Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Zeroth-order algorithms for smooth saddle-point problems

Abdurakhmon Sadiev[1], Aleksandr Beznosikov[1,2], Pavel Dvurechensky[3], Alexander Gasnikov[1,4]

submitted: April 12, 2021

[1] Moscow Institute of Physics and Technology
Institutskiy Pereulok, 9
Dolgoprudny
141701 Moscow Region
Russian Federation
E-Mail: sadiev.aa@phystech.edu
beznosikov.an@phystech.edu

[2] Higher School of Economics
Myasnitskaya street, 20
Moscow
101000 Moscow Region
Russian Federation
E-Mail: beznosikov.an@phystech.edu

[3] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: pavel.dvurechensky@wias-berlin.de

[4] Caucasus Mathematical Center
Adyghe State University
Pervomaiskaya street, 208
Maykop
385000 Republic of Adygea
Russian Federation
E-Mail: gasnikov@yandex.ru

No. 2827
Berlin 2021

2020 *Mathematics Subject Classification.* 90C30, 90C25, 68Q25.

Key words and phrases. Zeroth-order optimization, saddle-point problems, stochastic optimization.

The research of A. Sadiev, A. Beznosikov in Section 4 was supported by the Russian Science Foundation (project 21-71-30005). The research of A. Gasnikov in Section 5 was partially supported by RFBR, project number 18-29-03071 mk. This work was partially conducted while A. Sadiev and A. Beznosikov were on the project internship in Sirius University of Science and Technology.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Zeroth-order algorithms for smooth saddle-point problems

Abdurakhmon Sadiev, Aleksandr Beznosikov, Pavel Dvurechensky, Alexander Gasnikov

Abstract

Saddle-point problems have recently gained an increased attention from the machine learning community, mainly due to applications in training Generative Adversarial Networks using stochastic gradients. At the same time, in some applications only a zeroth-order oracle is available. In this paper, we propose several algorithms to solve stochastic smooth (strongly) convex-concave saddle-point problems using zeroth-order oracles, and estimate their convergence rate and its dependence on the dimension n of the variable. In particular, our analysis shows that in the case when the feasible set is a direct product of two simplices, our convergence rate for the stochastic term is only by a $\log n$ factor worse than for the first-order methods. We also consider a mixed setup and develop 1/2th-order methods which use zeroth-order oracle for the minimization part and first-order oracle for the maximization part. Finally, we demonstrate the practical performance of our zeroth-order and 1/2th-order methods on practical problems.

1 Introduction

Zeroth-order or derivative-free methods [39, 16, 6, 43, 11] are well known in optimization in application to problems with unavailable or computationally expensive gradients. In particular, the framework of derivative-free methods turned out to be very fruitful in application to different learning problems such as online learning in the bandit setup [7] and reinforcement learning [40, 10, 18], which can be considered as a particular case of simulation optimization [19, 42]. We study stochastic derivative-free methods in a two-point feedback situation, meaning that two observations of the objective per iteration are available. This setting was considered for optimization problems by [1, 14, 41] in the learning community and by [37, 45, 21, 22, 20, 15] in the optimization community.

In this paper we go beyond the setting of optimization problems and consider convex-concave saddle-point problems for which partial derivatives of the objective are not available, which forces to use derivative-free methods. Saddle-point problems are tightly connected with equilibrium [17] and game problems [2] in many applications, e.g., economics [33], with tractable reformulations of non-smooth optimization problems [36], and with variational inequalities [27]. Gradient methods for saddle-point problems are an area of intensive study in the machine learning community in application to training of Generative Adversarial Networks [24], and other adversarial models [32], as well as to robust reinforcement learning [38]. In the latter two applications, gradients are often unavailable, which motivates the application of zeroth-order methods to the respective saddle-point problems. Moreover, this also motivates 1/2th-order methods, when the training of the network is made via stochastic gradient method with backpropagation, and adversarial examples, which are generated to force the network to give incorrect prediction, are generated by zeroth-order methods. Another application area for zeroth-order methods are Adversarial Attacks [25, 46], in particular the Black-Box Adversarial Attacks [34]. The goal is not only to train the network, but to find also a perturbation of the data in such a way that the network outputs wrong prediction. Then the training is repeated to make the network robust to such adversarial examples. Since the attacking model does not have access to the architecture of the

DOI 10.20347/WIAS.PREPRINT.2827 Berlin 2021

main network, but only to the input and output of the network, the only available oracle for the attacker is the zeroth-order oracle for the loss function. As it is shown in [12, 48, 13], this approach allows to obtain the same quality of robust training as the more laborious methods of Adversarial Attacks, but faster in up to a factor of three in terms of the training time [9].

Gradient methods for saddle-point problems are a well studied area with the classical algorithm being the extra-gradient method [30]. It was later generalized to the non-Euclidean geometry in the form of Mirror Descent [3] and Mirror-Prox [36]. These methods are designed for a more general problem of solving variational inequalities. There are also direct methods for saddle-point problems such as gradient descent ascent [35] or primal-dual hybrid gradient method [8] for saddle-point problems with bilinear structure. On the contrary, the theory of zeroth-order methods for saddle-point problems seems to be underdeveloped in the literature. We give a more detailed overview of such methods and explain our contribution in comparison with the literature below.

1.1 Our contribution and related works

In the first part of the work, we present *zeroth-order* variants of Mirror-Descent [3] and Mirror-Prox [29] methods for *stochastic saddle-point problems* in convex-concave and strongly convex-concave cases. We consider various concepts of zeroth-order oracles and various concepts of noise. Also we introduce a new class of smooth saddle-point problems – firmly smooth.

In the particular case of deterministic problems, our methods have a linear rate in the smooth strongly-convex-strongly-concave case, and sublinear rate $\mathcal{O}(1/N)$ in the convex-concave case, where N is the number of iterations. One can note that in some estimates, there is a factor of the problem's dimension n, but somewhere $n^{2/q}$. This factor q depends on geometric setup of our problem and gives a benefit when we work in the Hölder, but non-Euclidean case (use non-Euclidean prox), i.e. $\|\cdot\| = \|\cdot\|_p$ and $p \in [1; 2]$, then $\|\cdot\|_* = \|\cdot\|_q$, where $1/p + 1/q = 1$. Then q takes values from 2 to ∞, in particular, in the Euclidean case $q = 2$, but when the optimization set is a simplex, $q = \infty$. (see Table 1 for a comparison of the oracle complexity with zeroth-order methods for saddle-point problems in the literature and provided by our methods).

Our theoretical analysis shows that the zeroth-order methods has the same sublinear convergence rate in the stochastic part as the first-order method: $\mathcal{O}(1/\sqrt{N})$ in convex-concave case and $\mathcal{O}(1/N)$ in strongly-convex-strongly-concave case. (see Table 2 for a comparison of the oracle complexity in the stochastic part for first-order methods and available zeroth-order methods for stochastic saddle-point problems).

The second part of the work is devoted to the use of a mixed order oracle, i.e. a zeroth-order oracle in one variable and a first-order oracle for the other. First, we analyze a special case when such an approach is appropriate - the Lagrange multiplier method. Then we also present a general approach for this setup. The idea of using such an oracle is found in the in literature [4], but for the composite optimization problem.

As mentioned above, all theoretical results are tested in practice on a classical bilinear problem.

Method	Assumptions	Complexity in deterministic setup
ZO−GDMSA [47]	NC-SC, UCst-Cst, S	$\tilde{\mathcal{O}}\left(\frac{n\kappa^2}{\varepsilon^2}\right)$
ZO−Min−Max [31]	NC-SC, Cst-Cst, S	$\tilde{\mathcal{O}}\left(\frac{n}{\varepsilon^6}\right)$
zoSPA [5]	C-C, Cst-Cst, BG	$\mathcal{O}\left(n^{2/q}\frac{M^2D^2}{\varepsilon^2}\right)$
[**Alg 1 and 3**]	SC-SC, Cst-Cst, S	$\tilde{\mathcal{O}}\left(\min\left[n^{2/q}\kappa^2, n\kappa\right]\cdot\log\left(\frac{1}{\varepsilon}\right)\right)$
[**Alg 2**]	C-C, Cst-Cst, S	$\tilde{\mathcal{O}}\left(n\frac{LD^2}{\varepsilon}\right)$
[**Alg 1**]	C-C, Cst-Cst, FS	$\tilde{\mathcal{O}}\left(n^{2/q}\frac{L^2D^2}{\varepsilon}\right)^*$

Table 1: Comparison of oracle complexity in deterministic setup of different zeroth-order methods with different assumptions on target function $f(x,y)$: C-C – convex-concave, SC-SC – strongly-convex-strongly-concave, NC-SC – nonconvex-strongly-concave; Cst – optimizaation set is constrained, UCst – unconstrained; S - smooth, FS - firmly smooth (see (9)), BG - bounded gradients. Here ε means the accuracy of the solution, D – the diameter of the optimization set, μ – strong convexity constant (see (7)), L – smoothness constant (see (8)), $\kappa = {}^{L}/_{\mu}$, M – bound of the gradient ($\|\nabla_x f(x,y)\|_2 \leq M$, $\|\nabla_y f(x,y)\|_2 \leq M$), n – the sum of the dimensions of the variables x and y, $q = 2$ for the Euclidean case and $q = \infty$ for setup of $\|\cdot\|_1$-norm. *convergence on $\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\|F(x_k,y_k) - F(x^*,y^*)\|_2^2\right]$, where $F(x,y) = (\nabla_x f(x,y), -\nabla_y f(x,y))$.

Method	Order	Assumptions	Complexity for stochastic part
EGMP [29]	1st	C-C, Cst-Cst, S	$\mathcal{O}\left(\frac{\sigma^2D^2}{\varepsilon^2}\right)$
PEG [28]	1st	SC-SC, Cst-Cst, S	$\mathcal{O}\left(\frac{\sigma^2}{\mu^2\varepsilon}\right)$
ZO−SGDMSA[47]	0th	NC-SC, UCst-Cst, S	$\tilde{\mathcal{O}}\left(\frac{\kappa^2n\sigma^2}{\varepsilon^4}\right)$
[**Alg 1**]	0th	SC-SC, Cst-Cst, S	$\mathcal{O}\left(\frac{n^{2/q}\sigma^2}{\mu^2\varepsilon}\right)$
[**Alg 2**]	0th	C-C, Cst-Cst, S	$\mathcal{O}\left(\frac{n\sigma^2D^2}{\varepsilon^2}\right)$
[**Alg 1**]	0th	C-C, Cst-Cst, FS	$\mathcal{O}\left(\frac{n^{2/q}\sigma^2D^2}{\varepsilon^2}\right)$

Table 2: Comparison of oracle complexity for stochastic part of different first- and zeroth-order methods with different assumptions on $f(x,y)$: see notation in Table 1. Here σ^2 – the bound of variance (see (3)).

2 Problem setup and assumptions

We consider a saddle-point problem:

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

where $\mathcal{X} \subset \mathbb{R}^{n_x}$ and $\mathcal{Y} \subset \mathbb{R}^{n_y}$ are convex compact sets. For simplicity, we introduce the set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $z = (x, y)$ and the operator F:

$$F(z) = F(x, y) = \begin{pmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{pmatrix}. \tag{2}$$

We focus on the case when we do not have access to the values of $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$, but we have access to the inexact zeroth-order oracle, i.e. inexact values of the objective $f(x, y)$. The inexactness in the zeroth-order oracle includes stochastic noise and unknown bounded noise, which can be of an adversarial nature. More precisely, we have access to the values $\tilde{f}(z, \xi)$ such that $\tilde{f}(z, \xi) = f(z, \xi) + \delta(z)$ and

$$\begin{aligned} \mathbb{E}[f(z, \xi)] = f(z), \;\; \mathbb{E}[F(z, \xi)] = F(z), \\ \mathbb{E}[\|F(z, \xi) - F(z)\|_2^2] \leq \sigma^2, \;\; |\delta(z)| \leq \Delta. \end{aligned} \tag{3}$$

We consider two types of approximations for $F(z)$ based on the available observations of $\tilde{f}(z, \xi)$.

Random direction oracle. In this strategy, the vectors e_x, e_y are generated uniformly on the unit Euclidean sphere, i.e. $e_x \in \mathcal{RS}^2_{n_x}(1)$ and $e_y \in \mathcal{RS}^2_{n_y}(1)$. And

$$g_d(z, e, \tau, \xi) = \frac{n}{\tau} \begin{pmatrix} \left(\tilde{f}(x + \tau e_x, y, \xi) - \tilde{f}(x, y, \xi)\right) e_x \\ \left(\tilde{f}(x, y, \xi) - \tilde{f}(x, y + \tau e_y, \xi)\right) e_y \end{pmatrix}, \tag{4}$$

where $\tau > 0$ is called smoothed parameter and $n = n_x + n_y + 1$.

Full coordinates oracle. Here we consider a standard orthonormal basis $\{h_1, \ldots, h_{n_x+n_y}\}$ and construct an approximation for the operator F in the following form:

$$\begin{aligned} g_f(z, h, \tau, \xi) \;\; = \;\; & \frac{1}{\tau} \sum_{i=1}^{n_x} \left(\tilde{f}(z + \tau h_i, \xi) - \tilde{f}(z, \xi)\right) h_i \\ & + \frac{1}{\tau} \sum_{i=n_x+1}^{n_x+n_y} \left(\tilde{f}(z, \xi) - \tilde{f}(z + \tau h_i, \xi)\right) h_i. \end{aligned} \tag{5}$$

In this concept, we need to call $\tilde{f}$ oracle $n_x + n_y + 1$ times, whereas in the previous case only 3 times.

3 Notation and Definitions

We use $\langle x, y \rangle \stackrel{\text{def}}{=} \sum_{i=1}^n x_i y_i$ 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 $\|x\|_2 \stackrel{\text{def}}{=} \sqrt{\langle x, x \rangle}$. We define ℓ_p-norms as $\|x\|_p \stackrel{\text{def}}{=} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for $p \in (1, \infty)$ and for $p = \infty$ we use $\|x\|_\infty \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} |x_i|$. The dual norm $\|\cdot\|_q$ for the norm $\|\cdot\|_p$ is defined in the following way: $\|y\|_q \stackrel{\text{def}}{=} \max\{\langle x, y \rangle \mid \|x\|_p \leq 1\}$. Operator $\mathbb{E}[\cdot]$ is full mathematical expectation and operator $\mathbb{E}_\xi[\cdot]$ express conditional mathematical expectation.

As stated above, during the course of the paper we will work in an arbitrary norm $\|\cdot\| = \|\cdot\|_p$, where $p \in [1; 2]$. And its conjugate $\|\cdot\|_* = \|\cdot\|_q$ with $q \in [2; +\infty)$ and $1/p + 1/q = 1$. Some assumptions will be made later in the Euclidean norm - we will write this explicitly $\|\cdot\|_2$.

Definition 1. 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 2. 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:

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

Definition 3. Let $V_z(w)$ Bregman divergence. For all $x \in \mathcal{Z}$ define prox-operator of ξ:

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

Next we present the assumptions that we will use in the convergence analysis.

Assumption 1. The set $\mathcal{Z}$ is bounded w.r.t $\|\cdot\|$ by constant D_p, i.e.

$$V_{z_1}(z_2) \leq D_p^2, \ \ \forall z_1, z_2 \in \mathcal{Z}. \tag{6}$$

Assumption 2. $f(x, y)$ is convex-concave. It means that $f(\cdot, y)$ is convex for all y and $f(x, \cdot)$ is concave for all x.

Assumption 2(s). $f(x, y)$ is strongly-convex-strongly-concave. It means that $f(\cdot, y)$ is strongly-convex for all y and $f(x, \cdot)$ is strongly-concave for all x w.r.t. $V_\cdot(\cdot)$, i.e. for all $x_1, x_2 \in \mathcal{X}$ and for all $y_1, y_2 \in \mathcal{Y}$ we have

$$\begin{aligned} f(x_1, y_2) &\geq f(x_2, y_2) + \langle \nabla_x f(x_2, y_2), x_1 - x_2 \rangle \\ &\quad + \frac{\mu}{2} \left(V_{(x_2, y_2)}(x_1, y_2) + V_{(x_1, y_2)}(x_2, y_2) \right), \\ -f(x_2, y_1) &\geq -f(x_2, y_2) + \langle -\nabla_y f(x_2, y_2), y_1 - y_2 \rangle \\ &\quad + \frac{\mu}{2} \left(V_{(x_2, y_2)}(x_2, y_1) + V_{(x_1, y_1)}(x_2, y_2) \right). \end{aligned} \tag{7}$$

Assumption 3. $f(x, y, \xi)$ is $L(\xi)$-Lipschitz continuous w.r.t $\|\cdot\|_2$, i.e. for all $x_1, x_2 \in \mathcal{X}, y_1, y_2 \in \mathcal{Y}$ and ξ

$$\left\| \begin{pmatrix} \nabla_x f(x_1, y_1, \xi) \\ -\nabla_y f(x_1, y_1, \xi) \end{pmatrix} - \begin{pmatrix} \nabla_x f(x_2, y_2, \xi) \\ -\nabla_y f(x_2, y_2, \xi) \end{pmatrix} \right\|_2 \leq L(\xi) \left\| \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \right\|_2. \tag{8}$$

Assumption 3(f). $f(x, y)$ s L-firmly Lipschitz continuous w.r.t $\|\cdot\|_2$, i.e. for all $x_1, x_2 \in \mathcal{X}, y_1, y_2 \in \mathcal{Y}$

$$\begin{aligned} &\left\| \begin{pmatrix} \nabla_x f(x_1, y_1, \xi) \\ -\nabla_y f(x_1, y_1, \xi) \end{pmatrix} - \begin{pmatrix} \nabla_x f(x_2, y_2, \xi) \\ -\nabla_y f(x_2, y_2, \xi) \end{pmatrix} \right\|_2^2 \\ &\qquad \leq L(\xi) \left\langle \begin{pmatrix} \nabla_x f(x_1, y_1, \xi) \\ -\nabla_y f(x_1, y_1, \xi) \end{pmatrix} - \begin{pmatrix} \nabla_x f(x_2, y_2, \xi) \\ -\nabla_y f(x_2, y_2, \xi) \end{pmatrix}, \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \right\rangle. \end{aligned} \tag{9}$$

For (8) and (9) we assume that exists L_2 such that $\mathbb{E}[L^2(\xi)] \leq L_2^2$. For deterministic case L_2 is equal to deterministic constant L (without ξ).

By Cauchy-Schwarz, (8) follows from (9). It is easy to see that the assumptions 4 and 4(f) above can be easily rewritten in a more compact form using $F(z)$. For assumption 3(s) it is more complicated:

Lemma 3.1. *If $f(x, y)$ is μ-strongly convex on x and μ-strongly concave on y w.r.t $V_\cdot(\cdot)$, then for $F(z)$ we have*

$$\langle F(z_1) - F(z_2), z_1 - z_2\rangle \geq \frac{\mu}{2}\left(V_{z_1}(z_2) + V_{z_2}(z_1)\right), \quad \forall z_1, z_2 \in \mathcal{Z}.$$

Hereinafter, we do not present the proofs of lemmas and theorems in the main part of the paper – see the corresponding parts of the appendix. And we can present some properties of oracles (4), (5):

Lemma 3.2. *Let $e \in \mathcal{RS}^2(1)$, i.e. uniformly distributed on the unit Euclidean sphere. Randomness comes from independent variables e, ξ and a point z. Norm $\|\cdot\|_* = \|\cdot\|_q$ satisfies $q \in [2; +\infty)$. We introduce the constant ρ_n:*

$$\rho_n = \min\{q - 1, 16\log(n) - 8\}.$$

Then under Assumption 3 or 3(f) the following statements hold:

- *for Random direction oracle*

$$\begin{aligned}
\mathbb{E}\left[\|g_d(z, e, \tau, \xi)\|_q^2\right] &\leq 48n^{2/q}\rho_n\mathbb{E}\left[\|F(z) - F(z^*)\|_2^2\right] + 48n^{2/q}\rho_n\|F(z^*)\|_2^2 \\
&\quad + 48n^{2/q}\rho_n\sigma^2 + 8n^{2/q+1}\rho_n L^2\tau^2 \\
&\quad + 16\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2}, \\
\|\mathbb{E}[g_d(z, e, \tau, \xi)] - F(z)\|_q &\leq 2n^{1/q+1/2}\sqrt{\rho_n}L\tau + 4n^{1/q+1/2}\sqrt{\rho_n}\frac{\Delta}{\tau};
\end{aligned}$$

- *for Full coordinates oracle*

$$\begin{aligned}
\mathbb{E}\left[\|g_f(z, \tau, \xi) - F(z)\|_q^2\right] &\leq 3\sigma^2 + 3nL_2^2\tau^2 + \frac{6n\Delta^2}{\tau^2}, \\
\|\mathbb{E}\left[g_f(z, \tau, \xi)\right] - F(z)\|_q &\leq \sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}.
\end{aligned}$$

4 Zeroth-Order Methods

In this part, we present methods for solving problem (1), which use only the zeroth-order oracle. First of all, we want to consider the classic version of the Mirror-Descent algorithm. For theoretical and practical analysis of this

Algorithm 1 `zoVIA`

Input: z_0, N, γ, τ.
Choose grad to be either g_d or g_f.
for $k = 0$ **to** N **do**
  Sample indep. e_k, ξ_k.
  $d_k = \text{grad}(z_k, e_k, \tau, \xi_k)$.
  $z_{k+1} = \text{prox}_{z_k}(\gamma \cdot d_k)$.
end for
Output: z_{N+1} or $\bar{z}_{N+1}$.

algorithm in the non-smooth case, but with a bounded gradient, see [3](first order), [5](zero order). The main problem of this approach is that it is difficult to analyze in the case when f is convex-concave and Lipschitz continuous (Assumptions 2 and 3). But in practice, this algorithm does not differ much from its counterparts, which will be given below. Let us analyze this algorithm in convex-concave and strongly-convex-strongly-concave cases with Random direction oracle:

Theorem 4.1. *By Algorithm 1 with Random direction oracle*

- *under Assumptions 1, 2, 3(f) and with* $\gamma \leq \frac{1}{48n^{2/q}\rho_n L}$, *we get*

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[\|F(z_k)-F(z^*)\|_2^2\right] &\leq \frac{2LD_p^2}{\gamma N}+48\gamma n^{2/q}\rho_n L\left(\|F(z^*)\|_2^2+\sigma^2\right)\\ &\quad+8\gamma n^{2/q+1}\rho_n L\left(L_2^2\tau^2+2\frac{\Delta^2}{\tau^2}\right)\\ &\quad+8n^{1/q+1/2}\sqrt{\rho_n}LD_p\left(L\tau+\frac{2\Delta}{\tau}\right);\end{aligned}$$

- *under Assumptions 1, 2(s), 3 and with* $\gamma \leq \frac{\mu}{96n^{2/q}\rho_n L^2}$*:*

$$\begin{aligned}\mathbb{E}\left[V_{z_{N+1}}(z^*)\right] &\leq V_{z_0}(z^*)\exp\left(-\frac{\mu^2 N}{400n^{2/q}\rho_n L^2}\right)+\\ &\quad+\frac{24n^{2/q}\rho_n}{\mu^2 N}\left(\|F(z^*)\|_2^2+\sigma^2\right)\\ &\quad+\frac{4n^{2/q+1}\rho_n}{\mu^2 N}\left(L_2^2\tau^2+2\frac{\Delta^2}{\tau^2}\right)\\ &\quad+\frac{4n^{1/q+1/2}\sqrt{\rho_n}D_p}{\gamma\mu^2 N}\left(L\tau+\frac{2\Delta}{\tau}\right).\end{aligned}$$

Remark. In the first statement of the Theorem, we used an unusual convergence criterion, it can be interpreted as follows: let as the output $\tilde{z}_N$ of the algorithm we choose a random point from z_0 to z_N. Then

$$\mathbb{E}\left[\|F(\tilde{z}_N)\|_2^2\right]=\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\left[\|F(z_k)\|_2^2\right].$$

In this theorem and below, we draw attention to the fact that in the main part of the convergence there is a deterministic constant L, and in the parts that are responsible for noise – L_2 (see (8),(9)).

Corollary 4.2. *For Algorithm 1*

- *under Assumptions 1, 2, 3(f) and with* $\gamma=\min\left\{\frac{1}{48n^{2/q}\rho_n L},\frac{D_p}{n^{1/q}\sqrt{\rho_n}\sigma\sqrt{N}}\right\}$,

$$\tau=\Theta\left(\min\left\{\frac{\varepsilon}{n^{1/q+1/2}\sqrt{\rho_n}L^2D_p},\max\left[\sqrt{\frac{\varepsilon}{nL_2^2}},\frac{\sigma}{\sqrt{n}L_2}\right]\right\}\right),\quad \Delta=\mathcal{O}\left(L_2\tau^2\right),$$

the oracle complexity (coincides with the number of iterations) to find ε*-solution (in terms of the convergence criterion from Theorem 1) is*

$$N=\mathcal{O}\left(\max\left\{\frac{n^{2/q}\rho_n L^2D_p^2}{\varepsilon},\frac{n^{2/q}\rho_n\sigma^2D_p^2}{\varepsilon^2}\right\}\right).$$

- *under Assumptions 1, 2(s), 3 and with* $\gamma=\frac{\mu}{96n^{2/q}\rho_n L^2}$,

$$\tau=\Theta\left(\min\left\{\max\left[\frac{\sqrt{\varepsilon}L}{L_2},\frac{\sigma}{\sqrt{n}L_2}\right],\max\left[\frac{\varepsilon\mu}{n^{1/q+1/2}\sqrt{\rho_n}LD_p},\frac{\sigma^2\mu}{n^{1/q+1/2}\sqrt{\rho_n}L^3D_p}\right]\right\}\right),$$

$\Delta = \mathcal{O}\left(L_2\tau^2\right)$, the oracle complexity (coincides with the number of iterations) to find ε-solution (in terms of the convergence criterion from Theorem 1) can be bounded by

$$N = \widetilde{\mathcal{O}}\left(\max\left\{\frac{n^{2/q}\rho_n L^2}{\mu^2}\log\left(\frac{1}{\varepsilon}\right), \frac{n^{2/q}\rho_n\sigma^2}{\mu^2\varepsilon}\right\}\right).$$

Remark. We analyze only Random direction oracle. The estimate of the oracle complexity with Full coordinate oracle has the same form with $q = 2$.

Next, we consider a standard algorithm for working with smooth saddle-point problem. It builds on the extra-gradient method [30]. The idea of using this approach for saddle-point problems is not new [29]. It has both heuristic advantages (we forestall the properties of the gradient) as well as purely mathematical ones (a more clear theoretical analysis). We use two versions of this approach: classic and single call version from [28].

Algorithm 2 `zoESVIA`

Input: z_0, N, γ, τ.
Choose oracle grad from g_d, g_f.
for $k = 0$ **to** N **do**
  Sample indep. e_k, $e_{k+1/2}$, ξ_k, $\xi_{k+1/2}$.
  $d_k = \text{grad}(z_k, e_k, \tau, \xi_k)$.
  $z_{k+1/2} = \text{prox}_{z_k}(\gamma \cdot d_k)$.
  $d_{k+1/2} = \text{grad}(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2})$.
  $z_{k+1} = \text{prox}_{z_k}(\gamma \cdot d_{k+1/2})$.
end for
Output: z_{N+1} or $\bar{z}_{N+1}$.

Algorithm 3 `zoscESVIA`

Input: z_0, N, γ, τ.
Choose oracle grad from g_d, g_f.
for $k = 0$ **to** N **do**
  Sample independent e_k, ξ_k.
  Take d_{k-1} from previous step.
  $z_{k+1/2} = \text{prox}_{z_k}(\gamma \cdot d_{k-1})$.
  $d_k = \text{grad}(z_{k+1/2}, e_{k+1/2}, \tau, \xi_k)$.
  $z_{k+1} = \text{prox}_{z_k}(\gamma \cdot d_k)$.
end for
Output: z_{N+1} or $\bar{z}_{N+1}$.

Here $\bar{z}_{N+1} = \frac{1}{N+1}\sum_{i=0}^{N} z_{i+1/2}$.

Next, we will deal with the theoretical analysis of convergence:

Theorem 4.3. ■ *By Algorithm 2 with Full coordinates oracle under Assumptions 1, 2, 3 and with $\gamma \leq 1/2L$, we have*

$$\begin{aligned}\mathbb{E}\left[\varepsilon_{sad}(\bar{z}_{N+1})\right] \leq\ & \frac{2D_p^2}{\gamma N} + 11\gamma\left(nL_2^2\tau^2 + \sigma^2 + 2\frac{n\Delta^2}{\tau^2}\right) \\ & + 2D_p\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right),\end{aligned}$$

where

$$\varepsilon_{sad}(\bar{z}_{N+1}) = \max_{y'\in\mathcal{Y}} f(\bar{x}_{N+1}, y') - \min_{x'\in\mathcal{X}} f(x', \bar{y}_{N+1}),$$

$\bar{x}_{N+1}$, $\bar{y}_{N+1}$ are defined the same way as $\bar{z}_{N+1}$.

■ *By Algorithm 3 with Full coordinates oracle under Assumptions 1, 2(s), 3 and with $p = 2$ ($V_x(y) = 1/2\|x - y\|_2^2$), $\gamma \leq 1/6L$:*

$$\begin{aligned}\mathbb{E}\left[\|z_{N+1} - z^*\|_2^2\right] \leq\ & \exp\left(-\frac{\mu N}{12L}\right)\left(\|z_0 - z^*\|_2^2 + \|g_f(z_0,\tau,\xi_0) - g_f(z_0,\tau,\xi_0)\|_2^2\right) \\ & + \frac{1}{\mu^2 N}12\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right) \\ & + \frac{1}{\mu^2 N}\frac{4D_2}{\gamma}\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right).\end{aligned}$$

Corollary 4.4. *Let ε – accuracy of the solution (in terms of the convergence criterion from Theorem 2).*

- *For Algorithm 2 with Full coordinates oracle under Assumptions 1, 2, 3 with $\gamma = \min\left\{1/2L, D_p/(\sigma\sqrt{N})\right\}$ and additionally*

$$\tau = \mathcal{O}\left(\min\left\{\frac{\varepsilon}{\sqrt{n}LD_2}, \max\left[\sqrt{\frac{\varepsilon L}{nL_2^2}}, \frac{\sigma}{\sqrt{n}L_2}\right]\right\}\right), \quad \Delta = \mathcal{O}\left(L_2\tau^2\right),$$

we have the number of iterations to find ε-solution

$$N = \mathcal{O}\left(\max\left\{\frac{LD_2^2}{\varepsilon}, \frac{\sigma^2 D_p^2}{\varepsilon^2}\right\}\right).$$

- *For Algorithm 3 with Full coordinates oracle under Assumptions 1, 2(s), 3, with $p = 2$ ($V_x(y) = 1/2\|x-y\|_2^2$), $\gamma = 1/6L$ and additionally*

$$\tau = \mathcal{O}\left(\min\left\{\max\left[\sqrt{\frac{\varepsilon\mu L}{L_2^2}}, \frac{\sigma}{\sqrt{n}L_2}\right], \max\left[\frac{\mu\varepsilon}{\sqrt{n}LD_2}, \frac{\sigma^2}{\sqrt{n}L^2D_2}\right]\right\}\right),$$

$\Delta = \mathcal{O}\left(L_2\tau^2\right)$*, the number of iterations to find ε-solution:*

$$N = \widetilde{\mathcal{O}}\left(\max\left\{\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right), \frac{\sigma^2}{\mu^2\varepsilon}\right\}\right).$$

Remark. The oracle complexity for the Full coordinate oracle is n times greater than the number of iterations.

The analysis is carried out only for the Full coordinate oracle. The main problem of using Random Direction is that their variance is tied to the norm of the gradient; therefore, using an extra step does not give any advantages over Algorithm 1. A possible way out of this situation is to use the same direction e within one iteration of Algorithm 2 – this idea is implemented in Appendix F and in Practice part. It is interesting how it work in practice, because in the non-smooth case [5] the gain by the factor $n^{2/q}$ can be obtained.

5 $1/2$-Order Methods

In this section, we have access to a first-order oracle in one of the variables, and in the other – only a zeroth-order oracle. For such a case, we suggest using an oracle of the form:

$$\widetilde{g}(z,\tau) = \begin{pmatrix} [grad(x,y)]_x \\ -\nabla_y f(x,y) \end{pmatrix},$$

where $[grad(x,y)]_x$ – one of the zeroth-order approximations on variable x: (4) or (5). Before proving the general case, we consider one illustrative example:

5.1 Lagrange multiplier method

Let $\mathcal{X} \subset \mathbb{R}^n$ be a convex, compact set and functions $f(x), g_1(x), \ldots, g_m(x)$ be convex, smooth. We solve the following optimization problem:

$$\min_{x \in \mathcal{X}} f(x),$$
$$\text{s.t. } g_i(x) \leq 0 \, \forall i \in 1, \ldots m.$$

A dual problem to the original one:

$$\max_{\lambda \in \perp_m} \min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda) = f(x) + \langle \lambda, g(x) \rangle,$$

where $\perp_m = \{y \in \mathbb{R}^m \mid y_i \geq 0\}$ – a positive orthant, $\mathcal{L}(x, \lambda)$ – a Lagrange function, λ – a Lagrange multiplier, $g(x) = (g_1(x), \ldots, g_m(x)))^T$. We got a saddle-point problem that we want to solve using the zeroth-order method, i.e. only function values are available. But it turns out that we have access to $\nabla_\lambda \mathcal{L}(x, \lambda) = g(x)$ completely free: when we build the "gradient" on x using finite differences, we call the value for $g(x)$ and immediately get the gradient λ.

For such a problem, the oracle of the zero and first orders can be called the same number of times. In general, it is unprofitable to calculate the gradient as many times as the zeroth-order oracles and a slightly different result is obtained:

5.2 Universal approach with Full gradient method

Define Mixed oracle:

$$\widetilde{g}_f(z, \tau) = \begin{pmatrix} [g_f(x, y)]_x \\ -\nabla_y f(x, y) \end{pmatrix},$$

then

Theorem 5.1. *By Algorithm 2 under assumption 1, 2, 3 with Mixed oracle $\widetilde{g}_f$ and $\gamma \leq 1/2L$, we get*

$$\begin{aligned} \mathbb{E}\left[\varepsilon_{sad}(\bar{z}_N)\right] \quad \leq \quad & \frac{D_p^2}{\gamma N} + 2D_p \left(\sqrt{n_x} L_2 \tau + \frac{2\sqrt{n_x}\Delta}{\tau} \right) \\ & + 9\gamma \left(\sigma^2 + n_x L_2^2 \tau^2 + \frac{2 n_x \Delta^2}{\tau^2} \right) \end{aligned}$$

Corollary 5.2. *To get accuracy ε (in terms of the convergence criterion from Theorem 2) in Algorithm 2 with Mixed oracle, under Assumptions 1, 2, 3, with $\gamma = \min\left\{1/2L, D_p/(\sigma\sqrt{N})\right\}$,*

$$\tau = \mathcal{O}\left(\min\left\{ \frac{\varepsilon}{\sqrt{n} L D_p}, \max\left[\sqrt{\frac{\varepsilon L}{n L_2^2}}, \frac{\sigma}{\sqrt{n} L_2} \right] \right\} \right), \quad \Delta = \mathcal{O}\left(L_2 \tau^2 \right),$$

we need to call Full coordinates oracle for x

$$N = \mathcal{O}\left(\max\left\{ \frac{L D_p^2}{\varepsilon}, \frac{\sigma^2 D_p^2}{\varepsilon^2} \right\} \right) \quad \textit{times.}$$

6 Practice part

The main goal of our experiments is to compare the Algorithms 1,2,3 and 4 (see Appendix F) described in this paper with Full coordinate and Random direction oracles. We consider the classical bilinear saddle-point problem on a probability simplex:

$$\min_{x\in\Delta_n}\max_{y\in\Delta_k}\left[y^TCx\right], \tag{10}$$

This problem is often referred to as a matrix game (see Part 5 in [3]). Two players X and Y are playing. The goal of player Y is to win as much as possible by correctly choosing an action from 1 to k, the goal of player X is to minimize the gain of player X using his actions from 1 to n. Each element of the matrix c_{ij} are interpreted as a winning, provided that player X has chosen the i-th strategy and player Y has chosen the j-th strategy.

Let consider the step of algorithm. 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

$$u=\mathsf{prox}_{z_k}(\gamma_k\mathsf{grad}(z_k,e_k,\tau,\xi_k))=z_k\exp(-\gamma_k\mathsf{grad}(z_k,e_k,\tau,\xi_k)),$$

by this entry we mean:

$$u_i=[z_k]_i\exp(-\gamma_k[\mathsf{grad}(z_k,e_k,\tau,\xi_k)]_i).$$

Using the Bregman projection onto the simplex in following way $P(x)={}^{x}/_{\|x\|_1}$, we have

$$[x_{k+1}]_i=\frac{[x_k]_i\exp(-\gamma_k[\mathsf{grad}_x(z_k,e_k,\tau,\xi_k)]_i)}{\sum\limits_{j=1}^n[x_k]_j\exp(-\gamma_k[\mathsf{grad}_x(z_k,e_k,\tau,\xi_k)]_j)},$$

$$[y_{k+1}]_i=\frac{[y_k]_i\exp(\gamma_k[\mathsf{grad}_y(z_k,e_k,\tau,\xi_k)]_i)}{\sum\limits_{j=1}^n[y_k]_j\exp(\gamma_k[\mathsf{grad}_y(z_k,e_k,\tau,\xi_k)]_j)},$$

where under g_x, g_y we mean parts of g which are responsible for x and for y.

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. The results of the experiment is on Figure 1.

From the experiment results, one can easily see the best approach in terms of oracle complexity.

7 Conclusion

In this paper, we presented various algorithms for optimizing smooth stochastic saddle point problems using zero-order oracles. For some oracles, we provide a theoretical analysis. We also compare the approaches covered in the work on a practical matrix game.

As a continuation of the work, we can distinguish the following areas: convergence estimates for Algorithm 4 (see the appendix), the study of gradient-free methods for saddle point problems already with a one-point approximation (in this work, we used a two-point one). We also highlight the acceleration of these methods.

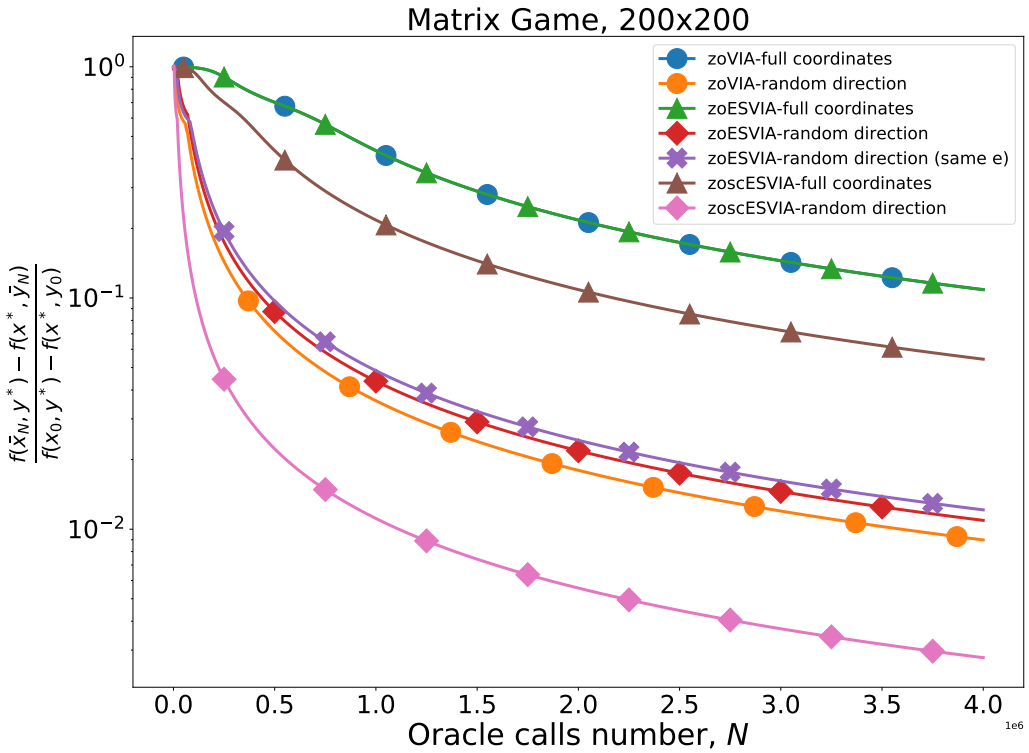

Figure 1: Different algorithms with Full coordinate and Random direction oracles applied to solve saddle-problem (10).

A General facts and technical lemmas

Lemma A.1. *For arbitrary integer $n \geq 1$ and arbitrary set of positive numbers $a_1, \ldots, a_n$ we have*

$$\left(\sum_{i=1}^{m} a_i\right)^2 \leq m \sum_{i=1}^{m} a_i^2. \tag{11}$$

Lemma A.2. *For $q \geq 2$ and for arbitrary vectors $a \in \mathbb{R}^n, b \in \mathbb{R}^m$ we have*

$$\left\| \begin{pmatrix} a \\ b \end{pmatrix} \right\|_q^2 \leq \|a\|_q^2 + \|b\|_q^2. \tag{12}$$

Lemma A.3 (Fact 5.3.2 from [3])**.** *Given norm $\|\cdot\|$ on space $\mathcal{Z}$ and prox-function $d(z)$, let $z \in \mathcal{Z}$, $w \in \mathbb{R}^n$ and $z_+ = \text{prox}_z(w)$. Then for all $u \in \mathcal{Z}$*

$$\langle w, z_+ - u\rangle \leqslant V_z(u) - V_{z_+}(u) - V_z(z_+). \tag{13}$$

Lemma A.4 (see Lemma 1 from [26])**.** *Let $e \in \mathcal{RS}^2(1)$, i.e. a random vector uniformly distributed on the surface of the unit Euclidean sphere in $\mathbb{R}^n$, $q \in [2; +\infty)$. Then, for $n \geq 8$,*

$$\mathbb{E}\left[\|e\|_q^2\right] \leq n^{2/q-1}\rho_n, \tag{14}$$

$$\mathbb{E}\left[\langle s, e\rangle^2 \|e\|_q^2\right] \leq 6n^{2/q-2}\rho_n \|s\|_2^2, \ \forall s \in \mathbb{R}^n, \tag{15}$$

where $\rho_n = \min\{q-1, 16\log n - 8\}$.

Lemma A.5 (see Lemma 2 from [44])**.** *Let consider non-negative sequence r_k:*

$$r_{k+1} \leq (1 - a\gamma)r_k + c\gamma^2,$$

where $a, c > 0$, $\gamma = \min\left(\frac{1}{d}, \frac{\log(\max(2, a^2 r_0 T/c)}{aN}\right)$. Then

$$r_{N+1} \leq r_0 \cdot \exp\left(-\frac{aN}{2d}\right) + \frac{c}{a^2 N}. \tag{16}$$

B Proof of Lemma 3.1

Lemma. If $f(x, y)$ is μ-strongly convex on x and μ-strongly concave on y w.r.t $V_\cdot(\cdot)$, then for $F(z)$ we have

$$\langle F(z_1) - F(z_2), z_1 - z_2\rangle \geq \frac{\mu}{2}\left(V_{z_1}(z_2) + V_{z_2}(z_1)\right), \ \forall z_1, z_2 \in \mathcal{Z}. \tag{17}$$

Proof. By definition of μ-strong convexity w.r.t $V_\cdot(\cdot)$:

$$f(x_1, y_2) \geq f(x_2, y_2) + \langle \nabla_x f(x_2, y_2), x_1 - x_2\rangle + \frac{\mu}{2}\left(V_{(x_2,y_2)}(x_1, y_2) + V_{(x_1,y_2)}(x_2, y_2)\right),$$

$$f(x_2, y_1) \geq f(x_1, y_1) + \langle \nabla_x f(x_1, y_1), x_2 - x_1\rangle + \frac{\mu}{2}\left(V_{(x_1,y_1)}(x_2, y_1) + V_{(x_2,y_1)}(x_1, y_1)\right),$$

$$-f(x_1, y_2) \geq -f(x_1, y_1) + \langle -\nabla_y f(x_1, y_1), y_2 - y_1\rangle + \frac{\mu}{2}\left(V_{(x_1,y_1)}(x_1, y_2) + V_{(x_1,y_2)}(x_1, y_1)\right),$$

$$-f(x_2,y_1) \geq -f(x_2,y_2) + \langle -\nabla_y f(x_2,y_2), y_1 - y_2\rangle + \frac{\mu}{2}\left(V_{(x_2,y_2)}(x_2,y_1) + V_{(x_1,y_1)}(x_2,y_2)\right).$$

Let introduce a new definition for sum of Bregman divergences:

$$\begin{aligned}\mathcal{V} &= V_{(x_2,y_2)}(x_1,y_2) + V_{(x_1,y_2)}(x_2,y_2) + V_{(x_1,y_1)}(x_2,y_1) + V_{(x_2,y_1)}(x_1,y_1)\\ &\quad + V_{(x_1,y_1)}(x_1,y_2) + V_{(x_1,y_2)}(x_1,y_1) + V_{(x_2,y_2}(x_2,y_1) + V_{(x_1,y_1)}(x_2,y_2).\end{aligned}$$

Using definition of Bregman divergence and 1-stronge convexity of prox-function d, we get:

$$\begin{aligned}\mathcal{V} &= \langle \nabla_x d(x_2,y_2) - \nabla_x d(x_1,y_2), x_2 - x_1\rangle\\ &\quad + \langle \nabla_x d(x_2,y_1) - \nabla_x d(x_1,y_1), x_2 - x_1\rangle\\ &\quad + \langle \nabla_y d(x_2,y_2) - \nabla_y d(x_2,y_1), y_2 - y_1\rangle\\ &\quad + \langle \nabla_y d(x_1,y_2) - \nabla_y d(x_1,y_1), y_2 - y_1\rangle\\ &= \langle \nabla d(z_2) - \nabla d(z_1), z_2 - z_1\rangle + \langle \nabla d(\tilde{z}_2) - \nabla d(\tilde{z}_1), \tilde{z}_2 - \tilde{z}_1\rangle\\ &\geq V_{z_1}(z_2) + V_{z_2}(z_1),\end{aligned}$$

where $\tilde{z}_2 = (x_2, y_1)$, $\tilde{z}_1 = (x_1, y_2)$ Thus, we have $\mathcal{V} \geq V_{z_1}(z_2) + V_{z_2}(z_1)$. Summming up:

$$\begin{aligned}&\langle \nabla_x f(x_2,y_2) - \nabla_x f(x_1,y_1), x_1 - x_2\rangle\\ &\qquad - \langle \nabla_y f(x_2,y_2) - \nabla_y f(x_1,y_1), y_1 - y_2\rangle + \frac{\mu\mathcal{V}}{2} \leq 0.\end{aligned}$$

Using $\mathcal{V} \geq V_{z_1}(z_2) + V_{z_2}(z_1)$, we have

$$\begin{aligned}\langle \nabla_x f(x_2,y_2) - \nabla_x f(x_1,y_1), x_1 - x_2\rangle - \langle \nabla_y f(x_2,y_2) - \nabla_y f(x_1,y_1), y_1 - y_2\rangle\\ + \frac{\mu}{2}(V_{z_1}(z_2) + V_{z_2}(z_1)) \leq 0,\end{aligned}$$

and

$$\begin{aligned}\langle F(z_1) - F(z_2), z_1 - z_2\rangle &= \langle \nabla_x f(x_2,y_2) - \nabla_x f(x_1,y_1), x_2 - x_1\rangle\\ &\quad - \langle \nabla_y f(x_2,y_2) - \nabla_y f(x_1,y_1), y_2 - y_1\rangle\\ &\geq \frac{\mu}{2}(V_{z_1}(z_2) + V_{z_2}(z_1)).\end{aligned}$$

□

C Proof of Lemma 3.2

Lemma. Let $e \in \mathcal{RS}^2(1)$, i.e. uniformly distributed on the unit Euclidean sphere. Randomness comes from independent variables e, ξ and a point z. Norm $\|\cdot\|_* = \|\cdot\|_q$ satisfies $q \in [2; +\infty)$. We introduce the constant ρ_n:

$$\rho_n = \min\{q - 1, 16\log(n) - 8\}.$$

Then under Assumption 3 or 3(f) the following statements hold:

- for Random direction oracle

$$\begin{aligned}\mathbb{E}\left[\|g_d(z,e,\tau,\xi)\|_q^2\right] &\leq 48n^{2/q}\rho_n\mathbb{E}\left[\|F(z) - F(z^*)\|_2^2\right] + 48n^{2/q}\rho_n\|F(z^*)\|_2^2\\ &\quad + 48n^{2/q}\rho_n\sigma^2 + 8n^{2/q+1}\rho_n L^2\tau^2\\ &\quad + 16\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2}, \qquad (18)\\ \|\mathbb{E}[g_d(z,e,\tau,\xi)] - F(z)\|_q &\leq 2n^{1/q+1/2}\sqrt{\rho_n}L\tau + 4n^{1/q+1/2}\sqrt{\rho_n}\frac{\Delta}{\tau}; \qquad (19)\end{aligned}$$

- for Full coordinates oracle

$$\mathbb{E}\left[\|g_f(z,\tau,\xi) - F(z)\|_q^2\right] \leq 3\sigma^2 + 3nL_2^2\tau^2 + \frac{6n\Delta^2}{\tau^2}, \tag{20}$$

$$\|\mathbb{E}\left[g_f(z,\tau,\xi)\right] - F(z)\|_q \leq \sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}. \tag{21}$$

Proof of (18).

$$\begin{aligned}
\mathbb{E}\left[\|g_d(z,e,\tau,\xi)\|_q^2\right] \overset{(11)}{\leq}\ & 4n^2\mathbb{E}\left[\left\|\begin{pmatrix} \langle \nabla_x f(x,y), e_x\rangle e_x \\ \langle -\nabla_y f(x,y), e_y\rangle e_y \end{pmatrix}\right\|_q^2\right] \\
& +4n^2\mathbb{E}\left[\left\|\begin{pmatrix} \langle \nabla_x f(x,y,\xi) - \nabla_x f(x,y), e_x\rangle e_x \\ \langle -\nabla_y f(x,y,\xi) + \nabla_y f(x,y), e_y\rangle e_y \end{pmatrix}\right\|_q^2\right] \\
& +4\frac{n^2}{\tau^2}\mathbb{E}\left[\left\|\begin{pmatrix} (f(x+\tau e_x,y,\xi) - f(x,y,\xi) - \langle \nabla_x f(x,y,\xi), \tau e_x\rangle) e_x \\ (f(x,y,\xi) - f(x,y+\tau e_y,\xi) + \langle \nabla_y f(x,y,\xi), \tau e_y\rangle) e_y \end{pmatrix}\right\|_q^2\right] \\
& +4\frac{n^2}{\tau^2}\mathbb{E}\left[\left\|\begin{pmatrix} (\delta(x+\tau e_x,y) - \delta(x,y)) e_x \\ (\delta(x,y) - \delta(x,y+\tau e_y)) e_y \end{pmatrix}\right\|_q^2\right] \\
\overset{(12)}{\leq}\ & 4n^2\mathbb{E}\left[\|\langle \nabla_x f(x,y), e_x\rangle e_x\|_q^2\right] + 4n^2\mathbb{E}\left[\|\langle -\nabla_y f(x,y), e_y\rangle e_y\|_q^2\right] \\
& +4n^2\mathbb{E}\left[\|\langle \nabla_x f(x,y,\xi) - \nabla_x f(x,y), e_x\rangle e_x\|_q^2\right] \\
& +4n^2\mathbb{E}\left[\|\langle -\nabla_y f(x,y,\xi) + \nabla_y f(x,y), e_y\rangle e_y\|_q^2\right] \\
& +4\frac{n^2}{\tau^2}\mathbb{E}\left[\left\|\left(\tilde{f}(x+\tau e_x,y,\xi) - \tilde{f}(x,y,\xi) - \langle \nabla_x f(x,y,\xi), \tau e_x\rangle\right) e_x\right\|_q^2\right] \\
& +4\frac{n^2}{\tau^2}\mathbb{E}\left[\left\|\left(\tilde{f}(x,y,\xi) - \tilde{f}(x,y+\tau e_y,\xi) + \langle \nabla_y f(x,y,\xi), \tau e_y\rangle\right) e_y\right\|_q^2\right] \\
& +4\frac{n^2}{\tau^2}\mathbb{E}\left[\|(\delta(x+\tau e_x,y) - \delta(x,y)) e_x\|_q^2\right] \\
& +4\frac{n^2}{\tau^2}\mathbb{E}\left[\|(\delta(x,y) - \delta(x,y+\tau e_y)) e_y\|_q^2\right].
\end{aligned}$$

From (8) we get $\|\nabla_x f(x_1,y,\xi) - \nabla_x f(x_2,y,\xi)\|_2 \leq L\|x_1 - x_2\|_2$ and $\|\nabla_y f(x,y_1,\xi) - \nabla_y f(x,y_2,\xi)\|_2 \leq L\|y_1 - y_2\|_2$ for all $x, x_1, x_2 \in \mathcal{X}$, $y, y_1, y_2 \in \mathcal{Y}$. It follows that functions $f(\cdot,y,\xi)$ and $f(x,\cdot,\xi)$ are $L(\xi)$-Lipschitz continuous. Then

$$
\begin{aligned}
\mathbb{E}\left[\|g_d(z,e,\tau,\xi)\|_q^2\right] &\leq 4n^2\mathbb{E}\left[\|\langle\nabla_x f(x,y),\tau e_x\rangle\, e_x\|_q^2\right] + 4n^2\mathbb{E}\left[\|\langle-\nabla_y f(x,y),\tau e_y\rangle\, e_y\|_q^2\right] \\
&\quad +4n^2\mathbb{E}\left[\|\langle\nabla_x f(x,y,\xi)-\nabla_x f(x,y),\tau e_x\rangle\, e_x\|_q^2\right] \\
&\quad +4n^2\mathbb{E}\left[\|\langle-\nabla_y f(x,y,\xi)+\nabla_y f(x,y),\tau e_y\rangle\, e_y\|_q^2\right] \\
&\quad +4n^2L_2^2\tau^2\mathbb{E}\left[\|e_x\|_q^2\right] + 4n^2L_2^2\tau^2\mathbb{E}\left[\|e_y\|_q^2\right] \\
&\quad +8\frac{n^2\Delta^2}{\tau^2}\mathbb{E}\left[\|e_x\|_q^2\right] + 8\frac{n^2\Delta^2}{\tau^2}\mathbb{E}\left[\|e_y\|_q^2\right].
\end{aligned}
$$

In the last inequality, we additionally use (3) + (11) and independence of e and ξ. With (14) and (15), one can get the following result:

$$
\begin{aligned}
\mathbb{E}\left[\|g_d(z,e,\tau,\xi)\|_q^2\right] &\leq 24n^{2/q}\rho_n\mathbb{E}\left[\|\nabla_x f(x,y)\|_2^2\right] + 24n^{2/q}\rho_n\mathbb{E}\left[\|-\nabla_y f(x,y)\|_2^2\right] \\
&\quad +24n^{2/q}\rho_n\mathbb{E}\left[\|\nabla_x f(x,y,\xi)-\nabla_x f(x,y)\|_2^2\right] \\
&\quad +24n^{2/q}\rho_n\mathbb{E}\left[\|-\nabla_y f(x,y,\xi)+\nabla_y f(x,y)\|_2^2\right] \\
&\quad +8n^{2/q+1}\rho_n L^2\tau^2 + 16\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2} \\
&\overset{(3)}{\leq} 24n^{2/q}\rho_n\mathbb{E}\left[\|F(z)\|_2^2\right] + 48n^{2/q}\rho_n\sigma^2 + 8n^{2/q+1}\rho_n L_2^2\tau^2 + 16\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2} \\
&\overset{(11)}{\leq} 48n^{2/q}\rho_n\mathbb{E}\left[\|F(z)-F(z^*)\|_2^2\right] + 48n^{2/q}\rho_n\|F(z^*)\|_2^2 \\
&\quad +48n^{2/q}\rho_n\sigma^2 + 8n^{2/q+1}\rho_n L_2^2\tau^2 + 16\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2}.
\end{aligned}
$$

Proof of (19) .

$$
\begin{aligned}
\|\mathbb{E}[g_d(z,e,\tau,\xi)]-F(z)\|_q &\leq \frac{n}{\tau}\left\|\mathbb{E}\left[\begin{pmatrix}(f(x+\tau e_x,y,\xi)-f(x,y,\xi)-\langle\nabla_x f(x,y,\xi),\tau e_x\rangle)\, e_x\\ (f(x,y,\xi)-f(x,y+\tau e_y,\xi)+\langle\nabla_y f(x,y,\xi),\tau e_y\rangle)\, e_y\end{pmatrix}\right]\right\|_q \\
&\quad +n\left\|\mathbb{E}\left[\begin{pmatrix}\langle\nabla_x f(x,y,\xi)-\nabla_x f(x,y),e_x\rangle\, e_x\\ \langle-\nabla_y f(x,y,\xi)+\nabla_y f(x,y),e_y\rangle\, e_y\end{pmatrix}\right]\right\|_q \\
&\quad +n\left\|\mathbb{E}\left[\begin{pmatrix}\langle\nabla_x f(x,y),e_x\rangle\, e_x\\ \langle-\nabla_y f(x,y),e_y\rangle\, e_y\end{pmatrix}\right]-\begin{pmatrix}\nabla_x f(x,y)\\ -\nabla_y f(x,y)\end{pmatrix}\right\|_q \\
&\quad +\frac{n}{\tau}\left\|\mathbb{E}\left[\begin{pmatrix}(\delta(x+\tau e_x,y)-\delta(x,y))\, e_x\\ (\delta(x,y)-\delta(x,y+\tau e_y))\, e_y\end{pmatrix}\right]\right\|_q.
\end{aligned}
$$

Taking into account the independence of e and ξ, as well as using their unbiasedness, we get

$$\begin{aligned}
\|\mathbb{E}[g_d(z,e,\tau,\xi)] - F(z)\|_q &\leq \frac{n}{\tau}\left\|\mathbb{E}\left[\begin{pmatrix} (f(x+\tau e_x,y) - f(x,y) - \langle \nabla_x f(x,y), \tau e_x\rangle)\, e_x \\ (f(x,y) - f(x,y+\tau e_y) + \langle \nabla_y f(x,y), \tau e_y\rangle)\, e_y \end{pmatrix}\right]\right\|_q \\
&\quad + \frac{n}{\tau}\left\|\mathbb{E}\left[\begin{pmatrix} (\delta(x+\tau e_x,y) - \delta(x,y))\, e_x \\ (\delta(x,y) - \delta(x,y+\tau e_y))\, e_y \end{pmatrix}\right]\right\|_q \\
&\overset{(12)}{\leq} \frac{n}{\tau}\left\|\mathbb{E}\left[(f(x+\tau e_x,y) - f(x,y) - \langle \nabla_x f(x,y), \tau e_x\rangle)\, e_x\right]\right\|_q \\
&\quad + \frac{n}{\tau}\left\|\mathbb{E}\left[(f(x,y) - f(x,y+\tau e_y) + \langle \nabla_y f(x,y), \tau e_y\rangle)\, e_y\right]\right\|_q \\
&\quad + \frac{n}{\tau}\left\|\mathbb{E}\left[(\delta(x+\tau e_x,y) - \delta(x,y))\, e_x\right]\right\|_q \\
&\quad + \frac{n}{\tau}\left\|\mathbb{E}\left[(\delta(x,y) - \delta(x,y+\tau e_y))\, e_y\right]\right\|_q.
\end{aligned}$$

Further, Jensen inequality gives

$$\begin{aligned}
\|\mathbb{E}[g_d(z,e,\tau,\xi)] - F(z)\|_q &\leq \frac{n}{\tau}\mathbb{E}\left[\left|f(x+\tau e_x,y) - f(x,y) - \langle \nabla_x f(x,y), \tau e_x\rangle\right| \|e_x\|_q\right] \\
&\quad + \frac{n}{\tau}\mathbb{E}\left[\left|f(x,y) - f(x,y+\tau e_y) + \langle \nabla_y f(x,y), \tau e_y\rangle\right| \|e_y\|_q\right] \\
&\quad + \frac{n}{\tau}\mathbb{E}\left[\left|\delta(x+\tau e_x,y) - \delta(x,y)\right| \|e_x\|_q\right] \\
&\quad + \frac{n}{\tau}\mathbb{E}\left[\left|\delta(x,y) - \delta(x,y+\tau e_y)\right| \|e_y\|_q\right].
\end{aligned}$$

It remains to use L-Lipschitz continuous of $f(\cdot,y)$ and $f(x,\cdot)$:

$$\begin{aligned}
\|\mathbb{E}[g_d(z,e,\tau,\xi)] - F(z)\|_q &\leq nL\tau\mathbb{E}\left[\|e_x\|_q\right] + nL\tau\mathbb{E}\left[\|e_y\|_q\right] \\
&\quad + \frac{n}{\tau}\mathbb{E}\left[\left(|\delta(x+\tau e_x,y)| + |\delta(x,y)|\right)\|e_x\|_q\right] \\
&\quad + \frac{n}{\tau}\mathbb{E}\left[\left(|\delta(x,y)| + |\delta(x,y+\tau e_y)|\right)\|e_y\|_q\right] \\
&\overset{(3),(14)}{\leq} 2n^{1/q+1/2}\sqrt{\rho_n}L\tau + 4n^{1/q+1/2}\sqrt{\rho_n}\frac{\Delta}{\tau}.
\end{aligned}$$

Proof of (20).

$$\begin{aligned}
\mathbb{E}\left[\|g_f(z,\tau,\xi)-F(z)\|_q^2\right] &\overset{(5),(11)}{\leq} 3\mathbb{E}\Bigg[\Bigg\|\frac{1}{\tau}\sum_{i=1}^{n_x}\left(f(z+\tau h_i,\xi)-f(z,\xi)\right)h_i \\
&\quad +\frac{1}{\tau}\sum_{i=n_x+1}^{n_x+n_y}\left(f(z,\xi)-f(z+\tau h_i,\xi)\right)h_i - F(z,\xi)\Bigg\|_2^2\Bigg] \\
&\quad +3\mathbb{E}\left[\|F(z,\xi)-F(z)\|_2^2\right] \\
&\quad +3\mathbb{E}\left[\left\|\sum_{i=1}^{n_x+n_y}\frac{(\delta(z+\tau h_i)-\delta(z))}{\tau}h_i\right\|_2^2\right] \\
&\overset{(3),(11)}{\leq} 3\mathbb{E}\left[\sum_{i=1}^{n_x+n_y}\left|\frac{(f(z+\tau h_i,\xi)-f(z,\xi))}{\tau}-\frac{\partial f(z,\xi)}{\partial z_i}\right|^2\right] \\
&\quad +3\sigma^2+6\frac{n\Delta^2}{\tau^2}.
\end{aligned}$$

By the mean value theorem we have that for some $|q_i|\leq|\tau|$:

$$\begin{aligned}
\mathbb{E}\left[\|g_f(z,\tau,\xi)-F(z)\|_*^2\right] &\leq 3\mathbb{E}\left[\sum_{i=1}^{n}\left|\frac{\partial f(z+q_ih_i,\xi)}{\partial z_i}-\frac{\partial f(z,\xi)}{\partial z_i}\right|^2\right] \\
&\quad +3\sigma^2+6\frac{n\Delta^2}{\tau^2} \\
&\leq 3\sum_{i=1}^{n}L_2^2q_i^2+3\sigma^2+6\frac{n\Delta^2}{\tau^2} \\
&\leq 3nL_2^2\tau^2+3\sigma^2+6\frac{n\Delta^2}{\tau^2}.
\end{aligned}$$

Proof of (21). Using unbiasedness of ξ:

$$\begin{aligned}
\left\|\mathbb{E}\left[g_f(z,\tau,\xi)\right]-F(z)\right\|_q &\leq \Bigg\|\frac{1}{\tau}\sum_{i=1}^{n_x}\left(f(z+\tau h_i)-f(z)\right)h_i \\
&\quad +\frac{1}{\tau}\sum_{i=n_x+1}^{n_x+n_y}\left(f(z)-f(z+\tau h_i,)\right)h_i - F(z)\Bigg\|_2 \\
&\quad +\left\|\sum_{i=1}^{n_x+n_y}\frac{(\delta(z+\tau h_i)-\delta(z))}{\tau}h_i\right\|_2 \\
&\overset{(3)}{\leq} \sqrt{\sum_{i=1}^{n}L^2q_i^2}+\frac{2\sqrt{n}\Delta}{\tau} \\
&\leq \sqrt{n}L\tau+\frac{2\sqrt{n}\Delta}{\tau}.
\end{aligned}$$

□

D Proof of Theorem 1

Lemma D.1. *Let $z, g \in \mathbb{R}^n$ and $\mathcal{Z} \subset \mathbb{R}^n$. Then for $z_1 = prox_z(g)$ and for all $u \in \mathcal{Z}$ we have*

$$\langle g, z - u\rangle \leq V_z(u) - V_{z_1}(u) + \frac{1}{2}\|g\|_q^2 \tag{22}$$

Proof. By (13), we have for all $u \in \mathcal{Z}$

$$\langle g, z_1 - u\rangle = \langle g, z_1 - z + z - u\rangle \leq V_z(u) - V_{z_1}(u) - V_z(z_1).$$

Making simple transformations:

$$\begin{aligned}\langle g, z - u\rangle &\leq \langle g, z - z_1\rangle + V_z(u) - V_{z_1}(u) - V_z(z_1)\\ &\leq \langle g, z - z_1\rangle + V_z(u) - V_{z_1}(u) - \frac{1}{2}\|z_1 - z\|_p^2.\end{aligned}$$

In last inequality we use the property of the Bregman divergence: $V_x(y) \geq \frac{1}{2}\|x - y\|_p^2$. With Hölder's inequality and the fact: $ab - {}^{b^2}/_2 \leqslant {}^{a^2}/_2$, we get

$$\begin{aligned}\langle g, z - u\rangle &\leq \|g\|_q\|z - z_1\|_p + V_z(u) - V_{z_1}(u) - \frac{1}{2}\|z_1 - z\|_p^2\\ &\leq V_z(u) - V_{z_1}(u) + \frac{1}{2}\|g\|_q^2.\end{aligned}$$

□

Theorem. By Algorithm 1 with Random direction oracle

- under Assumptions 1, 2, 3(f) and with $\gamma \leq \frac{1}{48n^{2/q}\rho_n L}$, we get

$$\begin{aligned}\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\left[\|F(z_k) - F(z^*)\|_2^2\right] &\leq \frac{2LD_p^2}{\gamma N} + 48\gamma n^{2/q}\rho_n L\left(\|F(z^*)\|_2^2 + \sigma^2\right)\\ &\quad +8\gamma n^{2/q+1}\rho_n L\left(L_2^2\tau^2 + 2\frac{\Delta^2}{\tau^2}\right)\\ &\quad +8n^{1/q+1/2}\sqrt{\rho_n}LD_p\left(L\tau + \frac{2\Delta}{\tau}\right);\end{aligned} \tag{23}$$

- under Assumptions 1, 2(s), 3 and with $\gamma \leq \frac{\mu}{96n^{2/q}\rho_n L^2}$:

$$\begin{aligned}\mathbb{E}\left[V_{z_{N+1}}(z^*)\right] &\leq V_{z_0}(z^*)\exp\left(-\frac{\mu^2 N}{400n^{2/q}\rho_n L^2}\right) +\\ &\quad +\frac{24n^{2/q}\rho_n}{\mu^2 N}\left(\|F(z^*)\|_2^2 + \sigma^2\right) &(24)\\ &\quad +\frac{4n^{2/q+1}\rho_n}{\mu^2 N}\left(L_2^2\tau^2 + 2\frac{\Delta^2}{\tau^2}\right) &(25)\\ &\quad +\frac{4n^{1/q+1/2}\sqrt{\rho_n}D_p}{\gamma\mu^2 N}\left(L\tau + \frac{2\Delta}{\tau}\right). &(26)\end{aligned}$$

□

Proof of (23)**.** We begin with descent lemma (22):

$$\gamma\langle g_d(z_k,e_k,\tau,\xi_k), z_k - u\rangle \;\leq\; V_{z_k}(u) - V_{z_{k+1}}(u) + \frac{\gamma^2}{2}\|g_d(z_k,e_k,\tau,\xi_k)\|_q^2.$$

Taking $u = z^*$ and using convexity - concavity of $f(x,y)$ in form $\langle F(z^*), z_k - z^*\rangle \geq 0$, we get

$$\begin{aligned}\gamma\langle F(z_k) - F(z^*), z_k \;-\; u\rangle &\leq V_{z_k}(z^*) - V_{z_{k+1}}(z^*)\\ &+ \gamma\langle F(z_k) - g_d(z_k,e_k,\tau,\xi_k), z_k - z^*\rangle + \frac{\gamma^2}{2}\|g_d(z_k,e_k,\tau,\xi_k)\|_q^2.\end{aligned}$$

With (9), this gives

$$\begin{aligned}\frac{\gamma}{L}\|F(z_k) - F(z^*)\|_2^2 \;\leq\; & V_{z_k}(z^*) - V_{z_{k+1}}(z^*)\\ &+\gamma\langle F(z_k) - g_d(z_k,e_k,\tau,\xi_k), z_k - u\rangle + \frac{\gamma^2}{2}\|g_d(z_k,e_k,\tau,\xi_k)\|_q^2.\end{aligned}$$

Taking full expectation and using Hölder's inequality, (18), (19), we have

$$\begin{aligned}\frac{\gamma}{L}\mathbb{E}\left[\|F(z_k) - F(z^*)\|_2^2\right] \;\leq\; & \mathbb{E}\left[V_{z_k}(z^*)\right] - \mathbb{E}\left[V_{z_{k+1}}(u)\right]\\ &+2\gamma\left(2n^{1/q+1/2}\sqrt{\rho_n}L\tau + 4n^{1/q+1/2}\sqrt{\rho_n}\frac{\Delta}{\tau}\right)D_p\\ &+\frac{\gamma^2}{2}\left(48n^{2/q}\rho_n\mathbb{E}\left[\|F(z_k) - F(z^*)\|_2^2\right] + 48n^{2/q}\rho_n\|F(z^*)\|_2^2\right)\\ &+\frac{\gamma^2}{2}\left(48n^{2/q}\rho_n\sigma^2 + 8n^{2/q+1}\rho_n L_2^2\tau^2 + 16\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2}\right).\end{aligned}$$

$\gamma \leq 1/48n^{\frac{2}{q}}\rho_n L$ gives

$$\begin{aligned}\frac{\gamma}{2L}\mathbb{E}\left[\|F(z_k) - F(z^*)\|_2^2\right] \;\leq\; & \mathbb{E}\left[V_{z_k}(z^*)\right] - \mathbb{E}\left[V_{z_{k+1}}(z^*)\right]\\ &+2\gamma\left(2n^{1/q+1/2}\sqrt{\rho_n}L\tau + 4n^{1/q+1/2}\sqrt{\rho_n}\frac{\Delta}{\tau}\right)D_p\\ &+\frac{\gamma^2}{2}\left(48n^{2/q}\rho_n\|F(z^*)\|_2^2 + 48n^{2/q}\rho_n\sigma^2\right)\\ &+\frac{\gamma^2}{2}\left(8n^{2/q+1}\rho_n L_2^2\tau^2 + 16\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2}\right).\end{aligned}$$

It remains to sum up from $k=0$ to $k=N$:

$$\begin{aligned}\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\left[\|F(z_k) - F(z^*)\|_2^2\right] \;\leq\; & \frac{2LD_p^2}{\gamma N} + 48\gamma n^{2/q}\rho_n L\left(\|F(z^*)\|_2^2 + \sigma^2\right)\\ &+8\gamma n^{2/q+1}\rho_n L\left(L_2^2\tau^2 + 2\frac{\Delta^2}{\tau^2}\right)\\ &+8n^{1/q+1/2}\sqrt{\rho_n}LD_p\left(L\tau + \frac{2\Delta}{\tau}\right).\end{aligned}$$

□

Proof of (24)**.** Similarly to the previous proof, we begin with descent lemma (22):

$$\gamma\langle g(z_k,e_k,\tau,\xi_k),z_k-u\rangle \;\leq\; V_{z_k}(u)-V_{z_{k+1}}(u)+\frac{\gamma^2}{2}\|g_d(z_k,e_k,\tau,\xi_k)\|_q^2.$$

Taking $u=z^*$ and using $\langle F(z^*),z_k-z^*\rangle\geq 0$, we get:

$$\begin{aligned}\gamma\langle F(z_k)-F(z^*),z_k &- z^*\rangle \leq V_{z_k}(z^*)-V_{z_{k+1}}(z^*)\\ &+ \gamma\langle F(z_k)-g_d(z_k,e_k,\tau,\xi_k),z_k-u\rangle+\frac{\gamma^2}{2}\|g(z_k,e_k,\tau,\xi_k)\|_q^2.\end{aligned}$$

With (17), it gives

$$\begin{aligned}\frac{\gamma\mu}{2}V_{z_k}(z^*) \;\leq\;& V_{z_k}(z^*)-V_{z_{k+1}}(z^*)\\ &+\gamma\langle F(z_k)-g_d(z_k,e_k,\tau,\xi_k),z_k-u\rangle+\frac{\gamma^2}{2}\|g_d(z_k,e_k,\tau,\xi_k)\|_q^2.\end{aligned}$$

Taking full expectation and using (18), (19), we have

$$\begin{aligned}\mathbb{E}\left[V_{z_{k+1}}(z^*)\right] \;\leq\;& \left(1-\frac{\gamma\mu}{2}\right)\mathbb{E}\left[V_{z_k}(z^*)\right]+2\gamma\left(2n^{1/q+1/2}\sqrt{\rho_n}L\tau+4n^{1/q+1/2}\sqrt{\rho_n}\frac{\Delta}{\tau}\right)D_p\\ &+\frac{\gamma^2}{2}\left(48n^{2/q}\rho_n\mathbb{E}\left[\|F(z_k)-F(z^*)\|_2^2\right]+48n^{2/q}\rho_n\|F(z^*)\|_2^2\right)\\ &+\frac{\gamma^2}{2}\left(48n^{2/q}\rho_n\sigma^2+8n^{2/q+1}\rho_nL_2^2\tau^2+16\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2}\right).\end{aligned}$$

Using (8) and assuming $\gamma\leq {}^{\mu}\!/\!\left(96n^{2/q}\rho_nL^2\right)$:

$$\begin{aligned}\mathbb{E}\left[V_{z_{k+1}}(z^*)\right] \;\leq\;& \left(1-\frac{\gamma\mu}{4}\right)\mathbb{E}\left[V_{z_k}(z^*)\right]+2\gamma^2\left(\frac{2n^{1/q+1/2}\sqrt{\rho_n}L\tau}{\gamma}+\frac{4n^{1/q+1/2}\sqrt{\rho_n}\Delta}{\gamma\tau}\right)D_p\\ &+\gamma^2\left(24n^{2/q}\rho_n\|F(z^*)\|_2^2+24n^{2/q}\rho_n\sigma^2\right)\\ &+\gamma^2\left(4n^{2/q+1}\rho_nL_2^2\tau^2+8\frac{n^{2/q+1}\rho_n\Delta^2}{\tau^2}\right).\end{aligned}$$

It remains to use (16) and get

$$\begin{aligned}\mathbb{E}\left[V_{z_{N+1}}(z^*)\right] \;\leq\;& V_{z_0}(z^*)\exp\left(-\frac{\mu^2N}{400n^{2/q}\rho_nL^2}\right)+\\ &+\frac{24n^{2/q}\rho_n}{\mu^2N}\left(\|F(z^*)\|_2^2+\sigma^2\right)\\ &+\frac{4n^{2/q+1}\rho_n}{\mu^2N}\left(L_2^2\tau^2+2\frac{\Delta^2}{\tau^2}\right)\\ &+\frac{4n^{1/q+1/2}\sqrt{\rho_n}D_p}{\gamma\mu^2N}\left(L\tau+\frac{2\Delta}{\tau}\right).\end{aligned}$$

□

E Proof of Theorem 2

Lemma E.1. *Let $z, g, g_{1/2} \in \mathbb{R}^n$ and $\mathcal{Z} \subset \mathbb{R}^n$. Then for $z_{1/2} = \text{prox}_z(g)$ and $z_1 = \text{prox}_z(g_{1/2})$ and for all $u \in \mathcal{Z}$ we have*

$$\langle g_{1/2}, z_{1/2} - u\rangle \leq V_z(u) - V_{z_1}(u) + \frac{1}{2}\|g - g_{1/2}\|_q^2 - V_z(z_{1/2}). \tag{27}$$

Proof. Using (13) with $z = z$, $z_+ = z_1$, $w = g_{1/2}$, $u = u$ and with $z = z$, $z_+ = z_{1/2}$, $w = g$, $u = z_1$:

$$\begin{aligned}\langle g_{1/2}, z_1 - u\rangle &\leq V_z(u) - V_{z_1}(u) - V_z(z_1),\\ \langle g, z_{1/2} - z_1\rangle &\leq V_z(z_1) - V_{z_{1/2}}(z_1) - V_z(z_{1/2}).\end{aligned}$$

By summing these two inequalities, we get

$$\begin{aligned}\langle g_{1/2}, z_{1/2} - u\rangle \leq\ & V_z(u) - V_{z_1}(u) + \langle g - g_{1/2}, z_1 - z_{1/2}\rangle\\ & -V_{z_{1/2}}(z_1) - V_z(z_{1/2}).\end{aligned}$$

Applying Cauchy-Schwartz inequality and property: $V_{z_{1/2}}(z_1) \geq 1/2\|z_{1/2} - z_1\|^2$, we have

$$\langle g_{1/2}, z_{1/2} - u\rangle \leq V_z(u) - V_{z_1}(u) + \frac{1}{2}\|g - g_{1/2}\|_q^2 - V_z(z_{1/2}).$$

□

Theorem.

- By Algorithm 2 with Full coordinates oracle under Assumptions 1, 2, 3 and with $\gamma \leq 1/2L$, we have

$$\begin{aligned}\mathbb{E}\left[\varepsilon_{sad}(\bar{z}_N)\right] \leq\ & \frac{2D_p^2}{\gamma(N+1)} + 11\gamma\left(nL_2^2\tau^2 + \sigma^2 + 2\frac{n\Delta^2}{\tau^2}\right)\\ & +2D_p\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right),\end{aligned} \tag{28}$$

  where

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

  $\bar{x}_N$, $\bar{y}_N$ are defined the same way as $\bar{z}_N$.

- By Algorithm 3 with Full coordinates oracle under Assumptions 1, 2(s), 3 and with $p = 2$ ($V_x(y) = 1/2\|x - y\|_2^2$), $\gamma \leq 1/6L$:

$$\begin{aligned}\mathbb{E}\left[\|z_{N+1} - z^*\|_2^2\right] \leq\ & \exp\left(-\frac{\mu N}{12L}\right)\left(\|z_0 - z^*\|_2^2 + \|g_f(z_0,\tau,\xi_0) - g_f(z_0,\tau,\xi_0)\|_2^2\right)\\ & +\frac{1}{\mu^2 N}12\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right)\\ & +\frac{1}{\mu^2 N}\frac{4D_2}{\gamma}\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right).\end{aligned} \tag{29}$$

Proof of (28)**.** We begin with (27) and taking $z = z_k, g = \gamma g_f(z_k, e_k, \tau, \xi_k), g_{1/2} = \gamma g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2})$, then $z_{1/2} = z_{k+1/2}$, $z_1 = z_{k+1}$ and have

$$
\begin{aligned}
\gamma\langle g_f(z_{k+1/2}, e_{k+1/2}, \tau, \quad \xi_{k+1/2} \quad), z_{k+1/2} - u\rangle & \\
&\leq \quad V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1/2}) \\
&\quad\; + \frac{\gamma^2}{2}\|g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - g_f(z_k, e_k, \tau, \xi_k)\|_q^2 \\
&\overset{(11)}{\leq} \quad V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1/2}) \\
&\quad\; + \frac{3\gamma^2}{2}\|F(z_{k+1/2}) - F(z_k)\|_q^2 \\
&\quad\; + \frac{3\gamma^2}{2}\|g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})\|_q^2 \\
&\quad\; + \frac{3\gamma^2}{2}\|g_f(z_k, e_k, \tau, \xi_k) - F(z_k)\|_q^2 \\
&\overset{(8)}{\leq} \quad V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1/2}) \\
&\quad\; + \frac{3\gamma^2 L^2}{2}\|z_{k+1/2} - z_k\|_2^2 \\
&\quad\; + \frac{3\gamma^2}{2}\|g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})\|_q^2 \\
&\quad\; + \frac{3\gamma^2}{2}\|g_f(z_k, e_k, \tau, \xi_k) - F(z_k)\|_q^2.
\end{aligned}
$$

Applying the property: $V_{z_k}(z_{k+1/2}) \geq 1/2\|z_{k+1/2} - z_k\|^2 \geq 1/2\|z_{k+1/2} - z_k\|_2^2$, with $\gamma \leq 1/2L$, we get

$$
\begin{aligned}
\gamma\langle g_f(z_{k+1/2}, e_{k+1/2}, \quad \tau, \quad \xi_{k+1/2}), z_{k+1/2} - u\rangle & \\
&\leq \quad V_{z_k}(u) - V_{z_{k+1}}(u) \\
&\quad\; + \frac{3\gamma^2}{2}\|g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})\|_q^2 \\
&\quad\; + \frac{3\gamma^2}{2}\|g_f(z_k, e_k, \tau, \xi_k) - F(z_k)\|_q^2,
\end{aligned}
$$

and

$$
\begin{aligned}
\gamma\langle F(z_{k+1/2}), z_{k+1/2} - u\rangle \quad &\leq \quad V_{z_k}(u) - V_{z_{k+1}}(u) \\
&\quad\; + \gamma\langle F(z_{k+1/2}) - g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}), z_{k+1/2} - u\rangle \\
&\quad\; + \frac{3\gamma^2}{2}\|g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})\|_q^2 \\
&\quad\; + \frac{3\gamma^2}{2}\|g_f(z_k, e_k, \tau, \xi_k) - F(z_k)\|_q^2.
\end{aligned}
$$

Summing over all k from 0 to N, one can have

$$\begin{aligned}\sum_{k=0}^{N}\langle F(z_{k+1/2}), z_{k+1/2}-u\rangle &\leq \frac{V_{z_0}(u)-V_{z_{K+1}}(u)}{\gamma}\\ &+\sum_{k=0}^{N}\langle F(z_{k+1/2})-g_f(z_{k+1/2},e_{k+1/2},\tau,\xi_{k+1/2}), z_{k+1/2}-u\rangle\\ &+\frac{3\gamma}{2}\sum_{k=0}^{N}\|g_f(z_{k+1/2},e_{k+1/2},\tau,\xi_{k+1/2})-F(z_{k+1/2})\|_q^2\\ &+\frac{3\gamma}{2}\sum_{k=0}^{N}\|g_f(z_k,e_k,\tau,\xi_k)-F(z_k)\|_q^2\\ &\leq \frac{D_p^2}{\gamma}+\sum_{k=0}^{N}\langle F(z_{k+1/2})-g_f(z_{k+1/2},e_{k+1/2},\tau,\xi_{k+1/2}), z_{k+1/2}-u\rangle\\ &+\frac{3\gamma}{2}\sum_{k=0}^{N}\|g_f(z_{k+1/2},e_{k+1/2},\tau,\xi_{k+1/2})-F(z_{k+1/2})\|_q^2\\ &+\frac{3\gamma}{2}\sum_{k=0}^{N}\|g_f(z_k,e_k,\tau,\xi_k)-F(z_k)\|_q^2. \end{aligned}\tag{30}$$

Next we need to connect $\sum\limits_{k=0}^{N}\langle F(z_{k+1/2}), z_{k+1/2}-u\rangle$ and $\varepsilon_{sad}(\bar{z}_{N+1})$. By the definition of $\bar{x}_N$ and $\bar{y}_N$, Jensen's inequality and convexity-concavity of f:

$$\begin{aligned}\varepsilon_{sad}(\bar{z}_{N+1}) &\leq \max_{y'\in\mathcal{Y}} f\left(\frac{1}{N+1}\left(\sum_{k=0}^{N}x_{k+1/2}\right), y'\right)-\min_{x'\in\mathcal{X}} f\left(x', \frac{1}{N+1}\left(\sum_{k=0}^{N}y_{k+1/2}\right)\right)\\ &\leq \max_{y'\in\mathcal{Y}}\frac{1}{N+1}\sum_{k=0}^{N}f(x_{k+1/2},y')-\min_{x'\in\mathcal{X}}\frac{1}{N+1}\sum_{k=0}^{N}f(x',y_{k+1/2}).\end{aligned}$$

Given the fact of linear independence of x' and y':

$$\varepsilon_{sad}(\bar{z}_N) \leq \max_{(x',y')\in\mathcal{Z}}\frac{1}{N+1}\sum_{k=0}^{N}\left(f(x_{k+1/2},y')-f(x',y_{k+1/2})\right).$$

Using convexity and concavity of the function f:

$$\begin{aligned}\varepsilon_{sad}(\bar{z}_N) &\leq \max_{(x',y')\in\mathcal{Z}}\frac{1}{N+1}\sum_{k=0}^{N}\left(f(x_{k+1/2},y')-f(x',y_{k+1/2})\right)\\ &= \max_{(x',y')\in\mathcal{Z}}\frac{1}{N+1}\sum_{k=0}^{N}\left(f(x_{k+1/2},y')-f(x_{k+1/2},y_{k+1/2})+f(x_{k+1/2},y_{k+1/2})-f(x',y_{k+1/2})\right)\\ &\leq \max_{(x',y')\in\mathcal{Z}}\frac{1}{N+1}\sum_{k=0}^{N}\left(\langle\nabla_y f(x_{k+1/2},y_{k+1/2}),y'-y_k\rangle+\langle\nabla_x f(x_{k+1/2},y_{k+1/2}),x_k-x'\rangle\right)\\ &\leq \max_{u\in\mathcal{Z}}\frac{1}{N+1}\sum_{k=0}^{N}\langle F(z_{k+1/2}), z_{k+1/2}-u\rangle.\end{aligned}\tag{31}$$

(31) together with (30) gives

$$\begin{aligned}\varepsilon_{sad}(\bar{z}_N) \;\leq\; & \frac{D_p^2}{\gamma(N+1)} + \frac{1}{N+1}\max_{u\in\mathcal{Z}}\left[\sum_{k=0}^{N}\langle F(z_{k+1/2}) - g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}), z_{k+1/2} - u\rangle\right] \\ & + \frac{3\gamma}{2(N+1)}\sum_{k=0}^{N}\|g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})\|_q^2 \\ & + \frac{3\gamma}{2(N+1)}\sum_{k=0}^{N}\|g_f(z_k, e_k, \tau, \xi_k) - F(z_k)\|_q^2.\end{aligned}$$

Taking the full expectation and using (20) with (6):

$$\begin{aligned}\mathbb{E}\left[\varepsilon_{sad}(\bar{z}_N)\right] \;\leq\; & \frac{D_p^2}{\gamma(N+1)} + \frac{1}{N+1}\mathbb{E}\left[\max_{u\in\mathcal{Z}}\left[\sum_{k=0}^{N}\langle F(z_{k+1/2}) - g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}), z_{k+1/2} - u\rangle\right]\right] \\ & + 9\gamma\left(nL_2^2\tau^2 + \sigma^2 + 2\frac{n\Delta^2}{\tau^2}\right). \qquad (32)\end{aligned}$$

To finish the proof it remains to estimate
$\mathbb{E}\left[\max\limits_{u\in\mathcal{Z}}\left[\sum\limits_{k=0}^{N}\langle F(z_{k+1/2}) - g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}), z_{k+1/2} - u\rangle\right]\right]$. Let define sequence v: $v_0 \overset{\text{def}}{=} z_{1/2}$, $v_{k+1} \overset{\text{def}}{=} \text{prox}_{v_k}(-\gamma\delta_k)$ with
$\delta_k = g_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})$:

$$\sum_{k=0}^{N}\langle -\delta_k, z_{k+1/2} - u\rangle = \sum_{k=0}^{N}\langle -\delta_k, z_{k+1/2} - v_k\rangle + \sum_{k=0}^{N}\langle -\delta_k, v_k - u\rangle. \qquad (33)$$

By the definition of v and an optimal condition for the prox-operator, we have for all $u \in \mathcal{Z}$

$$\langle -\gamma\delta_k - \nabla d(v_{k+1}) + \nabla d(v_{k+1}), u - v_{k+1}\rangle \geq 0.$$

Rewriting this inequality, we get

$$\begin{aligned}\langle -\gamma\delta_k, v_k - u\rangle \;&\leq\; \langle -\gamma\delta_k, v_k - v_{k+1}\rangle + \langle \nabla d(v_{k+1}) - \nabla d(v_k), u - v_{k+1}\rangle \\ &\leq\; \langle -\gamma\delta_k, v_k - v_{k+1}\rangle + V_{v_k}(u) - V_{v_{k+1}}(u) - V_{v_k}(v_{k+1}).\end{aligned}$$

Bearing in mind the Bregman divergence property $2V_x(y) \geq \|x - y\|_p^2$:

$$\langle -\gamma\delta_k, v_k - u\rangle \leq \langle -\gamma\delta_k, v_k - v_{k+1}\rangle + V_{v_k}(u) - V_{v_{k+1}}(u) - \frac{1}{2}\|v_{k+1} - v_k\|_p^2.$$

Using the definition of the conjugate norm:

$$\begin{aligned}\langle -\gamma\delta_k, v_k - u\rangle \;&\leq\; \|\gamma\delta_k\|_q \cdot \|v_k - v_{k+1}\|_p + V_{v_k}(u) - V_{v_{k+1}}(u) - \frac{1}{2}\|v_{k+1} - v_k\|_p^2 \\ &\leq\; \frac{\gamma^2}{2}\|\delta_k\|_q^2 + V_{v_k}(u) - V_{v_{k+1}}(u).\end{aligned}$$

Summing over k from 0 to N:

$$\begin{aligned}\gamma\sum_{k=0}^{N}\langle -\delta_k, v_k - u\rangle &\leq V_{v_0}(u) - V_{v_{N+1}}(u) + \frac{\gamma^2}{2}\sum_{k=0}^{N}\|\delta_k\|_q^2 \\ &\leq D_p^2 + \frac{\gamma^2}{2}\sum_{k=0}^{N}\|\delta_k\|_q^2. \end{aligned} \tag{34}$$

Substituting (34) into (33):

$$\sum_{k=0}^{N}\langle -\delta_k, z_{k+1/2} - u\rangle = \sum_{k=0}^{N}\langle \delta_k, v_k - z_{k+1/2}\rangle + \frac{D_p^2}{\gamma} + \frac{\gamma}{2}\sum_{k=0}^{N}\|\delta_k\|_q^2.$$

The right side is independent of u, then

$$\max_{u\in\mathcal{Z}}\sum_{k=0}^{N}\langle -\delta_k, z_{k+1/2} - u\rangle \leq \sum_{k=0}^{N}\langle \delta_k, v_k - z_{k+1/2}\rangle + \frac{D_p^2}{\gamma} + \frac{\gamma^2}{2}\sum_{k=0}^{N}\|\delta_k\|_q^2.$$

Taking the full expectation with independence $v_k - z_{k+1/2}, \xi_{k+1/2}, e_{k+1/2}$ and using (20), (21), we get

$$\begin{aligned}\mathbb{E}\left[\max_{u\in\mathcal{Z}}\sum_{k=0}^{N}\langle -\delta_k, z_{k+1/2} - u\rangle\right] &\leq \mathbb{E}\left[\sum_{k=0}^{N}\langle \delta_k, v_k - z_{k+1/2}\rangle\right] + \frac{D_p^2}{\gamma} + \frac{\gamma}{2}\sum_{k=0}^{N}\mathbb{E}\left[\|\delta_k\|_q^2\right] \\ &\leq \mathbb{E}\left[\sum_{k=0}^{N}\langle \mathbb{E}_{e_{k+1/2},\xi_{k+1/2}}\left[\delta_k\right], v_k - z_{k+1/2}\rangle\right] + \frac{D_p^2}{\gamma} + \frac{\gamma}{2}\sum_{k=0}^{N}\mathbb{E}\left[\|\delta_k\|_q^2\right] \\ &\leq 2(N+1)D_p\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right) + \frac{D_p^2}{\gamma} \\ &\quad + \frac{3\gamma(N+1)}{2}\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right). \end{aligned} \tag{35}$$

Connecting (32) and (35), we have

$$\begin{aligned}\mathbb{E}\left[\varepsilon_{sad}(\bar{z}_N)\right] &\leq \frac{2D_p^2}{\gamma(N+1)} + 11\gamma\left(nL_2^2\tau^2 + \sigma^2 + 2\frac{n\Delta^2}{\tau^2}\right) \\ &\quad + 2D_p\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right). \end{aligned}$$

Proof of (29)**.** Similarly to the previous proof, let begin with (27) and take full expectation:

$$\begin{aligned}\mathbb{E}\left[\|z_{k+1} - z^*\|_2^2\right] &\leq \mathbb{E}\left[\|z_k - z^*\|_2^2\right] - 2\gamma\mathbb{E}\left[\langle g_f(z_{k+1/2}, \tau, \xi_{k+1/2}), z_{k+1/2} - z^*\rangle\right] \\ &\quad + \gamma^2\mathbb{E}\left[\|g_f(z_{k+1/2}, \tau, \xi_{k+1/2}) - g_f(z_{k-1/2}, \tau, \xi_{k-1/2})\|_2^2\right] \\ &\quad - \mathbb{E}\left[\|z_{k+1/2} - z_k\|_2^2\right]. \end{aligned} \tag{36}$$

Next we work with $\mathbb{E}\left[\|g_f(z_{k+1/2},\tau,\xi_{k+1/2}) - g_f(z_{k-1/2},\tau,\xi_{k-1/2})\|_2^2\right]$:

$$
\begin{aligned}
\mathbb{E}\big[\|g_f(z_{k+1/2},\tau,\xi_{k+1/2}) - g_f \quad (z_{k-1/2}, \quad \tau,\xi_{k-1/2})\|_2^2\big] & \\
&\overset{(11)}{\leq} \quad 3\mathbb{E}\left[\|g_f(z_{k+1/2},\tau,\xi_{k+1/2}) - F(z_{k+1/2})\|_2^2\right] \\
&\qquad +3\mathbb{E}\left[\|g_f(z_{k-1/2},\tau,\xi_{k+1/2}) - F(z_{k-1/2})\|_2^2\right] \\
&\qquad +3\mathbb{E}\left[\|F(z_{k+1/2}) - F(z_{k-1/2})\|_2^2\right] \\
&\overset{(20),(8)}{\leq} \quad 3L^2\mathbb{E}\left[\|z_{k+1/2} - z_{k-1/2}\|_2^2\right] + 6\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right) \\
&\overset{(11)}{\leq} \quad 6L^2\mathbb{E}\left[\|z_{k+1/2} - z_k\|_2^2\right] + 6L^2\mathbb{E}\left[\|z_k - z_{k-1/2}\|_2^2\right] \\
&\qquad +6\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right) \\
&\leq \quad 6L^2\mathbb{E}\left[\|z_{k+1/2} - z_k\|_2^2\right] \\
&\qquad +6\gamma^2L^2\mathbb{E}\left[\|g_f(z_{k-1/2},\tau,\xi_{k-1/2}) - g_f(z_{k-3/2},\tau,\xi_{k-3/2})\|_2^2\right] \\
&\qquad +6\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right).
\end{aligned}
$$

In last inequality we use non-expansiveness of Euclidean prox operator. By simple transformation:

$$
\begin{aligned}
\mathbb{E}\big[\|g_f(z_{k+1/2}, \quad \tau, \quad \xi_{k+1/2}) - g_f(z_{k-1/2},\tau,\xi_{k-1/2})\|_2^2\big] & \\
&\leq \quad 12L^2\mathbb{E}\left[\|z_{k+1/2} - z_k\|_2^2\right] \\
&\qquad +12\gamma^2L^2\mathbb{E}\left[\|g_f(z_{k-1/2},\tau,\xi_{k-1/2}) - g_f(z_{k-3/2},\tau,\xi_{k-3/2})\|_2^2\right] \\
&\qquad -\mathbb{E}\big[\|g_f(z_{k+1/2},\tau,\xi_{k+1/2}) - g_f(z_{k-1/2},\tau,\xi_{k-1/2})\|_2^2\big] \\
&\qquad +12\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right).
\end{aligned}
$$

If $\gamma \leq 1/6L$, then $12\gamma^2L^2 \leq 1 - \mu\gamma$, and we can rewrite previous inequality:

$$
\begin{aligned}
\mathbb{E}\big[\|g_f(z_{k+1/2}, \quad \tau, \quad \xi_{k+1/2}) - g_f(z_{k-1/2},\tau,\xi_{k-1/2})\|_2^2\big] & \\
&\leq \quad 12L^2\mathbb{E}\left[\|z_{k+1/2} - z_k\|_2^2\right] \\
&\qquad +(1-\mu\gamma)\mathbb{E}\left[\|g_f(z_{k-1/2},\tau,\xi_{k-1/2}) - g_f(z_{k-3/2},\tau,\xi_{k-3/2})\|_2^2\right] \\
&\qquad -\mathbb{E}\big[\|g_f(z_{k+1/2},\tau,\xi_{k+1/2}) - g_f(z_{k-1/2},\tau,\xi_{k-1/2})\|_2^2\big] \\
&\qquad +12\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right). \qquad (37)
\end{aligned}
$$

Next we consider $-2\gamma\mathbb{E}\left[\langle g_f(z_{k+1/2},\tau,\xi_{k+1/2}), z_{k+1/2}-z^*\rangle\right]$:

$$\begin{aligned}
-2\gamma\mathbb{E}\big[\langle g_f(z_{k+1/2}, &\tau, &\xi_{k+1/2}), z_{k+1/2}-z^*\rangle\big] \\
&= -2\gamma\mathbb{E}\left[\langle F(z_{k+1/2}), z_{k+1/2}-z^*\rangle\right] \\
&\quad +2\gamma\mathbb{E}\left[\langle F(z_{k+1/2}) - g_f(z_{k+1/2},\tau,\xi_{k+1/2}), z_{k+1/2}-z^*\rangle\right] \\
&\le -2\gamma\mathbb{E}\left[\langle F(z_{k+1/2}), z_{k+1/2}-z^*\rangle\right] \\
&\quad +4\gamma\|\mathbb{E}\left[F(z_{k+1/2}) - g_f(z_{k+1/2},\tau,\xi_{k+1/2})\right]\|_2 D_2 \\
&\overset{(21)}{\le} -2\gamma\mathbb{E}\left[\langle F(z_{k+1/2}), z_{k+1/2}-z^*\rangle\right] \\
&\quad +4\gamma\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right) D_2 \\
&\overset{(17)}{\le} -2\gamma\mu\mathbb{E}\left[\|z_{k+1/2}-z^*\|_2^2\right] \\
&\quad +4\gamma\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right) D_2 \\
&\le -\gamma\mu\mathbb{E}\left[\|z_k-z^*\|_2^2\right] + 2\gamma\mu\mathbb{E}\left[\|z_{k+1/2}-z_k\|_2^2\right] \\
&\quad +4\gamma\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right) D_2. \qquad (38)
\end{aligned}$$

Combining (36), (37), and (38), we have

$$\begin{aligned}
\mathbb{E}\big[\|z_{k+1}-z^*\|_2^2\big] &+ \mathbb{E}\big[\|g_f(z_{k+1/2},\tau,\xi_{k+1/2}) - g_f(z_{k-1/2},\tau,\xi_{k-1/2})\|_2^2\big] \\
&\le (1-\gamma\mu)\left(\mathbb{E}\big[\|z_k-z^*\|_2^2\big] + \mathbb{E}\left[\|g_f(z_{k-1/2},\tau,\xi_{k-1/2}) - g_f(z_{k-3/2},\tau,\xi_{k-3/2})\|_2^2\right]\right) \\
&\quad +(2\gamma\mu + 12\gamma^2L^2 - 1)\mathbb{E}\left[\|z_{k+1/2}-z_k\|_2^2\right] \\
&\quad +\gamma^2\left[12\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right) + \frac{4D_2}{\gamma}\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right)\right].
\end{aligned}$$

With $\gamma \le 1/6L$ we have $12\gamma^2L^2 \le 1-2\mu\gamma$ and

$$\begin{aligned}
\mathbb{E}\big[\|z_{k+1}-z^*\|_2^2\big] &+ \mathbb{E}\big[\|g_f(z_{k+1/2},\tau,\xi_{k+1/2}) - g_f(z_{k-1/2},\tau,\xi_{k-1/2})\|_2^2\big] \\
&\le (1-\gamma\mu)\left(\mathbb{E}\big[\|z_k-z^*\|_2^2\big] + \mathbb{E}\left[\|g_f(z_{k-1/2},\tau,\xi_{k-1/2}) - g_f(z_{k-3/2},\tau,\xi_{k-3/2})\|_2^2\right]\right) \\
&\quad +\gamma^2\left[12\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right) + \frac{4D_2}{\gamma}\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right)\right].
\end{aligned}$$

It remains to apply (16) and then :

$$\begin{aligned}
\mathbb{E}\big[\|z_{N+1}-z^*\|_2^2\big] &\le \exp\left(-\frac{\mu N}{12L}\right)\left(\|z_0-z^*\|_2^2 + \|g_f(z_0,\tau,\xi_0) - g_f(z_0,\tau,\xi_0)\|_2^2\right) \\
&\quad +\frac{1}{\mu^2 N}\left[12\left(\sigma^2 + nL_2^2\tau^2 + \frac{2n\Delta^2}{\tau^2}\right) + \frac{4D_2}{\gamma}\left(\sqrt{n}L\tau + \frac{2\sqrt{n}\Delta}{\tau}\right)\right].
\end{aligned}$$

□

F Other approach for e in Algorithm 2

This algorithm is an easy modification of Algorithm 2. The only difference is that we use the same direction e and random variable ξ within one iteration

Algorithm 4 `zoESVIA (same direction)`

Input: z_0, N, γ, τ.
Choose oracle grad from G, g_d, g_f..
for $k = 0$ **to** N **do**
Sample indep. e_k, ξ_k.
$d_k = \text{grad}(z_k, e_k, \tau, \xi_k)$.
$z_{k+1/2} = \text{prox}_{z_k}(\gamma \cdot d_k)$.
$d_{k+1/2} = \text{grad}(z_{k+1/2}, e_k, \tau, \xi_k)$.
$z_{k+1} = \text{prox}_{z_k}(\gamma \cdot d_{k+1/2})$.
end for
Output: z_{N+1} or $\bar{z}_{N+1}$.

In this section we consider euclidean setup: $V_x(y) = 1/2\|x - y\|_2^2$. Used approach is based on [23].

Theorem.

By Algorithm 4 with Random direction oracle under Assumptions 1, 2, 3 and $\gamma \leq 1/2nL_2$, we get

$$\begin{aligned}\mathbb{E}[\varepsilon_{sad}(\bar{z}_N)] \quad \leq \quad & \frac{D_2^2}{\gamma N} + 210\gamma n^2 L_2^2 D_2^2 \\ & + 24\gamma\left(n^2L^2\tau^2 + \frac{n^2\Delta^2}{\tau^2}\right) + 12\left(nL\tau + \frac{n\Delta}{\tau}\right)D_2 \\ & + 200\gamma\left(n\mathbb{E}\left[\|F(z^*)\|_2^2\right] + \frac{n\sigma^2}{2}\right), \end{aligned} \tag{39}$$

where

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

Proof of (39) We begin with applying Lemma E.1

$$\begin{aligned}\|z_{k+1} - u\|_2^2 \quad \leq \quad & \|z_k - u\|_2^2 - 2\langle \gamma g_d(z_{k+1/2}, e_k, \tau, \xi_k), z_{k+1/2} - u\rangle \\ & + \gamma^2\|g_d(z_{k+1/2}, e_k, \tau, \xi_k) - g_d(z_k, e_k, \tau, \xi_k)\|_2^2 - \|z_{k+1/2} - z_k\|_2^2\end{aligned}$$

Next, using triangle inequality, we have

$$\begin{aligned}\|z_{k+1} - u\|_2^2 \quad \leq \quad & \|z_k - u\|_2^2 - 2\langle \gamma g_d(z_{k+1/2}, e_k, \tau, \xi_k), z_{k+1/2} - u\rangle \\ & + \gamma^2\|g_d(z_{k+1/2}, e_k, \tau, \xi_k) - n\langle F(z_{k+1/2}, \xi_k), e_k\rangle e_k\|_2^2 \\ & + \gamma^2\|g_d(z_k, e_k, \tau, \xi_k) - n\langle F(z_k, \xi_k), e_k\rangle e_k\|_2^2 \\ & + \gamma^2\|n\langle F(z_{k+1/2}, \xi_k), e_k\rangle e_k - n\langle F(z_k, \xi_k), e_k\rangle e_k\|_2^2 - \|z_{k+1/2} - z_k\|_2^2\end{aligned}$$

Using 8, we get

$$\begin{aligned}\|z_{k+1} - u\|_2^2 \quad \leq \quad & \|z_k - u\|_2^2 - 2\langle \gamma g_d(z_{k+1/2}, e_k, \tau, \xi_k), z_{k+1/2} - u\rangle \\ & + \gamma^2\|g_d(z_{k+1/2}, e_k, \tau, \xi_k) - n\langle F(z_{k+1/2}, \xi_k), e_k\rangle e_k\|_2^2 \\ & + \gamma^2\|g_d(z_k, e_k, \tau, \xi_k) - n\langle F(z_k, \xi_k), e_k\rangle e_k\|_2^2 \\ & + (\gamma^2 n^2 L^2(\xi_k) - 1)\|z_{k+1/2} - z_k\|_2^2\end{aligned}$$

By simple transformation we rewrite previous inequality

$$\begin{aligned}\langle F(z_{k+1/2}), z_{k+1/2}-u\rangle &\leq \|z_k-u\|_2^2-\|z_{k+1}-u\|_2^2\\ &-2\gamma\langle g_d(z_{k+1/2},e_k,\tau,\xi_k)-F(z_{k+1/2},\xi_k), z_{k+1/2}-u\rangle\\ &+\gamma^2\|g_d(z_{k+1/2},e_k,\tau,\xi_k)-n\langle F(z_{k+1/2},\xi_k),e_k\rangle e_k\|_2^2\\ &+\gamma^2\|g_d(z_k,e_k,\tau,\xi_k)-n\langle F(z_k,\xi_k),e_k\rangle e_k\|_2^2\\ &+(\gamma^2n^2L_2^2-1)\|z_{k+1/2}-z_k\|_2^2.\end{aligned}$$

We estimate some terms from the right side of the inequality.

$$\begin{aligned}&\|g_d(z_{k+1/2},e_k,\tau,\xi_k)-n\langle F(z_{k+1/2},\xi_k),e_k\rangle e_k\|_2^2\leq\\ &\frac{n^2}{\tau^2}\left\|\begin{pmatrix}\left(f(x_{k+1/2}+\tau e_{kx},y_{k+1/2},\xi_k)-f(x_{k+1/2},y_{k+1/2},\xi_k)-\langle\nabla_x f(x_{k+1/2},y_{k+1/2},\xi_k),\tau e_{kx}\rangle\right)e_{kx}\\ \left(f(x_{k+1/2},y_{k+1/2},\xi_k)-f(x_{k+1/2},y_{k+1/2}+\tau e_{ky},\xi_k)+\langle\nabla_y f(x_{k+1/2},y_{k+1/2},\xi_k),\tau e_{ky}\rangle\right)e_{ky}\end{pmatrix}\right\|_2^2\\ &+\frac{n^2}{\tau^2}\left\|\begin{pmatrix}\left(\delta(x_{k+1/2}+\tau e_{kx},y_{k+1/2})-\delta(x_{k+1/2},y_{k+1/2})\right)e_{kx}\\ \left(\delta(x_{k+1/2},y_{k+1/2})-\delta(x_{k+1/2},y_{k+1/2}+\tau e_{ky})\right)e_{ky}\end{pmatrix}\right\|_2^2.\end{aligned}$$

Using L-smoothness of function $f(\cdot)$ and 12, we note that

$$\begin{aligned}\|g_d(z_{k+1/2},e_k,\tau,\xi_k)-n\langle F(z_{k+1/2},\xi_k),e_k\rangle e_k\|_2^2 &\leq \frac{n^2}{\tau^2}\left(L^2\|\tau e_{kx}\|_2^2+L^2\|\tau e_{ky}\|_2^2\right)\\ &+4\frac{n^2\Delta^2}{\tau^2}\left(\|e_{kx}\|_2^2+\|e_{ky}\|_2^2\right)\\ &\leq 4\left(n^2L^2\tau^2+\frac{n^2\Delta^2}{\tau^2}\right)\end{aligned}$$

Similarly, we estimate the following value. Using L-smoothness of function $f(\cdot)$, we have

$$\begin{aligned}\|g_d(z_k,e_k,\tau,\xi_k)-n\langle F(z_k,\xi_k),e_k\rangle e_k\|_2^2 &\leq \frac{n^2}{\tau^2}\left(L^2\|\tau e_{kx}\|_2^2+L\|\tau e_{ky}\|_2^2\right)+4\frac{n^2\Delta^2}{\tau^2}\left(\|e_{kx}\|_2^2+\|e_{ky}\|_2^2\right)\\ &\leq 4\left(n^2L^2\tau^2+\frac{n^2\Delta^2}{\tau^2}\right)\end{aligned}$$

Substituting the previous inequality we have

$$\begin{aligned}\langle F(z_{k+1/2}), z_{k+1/2}-u\rangle &\leq \|z_k-u\|_2^2-\|z_{k+1}-u\|_2^2\\ &-2\gamma\langle g_d(z_{k+1/2},e_k,\tau,\xi_k)-F(z_{k+1/2}), z_{k+1/2}-u\rangle\\ &+8\gamma^2\left(n^2L^2\tau^2+\frac{n^2\Delta^2}{\tau^2}\right)\\ &+(\gamma^2n^2L_2^2-1)\|z_{k+1/2}-z_k\|_2^2\end{aligned}$$

Consider $\mathcal{G}=\langle g_d(z_{k+1/2},e_k,\tau,\xi_k)-F(z_{k+1/2}),u-z_{k+1/2}\rangle$, by simple transformations we get

$$\begin{aligned}\mathcal{G} &= \langle g_d(z_{k+1/2},e_k,\tau,\xi_k)-g_d(z_k,e_k,\tau,\xi_k),u-z_{k+1/2}\rangle+\langle F(z_k)-F(z_{k+1/2}),u-z_{k+1/2}\rangle\\ &+\langle g_d(z_k,e_k,\tau,\xi_k)-F(z_k),u-z_k\rangle+\langle g_d(z_k,e_k,\tau,\xi_k)-F(z_k),z_k-z_{k+1/2}\rangle\\ &\leq 2nL_2\|z_k-z_{k+1/2}\|_2\|u-z_{k+1/2}\|_2+\|g_d(z_k,e_k,\tau,\xi_k)-F(z_k)\|_2\|z_k-z_{k+1/2}\|_2\\ &+\|g_d(z_{k+1/2},e_k,\tau,\xi_k)-n\langle F(z_{k+1/2}),e_k\rangle e_k\|_2\|u-z_{k+1/2}\|_2\\ &+\|g_d(z_k,e_k,\tau)-n\langle F(z_k),e_k\rangle e_k\|_2\|u-z_{k+1/2}\|_2+\langle g_d(z_k,e_k,\tau,\xi_k)-F(z_k),u-z_k\rangle.\end{aligned}$$

Using $2\|a\|\|b\| \leq C\|a\|_2^2 + \frac{1}{C}\|b\|_2^2$

$$\begin{aligned} 2\gamma\mathcal{G} \quad \leq \quad & \frac{1}{2}\|z_k - z_{k+1/2}\|_2^2 + 8\gamma^2 n^2 L_2^2 \|u - z_{k+1/2}\|_2^2 + 4\gamma^2 \|g_d(z_k, e_k, \tau, \xi_k) - F(z_k)\|_2^2 \\ & + \frac{1}{4}\|z_k - z_{k+1/2}\|_2^2 \\ & + 4\gamma \left(nL\tau + \frac{n\Delta}{\tau} \right) \|u - z_{k+1/2}\|_2 + 2\gamma \langle g_d(z_k, e_k, \tau, \xi_k) - F(z_k), u - z_k \rangle. \end{aligned}$$

Summing up we get

$$\begin{aligned} \gamma\langle F(z_{k+1/2}), z_{k+1/2} - u\rangle \quad \leq \quad & \|z_k - u\|_2^2 - \|z_{k+1} - u\|_2^2 + 2\gamma\langle g_d(z_k, e_k, \tau, \xi_k) - F(z_k), u - z_k\rangle \\ & + 8\gamma^2 \left(n^2 L^2 \tau^2 + \frac{n^2\Delta^2}{\tau^2} \right) + 4\gamma \left(nL\tau + \frac{n\Delta}{\tau} \right) \|u - z_{k+1/2}\|_2 \\ & + (\gamma^2 n^2 L_2^2 - \frac{1}{4}) \|z_{k+1/2} - z_k\|_2^2 \\ & + 8\gamma^2 n^2 L_2^2 \|u - z_{k+1/2}\|_2^2 + 4\gamma^2 \|g_d(z_k, e_k, \tau, \xi_k) - F(z_k)\|_2^2 \end{aligned}$$

Assuming $\gamma \leq 1/2nL_2$, convexity-concavity of function $f(\dot{})$ and summing from $k = 1$ to $k = N$, we get

$$\begin{aligned} \frac{\gamma}{N+1} \sum_{k=0}^{N} \langle F(z_{k+1/2}), z_{k+1/2} - u\rangle \quad \leq \quad & \frac{D_2^2}{N} + \frac{2\gamma}{N} \sum_{k=1}^{N} \langle g_d(z_k, e_k, \tau, \xi_k) - F(z_k), u - z_k\rangle \\ & + 8\gamma^2 \left(n^2 L^2 \tau^2 + \frac{n^2\Delta^2}{\tau^2} \right) + 4\gamma \left(nL\tau + \frac{n\Delta}{\tau} \right) D_2 \\ & + 8\gamma^2 n^2 L_2^2 D_2^2 + \frac{4\gamma^2}{N} \sum_{k=1}^{N} \|g_d(z_k, e_k, \tau, \xi_k) - F(z_k)\|_2^2 \end{aligned}$$

Taking full expectation and using 18 $(q = 2)$, 31 and 8, we have

$$\begin{aligned} \gamma\mathbb{E}[\varepsilon_{sad}(\bar{z}_N)] \quad \leq \quad & \frac{D_2^2}{N} + 210\gamma^2 n, 2L_2^2 D_2^2 \\ & + 24\gamma^2 \left(n^2 L^2 \tau^2 + \frac{n^2\Delta^2}{\tau^2} \right) + 12\gamma \left(nL\tau + \frac{n\Delta}{\tau} \right) D_2 \\ & + 200\gamma^2 \left(n\mathbb{E}\left[\|F(z^*)\|_2^2 \right] + \frac{n\sigma^2}{2} \right) \end{aligned}$$

□

G Proof of Theorem 3

Theorem G.1. *By Algorithm 2 under assumption 1, 2, 3 with Mixed oracle $\widetilde{g}_f$ and $\gamma \leq 1/2L$, we get*

$$\begin{aligned} \mathbb{E}\left[\varepsilon_{sad}(\bar{z}_N)\right] \quad \leq \quad & \frac{D_p^2}{\gamma N} + 2D_p \left(\sqrt{n_x} L_2 \tau + \frac{2\sqrt{n_x}\Delta}{\tau} \right) \\ & + 9\gamma \left(\sigma^2 + n_x L_2^2 \tau^2 + \frac{2n_x\Delta^2}{\tau^2} \right). \end{aligned} \tag{40}$$

Proof of (40): We begin with (27) and taking $z = z_k, g = \gamma\widetilde{g}_f(z_k, e_k, \tau, \xi_k), g_{1/2} = \gamma\widetilde{g}_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2})$, then $z_{1/2} = z_{k+1/2}$, $z_1 = z_{k+1}$ and we get

$$\begin{aligned}\gamma\langle\widetilde{g}_f(z_{k+1/2}, e_{k+1/2}, \tau, \quad \xi_{k+1/2} \quad), z_{k+1/2} - u\rangle & \\ \leq \quad & V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1/2}) \\ & + \frac{\gamma^2}{2}\|\widetilde{g}_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - \widetilde{g}_f(z_k, e_k, \tau, \xi_k)\|_q^2 \\ \overset{(11)}{\leq} \quad & V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1/2}) \\ & + \frac{3\gamma^2}{2}\|F(z_{k+1/2}) - F(z_k)\|_q^2 \\ & + \frac{3\gamma^2}{2}\|\widetilde{g}_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})\|_q^2 \\ & + \frac{3\gamma^2}{2}\|\widetilde{g}_f(z_k, e_k, \tau, \xi_k) - F(z_k)\|_q^2\end{aligned}$$

With (8) it gives

$$\begin{aligned}\gamma\langle\widetilde{g}_f(z_{k+1/2}, e_{k+1/2}, \tau, \quad \xi_{k+1/2} \quad), z_{k+1/2} - u\rangle & \\ \leq \quad & V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1/2}) \\ & + \frac{3\gamma^2 L^2}{2}\|z_{k+1/2} - z_k\|_2^2 \\ & + \frac{3\gamma^2}{2}\|\widetilde{g}_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})\|_q^2 \\ & + \frac{3\gamma^2}{2}\|\widetilde{g}_f(z_k, e_k, \tau, \xi_k) - F(z_k)\|_q^2.\end{aligned}$$

Applying the property: $V_{z_k}(z_{k+1/2}) \geq 1/2\|z_{k+1/2} - z_k\|^2 \geq 1/2\|z_{k+1/2} - z_k\|_2^2$, with $\gamma \leq 1/2L$, we get

$$\begin{aligned}\gamma\langle\widetilde{g}_f(z_{k+1/2}, e_{k+1/2}, \quad \tau, \quad \xi_{k+1/2}), z_{k+1/2} - u\rangle \leq V_{z_k}(u) - V_{z_{k+1}}(u) & \\ + \frac{3\gamma^2}{2}\|\widetilde{g}_f(z_{k+1/2}, e_{k+1/2}, \tau, \xi_{k+1/2}) - F(z_{k+1/2})\|_q^2 & \\ + \frac{3\gamma^2}{2}\|\widetilde{g}_f(z_k, e_k, \tau, \xi_k) - F(z_k)\|_q^2. &\end{aligned}$$

Taking the full expectation and using (20), (21) with (6):

$$\begin{aligned}\mathbb{E}\left[\gamma\langle F(z_{k+1/2}), z_{k+1/2} - u\rangle\right] \quad \leq \quad & \mathbb{E}\left[V_{z_k}(u)\right] - \mathbb{E}\left[V_{z_{k+1}}(u)\right] \\ & + 2\gamma\left(\sqrt{n_x}L_2\tau + \frac{2\sqrt{n_x}\Delta}{\tau}\right)D_p \\ & + 3\gamma^2\left(3\sigma^2 + 3n_xL_2^2\tau^2 + \frac{6n_x\Delta^2}{\tau^2}\right).\end{aligned}$$

It remains to sum up from $k = 0$ to $k = N$ and use 31 and finish the proof of this theorem.

□

References

[1] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *COLT 2010 - The 23rd Conference on Learning Theory*, 2010.

[2] Tamer Basar and Geert Jan Olsder. *Dynamic Noncooperative Game Theory, 2nd Edition*. Society for Industrial and Applied Mathematics, 1998.

[3] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. 2019.

[4] Aleksandr Beznosikov, Eduard Gorbunov, and Alexander Gasnikov. Derivative-free method for decentralized distributed non-smooth optimization. *arXiv preprint arXiv:1911.10645*, 2019.

[5] Aleksandr Beznosikov, Abdurakhmon Sadiev, and Alexander Gasnikov. Gradient-free methods for saddle-point problem. *arXiv preprint arXiv:2005.05913*, 2020.

[6] R.P. Brent. *Algorithms for Minimization Without Derivatives*. Dover Books on Mathematics. Dover Publications, 1973.

[7] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[8] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[9] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security - AISec '17*, 2017.

[10] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 970–978, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[11] Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, 2009.

[12] Francesco Croce and Matthias Hein. A randomized gradient-free attack on relu networks, 2018.

[13] Francesco Croce, Jonas Rauber, and Matthias Hein. Scaling up the randomized gradient-free adversarial attack reveals overestimation of robustness using established attacks, 2019.

[14] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Trans. Information Theory*, 61(5):2788–2806, 2015. arXiv:1312.2139.

[15] Pavel Dvurechensky, Eduard Gorbunov, and Alexander Gasnikov. An accelerated directional derivative method for smooth stochastic convex optimization. *European Journal of Operational Research*, 2020.

[16] Vaclav Fabian. Stochastic approximation of minima with improved asymptotic speed. *Ann. Math. Statist.*, 38(1):191–200, 02 1967.

[17] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

[18] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1467–1476, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[19] Michael C Fu et al. *Handbook of simulation optimization*, volume 216. Springer, 2015.

[20] A. V. Gasnikov, A. A. Lagunovskaya, I. N. Usmanova, and F. A. Fedorenko. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77(11):2018–2034, Nov 2016. arXiv:1412.3890.

[21] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. arXiv:1309.5549.

[22] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016. arXiv:1308.6594.

[23] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks, 2018.

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[25] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

[26] Eduard Gorbunov, Pavel Dvurechensky, and Alexander Gasnikov. An accelerated method for derivative-free smooth stochastic convex optimization. *arXiv preprint arXiv:1802.09022*, 2018.

[27] Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1-3):161–220, 1990.

[28] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods, 2019.

[29] Anatoli Juditsky, Arkadii S. Nemirovskii, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm, 2008.

[30] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. 1976.

[31] Sijia Liu, Songtao Lu, Xiangyi Chen, Yao Feng, Kaidi Xu, Abdullah Al-Dujaili, Minyi Hong, and Una-May O'Reilly. Min-max optimization without gradients: Convergence and applications to adversarial ml, 2019.

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[33] Oskar Morgenstern and John Von Neumann. *Theory of games and economic behavior*. Princeton university press, 1953.

[34] Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *CVPR Workshops*, pages 1310–1318. IEEE Computer Society, 2017.

[35] A. Nedić and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, Jul 2009.

[36] Arkadi Nemirovski. Prox-method with rate of convergence o (1/ t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15:229–251, 01 2004.

[37] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527–566, April 2017. First appeared in 2011 as CORE discussion paper 2011/16.

[38] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. volume 70 of *Proceedings of Machine Learning Research*, pages 2817–2826, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[39] H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3(3):175–184, 1960.

[40] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv:1703.03864*, 2017.

[41] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18:52:1–52:11, 2017.

[42] Sara Shashaani, Fatemeh S. Hashemi, and Raghu Pasupathy. Astro-df: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM Journal on Optimization*, 28(4):3145–3176, 2018.

[43] James C. Spall. *Introduction to Stochastic Search and Optimization*. John Wiley & Sons, Inc., New York, NY, USA, 1 edition, 2003.

[44] Sebastian U. Stich. Unified optimal analysis of the (stochastic) gradient method, 2019.

[45] Sebastian U Stich, Christian L Muller, and Bernd Gartner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23(2):1284–1309, 2013.

[46] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses, 2017.

[47] Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex minimax problems with improved complexities, 2020.

[48] Haishan Ye, Zhichao Huang, Cong Fang, Chris Junchi Li, and Tong Zhang. Hessian-aware zeroth-order optimization for black-box adversarial attack, 2018.