max_{x \in X} \left\| x - \operatorname*{argmin}_{\hat{x} \in X^\alpha} \|x - \hat{x}\|_2 \right\|_2,$$
$$r_y = \max_{y \in Y} \left\| y - \operatorname*{argmin}_{\hat{y} \in Y^\alpha} \|y - \hat{y}\|_2 \right\|_2.$$

In the previous cases we found the upper bound $\alpha_x \le h_x(\varepsilon)$ from the condition that $r_x M \le \varepsilon/2$. Now let's take $\tilde{\alpha}_x$ and $\tilde{\alpha}_y$ so that $r_x M \le \varepsilon/4$ and $r_y M \le \varepsilon/4$. For this we need $\tilde{\alpha}_x \le h_x(\varepsilon/2)$, $\tilde{\alpha}_y \le h_y(\varepsilon/2)$. It means that if we take $\alpha = \min(\tilde{\alpha}_x, \tilde{\alpha}_y)$, then $(r_x + r_y)M \le \varepsilon/2$ for such α. For a simplex, an orthant and a ball the function h is linear, therefore the formula turns into a simpler expression: $\alpha = \min(\alpha_x, \alpha_y)/2$.

For the new parameter $\alpha = \min(\tilde{\alpha}_x, \tilde{\alpha}_y)$, we find $\tilde{\tau}_x = g_x(\alpha) = g_x(\min(\tilde{\alpha}_x, \tilde{\alpha}_y))$ and $\tilde{\tau}_y = g_y(\alpha) = g_y(\min(\tilde{\alpha}_x, \tilde{\alpha}_y))$. Then for any $x \in X^\alpha$, $\mathbf{e}_x \in \mathcal{RS}_2^{\dim X}(1)$, $y \in Y^\alpha$, $\mathbf{e}_y \in \mathcal{RS}_2^{\dim Y}(1)$, $x + \tilde{\tau}_x \mathbf{e}_x \in X$ and $y + \tilde{\tau}_y \mathbf{e}_y \in Y$. Hence, it is easy to see that for $\tau = \min(\tilde{\tau}_x, \tilde{\tau}_y)$ and the vector $\tilde{\mathbf{e}}_x$ of the first $\dim X$ components of $\mathbf{e} \in \mathcal{RS}_2^{\dim X + \dim Y}(1)$ and for the vector $\tilde{\mathbf{e}}_y$ of the remaining $\dim Y$ components, for any $x \in X^\alpha$, $y \in Y^\alpha$ it is true that $x + \tau\tilde{\mathbf{e}}_x \in X$ and $y + \tau\tilde{\mathbf{e}}_y \in Y$. We get $\tau = \min(\tilde{\tau}_x, \tilde{\tau}_y)$. In the previous cases that we analyzed (simplex, orthant and ball), the function g and h are linear therefore the formula turns into a simpler expression: $\tau = \min(\alpha_x, \alpha_y) \cdot \min(\tau_x/\alpha_x, \tau_y/\alpha_y)/2$.

Summarize the results of this part of the paper in Table 2.

One can note that in (8) τ is independent of n. According to Table 2, we need to take into account the dependence on n. In Table 3, we present the constraints on τ and Δ so that Corollary 1 remains satisfied. We consider three cases when

Table 2. Summary of the part 3.2

Set	α of "reduced" set	Bound of τ	**e**
probability simplex	$\frac{\varepsilon}{4nM}$	$\frac{\varepsilon}{4nM}$	see (9)
positive orthant	$\frac{\varepsilon}{2\sqrt{n}M}$	$\frac{\varepsilon}{2\sqrt{n}M}$	$\mathcal{RS}_2^n(1)$
ball in p-norm	$\frac{\varepsilon}{2n^{1/q}RM}$	$\frac{\varepsilon}{2\sqrt{n}M}$	$\mathcal{RS}_2^n(1)$
$X^\alpha \times Y^\alpha$	$\frac{\min(\alpha_x,\alpha_y)}{2}$	$\frac{\min(\alpha_x,\alpha_y)\cdot\min(\tau_x/\alpha_x,\tau_y/\alpha_y)}{2}$	$\mathcal{RS}_2^n(1)$

both sets X and Y are simplexes, orthants and balls with the same dimension $n/2$.

The second column of Table 3 means whether the functions are defined not only on the set itself, but also in some neighbourhood of it.

Table 3. τ and Δ in Corollary 1 in different cases

Set	Neigh-d?	τ	Δ
Probability simplex	✓	$\Theta\left(\frac{\varepsilon}{M}\right)$	$\mathcal{O}\left(\frac{\varepsilon^2}{M\Omega n a_q}\right)$
	✗	$\Theta\left(\frac{\varepsilon}{Mn}\right)$ and $\leq \frac{\varepsilon}{4nM}$	$\mathcal{O}\left(\frac{\varepsilon^2}{M\Omega n^2 a_q}\right)$
Positive orthant	✓	$\Theta\left(\frac{\varepsilon}{M}\right)$	$\mathcal{O}\left(\frac{\varepsilon^2}{M\Omega n a_q}\right)$
	✗	$\Theta\left(\frac{\varepsilon}{M\sqrt{n}}\right)$ and $\leq \frac{\varepsilon}{\sqrt{8n}M}$	$\mathcal{O}\left(\frac{\varepsilon^2}{M\Omega n^{3/2} a_q}\right)$
Ball in p-norm	✓	$\Theta\left(\frac{\varepsilon}{M}\right)$	$\mathcal{O}\left(\frac{\varepsilon^2}{M\Omega n a_q}\right)$
	✗	$\Theta\left(\frac{\varepsilon}{M\sqrt{n}}\right)$ and $\leq \frac{\varepsilon}{\sqrt{8n}M}$	$\mathcal{O}\left(\frac{\varepsilon^2}{M\Omega n^{3/2} a_q}\right)$

4 Numerical Experiments

In a series of our experiments, we compare zeroth-order Algorithm 1 (`zoSPA`) proposed in this paper with Mirror-Descent algorithm from [2] which uses a first-order oracle.

We consider the classical saddle-point problem on a probability simplex:

$$\min_{x\in\Delta_n}\max_{y\in\Delta_k}\left[y^T C x\right], \tag{11}$$

This problem has many different applications and interpretations, one of the main ones is a matrix game (see Part 5 in [2]), i.e. the element c_{ij} of the matrix are interpreted as a winning, provided that player X has chosen the ith strategy and player Y has chosen the jth strategy, the task of one of the players is to maximize the gain, and the opponent's task – to minimize.

We briefly describe how the step of algorithm should look for this case. The prox-function is $d(x) = \sum_{i=1}^{n} x_i \log x_i$ (entropy) and $V_x(y) = \sum_{i=1}^{n} x_i \log {x_i}/{y_i}$ (KL divergence). The result of the proximal operator is $u = \text{prox}_{z_k}(\gamma_k g(z_k, \xi_k^{\pm}, \mathbf{e}_k)) = z_k \exp(-\gamma_k g(z_k, \xi_k^{\pm}, \mathbf{e}_k))$, by this entry we mean: $u_i = [z_k]_i \exp(-\gamma_k [g(z_k, \xi_k^{\pm}, \mathbf{e}_k)]_i)$. Using the Bregman projection onto the simplex in following way $P(x) = {x}/{\|x\|_1}$, we have

$$[x_{k+1}]_i = \frac{[x_k]_i \exp(-\gamma_k [g_x(z_k, \xi_k^{\pm}, \mathbf{e}_k)]_i)}{\sum\limits_{j=1}^{n} [x_k]_j \exp(-\gamma_k [g_x(z_k, \xi_k^{\pm}, \mathbf{e}_k)]_j)},$$

$$[y_{k+1}]_i = \frac{[y_k]_i \exp(\gamma_k [g_y(z_k, \xi_k^{\pm}, \mathbf{e}_k)]_i)}{\sum\limits_{j=1}^{n} [x_k]_j \exp(\gamma_k [g_y(z_k, \xi_k^{\pm}, \mathbf{e}_k)]_j)},$$

where under g_x, g_y we mean parts of g which are responsible for x and for y. From theoretical results one can see that in our case, the same step must be used in Algorithm 1 and Mirror Descent from [2], because $n^{1/q} = 1$ for $q = \infty$.

In the first part of the experiment, we take matrix 200×200. All elements of the matrix are generated from the uniform distribution from 0 to 1. Next, we select one row of the matrix and generate its elements from the uniform from 5 to 10. Finally, we take one element from this row and generate it uniformly from 1 to 5. Then we take the same matrix, but now at each iteration we add to elements of the matrix a normal noise with zero expectation and variance of 10, 20, 30, 40% of the value of the matrix element. The results of the experiment is on Fig. 1.

According to the results of the experiments, one can see that for the considered problems, the methods with the same step work either as described in the theory (slower n times or $\log n$ times) or generally the same as the full-gradient method.

5 Possible Generalizations

In this paper we consider non-smooth cases. Our results can be generalized for the case of strongly convex functions by using restart technique (see for example [7]). It seems that one can do it analogously.[1] Generalization of the results of

[1] To say in more details this can be done analogously for deterministic set up. As for stochastic set up we need to improve the estimates in this paper by changing the Bregman diameters of the considered convex sets Ω by Bregman divergence between starting point and solution. This requires more accurate calculations (like in [11]) and doesn't include in this paper. Note that all the constants, that characterized smoothness, stochasticity and strong convexity in all the estimates in this paper can be determine on the intersection of considered convex sets and Bregman balls around the solution of a radii equals to (up to a logarithmic factors) the Bregman divergence between the starting point and the solution.

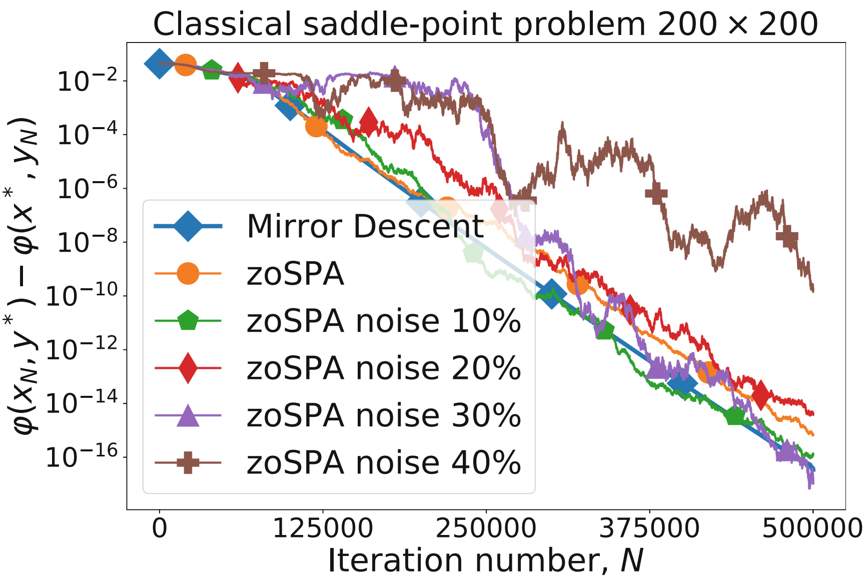

Fig. 1. `zoSPA` with 0–40% noise and `Mirror Descent` applied to solve saddle-problem (11).

[6,11,18] and [1,14] for the gradient-free saddle-point set-up is more challenging. Also, based on combinations of ideas from [1,12] it'd be interesting to develop a mixed method with a gradient oracle for x (outer minimization) and a gradient-free oracle for y (inner maximization).

References

1. Alkousa, M., Dvinskikh, D., Stonyakin, F., Gasnikov, A., Kovalev, D.: Accelerated methods for composite non-bilinear saddle point problem. arXiv preprint arXiv:1906.03620 (2019)
2. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. Society for Industrial and Applied Mathematics, Philadelphia (2019)
3. Beznosikov, A., Gorbunov, E., Gasnikov, A.: Derivative-free method for composite optimization with applications to decentralized distributed optimization. arXiv preprint arXiv:1911.10645 (2019)
4. Beznosikov, A., Sadiev, A., Gasnikov, A.: Gradient-free methods for saddle-point problem. arXiv preprint arXiv:2005.05913 (2020)
5. Duchi, J.C., Jordan, M.I., Wainwright, M.J., Wibisono, A.: Optimal rates for zero-order convex optimization: the power of two function evaluations (2013)
6. Dvurechensky, P., Gorbunov, E., Gasnikov, A.: An accelerated directional derivative method for smooth stochastic convex optimization. arXiv preprint arXiv:1804.02394 (2018)
7. Gasnikov, A.: Universal gradient descent. arXiv preprint arXiv:1711.00394 (2017)

8. Gasnikov, A.V., Krymova, E.A., Lagunovskaya, A.A., Usmanova, I.N., Fedorenko, F.A.: Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case. Autom. Remote Control **78**(2), 224–234 (2017). https://doi.org/10.1134/S0005117917020035
9. Gasnikov, A.V., Lagunovskaya, A.A., Usmanova, I.N., Fedorenko, F.A.: Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. Autom. Remote Control **77**(11), 2018–2034 (2016). https://doi.org/10.1134/S0005117916110114
10. Goodfellow, I.: Nips 2016 tutorial: generative adversarial networks. arXiv preprint arXiv:1701.00160 (2016)
11. Gorbunov, E., Dvurechensky, P., Gasnikov, A.: An accelerated method for derivative-free smooth stochastic convex optimization. arXiv preprint arXiv:1802.09022 (2018)
12. Ivanova, A., et al.: Oracle complexity separation in convex optimization. arXiv preprint arXiv:2002.02706 (2020)
13. Langley, P.: Crafting papers on machine learning. In: Langley, P. (ed.) Proceedings of the 17th International Conference on Machine Learning. (ICML 2000), Stanford, CA, pp. 1207–1216. Morgan Kaufmann (2000)
14. Lin, T., Jin, C., Jordan, M., et al.: Near-optimal algorithms for minimax optimization. arXiv preprint arXiv:2002.02417 (2020)
15. Nesterov, Y., Spokoiny, V.G.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**(2), 527–566 (2017)
16. Shamir, O.: An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. J. Mach. Learn. Res. **18**(52), 1–11 (2017)
17. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)
18. Vorontsova, E.A., Gasnikov, A.V., Gorbunov, E.A., Dvurechenskii, P.E.: Accelerated gradient-free optimization methods with a non-Euclidean proximal operator. Autom. Remote Control **80**(8), 1487–1501 (2019)