

# Gradient-Free Methods for Non-Smooth Convex Stochastic Optimization with Heavy-Tailed Noise on Convex Compact

Nikita Kornilov<sup>1</sup>, Alexander Gasnikov<sup>1,4,5</sup>, Pavel Dvurechensky<sup>2</sup>,  
Darina Dvinskikh<sup>3</sup>

<sup>1</sup>Moscow Institute of Physics and Technology, Dolgoprudny, Russia.

<sup>2</sup>Weierstrass Institute for Applied Analysis and Stochastics, Berlin,  
Germany.

<sup>3</sup>HSE University, Moscow, Russia.

<sup>4</sup>Skoltech, Moscow, Russia.

<sup>5</sup>ISP RAS Research Center for Trusted Artificial Intelligence, Moscow,  
Russia.

Contributing authors: [kornilov.nm@phystech.edu](mailto:kornilov.nm@phystech.edu); [gasnikov@yandex.ru](mailto:gasnikov@yandex.ru);  
[pavel.dvurechensky@wias-berlin.de](mailto:pavel.dvurechensky@wias-berlin.de); [dmdvinskikh@hse.ru](mailto:dmdvinskikh@hse.ru);

We present two easy-to-implement gradient-free/zeroth-order methods to optimize a stochastic non-smooth function accessible only via a black-box. The methods are built upon efficient first-order methods in the heavy-tailed case, i.e., when the gradient noise has infinite variance but bounded  $(1 + \kappa)$ -th moment for some  $\kappa \in (0, 1]$ . The first algorithm is based on the stochastic mirror descent with a particular class of uniformly convex mirror maps which is robust to heavy-tailed noise. The second algorithm is based on the stochastic mirror descent and gradient clipping technique. Additionally, for the objective functions satisfying the  $r$ -growth condition, faster algorithms are proposed based on these methods and the restart technique.

## 1 Introduction

We consider stochastic non-smooth convex minimization problem

$$\min_{x \in \mathcal{X} \subset \mathbb{R}^d} f(x) \stackrel{\text{def}}{=} \mathbb{E}_{\xi \sim \mathcal{D}} [f(x, \xi)], \quad (1)$$

where function  $f(x, \xi)$  is  $M_2(\xi)$ -Lipschitz continuous in  $x$  w.r.t. the Euclidean norm,  $\mathcal{X}$  is a compact convex, and the expectation  $\mathbb{E}_{\xi \sim \mathcal{D}} [f(x, \xi)]$  is w.r.t. random variable  $\xi$  with unknown distribution  $\mathcal{D}$ . We suppose that stochastic realizations of the function values  $f(x)$  are available only through a zeroth-order oracle corrupted by some deterministic (probably adversarial) noise  $\delta(x)$

$$\phi(x, \xi) \stackrel{\text{def}}{=} f(x, \xi) + \delta(x). \quad (2)$$

We consider two-point zeroth-order oracle setting meaning that for two query points  $x, y \in \mathcal{X}$  we can evaluate two outputs  $\phi(x, \xi)$  and  $\phi(y, \xi)$  with the same  $\xi$ . Function  $\phi(x, \xi)$  can be considered as a noisy approximation of a Lipschitz function  $f(x, \xi)$ .

Stochastic optimization problems implies that functions  $f(x, \xi)$  must have finite mathematical expectation for all  $x \in \mathcal{X}$ .

Zeroth-order methods were studied in a wide range of works, see e.g., [1, 2] and the references therein. Particularly, under different assumptions on black-box oracle (in the noisy or noiseless setup) the optimal oracle complexity was obtained [3–8]. This bound is proportional to  $d\varepsilon^{-2}$ , where  $\varepsilon$  is the desired precision to solve problem (1) in terms of the function values. For saddle point problems, we refer to papers [9, 10] obtaining the same bound. This result is quite expected since the above complexity is  $d$  times larger than the complexity of optimal stochastic gradient procedures. Factor  $d$  has a natural interpretation since to approximate (stochastic) gradient it suffices to use  $d+1$  function values.<sup>1</sup> This is obvious in the smooth case (see e.g. [11]), and is not so trivial in the non-smooth case [7]. This result was obtained in the classical setting of a finite variance of stochastic gradients:  $\mathbb{E}_{\xi} [M(\xi)^2] < \infty$ . However, in modern learning problems, this condition may be violated. To this end, we aim to relax this assumption and consider heavy-tailed noise with bounded  $(1+\kappa)$ -th moment for some  $\kappa \in (0, 1]$ , i.e., we suppose  $\mathbb{E}_{\xi} [M(\xi)^{1+\kappa}] < \infty$ . Under this assumption, for the first-order stochastic methods, the optimal oracle complexity is proportional to  $\varepsilon^{-\frac{1+\kappa}{\kappa}}$  [12]. Thus for zeroth-order oracle we may expect the bound  $d\varepsilon^{-\frac{1+\kappa}{\kappa}}$ . In this paper, we obtain the bound  $(\sqrt{d}/\varepsilon)^{\frac{1+\kappa}{\kappa}}$  matching the expected bound only for  $\kappa = 1$ . To the best of our knowledge, this poses the following open problem: is the bound  $(\sqrt{d}/\varepsilon)^{\frac{1+\kappa}{\kappa}}$  optimal in terms of the dependence on  $d$ ? For smooth stochastic convex optimization problems with  $(d+1)$ -points stochastic zeroth-order oracle the answer is negative and the optimal bound is proportional to  $d\varepsilon^{-\frac{1+\kappa}{\kappa}}$ . Thus, for  $\kappa \in (0, 1)$  our results are somewhat surprising since the dependence on  $d$  in our bound is very different from the known results for the case  $\kappa = 1$ . To the best of our knowledge, this paper provides the first known result for gradient-free methods without assuming a finite variance of the stochastic noise. Since we give an accurate analysis, including high-probability bounds,<sup>2</sup> our results could be of interest even in a very particular case of  $\kappa = 1$ . In this case, the high-probability bound was previously known only for compactly supported distributions of

---

<sup>1</sup>To say more precisely, it suffices to use  $d+1$  values of  $f(x, \xi)$  with the same  $\xi$  and different  $(d+1)$  points  $x$ .

<sup>2</sup>We emphasize that these bounds were obtained without any probabilistic assumptions, except  $\mathbb{E}_{\xi} [M^{1+\kappa}(\xi)] < \infty$ !

$f(x, \xi)$  [10]. That is, even for sub-Gaussian tails [13] it was an open question to obtain high-probability bounds for gradient-free methods. The main challenge in obtaining our results is in the combination of the auxiliary gradient-free randomization and the original stochasticity of the oracle in the problem. The known inequalities on measure concentration do not allow obtaining the desired sub-Gaussian concentration for the output of the algorithm.

Gradient clipping technique has become increasingly popular for obtaining convergence guarantees in terms of high probability [14–16]. Starting with the work [14] (see also [15, 16]) one can observe an increased interest of researchers in algorithms that use gradient clipping to be able to obtain high-probability convergence guarantees in stochastic optimization problems with heavy-tailed noise. In particular, only in the last two years optimal first-order algorithms were proposed and the following results were obtained for their convergence guarantees: **1.** in the expectation for general proximal setup and non-smooth stochastic convex optimization problems with infinite variance [17]; **2.** in high-probability for general proximal setup and non-smooth online stochastic convex optimization problems with infinite variance [18]; **3.** in high-probability for the Euclidean proximal setup and smooth and non-smooth stochastic convex optimization problems and variational inequalities with infinite variance [19–21]; **4.** in high-probability for convergence of optimal variance-adaptive algorithm in the Euclidean proximal setup for non-smooth stochastic (strongly) convex optimization problems with infinite variance [22]. Since the aforementioned results are strongly correlated with each other, in this paper, we depart from the works [17, 18] to incorporate zero-order oracle into their algorithms. The developed technique, which reduces randomization caused by the gradient-free nature of the oracle to the original stochasticity, allows generalizing the results of other papers considered above in a similar manner. The idea of this reduction is not new and has already been used many times, see e.g. [3, 4, 6, 7]. But, all these works are significantly based on the assumption of finite variance of the stochastic noise. For the infinite noise variance setting, the technique requires significant generalizations, which we make in this paper. We expect, that based on these results it is possible to obtain new results for zero-order algorithms in the smooth setting and also in the setting of one-point feedback.

### *Contribution*

1. For  $d$ -dimensional optimization, we propose two algorithms with oracle complexity proportional to  $\left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ . This upper bound is valid under the maximal admissible level of adversarial noise proportional to  $\varepsilon^2/\sqrt{d}$ . For the first algorithm the convergence results hold in expectation whereas for the second algorithm the results are valid with high probability.
2. If additionally the objective satisfies the  $r$ -growth condition ( this includes strongly convex problems and problems with a sharp minimum), the restart technique for these algorithms gives oracle complexity proportional to  $\left(\sqrt{d}/\varepsilon^{\frac{(r-1)}{r}}\right)^{\frac{1+\kappa}{\kappa}}$ . This upper bound is valid under the maximal level of adversarial noise proportional to  $\varepsilon^{(2-\frac{1}{r})}/\sqrt{d}$ .

### Organization

This paper is organized as follows. Section 2 presents the main objects and notions that are used to construct gradient-free algorithms. In Section 3, we present our first gradient-free algorithm which is based on mirror descent. In Section 4 we present our second gradient-free algorithm based on gradient clipping. Finally, in Section 5 for the objective functions satisfying the  $r$ -growth condition, we propose a faster algorithm using the restart technique.

## 2 Preliminaries

### Notations

For  $p \in [1, 2]$ , we use the  $l_p$ -norm, i.e.  $\|x\|_p = \left(\sum_{k=1}^d |x_k|^p\right)^{1/p}$ . The corresponding dual norm is  $\|y\|_q = \max_x \{\langle x, y \rangle \mid \|x\|_p \leq 1\}$ , where  $q$  is defined by the equality  $1/q + 1/p = 1$ . We use  $\langle x, y \rangle = \sum_{k=1}^d x_k y_k$  to denote the inner product of  $x, y \in \mathbb{R}^d$ . Let  $B_{p'}^d = \{x \in \mathbb{R}^d \mid \|x\|_{p'} \leq 1\}$  and  $S_{p'}^d = \{x \in \mathbb{R}^d \mid \|x\|_{p'} = 1\}$  be the unit  $l_{p'}$ -ball and the unit  $l_{p'}$ -sphere with center at 0, correspondingly. The full expectation of a random variable  $X$  is denoted by  $\mathbb{E}[X]$ . The expectation w.r.t. random variables  $Y_1, \dots, Y_n$  is denoted by  $\mathbb{E}_{Y_1, \dots, Y_n}[X]$ . The condition expectation w.r.t.  $x_k, \dots, x_1$  is referred to as  $\mathbb{E}[\cdot \mid x_k, \dots, x_1] \stackrel{\text{def}}{=} \mathbb{E}_{|\leq k}[\cdot]$  for brevity.

### 2.1 Assumptions

For a convex set  $\mathcal{X} \subset \mathbb{R}^d$  and  $\tau > 0$ , let us introduce  $\mathcal{X}_\tau = \mathcal{X} + \tau B_2^d$ .

**Assumption 1** (Convexity). *There exists  $\tau > 0$  such that function  $f(x, \xi)$  is convex w.r.t.  $x$  for any  $\xi$  on  $\mathcal{X}_\tau$ .*

This assumption implies that  $f(x)$  is convex on  $\mathcal{X}$ .

**Assumption 2** (Lipschitz continuity and boundedness of  $(1 + \kappa)$ -th moment). *There exists  $\tau > 0$  such that function  $f(x, \xi)$  is  $M_2(\xi)$ -Lipschitz continuous w.r.t.  $x$  in the  $l_2$ -norm, i.e., for all  $x_1, x_2 \in \mathcal{X}_\tau$*

$$|f(x_1, \xi) - f(x_2, \xi)| \leq M_2(\xi) \|x_1 - x_2\|_2.$$

Moreover, there exist  $\kappa \in (0, 1]$  and  $M_2$  such that  $\mathbb{E}_\xi[M_2(\xi)^{1+\kappa}] \leq M_2^{1+\kappa}$ .

**Lemma 2.1.** *Assumption 2 implies that  $f(x)$  is  $M_2$ -Lipschitz on  $\mathcal{X}$ .*

The proof can be found in Section 8 (Lemma 8.2).

**Assumption 3** (Boundedness of noise). *There exists a constant  $\Delta > 0$  such that  $|\delta(x)| \leq \Delta$  for all  $x \in Q$ .*

### Randomized smoothing.

The main scheme that allows us to develop gradient-free methods for non-smooth convex problems is randomized smoothing [1, 5, 12, 23, 24] of a non-smooth function  $f(x, \xi)$ . The smooth approximation to a non-smooth function  $f(x, \xi)$  is defined as

$$\hat{f}_\tau(x) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{u}, \xi}[f(x + \tau \mathbf{u}, \xi)], \quad (3)$$

where  $\mathbf{u} \sim U(B_2^d)$  is a random vector uniformly distributed on the Euclidean unit ball.

The next lemma gives estimates for the quality of this approximation. In contrast to  $f(x)$ , function  $\hat{f}_\tau(x)$  is smooth and has several useful properties.

**Lemma 2.2.** [24, Theorem 2.1] *Let Assumptions 1,2 hold. Then,*

1. *Function  $\hat{f}_\tau(x)$  is convex, Lipschitz with constant  $M_2$  on  $\mathcal{X}$ , and satisfies*

$$\sup_{x \in \mathcal{X}} |\hat{f}_\tau(x) - f(x)| \leq \tau M_2.$$

2. *Function  $\hat{f}_\tau(x)$  is differentiable on  $\mathcal{X}$  with the following gradient*

$$\nabla \hat{f}_\tau(x) = \mathbb{E}_{\mathbf{e}} \left[ \frac{d}{\tau} f(x + \tau \mathbf{e}) \mathbf{e} \right],$$

where  $\mathbf{e} \sim U(S_2^d)$  is a random vector uniformly distributed on the Euclidean unit sphere.

### Gradient estimate.

To employ first-order algorithms in the zero-order oracle setting, we use the following gradient estimate

$$\begin{aligned} g(x, \xi, \mathbf{e}) &= \frac{d}{2\tau} (\phi(x + \tau \mathbf{e}, \xi) - \phi(x - \tau \mathbf{e}, \xi)) \mathbf{e} \\ &= \frac{d}{2\tau} (f(x + \tau \mathbf{e}, \xi) + \delta(x + \tau \mathbf{e}) - f(x - \tau \mathbf{e}, \xi) - \delta(x - \tau \mathbf{e})) \mathbf{e}. \end{aligned} \quad (4)$$

We can notice that this vector will be an unbiased estimate of the gradient of  $\hat{f}_\tau(x)$  if there is no adversarial noise  $\Delta = 0$ . Moreover, this vector has bounded  $(1 + \kappa)$ -th moment, see the next lemma.

**Lemma 2.3.** *Under Assumptions 1, 2 and 3, for  $q \in [2, +\infty)$ , we have*

$$\mathbb{E}_{\xi, \mathbf{e}} [\|g(x, \xi, \mathbf{e})\|_q^{1+\kappa}] \leq 2^\kappa \left( \frac{\sqrt{d}}{2^{1/4}} a_q M_2 \right)^{1+\kappa} + 2^\kappa \left( \frac{d a_q \Delta}{\tau} \right)^{1+\kappa} \stackrel{\text{def}}{=} \sigma_q^{1+\kappa},$$

where  $a_q \stackrel{\text{def}}{=} d^{\frac{1}{q} - \frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q - 1}\}$ .

The proof can be found in the Appendix 8.

## 3 First Algorithm: ZO-RSMD

In this section, we present our first gradient-free algorithm which is built upon mirror descent algorithm with uniformly convex mirror map from [17]. Our algorithm as well as algorithm from [17] is robust to heavy-tailed noise. Firstly we provide mirror descent algorithm with uniformly convex mirror map from [17] and then we present its zeroth-order version.

### 3.1 Robust Stochastic Mirror Descent (RSMD)

Now we present convergence results for first-order algorithm from [17] called stochastic mirror descent algorithm with uniformly convex mirror map (RSMD). It is based on stochastic mirror descent algorithm [12] and the notion of uniform convexity (to be determined further).

**Definition 3.1** (Uniform convexity). *Consider a differentiable convex function  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ , an exponent  $r \geq 2$ , and a constant  $K > 0$ . Then,  $\psi$  is called  $(K, r)$ -uniformly convex w.r.t. the  $\ell_p$ -norm if, for any  $x, y \in \mathbb{R}^d$ ,*

$$\psi(y) - \psi(x) - \langle \nabla \psi(x), y - x \rangle \geq \frac{K}{r} \|x - y\|_p^r. \quad (5)$$

When  $r = 2$  the definition of  $(K, r)$ -uniform convexity is equivalent to  $K$ -strongly convexity. Examples of functions when  $r > 2$  can be obtained from the next lemma.

**Lemma 3.1.** *For  $\kappa \in (0, 1]$ ,  $q \in [1 + \kappa, \infty)$  and  $p$  such that  $1/q + 1/p = 1$ , we define*

$$K_q \stackrel{\text{def}}{=} 10 \max \left\{ 1, (q - 1)^{\frac{1+\kappa}{2}} \right\}. \quad (6)$$

Then,

$$\phi_p(x) \stackrel{\text{def}}{=} \frac{\kappa}{1 + \kappa} \|x\|_p^{\frac{1+\kappa}{\kappa}} \quad (7)$$

is  $\left( K_q^{-\frac{1}{\kappa}}, \frac{1+\kappa}{\kappa} \right)$ -uniformly convex w.r.t. the  $\ell_p$ -norm.

Now we describe robust stochastic mirror descent (RSMD) algorithm [17]. Let function  $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}$  be  $(K, r)$ -uniformly convex w.r.t. the  $\ell_p$ -norm. We denote its Fenchel conjugate and its Bregman divergence respectively as

$$\Psi^*(y) = \sup_{x \in \mathbb{R}^d} \{ \langle x, y \rangle - \Psi(x) \} \quad \text{and} \quad D_\Psi(y, x) = \Psi(y) - \Psi(x) - \langle \nabla \Psi(x), y - x \rangle.$$

For a given stepsize  $\nu$  and gradient  $g_{k+1}$ , the updates of RSMD are defined as follows:

$$y_{k+1} = \nabla(\Psi^*)(\nabla \Psi(x_k) - \nu g_{k+1}), \quad x_{k+1} = \arg \min_{x \in \mathcal{X}} D_\Psi(x, y_{k+1}). \quad (8)$$

Using the assumptions on the function  $\Psi$ , it can be proved that the updates are well-defined and that  $(\nabla \Psi)^{-1} = \nabla \Psi^*$ . The map  $\nabla \Psi$  is referred to as the mirror map. The next theorem presents the convergence guarantee for the RSMD. Let

$$x^* = \arg \min_{x \in \mathcal{X}} f(x).$$

**Theorem 3.2.** [17, Theorem 6] *Consider some  $\kappa \in (0, 1]$ ,  $p \in [1, \infty]$  and prox-function  $\Psi_p$  which is  $\left( 1, \frac{1+\kappa}{\kappa} \right)$ -uniformly convex w.r.t.  $p$  norm. Then, for the SMD Algorithm outlined in (8), after  $T$  iterations with any  $g_k \in \mathbb{R}^d, k \in \overline{1, T}$  and starting point*

$x_0 = \arg \min_{x \in \mathcal{X}} \Psi_p(x)$  we have

$$\frac{1}{T} \sum_{k=0}^{T-1} \langle g_{k+1}, x_k - x^* \rangle \leq \frac{\kappa}{\kappa+1} \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa} \frac{1}{T} \sum_{k=0}^{T-1} \|g_{k+1}\|_q^{1+\kappa}, \quad (9)$$

where  $R_0^{\frac{1+\kappa}{\kappa}} \stackrel{\text{def}}{=} \frac{1+\kappa}{\kappa} D_{\Psi_p}(x^*, x_0)$  is the distance between starting point  $x_0$  and solution  $x^*$ .

### 3.2 Zeroth order version of RSMD (ZO-RSMD)

Next, we present our first zeroth-order algorithm called ZO-RSMD (zeroth-order version of robust SMD algorithm). The main idea of the proposed ZO-RSMD (zeroth-order version of robust SMD algorithm) is to combine the above RSMD algorithm (8) with the two-point gradient approximation (4).

---

#### Algorithm 1 ZO-RSMD

---

```

1: procedure ZO-RSMD (number of iterations  $T$ , stepsize  $\nu$ , prox-function  $\Psi_p$ ,
   smoothing constant  $\tau$ )
2:    $x_0 \leftarrow \arg \min_{x \in \mathcal{X}} \Psi_p(x)$ 
3:   for  $k = 0, 1, \dots, T-1$  do
4:     Sample  $\mathbf{e}_k \sim \text{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$  independently
5:     Sample  $\xi_k$  independently
6:     Calculate  $g_{k+1} = \frac{d}{2\tau} (\phi(x_k + \tau \mathbf{e}_k, \xi_k) - \phi(x_k - \tau \mathbf{e}_k, \xi_k)) \mathbf{e}_k$ 
7:     Calculate  $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla \Psi_p(x_k) - \nu g_{k+1})$ 
8:     Calculate  $x_{k+1} \leftarrow \arg \min_{x \in \mathcal{X}} D_{\Psi_p}(x, y_{k+1})$ 
9:   end for
10:  return  $\bar{x}_T \leftarrow \frac{1}{T} \sum_{k=0}^{T-1} x_k$ 
11: end procedure

```

---

The next theorem provides a convergence guarantee for ZO-RSMD (see Algorithm 1). **Theorem 3.3.** *Let function  $f$  satisfy Assumptions 1, 2, 3,  $q \in [1 + \kappa, \infty]$ . Let  $\Psi_p(x)$  be a prox-function which is  $(1, \frac{1+\kappa}{\kappa})$ -uniformly convex w.r.t. the  $l_p$ -norm (e.g.,  $\Psi_p(x) = K_q^{1/\kappa} \phi_p(x)$ , where  $K_q, \phi_p$  are defined in (6) and (7) respectively). Let step-size  $\nu = \frac{R_0^{1/\kappa}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$  with  $\sigma_q$  given in Lemma 2.3, distance between starting point  $x_0$  and solution  $x^*$   $R_0^{\frac{1+\kappa}{\kappa}} \stackrel{\text{def}}{=} \frac{1+\kappa}{\kappa} D_{\Psi_p}(x^*, x_0)$  and diameter  $\mathcal{D}_{\Psi}^{\frac{1+\kappa}{\kappa}} \stackrel{\text{def}}{=} \frac{1+\kappa}{\kappa} \sup_{x, y \in \mathcal{X}} D_{\Psi_p}(x, y)$ . Then for the output  $\bar{x}_T$  of the Algorithm 1 the following holds*

1.

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau} \mathcal{D}_{\Psi} + \frac{R_0\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \quad (10)$$

where  $\sigma_q^{1+\kappa} = 2^\kappa \left( \frac{\sqrt{d}}{2^{1/4}} a_q M_2 \right)^{1+\kappa} + 2^\kappa \left( \frac{da_q \Delta}{\tau} \right)^{1+\kappa}$ .

2. Moreover, with optimal  $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0 da_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$ , we have

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0da_q\Delta}{T^{\frac{\kappa}{1+\kappa}}}} + \frac{2\sqrt{d}a_qM_2R_0}{T^{\frac{\kappa}{1+\kappa}}}. \quad (11)$$

*Sketch of the Proof of Theorem 3.3*. the proof is based on Theorem 3.2 and inequality (9) which give

$$\mathbb{E} \left[ \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \langle g_{k+1}, x_k - x^* \rangle}_{\textcircled{1}} \right] \leq \mathbb{E} \left[ \underbrace{\frac{\kappa}{\kappa+1} \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T}}_{\textcircled{2}} \right] + \mathbb{E} \left[ \underbrace{\frac{\nu^\kappa}{1+\kappa} \frac{1}{T} \sum_{k=0}^{T-1} \|g_{k+1}\|_q^{1+\kappa}}_{\textcircled{3}} \right]. \quad (12)$$

$\textcircled{1}$  term in (12) due to convexity and approximation properties of  $\hat{f}_\tau(x)$  in Lemma 2.2 and measure concentration Lemma 8.6 can be bounded as

$$\textcircled{1} \geq \mathbb{E}[f(\bar{x}_T)] - f(x^*) - 2M_2\tau - \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi.$$

$\textcircled{3}$  term in (12) can be bounded by Lemma 2.3 as

$$\textcircled{3} \leq \frac{\nu^\kappa}{1+\kappa} \sigma_q^{1+\kappa}.$$

Combining these bounds together, we get

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa} \sigma_q^{1+\kappa}.$$

Next we choose optimal stepsize  $\nu = \frac{R_0^{1/\kappa}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$ ,  $\tau$  and finish the proof.  $\square$

For the complete proof we refer to Section 9.

### 3.3 Discussion

#### *Maximum admissible level of adversarial noise*

Let  $\varepsilon > 0$  be a desired accuracy in terms of the function value, i.e., our goal is to guarantee  $\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq \varepsilon$ . According to Theorem 3.3 in the case of absence of the adversarial noise, i.e., when  $\Delta = 0$ , the iteration complexity to reach accuracy  $\varepsilon$  is  $T = O\left(\left(\frac{R_0\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}\right)$  if  $\tau$  is chosen sufficiently small. This complexity is optimal according to [12] in terms of  $\varepsilon$  dependency. In order to obtain the same complexity in



the case when  $\Delta > 0$ , we need to choose an appropriate value of  $\tau$  and ensure that  $\Delta$  is sufficiently small. Thus, the terms  $2M_2\tau$  and  $\frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi$  in (10) should be of the order  $\varepsilon$ . These conditions also make negligible the  $\tau$ -dependent term in  $\sigma_q$ . One can choose  $\tau = \frac{\varepsilon}{M_2}$  rather than optimal  $\tau$  proposed in Theorem 3.3 in order to get easier calculations. Consequently, when  $\tau = \frac{\varepsilon}{M_2}$  and  $\Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$ , we have

$$T = O\left(\left(\frac{R_0\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}\right).$$

According to [25, 26] bound  $\Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$  exactly matches the upper bound of admissible adversarial noise for non-smooth zeroth-order optimization.

#### *Dependency of the bounds on $q$ and $d$*

In Algorithm 1, we can freely choose  $p \in [1, 2]$  and  $\Psi_p$ , which lead to different values of  $\mathcal{D}_\Psi, R_0, a_q$  depending on the compact convex set  $\mathcal{X}$ . It is desirable to reduce  $a_q, \mathcal{D}_\Psi$  simultaneously, that would allow us to increase maximal noise level  $\Delta$  and converge faster without changing the rate according to (10). Yet, unlike the well-studied SMD Algorithm [12] with strongly convex prox-functions  $\Psi_p$ , there are only a few examples of effective choices of uniformly-convex prox-functions  $\Psi_p$ .

## 4 Second Algorithm: ZO-Clip-SMD

In this section, we present our second algorithm which is based on the mirror descent and gradient clipping technique.

An alternative approach for dealing with heavy-tailed noise distributions in stochastic optimization is based on the gradient clipping technique, see e.g., [27]. Given a constant  $c > 0$ , the clipping operator applied to a vector  $g$  is given by

$$\hat{g} = \begin{cases} \frac{g}{\|g\|} \min(\|g\|, c), & g \neq 0, \\ 0, & g = 0. \end{cases}$$

Clipped gradient has several useful properties for further proofs.

**Lemma 4.1.** *For  $c > 0$  and stochastic vector  $g = g(x, \xi, \mathbf{e})$  we define  $\hat{g} = \frac{g}{\|g\|_q} \min(\|g\|_q, c)$ . Then we have*

1.

$$\|\hat{g} - \mathbb{E}[\hat{g}]\|_q \leq 2c. \quad (13)$$

2. Also if  $\mathbb{E}[\|g(x, \xi, \mathbf{e})\|_q^{1+\kappa}] \leq \sigma_q^{1+\kappa}$ , then we have

(a)

$$\mathbb{E}[\|\hat{g}\|_q^2] \leq \sigma_q^{1+\kappa} c^{1-\kappa}, \quad (14)$$

(b)

$$\mathbb{E}[\|\hat{g} - \mathbb{E}[\hat{g}]\|_q^2] \leq 4\sigma_q^{1+\kappa} c^{1-\kappa}, \quad (15)$$

(c)

$$\|\mathbb{E}[g] - \mathbb{E}[\hat{g}]\|_q \leq \frac{\sigma_q^{1+\kappa}}{c^\kappa}. \quad (16)$$

The clipping constant  $c$  allows playing with the trade-off between the faster convergence due to bounded second moment of  $\hat{g}$  and bias  $\|\mathbb{E}[\hat{g} - g]\|$  when  $c \rightarrow 0$ .

---

**Algorithm 2** ZO-CLIP-SMD

---

```

1: procedure ZO-CLIP-SMD (Number of iterations  $T$ , stepsize  $\nu$ , clipping constant
    $c$ , prox-function  $\Psi_p$ , smoothing constant  $\tau$ )
2:    $x_0 \leftarrow \arg \min_{x \in \mathcal{X}} \Psi_p(x)$ 
3:   for  $k = 0, 1, \dots, T - 1$  do
4:     Sample  $\mathbf{e}_k \sim \text{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$  independently
5:     Sample  $\xi_k$  independently
6:     Calculate  $g_{k+1} = \frac{d}{2\tau}(\phi(x_k + \tau\mathbf{e}_k, \xi_k) - \phi(x_k - \tau\mathbf{e}_k, \xi_k))\mathbf{e}_k$ 
7:     Calculate  $\hat{g}_{k+1} = \frac{g_{k+1}}{\|g_{k+1}\|_q} \min(\|g_{k+1}\|_q, c)$ 
8:     Calculate  $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla\Psi_p(x_k) - \nu\hat{g}_{k+1})$ 
9:     Calculate  $x_{k+1} \leftarrow \arg \min_{x \in \mathcal{X}} D_{\Psi_p}(x, y_{k+1})$ 
10:  end for
11:  return  $\bar{x}_T \leftarrow \frac{1}{T} \sum_{k=0}^{T-1} x_k$ 
12: end procedure

```

---

The next theorem presents convergence rates for ZO-CLIP-SMD (Algorithm 2) in terms of the expectation of the suboptimality gap.

**Theorem 4.2.** *Let function  $f$  satisfy Assumptions 1, 2, 3,  $q \in [2, \infty]$ , arbitrary number of iterations  $T$ , smoothing constant  $\tau > 0$  be given. Let  $\Psi_p(x)$  be a prox-function which is 1-strongly convex w.r.t. the  $p$ -norm. Let the stepsize  $\nu = \left(\frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$  with  $\sigma_q$  given in Lemma 2.3, distance between starting point  $x_0$  and solution  $x^*$   $R_0^{\frac{1+\kappa}{\kappa}} \stackrel{\text{def}}{=} \frac{1+\kappa}{\kappa} D_{\Psi_p}(x^*, x_0)$ , diameter  $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \stackrel{\text{def}}{=} \frac{1+\kappa}{\kappa} \sup_{x, y \in \mathcal{X}} D_{\Psi_p}(x, y)$ , and the clipping constant  $c = \frac{2\kappa\mathcal{D}_\Psi}{(1-\kappa)\nu}$ . Then for the output  $\bar{x}_T$  of Algorithm 2 the following holds*

1. Then, we have

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) = 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \quad (17)$$

$$\text{where } \sigma_q^{1+\kappa} = 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}}a_qM_2\right)^{1+\kappa} + 2^\kappa \left(\frac{da_q\Delta}{\tau}\right)^{1+\kappa}.$$

2. Moreover, with the optimal  $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}} da_q\Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$ , we have

$$\begin{aligned} \mathbb{E}[f(\bar{x}_T)] - f(x^*) &\leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}} da_q\Delta}{T^{\frac{\kappa}{1+\kappa}}}} \\ &\quad + \frac{2\sqrt{d}a_qM_2R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \end{aligned} \quad (18)$$

*Sketch of the Proof of Theorem 4.2.* The proof is based on Theorem 3.2 and inequality (9) for 1-strongly convex  $\Psi_p$ , which give

$$\mathbb{E} \left[ \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1}, x_k - x^* \rangle}_{\textcircled{1}} \right] \leq \mathbb{E} \left[ \frac{1}{2} \frac{R_0^2}{\nu T} \right] + \mathbb{E} \left[ \underbrace{\frac{\nu}{2} \frac{1}{T} \sum_{k=0}^{T-1} \|\hat{g}_{k+1}\|_q^2}_{\textcircled{2}} \right]. \quad (19)$$

$\textcircled{1}$  term in (19) due to convexity and approximation properties of  $\hat{f}_\tau(x)$  in Lemma 2.2, measure concentration Lemma 8.6 and clipping properties in Lemma 4.1 can be bounded as

$$\textcircled{1} \geq \mathbb{E}[f(\bar{x}_T)] - f(x^*) - 2M_2\tau - \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi - \frac{\mathcal{D}_\Psi\sigma_q^{1+\kappa}}{c^\kappa}.$$

$\textcircled{2}$  term in (19) can be bounded by Lemma 4.1 as

$$\textcircled{2} \leq \frac{\nu}{2}c^{1-\kappa}\sigma_q^{1+\kappa}.$$

Combining these bounds together, we get

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{1}{2} \frac{R_0^2}{\nu T} + \frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa} + \left( \frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta \frac{\sqrt{d}}{\tau} \right) \mathcal{D}_\Psi.$$

Next, we choose optimal clipping constant  $c = \frac{2\kappa\mathcal{D}_\Psi}{(1-\kappa)\nu}$ . Then, the optimal stepsize  $\nu = \left( \frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}} \right)^{\frac{1}{1+\kappa}}$  and smoothing parameter  $\tau$  finish the proof.  $\square$

For the complete proof we refer to Appendix 10.

The next theorem present the convergence rates of for ZO-Clip-SMD (Algorithm 2) with high probability rather than in expectation. We will use the  $\tilde{O}(\cdot)$ -notation to hide polynomial factors of  $\log \frac{1}{\delta}$ .

**Theorem 4.3.** *Let function  $f$  satisfy Assumptions 1, 2, 3,  $q \in [2, \infty]$ , arbitrary number of iterations  $T$ , smoothing constant  $\tau > 0$  be given. Let  $\Psi_p(x)$  be a 1-strongly convex w.r.t. the  $p$ -norm prox-function. Let the clipping constant  $c = T^{\frac{1}{1+\kappa}}\sigma_q$  with*

$\sigma_q$  given in Lemma 2.3, the stepsize  $\nu = \frac{\mathcal{D}_\Psi}{c}$  with diameter  $\mathcal{D}_\Psi^2 \stackrel{\text{def}}{=} 2 \sup_{x, y \in \mathcal{X}} D_{\Psi_p}(x, y)$ .

Then for the output  $\bar{x}_T$  of the Algorithm 2 the following holds

1. Then, with probability at least  $1 - \delta$ , we have

$$f(\bar{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi + \tilde{O}\left(\frac{\mathcal{D}_\Psi\sigma_q}{T^{1+\kappa}}\right), \quad (20)$$

$$\text{where } \sigma_q^{1+\kappa} = 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}}a_qM_2\right)^{1+\kappa} + 2^\kappa \left(\frac{da_q\Delta}{\tau}\right)^{1+\kappa}.$$

2. Moreover, with the optimal  $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4\mathcal{D}_\Psi da_q\Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$ , we have

$$f(\bar{x}_T) - f(x^*) = \tilde{O}\left(\sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2\mathcal{D}_\Psi da_q\Delta}{T^{\frac{\kappa}{1+\kappa}}}} + \frac{2\sqrt{d}a_qM_2\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\right). \quad (21)$$

*Sketch of the Proof of Theorem 4.3.* To bound variables with probability at least  $1 - \delta$  we use the classical Bernstein inequality for the sum of martingale differences (i.e.  $\mathbb{E}[X_i|X_{j < i}] = 0$ , for all  $i \geq 1$ ) (Lemma 11.1) and the sum of squares of random variables (Lemma 11.2).

The proof is based on Theorem 3.2 and inequality (9) for 1-strongly convex  $\Psi_p$  which give

$$\frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1}, x_k - x^* \rangle \leq \frac{1}{2} \frac{R_0^2}{\nu T} + \underbrace{\frac{\nu}{2} \frac{1}{T} \sum_{k=0}^{T-1} \|\hat{g}_{k+1}\|_q^2}_{\textcircled{1}}. \quad (22)$$

Adding  $\pm \mathbb{E}_{|\leq k}[\hat{g}_{k+1}]$  and  $\pm \hat{f}_\tau(x_k)$  to the left part of (22), we obtain

$$\begin{aligned} \frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1}, x_k - x^* \rangle &= \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle}_{\textcircled{2}} \\ &+ \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \langle \mathbb{E}_{|\leq k}[\hat{g}_{k+1}] - \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle}_{\textcircled{3}}, \\ &+ \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle}_{\textcircled{4}}. \end{aligned}$$

We bound ① term in (22) using Lemma 11.2 and ② as the sum of martingale differences using Lemma 11.1:

$$\begin{aligned}\textcircled{1} &= \tilde{O}\left(\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} + \frac{1}{T}c^2\right) \\ \textcircled{2} &= \tilde{O}\left(\frac{4c\mathcal{D}_\Psi}{T} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\right).\end{aligned}$$

Next, we bound ④ using the convexity of  $\hat{f}_\tau(x)$  in Lemma 2.2 and ③ using the measure concentration Lemma 8.6 and clipping properties in Lemma 4.1:

$$\textcircled{3} \leq \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi,$$

$$\textcircled{4} \geq f(\bar{x}_T) - f(x^*) - 2M_2\tau.$$

Combining these bounds together, we get

$$\begin{aligned}f(\bar{x}_T) - f(x^*) &\leq 2M_2\tau + \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi + \frac{1}{2}\frac{R_0^2}{\nu T} \\ &\quad + \tilde{O}\left(\frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa} + \frac{\nu}{2}\frac{1}{T}c^2 + \frac{4c\mathcal{D}_\Psi}{T} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\right).\end{aligned}$$

Next, we choose the stepsize  $\nu = \frac{\mathcal{D}_\Psi}{c}$ , clipping constant  $c = T^{\frac{1}{1+\kappa}}\sigma_q$ , smoothing parameter  $\tau$ , and finish the proof.  $\square$

For the complete proof we refer to Section 11.

## 4.1 Discussion

### *Maximum admissible level of adversarial noise*

Let  $\varepsilon > 0$  be a desired accuracy in terms of the function value, i.e., with probability at least  $1 - \delta$  we have  $f(\bar{x}_T) - f(x^*) \leq \varepsilon$ . In Theorem 4.3 if there is no adversarial noise, i.e.,  $\Delta = 0$ , then the number of iterations  $T$  to reach this accuracy is given by  $T = \tilde{O}\left(\left(\frac{\mathcal{D}_\Psi\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}\right)$  when  $\tau \rightarrow 0$ . This bound is optimal in terms of  $\varepsilon$  dependency according to [12]. In order to keep the same complexity when  $\Delta > 0$ , the terms  $2M_2\tau$  and  $\frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi$  should be of the order  $\varepsilon$ . These conditions also make negligible the  $\tau$ -depending term in  $\sigma_q$ . One can choose  $\tau = \frac{\varepsilon}{M_2}$  rather than optimal  $\tau$  proposed in Theorem 4.3 in order to get easier calculations. Consequently, if  $\tau = \frac{\varepsilon}{M_2}$

and  $\Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$  then

$$T = \tilde{O} \left( \left( \frac{\mathcal{D}_\Psi \sqrt{d} a_q M_2}{\varepsilon} \right)^{\frac{1+\kappa}{\kappa}} \right).$$

According to [25, 26] bound  $\Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$  exactly matches the upper bound of admissible adversarial noise for non-smooth zeroth-order optimization.

### **Recommendations for choosing $\Psi_p$**

In Algorithm 2, we can freely choose  $p \in [1, 2]$  and  $\Psi_p$ , which, depending on the compact convex set  $\mathcal{X}$ , will change  $\mathcal{D}_\Psi, R_0, a_q$ . The main task is to reduce  $a_q, \mathcal{D}_\Psi$  simultaneously, which will allow us to increase maximal noise  $\Delta$  and converge faster without changing the rate according to (20).

Next, we discuss some standard sets  $\mathcal{X}$  and prox-functions  $\Psi_p$  taken from [28]. The two main setups are given by

1. Ball setup:

$$p = 2, \Psi_p(x) = \frac{1}{2} \|x\|_2^2, \quad (23)$$

2. Entropy setup:

$$p = 1, \Psi_p(x) = (1 + \gamma) \sum_{i=1}^d (x_i + \gamma/d) \log(x_i + \gamma/d), \gamma > 0. \quad (24)$$

We consider unit balls  $B_p^d$  and standard simplex  $\Delta_+^d = \{x \in \mathbb{R}^d : x \geq 0, \sum_i x_i = 1\}$  as  $\mathcal{X}$ . By Lemma 2.3 constant  $a_q$  equals  $d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q - 1}\}$ . The next tables collect the iteration complexity  $T^{\frac{\kappa}{1+\kappa}}$  and maximum feasible noise level  $\Delta$  up to  $O(\log \frac{1}{\delta})$  factor for each setup (row) and set (column).

**Table 1**  $T^{\frac{\kappa}{1+\kappa}}$  up to  $O(\log \frac{1}{\delta})$  factor for Algorithm 2

	$\Delta_+^d$	$B_1^d$	$B_2^d$	$B_\infty^d$
Ball	$\sqrt{d}M_2/\varepsilon$	$\sqrt{d}M_2/\varepsilon$	$\sqrt{d}M_2/\varepsilon$	$dM_2/\varepsilon$
Entropy	$\ln dM_2/\varepsilon$	$\ln dM_2/\varepsilon$	$\sqrt{d \ln d}M_2/\varepsilon$	$d \ln dM_2/\varepsilon$

**Table 2** Maximum feasible noise level  $\Delta$  up to  $O(1)$  factor for Algorithm 2

	$\Delta_+^d$	$B_1^d$	$B_2^d$	$B_\infty^d$
Ball	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(dM_2)$
Entropy	$\varepsilon^2/(\sqrt{d \ln d}M_2)$	$\varepsilon^2/(\sqrt{d \ln d}M_2)$	$\varepsilon^2/(d\sqrt{\ln d}M_2)$	$\varepsilon^2/(\sqrt{d^3 \ln d}M_2)$

From these tables, we see that for  $\mathcal{X} = \Delta_+^d$  or  $B_1^d$ , the Entropy setup is preferable, while the Ball setup allows maximum feasible noise level  $\Delta$  to be up to  $\sqrt{\ln d}$  greater.

Meanwhile, for  $\mathcal{X} = B_2^d$  or  $B_\infty^d$ , the Ball setup is better in terms of both convergence rate and noise robustness.

*Comparison of two algorithms: ZO-RSMD and ZO-Clip-SMD*

Despite the fact that both algorithms have the same convergence rates, ZO-Clip-SMD is more flexible due to the greater freedom of choice of prox-functions  $\Psi_p$ . However, its convergence dramatically depends on the clipping constant  $c$  which must be carefully chosen.

## 5 Algorithms with Restarts: ZO-Restarts

In this section, we assume the objective function satisfies the  $r$ -growth condition [13]. In this case, optimization algorithms can be accelerated by using the restart technique [29].

**Assumption 4.** *Function  $f$  is  $r$ -growth function if there are  $r \geq 1$  and  $\mu_r \geq 0$  such that for all  $x$*

$$\frac{\mu_r}{2} \|x - x^*\|_p^r \leq f(x) - f(x^*),$$

where  $x^*$  is problem solution.

In particular, the condition of  $\mu$ -strong convexity w.r.t. the  $\ell_p$ -norm is the 2-growth condition. The restart technique works if  $\Delta$  is small enough to keep the optimality of Algorithms 1 and 2. The general scheme of the restart algorithm is presented below.

---

**Algorithm 3** ZO-Restarts

---

- 1: **procedure** ZO-RESTARTS (Algorithm type  $\mathcal{A}$ , number of restarts  $N$ , sequence of number of steps  $\{T_k\}_{k=1}^N$ , sequence of smoothing constants  $\{\tau_k\}_{k=1}^N$ , sequence of stepsizes  $\{\nu_k\}_{k=1}^N$ , sequence of clipping constants  $\{c_k\}_{k=1}^N$  (if necessary), prox-function  $\Psi_p$ )
  - 2:      $x_0 \leftarrow \arg \min_{x \in \mathcal{X}} \Psi_p(x)$  or randomly
  - 3:     **for**  $k = 0, 1, \dots, N$  **do**
  - 4:         Set parameters  $\nu_k, (c_k), \Psi_p, \tau_k$  of the Algorithm  $\mathcal{A}$
  - 5:         Run  $T_k$  iterations of the Algorithm  $\mathcal{A}$  with starting point  $x_0$  and get  $x_{\text{final}}$
  - 6:          $x_0 \leftarrow x_{\text{final}}$
  - 7:     **end for**
  - 8:     **return**  $x_{\text{final}}$
  - 9: **end procedure**
- 

The next theorem provides the convergence guarantee for Algorithm ZO-Restarts run with ZO-RSMD.

**Theorem 5.1.** *Let function  $f$  satisfy Assumptions 1, 2. Let  $\varepsilon > 0$  be a fixed accuracy and the  $r$ -growth Assumption 4 holds with  $r \geq \frac{1+\kappa}{\kappa}$ .*

*Set  $R_0 \stackrel{\text{def}}{=} \sup_{x,y \in \mathcal{X}} \left( \frac{1+\kappa}{\kappa} D_{\Psi_p}(x,y) \right)^{\frac{\kappa}{1+\kappa}}$  and  $R_k = R_0/2^k$ .*

Set the number of restarts  $N = \tilde{O}\left(\frac{1}{r} \log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right)$ , sequence of number of steps  $\{T_k\}_{k=1}^N = \left\{\tilde{O}\left(\left[\frac{\sigma_q 2^{(1+r)}}{\mu_r R_k^{r-1}}\right]^{\frac{1+\kappa}{\kappa}}\right)\right\}_{k=1}^N$ , sequence of smoothing constants  $\{\tau_k\}_{k=1}^N = \left\{\frac{\sigma_q R_k}{M_2 T_k^{\frac{1+\kappa}{\kappa}}}\right\}_{k=1}^N$  and sequence of stepsizes  $\{\nu_k\}_{k=1}^N = \left\{\frac{R_k^{1/\kappa}}{\sigma_q} T_k^{-\frac{1}{1+\kappa}}\right\}_{k=1}^N$ , where  $\sigma_q$  is given in Lemma 2.3. Finally, let Assumption 3 hold with

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2 \sqrt{d}} \frac{1}{2^{k(2r-1)}}\right), \quad 1 \leq k \leq N.$$

If  $x_{\text{final}}$  is the final output of Algorithm 3 with ZO-RSMD (Algorithm 1) as  $\mathcal{A}$  and with the above parameters, then

$$\mathbb{E}[f(x_{\text{final}})] - f(x^*) \leq \varepsilon,$$

and the total number of steps is

$$T = \tilde{O}\left(\left[\frac{a_q M_2 \sqrt{d}}{\mu_r^{1/r}} \cdot \frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad a_q \stackrel{\text{def}}{=} d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q - 1}\},$$

and on the last restart the maximum  $\Delta$  threshold is

$$\Delta_N = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2 \sqrt{d}} \varepsilon^{(2-1/r)}\right).$$

The next theorem provides the convergence guarantee for Algorithm ZO-Restarts run with ZO-Clip-SMD.

**Theorem 5.2.**<sup>3</sup> Let function  $f$  satisfy Assumptions 1, 2. Let  $\varepsilon > 0$  be a fixed accuracy and  $r$ -growth Assumption 4 holds with  $r \geq 2$  for in expectation estimate or  $r \geq 1$  for in high probability estimate. Set  $R_0 \stackrel{\text{def}}{=} \sup_{x,y \in \mathcal{X}} (2D_{\Psi_p}(x,y))^{\frac{1}{2}}$  and  $R_k = R_0/2^k$ . Set the number of restarts  $N = \tilde{O}\left(\frac{1}{r} \log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right)$ , sequence of number of steps  $\{T_k\}_{k=1}^N = \left\{\tilde{O}\left(\left[\frac{\sigma_q 2^{(1+r)}}{\mu_r R_k^{r-1}}\right]^{\frac{1+\kappa}{\kappa}}\right)\right\}_{k=1}^N$ , sequence of smoothing constants  $\{\tau_k\}_{k=1}^N = \left\{\frac{\sigma_q R_k}{M_2 T_k^{\frac{1+\kappa}{\kappa}}}\right\}_{k=1}^N$ , sequence of clipping constants  $\{c_k\}_{k=1}^N = \left\{T_k^{\frac{1}{(1+\kappa)}} \sigma_q\right\}_{k=1}^N$  and sequence of stepsizes  $\{\nu_k\}_{k=1}^N = \left\{\frac{R_k}{c_k}\right\}_{k=1}^N$ , where  $\sigma_q$  is given in Lemma 2.3.

<sup>3</sup>In this theorem  $\tilde{O}(\cdot)$  denotes  $\log d$  factor for in expectation bounds and  $\log d, \log \frac{1}{\delta}$  factors for in high probability bounds. More explicit formulas are provided in the full proof.



Finally, let Assumption 3 hold with

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2 \sqrt{d}} \frac{1}{2^{k(2r-1)}}\right), \quad 1 \leq k \leq N.$$

If  $x_{\text{final}}$  is the final output of Algorithm 3 with ZO-Clip-SMD (Algorithm 2) as  $\mathcal{A}$  and with the above parameters, then

$$\mathbb{E}[f(x_{\text{final}})] - f(x^*) \leq \varepsilon,$$

or with probability at least  $1 - \delta$

$$f(x_{\text{final}}) - f(x^*) \leq \varepsilon.$$

The total number of steps is

$$T = \begin{cases} \tilde{O}\left(\left[\frac{a_q M_2 \sqrt{d}}{\mu_r^{1/r}} \cdot \frac{1}{\varepsilon^{\frac{1}{r-1}}}\right]^{\frac{1+\kappa}{\kappa}}\right), & r > 1 \\ \tilde{O}\left(\left[\frac{a_q M_2 \sqrt{d}}{\mu_r}\right]^{\frac{1+\kappa}{\kappa}} \log_2\left(\frac{\mu_r R_0}{2\varepsilon}\right)\right), & r = 1 \end{cases},$$

$$a_q \stackrel{\text{def}}{=} d^{\frac{1}{q} - \frac{1}{2}} \min\{\sqrt{32 \ln d} - 8, \sqrt{2q - 1}\},$$

and on the last restart the maximum  $\Delta$  threshold is

$$\Delta_N = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2 \sqrt{d}} \varepsilon^{(2-1/r)}\right).$$

For the complete proofs of Theorems 5.1, 5.2 we refer to the Appendix 12.

## 5.1 Discussion

### Maximum admissible level of adversarial noise

Next, we compare the maximum value of adversarial noise allowed in ZO-RSMD, ZO-Clip-SMD and ZO-Restarts,

$$\begin{aligned} \text{ZO-RSMD (1) or ZO-Clip-SMD (2)} : \quad \Delta &= O\left(\frac{\varepsilon^2}{M_2 \sqrt{d} \mathcal{D}_\Psi}\right), \\ \text{ZO-Restarts (3)} : \quad \Delta &= \tilde{O}\left(\frac{\mu_r^{1/r} \varepsilon^{(2-1/r)}}{M_2 \sqrt{d}}\right). \end{aligned}$$

We notice  $r \geq 1$ . When  $r = 1$ , the first bound depends on  $\varepsilon$  quadratically whereas in the second bound this dependence is linear. When  $r$  tends to infinity, the results are the same. Also, in the beginning  $\Delta_k$  can be much bigger and it starts to decrease as  $\Delta_k = \frac{\Delta_1}{2^{k(2r-1)}}$  only on subsequent restarts to reach the required accuracy.

*q, d, ε dependencies*

Next, we compare the oracle complexity of **ZO-RSMD**, **ZO-Clip-SMD** and **ZO-Restarts**. Again, **ZO-Restarts** guarantees a better dependence on  $\varepsilon$ . Below we state the results in expectation for  $r > 1$

$$\begin{aligned} \text{ZO-RSMD (1) or ZO-Clip-SMD (2)} : \quad T &= O \left( \left[ \frac{\sqrt{d}M_2\mathcal{D}_\Psi a_q}{\varepsilon} \right]^{\frac{1+\kappa}{\kappa}} \right), \\ \text{ZO-Restarts (3)} : \quad T &= \tilde{O} \left( \left[ \frac{\sqrt{d}M_2 a_q}{\mu r^{1/r} \varepsilon^{\frac{(r-1)}{r}}} \right]^{\frac{1+\kappa}{\kappa}} \right). \end{aligned}$$

In case of  $r = 1$  **ZO-Restarts** achieves linear convergence.

## 6 Conclusion and Future Work

In this paper, we proposed and theoretically studied new zeroth-order algorithms to solve non-smooth optimization problems on a convex compact set with zeroth-order oracle corrupted by heavy-tailed stochastic noise (random noise with  $(1 + \kappa)$ -th bounded moment) and adversarial noise. We believe that the convergence rates can be improved with the following possible modifications:

1. different sampling strategy for estimating  $g_k$ , namely uniform sampling from the unit  $\ell_1$ -sphere  $\{\mathbf{e} : \|\mathbf{e}\|_1 = 1\}$ , see, e.g., [30], [31].
2. different assumption about adversarial noise, namely, Lipschitz continuity

$$|\delta(x_1) - \delta(x_2)| \leq M\|x_1 - x_2\|_2, \quad \forall x_1, x_2 \in \mathcal{X}$$

see, e.g., [10].

3. adaptive strategies and heuristic methods for choosing input parameters of the algorithm, such as stepsize  $\nu$ , smoothing constant  $\tau$ , etc. In practice, these constants are difficult to estimate.

We leave their implementation for future work. We believe that the technique developed in this paper is rather general and makes it possible to use other stochastic gradient methods to obtain new complexity bounds for zeroth-order algorithms.

Also our results can be generalized to obtain the same complexity bounds for non-smooth convex-concave saddle-point problems in terms of the duality gap used in [9] (rather than the gap used in [10]).<sup>4</sup> We leave this for future work.

## 7 Acknowledgments

The work of A. Gasnikov was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier

---

<sup>4</sup>See the full version of the paper [10].

000000D730321P5Q0002 ) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

## References

- [1] Spall, J.C.: Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. John Wiley & Sons, Chichester (2005)
- [2] Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-free Optimization. SIAM, Montreal (2009)
- [3] Duchi, J.C., Jordan, M.I., Wainwright, M.J., Wibisono, A.: Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory* **61**(5), 2788–2806 (2015)
- [4] Gasnikov, A.V., Lagunovskaya, A.A., Usmanova, I.N., Fedorenko, F.A.: Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control* **77**, 2018–2034 (2016)
- [5] Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics* **17**, 527–566 (2017)
- [6] Gasnikov, A.V., Krymova, E.A., Lagunovskaya, A.A., Usmanova, I.N., Fedorenko, F.A.: Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case. *Automation and remote control* **78**, 224–234 (2017)
- [7] Shamir, O.: An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research* **18**(1), 1703–1713 (2017)
- [8] Bayandina, A.S., Gasnikov, A.V., Lagunovskaya, A.A.: Gradient-free two-point methods for solving stochastic nonsmooth convex optimization problems with small non-random noises. *Automation and Remote Control* **79**, 1399–1408 (2018)
- [9] Beznosikov, A., Sadiev, A., Gasnikov, A.: Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In: *Mathematical Optimization Theory and Operations Research: 19th International Conference, MOTOR 2020, Novosibirsk, Russia, July 6–10, 2020, Revised Selected Papers 19*, pp. 105–119 (2020). Springer
- [10] Dvinskikh, D., Tominin, V., Tominin, Y., Gasnikov, A.: Gradient-free optimization for non-smooth minimax problems with maximum value of adversarial noise. arXiv preprint arXiv:2202.06114 (2022)

- [11] Gasnikov, A., Dvinskikh, D., Dvurechensky, P., Gorbunov, E., Beznosikov, A., Lobanov, A.: Randomized gradient-free methods in convex optimization. arXiv preprint arXiv:2211.13566 (2022)
- [12] Nemirovskij, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization (1983)
- [13] Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, Philadelphia (2021)
- [14] Nazin, A.V., Nemirovsky, A.S., Tsybakov, A.B., Juditsky, A.B.: Algorithms of robust stochastic optimization based on mirror descent method. Automation and Remote Control **80**, 1607–1627 (2019)
- [15] Davis, D., Drusvyatskiy, D., Xiao, L., Zhang, J.: From low probability to high confidence in stochastic convex optimization. The Journal of Machine Learning Research **22**(1), 2237–2274 (2021)
- [16] Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., Gasnikov, A.: Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. arXiv preprint arXiv:2106.05958 (2021)
- [17] Vural, N.M., Yu, L., Balasubramanian, K., Volgushev, S., Erdogdu, M.A.: Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In: Conference on Learning Theory, pp. 65–102 (2022). PMLR
- [18] Zhang, J., Cutkosky, A.: Parameter-free regret in high probability with heavy tails. arXiv preprint arXiv:2210.14355 (2022)
- [19] Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., Richtárik, P.: High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. arXiv preprint arXiv:2302.00999 (2023)
- [20] Nguyen, T.D., Nguyen, T.H., Ene, A., Nguyen, H.L.: High probability convergence of clipped-sgd under heavy-tailed noise. arXiv preprint arXiv:2302.05437 (2023)
- [21] Nguyen, T.D., Ene, A., Nguyen, H.L.: Improved convergence in high probability of clipped gradient methods with heavy tails. arXiv preprint arXiv:2304.01119 (2023)
- [22] Liu, Z., Zhou, Z.: Stochastic nonsmooth convex optimization with heavy-tailed noises. arXiv preprint arXiv:2303.12277 (2023)
- [23] Ermoliev, Y.: Stochastic programming methods. Nauka (1976)
- [24] Gasnikov, A., Novitskii, A., Novitskii, V., Abdukhakimov, F., Kamzolov, D.,

- Beznosikov, A., Takáč, M., Dvurechensky, P., Gu, B.: The power of first-order smooth optimization for black-box non-smooth problems. arXiv preprint arXiv:2201.12289 (2022)
- [25] Pasechnyuk, D.A., Lobanov, A., Gasnikov, A.: Upper bounds on maximum admissible noise in zeroth-order optimisation. arXiv preprint arXiv:2306.16371 (2023)
- [26] Risteski, A., Li, Y.: Algorithms and matching lower bounds for approximately-convex optimization. Advances in Neural Information Processing Systems **29** (2016)
- [27] Zhang, J., Karimireddy, S.P., Veit, A., Kim, S., Reddi, S., Kumar, S., Sra, S.: Why are adaptive methods good for attention models? Advances in Neural Information Processing Systems **33**, 15383–15393 (2020)
- [28] Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. SIAM, Philadelphia (2001)
- [29] Juditsky, A., Nesterov, Y.: Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. Stochastic Systems **4**(1), 44–80 (2014)
- [30] Akhavan, A., Chzhen, E., Pontil, M., Tsybakov, A.B.: A gradient estimator via  $l_1$ -randomization for online zero-order optimization with two point feedback. arXiv preprint arXiv:2205.13910 (2022)
- [31] Lobanov, A., Alashqar, B., Dvinskikh, D., Gasnikov, A.: Gradient-free federated learning methods with  $l_1$  and  $l_2$ -randomization for non-smooth convex stochastic optimization problems. arXiv preprint arXiv:2211.10783 (2022)
- [32] Ledoux, M.: The concentration of measure phenomenon. ed. by peter landweber et al. vol. 89. Mathematical Surveys and Monographs. Providence, Rhode Island: American Mathematical Society, 181 (2005)
- [33] Gorbunov, E., Vorontsova, E.A., Gasnikov, A.V.: On the upper bound for the expectation of the norm of a vector uniformly distributed on the sphere and the phenomenon of concentration of uniform measure on the sphere. Mathematical Notes **106** (2019)
- [34] Gasnikov, A.V., Nesterov, Y.E.: Universal method for stochastic composite optimization problems. Computational Mathematics and Mathematical Physics **58**, 48–64 (2018)

## 8 Proofs of Lemmas

### 8.1 General results

**Lemma 8.1.** 1. For all  $x, y \in \mathbb{R}^d$  and  $\kappa \in (0, 1]$ :

$$\|x - y\|_q^{1+\kappa} \leq 2^\kappa \|x\|_q^{1+\kappa} + 2^\kappa \|y\|_q^{1+\kappa}, \quad (25)$$

2.

$$\forall x, y \geq 0, \kappa \in [0, 1] : (x + y)^\kappa \leq x^\kappa + y^\kappa. \quad (26)$$

*Proof.* 1. By Jensen's inequality for convex  $\|\cdot\|_q^{1+\kappa}$  with  $1 + \kappa > 1$

$$\|x - y\|_q^{1+\kappa} = 2^{1+\kappa} \|x/2 - y/2\|_q^{1+\kappa} \leq 2^\kappa \|x\|_q^{1+\kappa} + 2^\kappa \|y\|_q^{1+\kappa}.$$

2. Proposition 9 from [17]. □

**Lemma 8.2.** Assumption 2 implies that  $f(x)$  is  $M_2$  Lipschitz on  $\mathcal{X}$ .

*Proof.* For all  $x, y \in \mathcal{X}$

$$\begin{aligned} |f(x) - f(y)| &= |\mathbb{E}_\xi[f(x, \xi) - f(y, \xi)]| \stackrel{\text{Jensen's inq}}{\leq} \mathbb{E}_\xi[|f(x, \xi) - f(y, \xi)|] \\ &\leq \mathbb{E}_\xi[M_2] \|x - y\|_2 \stackrel{\text{Jensen's inq}}{\leq} \mathbb{E}_\xi[M_2^{(1+\kappa)}]^{1/(1+\kappa)} \|x - y\|_2 \\ &\leq M_2 \|x - y\|_2. \end{aligned}$$

□

### 8.2 Smoothing

**Lemma 8.3.** Let  $f(x)$  be  $M_2$  Lipschitz continuous function w.r.t  $\|\cdot\|_2$ . If  $\mathbf{e}$  is random and uniformly distributed on the Euclidean sphere and  $\kappa \in (0, 1]$ , then

$$\mathbb{E}_{\mathbf{e}} \left[ (f(\mathbf{e}) - \mathbb{E}_{\mathbf{e}}[f(\mathbf{e})])^{2(1+\kappa)} \right] \leq \left( \frac{bM_2^2}{d} \right)^{1+\kappa}, \quad b = \frac{1}{\sqrt{2}}.$$

*Proof.* A standard result of the measure concentration on the Euclidean unit sphere implies that for all  $t > 0$

$$\Pr(|f(\mathbf{e}) - \mathbb{E}[f(\mathbf{e})]| > t) \leq 2 \exp(-b't^2/M_2^2), \quad b' = 2 \quad (27)$$

(see the proof of Proposition 2.10 and Corollary 2.6 in [32]). Therefore,

$$\mathbb{E}_{\mathbf{e}} \left[ (f(\mathbf{e}) - \mathbb{E}_{\mathbf{e}}[f(\mathbf{e})])^{2(1+\kappa)} \right] = \int_{t=0}^{\infty} \Pr(|f(\mathbf{e}) - \mathbb{E}_{\mathbf{e}}[f(\mathbf{e})]|^{2(1+\kappa)} > t) dt$$

$$\begin{aligned}
&= \int_{t=0}^{\infty} \Pr \left( |f(\mathbf{e}) - \mathbb{E}_{\mathbf{e}}[f(\mathbf{e})]| > t^{\frac{1}{2(1+\kappa)}} \right) dt \\
&\leq \int_{t=0}^{\infty} 2 \exp \left( -b' dt^{\frac{1}{(1+\kappa)}} / M_2^2 \right) dt \leq \left( \frac{bM_2^2}{d} \right)^{1+\kappa}.
\end{aligned}$$

□

The following lemma gives some useful facts about the measure concentration on the Euclidean unit sphere.

**Lemma 8.4.** For  $q \geq 2, \kappa \in (0, 1]$  we get

$$\mathbb{E}_{\mathbf{e}} \left[ \|\mathbf{e}\|_q^{2(1+\kappa)} \right] \leq a_q^{2(1+\kappa)} \stackrel{def}{=} d^{\frac{1}{q} - \frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q - 1}\}.$$

This lemma is generalization of Lemma from [33] for  $\kappa < 1$ .

*Proof.* We use Lemma 1 from Theorem 1 from [33] which states that

1. Let  $e_k$  be  $k$ -th component of  $\mathbf{e}$  then next inequality holds true

$$\mathbb{E} [|e_k|^q] \leq \left( \frac{q-1}{d} \right)^{\frac{q}{2}}, \quad q \geq 2. \quad (28)$$

2. For any  $x \in \mathbb{R}^d$  and  $q_1 \geq q_2$  we get

$$\|x\|_{q_1} \leq \|x\|_{q_2}, \quad (29)$$

We rewrite our objective value as

$$\mathbb{E}_{\mathbf{e}} \left[ \|\mathbf{e}\|_q^{2(1+\kappa)} \right] = \mathbb{E}_{\mathbf{e}} \left[ \left( \left( \sum_{k=1}^d |e_k|^q \right)^2 \right)^{\frac{1+\kappa}{q}} \right].$$

Due to Jensen's inequality and equally distributed  $e_k$  we obtain

$$\mathbb{E}_{\mathbf{e}} \left[ \left( \left( \sum_{k=1}^d |e_k|^q \right)^2 \right)^{\frac{1+\kappa}{q}} \right] \leq \left( \mathbb{E}_{\mathbf{e}} \left[ \left( \sum_{k=1}^d |e_k|^q \right)^2 \right] \right)^{\frac{1+\kappa}{q}}.$$

We use fact that for all  $x_k \geq 0, k = \overline{1, d}$

$$d \sum_{k=1}^d x_k^2 \geq \left( \sum_{k=1}^d x_k \right)^2.$$

Therefore, we estimate

$$\left( \mathbb{E}_{\mathbf{e}} \left[ \left( \sum_{k=1}^d |e_k|^q \right)^2 \right] \right)^{\frac{1+\kappa}{q}} \leq \left( d \mathbb{E}_{\mathbf{e}} \left[ \sum_{k=1}^d |e_k|^{2q} \right] \right)^{\frac{1+\kappa}{q}} = (d^2 \mathbb{E}_{\mathbf{e}}[|e_k|^{2q}])^{\frac{1+\kappa}{q}}.$$

Using (28) with  $2q$  we continue chain of previous inequalities

$$(d^2 \mathbb{E}_{\mathbf{e}}[|e_2|^{2q}])^{\frac{1+\kappa}{q}} \leq d^{\frac{2(1+\kappa)}{q}} \left( \frac{2q-1}{d} \right)^{1+\kappa} = \left( d^{\frac{2}{q}-1} (2q-1) \right)^{1+\kappa}.$$

Thus, by definition of  $a_q$  and obtained estimates we conclude

$$a_q = \sqrt{d^{\frac{2}{q}-1} (2q-1)}.$$

With fixed  $d$  and large  $q$  more precise upper bound can be obtained. We define function  $h_d(q)$  and find its minimum with fixed  $d$ .

$$h_d(q) = \ln \left( \sqrt{d^{\frac{2}{q}-1} (2q-1)} \right) = \left( \frac{1}{q} - \frac{1}{2} \right) \ln(d) + \frac{1}{2} \ln(2q-1),$$

$$\begin{aligned} \frac{dh_d(q)}{dq} &= \frac{-\ln(d)}{q^2} + \frac{1}{2q-1} = 0, \\ q^2 - 2 \ln(d)q + \ln(d) &= 0. \end{aligned}$$

When  $d \geq 3$  minimal point  $q_0$  lies in  $[2, +\infty)$

$$q_0 = (\ln d) \left( 1 + \sqrt{1 - \frac{1}{\ln d}} \right), \quad \ln d \leq q_0 \leq 2 \ln d.$$

When  $q \geq q_0$  we obtain from (29)

$$\begin{aligned} a_q &< a_{q_0} = \sqrt{d^{\frac{2}{q_0}-1} (2q_0-1)} \leq d^{\frac{1}{\ln d} - \frac{1}{2}} \sqrt{4 \ln d - 1} \\ &= \frac{e}{\sqrt{d}} \sqrt{4 \ln d - 1} \leq d^{\frac{1}{q} - \frac{1}{2}} \sqrt{32 \ln d - 8}, \end{aligned}$$

Consequently, we get

$$a_q = d^{\frac{1}{q} - \frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q-1}\}.$$

□



**Lemma 8.5.** For the random vector  $\mathbf{e}$  uniformly distributed on the Euclidean sphere  $\{\mathbf{e} \in \mathbb{R}^d : \|\mathbf{e}\|_2 = 1\}$  and for any  $r \in \mathbb{R}^d$ , we have

$$\mathbb{E}_{\mathbf{e}}[\langle \mathbf{e}, r \rangle] \leq \frac{\|r\|_2}{\sqrt{d}}.$$

**Lemma 8.6.** Let  $g(x, \xi, \mathbf{e})$  be defined in (4) and  $\hat{f}_\tau(x)$  be defined in (3). Then, the following holds under Assumption 3:

$$\mathbb{E}_{\xi, \mathbf{e}}[\langle g(x, \xi, \mathbf{e}), r \rangle] \geq \langle \nabla \hat{f}_\tau(x), r \rangle - \frac{d\Delta}{\tau} \mathbb{E}_{\mathbf{e}}[\langle \mathbf{e}, r \rangle]$$

for any  $r \in \mathbb{R}^d$ .

*Proof.* We remind that by definition (4) of estimated gradient  $g$

$$g(x, \xi, \mathbf{e}) = \frac{d}{2\tau} (f(x + \tau\mathbf{e}, \xi) + \delta(x + \tau\mathbf{e}) - f(x - \tau\mathbf{e}, \xi) - \delta(x - \tau\mathbf{e}))\mathbf{e}.$$

Then multiplying  $g$  on arbitrary  $r$  and taking full expectation from both sides we get

$$\begin{aligned} \mathbb{E}_{\xi, \mathbf{e}}[\langle g(x, \xi, \mathbf{e}), r \rangle] &= \frac{d}{2\tau} \mathbb{E}_{\xi, \mathbf{e}}[\langle (f(x + \tau\mathbf{e}, \xi) - f(x - \tau\mathbf{e}, \xi))\mathbf{e}, r \rangle] \\ &\quad + \frac{d}{2\tau} \mathbb{E}_{\xi, \mathbf{e}}[\langle (\delta(x + \tau\mathbf{e}) - \delta(x - \tau\mathbf{e}))\mathbf{e}, r \rangle]. \end{aligned}$$

In the first term we use fact that  $\mathbf{e}$  symmetrically distributed

$$\begin{aligned} \frac{d}{2\tau} \mathbb{E}_{\xi, \mathbf{e}}[\langle (f(x + \tau\mathbf{e}, \xi) - f(x - \tau\mathbf{e}, \xi))\mathbf{e}, r \rangle] &= \frac{d}{\tau} \mathbb{E}_{\xi, \mathbf{e}}[\langle f(x + \tau\mathbf{e}, \xi)\mathbf{e}, r \rangle] \\ &= \frac{d}{\tau} \mathbb{E}_{\mathbf{e}}[\langle \mathbb{E}_{\xi}[f(x + \tau\mathbf{e}, \xi)]\mathbf{e}, r \rangle] = \frac{d}{\tau} \langle \mathbb{E}_{\mathbf{e}}[f(x + \tau\mathbf{e})\mathbf{e}], r \rangle. \end{aligned} \quad (30)$$

Using Lemma 2.2 in (30) we take expectation

$$\frac{d}{\tau} \langle \mathbb{E}_{\mathbf{e}}[f(x + \tau\mathbf{e})\mathbf{e}], r \rangle = \langle \nabla \hat{f}_\tau(x), r \rangle.$$

In the second term we use Assumption 3

$$\frac{d}{2\tau} \mathbb{E}_{\xi, \mathbf{e}}[\langle (\delta(x + \tau\mathbf{e}) - \delta(x - \tau\mathbf{e}))\mathbf{e}, r \rangle] \geq -\frac{d\Delta}{\tau} \mathbb{E}_{\mathbf{e}}[\langle \mathbf{e}, r \rangle].$$

Adding two terms together we get necessary result. □

*Proof of Lemma 2.3.* By definition (4) of estimated gradient  $g$  we obtain next chain of inequalities

$$\begin{aligned}
& \mathbb{E}_{\xi, \mathbf{e}} [\|g(x, \xi, \mathbf{e})\|_q^{1+\kappa}] = \mathbb{E}_{\xi, \mathbf{e}} \left[ \left\| \frac{d}{2\tau} (\phi(x + \tau\mathbf{e}, \xi) - \phi(x - \tau\mathbf{e}, \xi)) \mathbf{e} \right\|_q^{1+\kappa} \right] \\
& = \left( \frac{d}{2\tau} \right)^{1+\kappa} \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |(f(x + \tau\mathbf{e}, \xi) - f(x - \tau\mathbf{e}, \xi) + \delta(x + \tau\mathbf{e}) - \delta(x - \tau\mathbf{e}))|^{1+\kappa}] \\
& \stackrel{(25)}{\leq} 2^\kappa \left( \frac{d}{2\tau} \right)^{1+\kappa} \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |f(x + \tau\mathbf{e}, \xi) - f(x - \tau\mathbf{e}, \xi)|^{1+\kappa}] \tag{31}
\end{aligned}$$

$$+ 2^\kappa \left( \frac{d}{2\tau} \right)^{1+\kappa} \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |\delta(x + \tau\mathbf{e}) - \delta(x - \tau\mathbf{e})|^{1+\kappa}]. \tag{32}$$

Lets deal with (31) term. Adding  $\pm\alpha(\xi)$  for any  $\alpha(\xi)$  in (31) we get

$$\begin{aligned}
& \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |f(x + \tau\mathbf{e}, \xi) - f(x - \tau\mathbf{e}, \xi)|^{1+\kappa}] \\
& \leq \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |(f(x + \tau\mathbf{e}, \xi) - \alpha) - (f(x - \tau\mathbf{e}, \xi) - \alpha)|^{1+\kappa}] \\
& \stackrel{(25)}{\leq} 2^\kappa \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |f(x + \tau\mathbf{e}, \xi) - \alpha|^{1+\kappa}] + 2^\kappa \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |f(x - \tau\mathbf{e}, \xi) - \alpha|^{1+\kappa}]. \tag{33}
\end{aligned}$$

We consider that distribution of  $\mathbf{e}$  is symmetric,

$$(33) \leq 2^{\kappa+1} \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |f(x + \tau\mathbf{e}, \xi) - \alpha|^{1+\kappa}]. \tag{34}$$

Let  $\alpha(\xi) = \mathbb{E}_{\mathbf{e}}[f(x + \tau\mathbf{e}, \xi)]$ , then because of Cauchy-Schwartz inequality and conditional expectation properties,

$$\begin{aligned}
(34) & \leq 2^{\kappa+1} \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |f(x + \tau\mathbf{e}, \xi) - \alpha|^{1+\kappa}] \\
& = 2^{\kappa+1} \mathbb{E}_{\xi} [\mathbb{E}_{\mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |f(x + \tau\mathbf{e}, \xi) - \alpha|^{1+\kappa}]] \\
& \leq 2^{\kappa+1} \mathbb{E}_{\xi} \left[ \sqrt{\mathbb{E}_{\mathbf{e}} [\|\mathbf{e}\|_q^{2(1+\kappa)}] \mathbb{E}_{\mathbf{e}} [|f(x + \tau\mathbf{e}, \xi) - \mathbb{E}_{\mathbf{e}}[f(x + \tau\mathbf{e}, \xi)]|^{2(1+\kappa)}]} \right] \tag{35}
\end{aligned}$$

Next, we use  $\mathbb{E}_{\mathbf{e}} [\|\mathbf{e}\|_q^{2(1+\kappa)}] \leq a_q^{2(1+\kappa)}$  and Lemma 8.3 for  $f(x + \tau\mathbf{e}, \xi)$  with fixed  $\xi$  and Lipschitz constant  $M_2(\xi)\tau$ ,

$$\begin{aligned}
(35) & \leq 2^{\kappa+1} a_q^{1+\kappa} \mathbb{E}_{\xi} \left[ \sqrt{\left( \frac{2^{-1/2} \tau^2 M_2^2(\xi)}{d} \right)^{1+\kappa}} \right] \\
& = 2^{\kappa+1} a_q^{1+\kappa} \left( \frac{\tau^2 2^{-1/2}}{d} \right)^{(1+\kappa)/2} \mathbb{E}_{\xi} [M_2^{1+\kappa}(\xi)]
\end{aligned}$$

$$\leq 2^{\kappa+1} \left( \sqrt{\frac{2^{-1/2}}{d}} a_q M_2 \tau \right)^{1+\kappa}. \quad (36)$$

Lets deal with (32) term. We use the Cauchy-Schwartz inequality, bounded noise Assumption 3 and inequality  $\mathbb{E}_{\mathbf{e}} [\|\mathbf{e}\|_q^{2(1+\kappa)}] \leq a_q^{2(1+\kappa)}$  that follows from the definition of  $a_q$

$$\begin{aligned} & \mathbb{E}_{\xi, \mathbf{e}} [\|\mathbf{e}\|_q^{1+\kappa} |\delta(x + \tau \mathbf{e}) - \delta(x - \tau \mathbf{e})|^{1+\kappa}] \\ & \leq \sqrt{\mathbb{E}_{\mathbf{e}} [\|\mathbf{e}\|_q^{2(1+\kappa)}] \mathbb{E}_{\mathbf{e}} [|\delta(x + \tau \mathbf{e}) - \delta(x - \tau \mathbf{e})|^{2(1+\kappa)}]} \\ & \leq a_q^{1+\kappa} 2^{1+\kappa} \Delta^{1+\kappa} = (2a_q \Delta)^{1+\kappa}. \end{aligned} \quad (37)$$

Adding(36) and (37) we get final result

$$\begin{aligned} \mathbb{E}_{\xi, \mathbf{e}} [\|g(x, \xi, \mathbf{e})\|_q^{1+\kappa}] & \leq \frac{1}{2} \left( \frac{d}{\tau} \right)^{1+\kappa} \left( 2^{1+\kappa} \left( \sqrt{\frac{2^{-1/2}}{d}} a_q \tau M_2 \right)^{1+\kappa} + (2a_q \Delta)^{1+\kappa} \right) = \\ & = 2^\kappa \left( \frac{\sqrt{d}}{2^{1/4}} a_q M_2 \right)^{1+\kappa} + 2^\kappa \left( \frac{da_q \Delta}{\tau} \right)^{1+\kappa}. \end{aligned}$$

□

## 9 Proof of Z0-RSMD in Expectation Convergence

*Proof of Theorem 3.3.* By definition  $x_* \in \arg \min_{x \in \mathcal{X}} f(x)$ .

We use Convergence Theorem 3.2 for Robust SMD Algorithm and set of update vectors  $g_k(x_k, \xi_k, \mathbf{e}_k)$

$$\frac{1}{T} \sum_{k=0}^{T-1} \langle g_{k+1}, x_k - x_* \rangle \leq \frac{\kappa}{\kappa+1} \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa} \frac{1}{T} \sum_{k=0}^{T-1} \|g_{k+1}\|_q^{1+\kappa}. \quad (38)$$

Then we take full expectation  $\mathbb{E}$  from both sides of (38)

$$\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\langle g_{k+1}, x_k - x_* \rangle] \leq \frac{\kappa}{\kappa+1} \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa} \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\|g_{k+1}\|_q^{1+\kappa}]. \quad (39)$$

Using boundness of estimated gradient  $(1+\kappa)$ -th moment from Lemma 2.3 for the right part of inequality (39) we get

$$\frac{\nu^\kappa}{1+\kappa} \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\|g_{k+1}\|_q^{1+\kappa}] \leq \frac{\nu^\kappa}{1+\kappa} \frac{1}{T} \sum_{k=0}^{T-1} \sigma_q^{1+\kappa} \leq \frac{\nu^\kappa}{1+\kappa} \sigma_q^{1+\kappa}. \quad (40)$$

Using conditional math expectation and Lemma 8.6 for the left part of inequality (39) we estimate

$$\begin{aligned} \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\langle g_{k+1}, x_k - x^* \rangle] &= \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\mathbb{E}_{|\leq k} [\langle g_{k+1}, x_k - x^* \rangle]] \\ &\geq \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle] - \frac{1}{T} \sum_{k=0}^{T-1} \frac{d\Delta}{\tau} \mathbb{E} [\mathbb{E}_{\mathbf{e}_k|\leq k} [|\langle \mathbf{e}_k, x_k - x^* \rangle|]]. \end{aligned} \quad (41)$$

1. For the first term of (41) by convexity of  $\hat{f}_\tau(x)$  we obtain

$$\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle] \geq \frac{1}{T} \sum_{k=0}^{T-1} \left( \mathbb{E} [\hat{f}_\tau(x_k)] - \hat{f}_\tau(x_*) \right).$$

Then we define  $\bar{x}_T = \frac{1}{T} \sum_{k=0}^{T-1} x_k$  and use Jensen's inequality

$$\frac{1}{T} \sum_{k=0}^{T-1} \left( \mathbb{E} [\hat{f}_\tau(x_k)] - \hat{f}_\tau(x_*) \right) \geq \mathbb{E} [\hat{f}_\tau(\bar{x}_T)] - \hat{f}_\tau(x^*).$$

Finally, we apply approximation property of  $\hat{f}_\tau(x)$  from Lemma 2.2

$$\mathbb{E} [\hat{f}_\tau(\bar{x}_T)] - \hat{f}_\tau(x^*) \geq \mathbb{E} [f(\bar{x}_T)] - f(x^*) - 2M_2\tau. \quad (42)$$

2. For the second term of (41) we use concentration measure property from Lemma 8.5 and estimate

$$\begin{aligned} -\frac{d\Delta}{T\tau} \sum_{k=0}^{T-1} \mathbb{E}_{\mathbf{e}_k|\leq k} [|\langle \mathbf{e}_k, x_k - x^* \rangle|] &\geq -\frac{d\Delta}{T\tau} \sum_{k=0}^{T-1} \frac{1}{\sqrt{d}} \|x_k - x^*\|_2 \\ &\stackrel{p \leq 2}{\geq} -\frac{d\Delta}{T\tau} \sum_{k=0}^{T-1} \frac{1}{\sqrt{d}} \|x_k - x^*\|_p. \end{aligned} \quad (43)$$

Let's notice that  $\Psi_p$  is  $(1, \frac{1+\kappa}{\kappa})$ -uniformly convex function w.r.t.  $p$  norm. Then by definition (5) we bound  $\|x_k - x^*\|_p$

$$\|x_k - x^*\|_p \leq \left( \frac{1+\kappa}{\kappa} D_{\Psi_p}(x_k, x^*) \right)^{\frac{\kappa}{1+\kappa}} \leq \sup_{x, y \in \mathcal{X}} \left( \frac{1+\kappa}{\kappa} D_{\Psi_{q^*}}(x, y) \right)^{\frac{\kappa}{1+\kappa}} = D_\Psi$$

Hence, after this bound we get

$$(43) \geq -\frac{d\Delta}{T\tau} \sum_{k=0}^{T-1} \frac{1}{\sqrt{d}} \|x_k - x^*\|_p \geq -\frac{\sqrt{d}\Delta}{\tau} \mathcal{D}_\Psi. \quad (44)$$

Next, we combine (40), (42), (44) together to obtain final estimate

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau} \mathcal{D}_\Psi + \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa} \sigma_q^{1+\kappa}. \quad (45)$$

Now we select good parameters of the Algorithm to lower right part of (45). By choosing optimal  $\nu = \frac{R_0^{\frac{1}{\kappa}}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$  we get

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau} \mathcal{D}_\Psi + 2R_0\sigma_q T^{-\frac{\kappa}{1+\kappa}}.$$

Finally, we get explicit bound of  $\sigma_q$  using Lemma 8.1

$$\sigma_q \leq 2 \left( \frac{\sqrt{d}}{2^{1/4}} a_q M_2 \right) + 2 \left( \frac{da_q \Delta}{\tau} \right).$$

And set optimal  $\tau$

$$\tau = \sqrt{\frac{\sqrt{d}\Delta \mathcal{D}_\Psi + 4R_0 da_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}.$$

□

## 10 Proof of Z0-Clip-SMD in Expectation Convergence

First, we prove some useful statements about clipped gradient vector properties. Similar proof can be found in [18].

*Proof of Lemma 4.1.* 1. By Jensen's inequality for  $\|\cdot\|_q$  and definition of  $\hat{g}$  we estimate

$$\begin{aligned} \|\hat{g} - \mathbb{E}[\hat{g}]\|_q &\leq \|\hat{g}\|_q + \|\mathbb{E}[\hat{g}]\|_q \\ &\leq \left\| \frac{g}{\|g\|_q} \min(\|g\|_q, c) \right\|_q + \mathbb{E} \left[ \left\| \frac{g}{\|g\|_q} \min(\|g\|_q, c) \right\|_q \right] \\ &= \min(\|g\|_q, c) + \mathbb{E}[\min(\|g\|_q, c)] \\ &\leq c + c = 2c. \end{aligned}$$

2.(a) Considering  $\mathbb{E}[\|g(x, \xi, \mathbf{e})\|_q^{1+\kappa}] \leq \sigma_q^{1+\kappa}$  and  $\|\hat{g}\|_q \leq c$  we get

$$\mathbb{E}[\|\hat{g}\|_q^{1+\kappa} \|\hat{g}\|_q^{1-\kappa}] \leq \sigma_q^{1+\kappa} c^{1-\kappa}.$$

(b) By Jensen's inequality for  $\|\cdot\|_q$  we obtain

$$\begin{aligned}\mathbb{E}[\|\hat{g} - \mathbb{E}[\hat{g}]\|_q^2] &\leq 2\mathbb{E}[\|\hat{g}\|_q^2] + 2\|\mathbb{E}[\hat{g}]\|_q^2 \\ &\leq 2\mathbb{E}[\|\hat{g}\|_q^2] + 2\mathbb{E}[\|\hat{g}\|_q^2] \\ &\stackrel{(14)}{\leq} 2\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} + 2\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} \leq 4\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa}.\end{aligned}$$

(c) Due to convexity of norm function and Jensen's inequality we estimate

$$\|\mathbb{E}[g] - \mathbb{E}[\hat{g}]\|_q \leq \mathbb{E}[\|g - \hat{g}\|_q] \leq \mathbb{E}[\|g\|_q \mathbf{1}_{\{\|g\|_q > c\}}].$$

Final result follows from  $\|g\|_q^{1+\kappa} \mathbf{1}_{\{\|g\|_q > c\}} \geq \|g\|_q c^\kappa \mathbf{1}_{\{\|g\|_q > c\}}$

$$\mathbb{E}[\|g\|_q \mathbf{1}_{\{\|g\|_q > c\}}] \leq \mathbb{E}[\|g\|_q^{1+\kappa} \mathbf{1}_{\{\|g\|_q > c\}}] \leq \frac{\sigma_{q,\kappa}^{1+\kappa}}{c^\kappa}.$$

□

*Proof of the Theorem 4.2.* Let us note from first term of (41) in the proof of Theorem 3.3 that for any  $x_k$

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle + 2M_2\tau. \quad (46)$$

Next, we define functions

$$l_k(x) \stackrel{\text{def}}{=} \langle \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x - x^* \rangle.$$

Note that  $l_k(x)$  is convex for any  $k$  and  $\nabla l_k(x) = \mathbb{E}_{|\leq k}[\hat{g}_{k+1}]$ . Therefore sampled estimation gradient is unbiased. With these functions we can rewrite the right part of (46) as follows

$$\begin{aligned}&\frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle + 2M_2\tau \\ &= \frac{1}{T} \sum_{k=0}^{T-1} \left( \langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle \right) + \frac{1}{T} \sum_{k=0}^{T-1} (l_k(x_k) - l_k(x^*)) + 2M_2\tau.\end{aligned} \quad (47)$$

Then we take full expectation from both sides of (47)

$$\mathbb{E} \left[ \frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle \right] + 2M_2\tau$$

$$\begin{aligned}
&= \mathbb{E} \left[ \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \left( \langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle \right)}_D \right] \\
&+ \mathbb{E} \left[ \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} (l_k(x_k) - l_k(x^*))}_E \right] + 2M_2\tau.
\end{aligned}$$

We add  $\pm \mathbb{E}_{|\leq k}[g_{k+1}]$  in D term and get

$$\begin{aligned}
&\mathbb{E} \left[ \frac{1}{T} \sum_{k=0}^{T-1} \left( \langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle \right) \right] \\
&= \mathbb{E} \left[ \frac{1}{T} \sum_{k=0}^{T-1} \langle \mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle \right] \\
&+ \mathbb{E} \left[ \frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[g_{k+1}], x_k - x^* \rangle \right]. \tag{48}
\end{aligned}$$

In order to bound the first term of (48) let's notice that  $\Psi_p$  is  $(1, 2)$ -uniformly convex function w.r.t.  $p$  norm. Then by definition (5) we bound  $\|x_k - x^*\|_p$

$$\|x_k - x^*\|_p \leq (2D_{\Psi_p}(x_k, x^*))^{\frac{1}{2}} \leq \sup_{x, y \in \mathcal{X}} (2D_{\Psi_p}(x, y))^{\frac{1}{2}} = \mathcal{D}_\Psi,$$

and estimate  $\|x_k - u\|_p \leq \mathcal{D}_\Psi$ , for all  $u \in \mathcal{X}$ .

We apply the Cauchy-Schwarz inequality to inner product in the first term of (48)

$$\begin{aligned}
&\mathbb{E} \left[ \frac{1}{T} \sum_{k=0}^{T-1} \left( \langle \mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle \right) \right] \\
&\leq \frac{1}{T} \sum_{k=0}^{T-1} \left( \mathbb{E} \left[ \mathbb{E}_{|\leq k} \left[ \|\mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}]\|_q \|x_k - x^*\|_p \right] \right] \right) \stackrel{(16)}{\leq} \mathcal{D}_\Psi \frac{\sigma_q^{1+\kappa}}{c^\kappa}. \tag{49}
\end{aligned}$$

To bound the second term in (48) we use Lemma 8.6 and Lemma 8.5

$$\begin{aligned}
&\mathbb{E} \left[ \frac{1}{T} \sum_{k=0}^{T-1} \left( \langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[g_{k+1}], x_k - x^* \rangle \right) \right] \\
&\leq \frac{1}{T} \sum_{k=0}^{T-1} \frac{d\Delta}{\tau} \mathbb{E} \left[ \mathbb{E}_{\mathbf{e}|\leq k} [|\langle \mathbf{e}, x_k - x^* \rangle|] \right] \\
&\leq \frac{1}{T} \sum_{k=0}^{T-1} \frac{d\Delta}{\tau} \frac{1}{\sqrt{d}} \mathbb{E} [\|x_k - x^*\|_2]
\end{aligned}$$

$$\stackrel{p \leq 2}{\geq} \frac{1}{T} \sum_{k=0}^{T-1} \frac{d\Delta}{\tau} \frac{1}{\sqrt{d}} \mathbb{E}[\|x_k - x^*\|_p] \leq \frac{\Delta\sqrt{d}}{\tau} \mathcal{D}_\Psi. \quad (50)$$

Next, we bound E term. First of all, we rewrite it as

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{T} \sum_{k=0}^{T-1} (l_k(x_k) - l_k(x^*)) \right] &= \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\mathbb{E}_{|\leq k}[\langle \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle]] \\ &= \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\mathbb{E}_{|\leq k}[\langle \hat{g}_{k+1}, x_k - x^* \rangle]]. \end{aligned}$$

For the Robust SMD Algorithm with update vectors  $\hat{g}_k$  by Convergence Theorem 3.2 with bounded second moment next inequality holds true

$$\frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1}, x_k - x^* \rangle \leq \frac{1}{2} \frac{R_0^2}{\nu T} + \frac{\nu}{2} \frac{1}{T} \sum_{k=0}^{T-1} \|\hat{g}_{k+1}\|_q^2. \quad (51)$$

Taking  $\mathbb{E}$  from both sides of (51) we get

$$\begin{aligned} \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E}[\langle \hat{g}_{k+1}, x_k - x^* \rangle] &= \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\mathbb{E}_{|\leq k}[\langle \hat{g}_{k+1}, x_k - x^* \rangle]] \\ &\leq \frac{1}{2} \frac{R_0^2}{\nu T} + \frac{\nu}{2} \frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E} [\mathbb{E}_{|\leq k}[\|\hat{g}_{k+1}\|_q^2]]. \end{aligned}$$

By (14) from Lemma 4.1 we bound second moment of clipped gradient

$$\mathbb{E}_{|\leq k}(\|\hat{g}_{k+1}\|_q^2) \leq \sigma_q^{1+\kappa} c^{1-\kappa},$$

And hence get,

$$\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E}[\langle \hat{g}_{k+1}, x_k - x^* \rangle] \leq \frac{1}{2} \frac{R_0^2}{\nu T} + \frac{\nu}{2} \sigma_q^{1+\kappa} c^{1-\kappa}. \quad (52)$$

Combining bounds (49), (50), (52) together, we obtain final estimate

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{1}{2} \frac{R_0^2}{\nu T} + \frac{\nu}{2} \sigma_q^{1+\kappa} c^{1-\kappa} + \left( \frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta \frac{\sqrt{d}}{\tau} \right) \mathcal{D}_\Psi.$$

In order to get minimal upper bound we find optimal parameters. First, we choose  $c$  by finding minimum of

$$\min_{c>0} \sigma_q^{1+\kappa} \left( \frac{1}{c^\kappa} \mathcal{D}_\Psi + \frac{\nu}{2} c^{1-\kappa} \right) = \min_c \sigma_q^{1+\kappa} h_1(c)$$



$$h'_1(c) = \frac{\nu}{2}(1-\kappa)c^{-\kappa} - \kappa \frac{1}{c^{1+\kappa}} \mathcal{D}_\Psi = 0 \Rightarrow c^* = \frac{2\kappa \mathcal{D}_\Psi}{(1-\kappa)\nu}.$$

$$\begin{aligned} \mathbb{E}[f(\bar{x}_T)] - f(x^*) &\leq 2M_2\tau + \frac{1}{2} \frac{R_0^2}{\nu T} + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi \\ &\quad + \sigma_q^{1+\kappa} \left( \mathcal{D}^{1-\kappa} 2^{-\kappa} \nu^\kappa \left[ \frac{(1-\kappa)^\kappa}{\kappa^\kappa} + \frac{\kappa^{(1-\kappa)}}{(1-\kappa)^{(1-\kappa)}} \right] \right). \end{aligned} \quad (53)$$

Considering bound of  $\kappa \in [0, 1]$  and as consequence

$$\left[ \frac{(1-\kappa)^\kappa}{\kappa^\kappa} + \frac{\kappa^{(1-\kappa)}}{(1-\kappa)^{(1-\kappa)}} \right] \leq 2,$$

we simplify (53)

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{1}{2} \frac{R_0^2}{\nu T} + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi + \sigma_q^{1+\kappa} (2\mathcal{D}_\Psi^{1-\kappa} \nu^\kappa). \quad (54)$$

Choosing optimal  $\nu^*$  similarly we get

$$\nu^* = \left( \frac{R_0^2}{4T\kappa\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}} \right)^{\frac{1}{1+\kappa}}$$

And

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi + \frac{R_0^{\frac{2\kappa}{1+\kappa}} \mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}} \sigma_q}{T^{\frac{\kappa}{1+\kappa}}} 2 \left[ \kappa^{\frac{1}{1+\kappa}} + \kappa^{-\frac{\kappa}{1+\kappa}} \right].$$

Considering bound of  $\kappa \in [0, 1]$  next inequality holds true

$$\left[ \kappa^{\frac{1}{1+\kappa}} + \kappa^{-\frac{\kappa}{1+\kappa}} \right] \leq 2.$$

Thus, we can simplify upper bound even more

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi + 2 \frac{R_0^{\frac{2\kappa}{1+\kappa}} \mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}} \sigma_q}{T^{\frac{\kappa}{1+\kappa}}}. \quad (55)$$

In order to avoid  $\nu \rightarrow \infty$  when  $\kappa \rightarrow 0$  one can also choose  $\nu^* = \left( \frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}} \right)^{\frac{1}{1+\kappa}}$ .

Estimation (55) does not change.

Finally, we get explicit bound  $\sigma_q$  with Lemma 8.1

$$\sigma_q \leq 2 \left( \frac{\sqrt{d}}{2^{1/4}} a_q M_2 \right) + 2 \left( \frac{da_q \Delta}{\tau} \right),$$

And set optimal  $\tau$

$$\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}da_q\Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}.$$

□

## 11 Proof of Z0-Clip-SMD in High Probability Convergence

Before we turn to the proof of Theorem 4.3, we present two lemmas representing the Bernstein inequality for the sum of martingale differences and the sum of squares of bounded random variables. These are some classical results on measure concentration.

**Lemma 11.1** (Lemma 23 from [18]). *Let  $\{X_i\}_{i \geq 1}$  be martingale difference sequence, i.e.  $\mathbb{E}[X_i|X_{i-1}, \dots, X_1] = 0$  for all  $i \geq 1$ . Also  $b, \sigma$  is such deterministic constants that  $|X_i| < b$  almost surely and  $\mathbb{E}[X_i^2|X_{i-1}, \dots, X_1] < \sigma^2$  for  $i \geq 1$ . Then for arbitrary fixed number  $\mu$  and for all  $T$  with probability at least  $1 - \delta$  next inequality holds true*

$$\left| \sum_{i=1}^t \mu X_i \right| \leq 2b|\mu| \log \frac{1}{\delta} + \sigma|\mu| \sqrt{2T \log \frac{1}{\delta}}.$$

**Lemma 11.2** (Theorem 20 from [18]). *Let  $Z_i$  be a sequence of random variables adapted to a filtration  $\mathcal{F}_i$ . Further, suppose  $|Z_i| < b$  almost surely and  $\mathbb{E}[Z_i^2] \leq \sigma^2$ . Then for any  $\mu > 0$  with probability at least  $1 - \delta$  next inequality holds true*

$$\begin{aligned} \sum_{k=1}^T Z_k^2 &\leq 3T\sigma^2 \log \left( \frac{4}{\delta} \left[ \log \left( \sqrt{\frac{\sigma^2 T}{\mu^2}} \right) + 2 \right]^2 \right) \\ &\quad + 20 \max(\mu^2, b^2) \log \left( \frac{112}{\delta} \left[ \log \left( \frac{2 \max(\mu, b)}{\mu} \right) + 1 \right]^2 \right). \end{aligned}$$

By choosing  $\mu = b \geq \sigma$  we simplify

$$\sum_{k=1}^T Z_k^2 \leq 3T\sigma^2 \log \left( \frac{4}{\delta} \left[ \log(\sqrt{T}) + 2 \right]^2 \right) + 20b^2 \log \left( \frac{12}{\delta} \right).$$

*Proof of the Theorem 4.3.* Lets notice from the first term of (41) in the proof of Theorem 3.3 that for any  $x_k$  next inequality holds true

$$f(\bar{x}_T) - f(x^*) \leq \frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle + 2M_2\tau. \quad (56)$$

For the Robust SMD Algorithm with update vectors  $\hat{g}_k$  Convergence Theorem 3.2 with the bounded second moment guarantees that

$$\frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1}, x_k - x^* \rangle \leq \frac{1}{2} \frac{R_0^2}{\nu T} + \frac{\nu}{2} \frac{1}{T} \sum_{k=0}^{T-1} \|\hat{g}_{k+1}\|_q^2. \quad (57)$$

Let's define random variable  $Z_k = \|\hat{g}_{k+1}\|_q$  and notice that  $|Z_k| \leq c$  by definition of clipping and  $\mathbb{E}[Z_i^2] \leq 4\sigma_{q,\kappa}^{1+\kappa} c^{1-\kappa}$  by (15) from clipped gradient properties Lemma 4.1. Thus, we can apply Lemma 11.2 and with probability at least  $1 - \delta$  bound mean sum of the clipped gradients second moments

$$\frac{1}{T} \sum_{k=0}^{T-1} \|\hat{g}_{k+1}\|_q^2 \leq 12\sigma_{q,\kappa}^{1+\kappa} c^{1-\kappa} \log\left(\frac{4}{\delta} \left[\log(\sqrt{T}) + 2\right]^2\right) + \frac{20}{T} c^2 \log\left(\frac{12}{\delta}\right). \quad (58)$$

The left part of (57) can be rewritten as

$$\begin{aligned} \frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1}, x_k - x^* \rangle &= \frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1} - \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle + \frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle \\ &= \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \langle \hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle}_{\textcircled{1}} + \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \langle \mathbb{E}_{|\leq k}[\hat{g}_{k+1}] - \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle}_{\textcircled{2}} \\ &\quad + \underbrace{\frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^* \rangle}_{\textcircled{3}}. \end{aligned}$$

In the  $\textcircled{1}$  term we can proof that this is the sum of the martingale sequence difference. Indeed, we notice that  $x_k$  is fixed when we take  $\mathbb{E}_{|\leq k}$  and martingale property holds true, i.e.

$$\mathbb{E}_{|\leq k}[\langle \hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle] = 0.$$

By Lemma 4.1 we bound each element of martingale sequence

$$|\langle \hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle| \leq \|\hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}]\|_q \|x_k - x^*\|_p \leq 2c \cdot \|x_k - x^*\|_p.$$

And by (15) from Lemma 4.1 we bound expectation of square of each element

$$\mathbb{E} [|\langle \hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle|^2] \leq 4\sigma_q^{1+\kappa} c^{1-\kappa} \cdot \|x_k - x^*\|_p^2.$$

Lets notice that  $\Psi_p$  is (1,2)-uniformly convex function w.r.t.  $p$  norm. Then by definition (5) we bound

$$\|x_k - x^*\|_p \leq (2D_{\Psi_p}(x_k, x^*))^{\frac{1}{2}} \leq \sup_{x, y \in \mathcal{X}} (2D_{\Psi_p}(x, y))^{\frac{1}{2}} = \mathcal{D}_{\Psi},$$

and estimate  $\|x_k - u\|_p \leq \mathcal{D}$  for all  $u \in \mathcal{X}$ . Hence, we can apply Bernstein's inequality Lemma 11.1 and get with probability at least  $1 - \delta$  and  $\mu = \frac{1}{T}$  that

$$\frac{1}{T} \sum_{k=0}^{T-1} |\langle \hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^* \rangle| \leq \frac{4c\mathcal{D}_{\Psi}}{T} \log \frac{1}{\delta} + \frac{\sqrt{4\sigma_q^{1+\kappa} c^{1-\kappa}}}{\sqrt{T}} \mathcal{D}_{\Psi}^2 \sqrt{2 \log \frac{1}{\delta}}. \quad (59)$$

For the (2) we use bound of D term from (48) in the proof of Theorem 4.2

$$|\langle \mathbb{E}_{|\leq k}[\hat{g}_{k+1}] - \nabla \hat{f}_{\tau}(x_k), x_k - x^* \rangle| \leq \left( \frac{\sigma_q^{1+\kappa}}{c^{\kappa}} + \Delta \frac{\sqrt{d}}{\tau} \right) \mathcal{D}_{\Psi}. \quad (60)$$

For the (3) we use already obtained bound (56)

$$f(\bar{x}_T) - f(x^*) - 2M_2\tau \leq \frac{1}{T} \sum_{k=0}^{T-1} \langle \nabla \hat{f}_{\tau}(x_k), x_k - x^* \rangle. \quad (61)$$

Putting (58), (59), (60), (61) in (57), we get with probability at least  $1 - \delta$  that

$$\begin{aligned} f(\bar{x}_T) - f(x^*) &\leq 2M_2\tau + \left( \frac{\sigma_q^{1+\kappa}}{c^{\kappa}} + \Delta \frac{\sqrt{d}}{\tau} \right) \mathcal{D}_{\Psi} + \frac{1}{2} \frac{R_0^2}{\nu T} \\ &\quad + \frac{\nu}{2} \left[ 12\sigma_q^{1+\kappa} c^{1-\kappa} \log \left( \frac{4}{\delta} \left[ \log(\sqrt{T}) + 2 \right]^2 \right) \right] \\ &\quad + \frac{\nu}{2} \frac{20}{T} c^2 \log \left( \frac{12}{\delta} \right) + \frac{4c\mathcal{D}_{\Psi}}{T} \log \frac{1}{\delta} + \frac{\sqrt{4\sigma_q^{1+\kappa} c^{1-\kappa}}}{\sqrt{T}} \mathcal{D}_{\Psi}^2 \sqrt{2 \log \frac{1}{\delta}}. \end{aligned} \quad (62)$$

Next we select optimal parameter in order to minimize upper bound. Choosing  $c = T^{\frac{1}{1+\kappa}} \sigma_q$  and putting it in (62), we get

$$\begin{aligned} f(\bar{x}_T) - f(x^*) &\leq 2M_2\tau + \left( \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} + \Delta \frac{\sqrt{d}}{\tau} \right) \mathcal{D}_{\Psi} + \frac{1}{2} \frac{R_0^2}{\nu T} \\ &\quad + \frac{\nu}{2} \left[ 12\sigma_q^2 T^{\frac{1-\kappa}{1+\kappa}} \log \left( \frac{4}{\delta} \left[ \log(\sqrt{T}) + 2 \right]^2 \right) \right] \\ &\quad + \frac{\nu}{2} \frac{20\sigma_q^2}{T^{\frac{\kappa-1}{1+\kappa}}} \log \left( \frac{12}{\delta} \right) + \frac{4\sigma_q \mathcal{D}_{\Psi}}{T^{\frac{\kappa}{1+\kappa}}} \log \frac{1}{\delta} + \frac{2\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} \mathcal{D}_{\Psi} \sqrt{2 \log \frac{1}{\delta}}. \end{aligned} \quad (63)$$

Then we define  $\tilde{\delta}^{-1} = \frac{4}{\delta} \left[ \log(\sqrt{T}) + 2 \right]^2$ , choose  $\nu = \frac{\mathcal{D}_\Psi}{c}$ , put it in (63) and obtain

$$\begin{aligned} f(\bar{x}_T) - f(x^*) &\leq 2M_2\tau + \left( \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} + \Delta \frac{\sqrt{d}}{\tau} \right) \mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi \sigma_q}{2T^{\frac{\kappa}{1+\kappa}}} \left[ 1 + 12 \log \frac{1}{\tilde{\delta}} + 20 \log \frac{4}{\tilde{\delta}} \right] \\ &\quad + \frac{4\sigma_q \mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}} \log \frac{1}{\tilde{\delta}} + \frac{2\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} \mathcal{D}_\Psi \sqrt{2 \log \frac{1}{\tilde{\delta}}}. \end{aligned} \quad (64)$$

Simplifying (64), we get

$$\begin{aligned} f(\bar{x}_T) - f(x^*) &\leq 2M_2\tau + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi \\ &\quad + \frac{\mathcal{D}_\Psi \sigma_q}{2T^{\frac{\kappa}{1+\kappa}}} \left[ 3 + 8 \log \frac{1}{\tilde{\delta}} + 12 \log \frac{1}{\tilde{\delta}} + 20 \log \frac{4}{\tilde{\delta}} + 4 \sqrt{2 \log \frac{1}{\tilde{\delta}}} \right]. \end{aligned}$$

Finally, we get explicit bound of  $\sigma_q$  with Lemma 8.1

$$\sigma_q \leq 2 \left( \frac{\sqrt{d}}{2^{1/4}} a_q M_2 \right) + 2 \left( \frac{d a_q \Delta}{\tau} \right),$$

And set optimal  $\tau$

$$\tau = \sqrt{\frac{\sqrt{d} \Delta \mathcal{D}_\Psi + 2\beta \mathcal{D}_\Psi d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}.$$

□

## 12 Sketch of Proof of Z0-Restarts Convergence

*Proof of Theorems 5.1, 5.2.* In this proof  $\tilde{O}(\cdot)$  denotes  $\log d$  factor.

### Step 1: Z0-RSMD in Expectation.

Now  $x_0$  in Algorithm 1 can be chosen in stochastic way.

Similarly to proof of Theorem 3.3 but with  $\nu = \frac{\mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{1}{1+\kappa}}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$  and bound  $R_0 \leq \mathcal{D}_\Psi$  one can get from (45)

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau} \mathcal{D}_\Psi + 2\mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{\kappa}{1+\kappa}} \sigma_q T^{-\frac{\kappa}{1+\kappa}}. \quad (65)$$

Under obligatory condition  $\Delta \leq \frac{\sigma_q^2 \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{\kappa}{1+\kappa}}}{M_2 \sqrt{dT}^{\frac{2\kappa}{1+\kappa}}}$  picking  $\tau = \frac{\sigma_q \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{\kappa}{1+\kappa}}}{M_2 T^{\frac{\kappa}{1+\kappa}}}$ , we obtain from (65) estimate

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq (2 + 1 + 2) \frac{\sigma_q \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \quad (66)$$

In  $\sigma_q$   $\tau$ -depending term has  $T^{\frac{-2\kappa}{1+\kappa}}$  decreasing rate, so we neglect it. Next, let's use fact that  $D_{\Psi_p}(x^*, x_0) = \tilde{O}(\|x_0 - x^*\|_p^{\frac{1+\kappa}{\kappa}})$  from [34](Remark 3) and denote  $R_k = \mathbb{E} \left[ \|\bar{x}_k - x^*\|_p^{\frac{1+\kappa}{\kappa}} \right]^{\frac{\kappa}{1+\kappa}}$ .

Under  $r$ -growth Assumption 4 we bound  $\mathbb{E}[f(\bar{x}_T)] - f(x^*)$  from both sides

$$\frac{\mu_r}{2} \mathbb{E} [\|\bar{x}_T - x^*\|_p^r] \leq \mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq \tilde{O} \left( R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} \right). \quad (67)$$

Due to Jensen's inequality which we can apply since  $r \geq \frac{1+\kappa}{\kappa}$  we rewrite (67) in order to obtain  $R_1$  in it as

$$\frac{\mu_r}{2} \mathbb{E} \left[ \|\bar{x}_T - x^*\|_p^{\frac{1+\kappa}{\kappa}} \right]^{r/\frac{1+\kappa}{\kappa}} \leq \frac{\mu_r}{2} \mathbb{E} [\|\bar{x}_T - x^*\|_p^r] \leq \tilde{O} \left( R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} \right). \quad (68)$$

Let's find out after how many iterations  $R_0$  value halves

$$\frac{\mu_r}{2} R_1^r \leq \tilde{O} \left( R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} \right) \leq \frac{\mu_r}{2} \left( \frac{R_0}{2} \right)^r. \quad (69)$$

From right inequality of (69) we obtain number of iterations for one stage

$$T_1 \geq \tilde{O} \left( \left( \frac{2^{(1+r)} \sigma_q}{\mu_r} \right)^{\frac{1+\kappa}{\kappa}} \frac{1}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}} \right).$$

For convenience we define  $A \stackrel{\text{def}}{=} \frac{2^{(1+r)} \sigma_q}{\mu_r}$ .

After  $T_1$  iterations we restart algorithm with starting point  $x_0 = \bar{x}_{T_1}$  and  $R_k = R_{k-1}/2 = R_0/2^k$ .

After  $N$  restarts total number of iterations  $T$  will be

$$\begin{aligned} T &= \sum_{k=1}^N T_k = \tilde{O} \left( \frac{A^{\frac{1+\kappa}{\kappa}}}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}} \sum_{k=0}^{N-1} 2^{k \left( \frac{(r-1)(1+\kappa)}{\kappa} \right)} \right) \\ &= \tilde{O} \left( \frac{A^{\frac{(1+\kappa)}{\kappa}}}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}} \left[ 2^{N \left( \frac{(r-1)(1+\kappa)}{\kappa} \right)} - 1 \right] \right). \end{aligned} \quad (70)$$

On the last stage we can get bound with number of restarts  $N$  in it

$$\begin{aligned} \mathbb{E}[f(x_{\text{final}})] - f(x^*) &\leq \varepsilon = \tilde{O} \left( R_{N-1} \frac{\sigma_q}{T_N^{\frac{\kappa}{1+\kappa}}} \right) \\ &\leq \tilde{O} \left( \frac{\mu_r}{2} \left( \frac{R_{N-1}}{2} \right)^r \right) \leq \tilde{O} \left( \frac{\mu_r}{2} \frac{R_0^r}{2^{(N-1)r}} \right). \end{aligned}$$

Consequently, in order to get  $\varepsilon$  accuracy we need  $N$  restarts and total number of iterations  $T$ , where

$$N = \tilde{O}\left(\frac{1}{r} \log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right), \quad (71)$$

$$T = \tilde{O}\left(\left[\frac{2^{\frac{r^2+1}{r}} \sigma_q}{\mu_r^{1/r}} \cdot \frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad T_k = \tilde{O}\left(\left[\frac{\sigma_q 2^{(1+r)}}{\mu_r R_0^{r-1}} 2^{k(r-1)}\right]^{\frac{1+\kappa}{\kappa}}\right). \quad (72)$$

In each restart section we get different bounds for noise absolute value. From  $T_k$  formula from (70) we get bound

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2 \sqrt{d}} \frac{1}{2^{k(2r-1)}}\right). \quad (73)$$

Hence,  $\Delta_k$  will be the smallest on the last iteration, when  $k = N$ , i.e.

$$\Delta_N = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2 \sqrt{d}} \varepsilon^{(2-1/r)}\right).$$

### Step 2: ZO-Clip-SMD in Expectation.

Now  $x_0$  in Algorithm 2 can be chosen in stochastic way.

Similarly to proof of Theorem 4.2 but with  $\nu^* = \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{1}{2}} \left(\frac{1}{4T\sigma_q^{1+\kappa}}\right)^{\frac{1}{1+\kappa}}$ ,  $c^* = \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{1}{2}} / \nu^*$  one can get from (54)

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq 2M_2\tau + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_{\Psi} + 2 \frac{\sigma_q \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}. \quad (74)$$

Under obligatory condition  $\Delta \leq \frac{\sigma_q^2 \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{1}{2}}}{M_2 \sqrt{dT}^{\frac{2\kappa}{1+\kappa}}}$  picking  $\tau = \frac{\sigma_q \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{1}{2}}}{M_2 T^{\frac{\kappa}{1+\kappa}}}$ , we obtain from (74) estimate

$$\mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq (2 + 1 + 2) \frac{\sigma_q \mathbb{E}[D_{\Psi_p}(x^*, x_0)]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}.$$

In  $\sigma_q$   $\tau$ -depending term has  $T^{-\frac{2\kappa}{1+\kappa}}$  decreasing rate, so we neglect it. Next, let's use fact that  $D_{\Psi_p}(x^*, x_0) = \tilde{O}(\|x_0 - x^*\|_p^2)$  from [34](Remark 3) and denote  $R_k = \mathbb{E}[\|\bar{x}_k - x^*\|_p^2]^{\frac{1}{2}}$ .

Under  $r$ -growth Assumption 4 we bound  $\mathbb{E}[f(\bar{x}_T)] - f(x^*)$  from both sides

$$\frac{\mu_r}{2} \mathbb{E}[\|\bar{x}_T - x^*\|_{q^*}^r] \leq \mathbb{E}[f(\bar{x}_T)] - f(x^*) \leq \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right).$$

Due to Jensen's inequality which we can apply since  $r \geq 2$  we obtain

$$\frac{\mu_r}{2} \mathbb{E} [\|\bar{x}_T - x^*\|_{q^*}^2]^{r/2} \leq \frac{\mu_r}{2} \mathbb{E} [\|\bar{x}_T - x^*\|_{q^*}^r] \leq \tilde{O} \left( R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} \right).$$

Next part of the proof is the same from **Step 1** starting from (68). Analogically, we get the same  $T_2, N_2$  and noise bounds from (72), (71) and (73) correspondingly.

**Step 3: ZO-Clip-SMD in High Probability.**

Now  $x_0$  in Algorithm 2 can be chosen in stochastic way.

Important moment about convergence in high probability in restart setup is to control final probability. Let number of restarts be  $N_3$ , if each restart has probability to be in bounds at least  $1 - \delta/N_3$  then final probability to be in bounds will be greater than  $1 - \delta$  which is probability of 'all restarts to be in bounds'. Usually  $N_3 \sim \log(\frac{1}{\delta})$ , thus

$$\log \frac{N_3}{1} = \log \log \frac{1}{\delta} \ll \log \frac{1}{\delta} \frac{1}{\varepsilon^{\frac{1+\kappa}{\kappa}}}.$$

It means that we can use  $\log \frac{1}{\delta}$  instead of  $\log \frac{N_3}{\delta}$ .

Similarly to proof of Theorem 4.3 but  $\nu^* = [D_{\Psi_p}(x^*, x_0)]^{1/2} \left( \frac{1}{T \sigma_q^{\frac{\kappa}{1+\kappa}}} \right)^{\frac{1}{1+\kappa}}$ ,  $c^* = \mathbb{E} [D_{\Psi_p}(x^*, x_0)]^{\frac{1}{2}} / \nu^*$  one can get from (63) with probability at least  $1 - \delta/N_3$

$$\begin{aligned} f(\bar{x}_T) - f(x^*) &\leq 2M_2\tau + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_{\Psi} \\ &\quad + \frac{[D_{\Psi_p}(x^*, x_0)]^{1/2} \sigma_q}{2T^{\frac{\kappa}{1+\kappa}}} \left[ 3 + 8 \log \frac{1}{\delta} + 12 \log \frac{1}{\delta} + 20 \log \frac{4}{\delta} + 4 \sqrt{2 \log \frac{1}{\delta}} \right]. \end{aligned}$$

Denote  $\tilde{\delta}^{-1} = \frac{4}{\delta} \left[ \log(\sqrt{T}) + 2 \right]^2$ ,  $\beta = \left[ 3 + 8 \log \frac{1}{\delta} + 12 \log \frac{1}{\delta} + 20 \log \frac{4}{\delta} + 4 \sqrt{2 \log \frac{1}{\delta}} \right]$ .

Under obligatory condition  $\Delta \leq \frac{\beta^2 \sigma_q^2 D_{\Psi_p}^{\frac{1}{2}}(x^*, x_0)}{M_2 \sqrt{dT}^{\frac{2\kappa}{1+\kappa}}}$  picking  $\tau = \frac{\beta \sigma_q D_{\Psi_p}^{\frac{1}{2}}(x^*, x_0)}{M_2 T^{\frac{\kappa}{1+\kappa}}}$ , we obtain estimate

$$f(\bar{x}_T) - f(x^*) \leq (2 + 1 + 1) \frac{\sigma_q \beta [D_{\Psi_p}(x^*, x_0)]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}.$$

In  $\sigma_q$   $\tau$ -depending term has  $T^{\frac{-2\kappa}{1+\kappa}}$  decreasing rate, so we neglect it. Next, let's use fact that  $D_{\Psi_p}(x^*, x_0) = \tilde{O}(\|x_0 - x^*\|_p^2)$  from [34](Remark 3) and denote  $R_k = \|\bar{x}_k - x^*\|_p$ .

Under  $r$ -growth Assumption 4 we get

$$\frac{\mu_r}{2} \|\bar{x}_T - x^*\|_p^r \leq f(\bar{x}_T) - f(x^*) \leq \tilde{O} \left( R_0 \frac{\sigma_q \beta}{T^{\frac{\kappa}{1+\kappa}}} \right).$$



For  $r > 1$  next part of the proof is the same from **Step 1** starting from (68) with

$$A \stackrel{\text{def}}{=} \frac{2^{(1+r)}\beta\sigma_q}{\mu_r}.$$

Analogically, we get  $T_3, N_3$  and noise bounds from (72), (71) and (73) correspondingly.

$$N = \tilde{O}\left(\frac{1}{r} \log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right),$$

$$T = \tilde{O}\left(\left[\frac{2^{\frac{r^2+1}{r}}\sigma_q\beta}{\mu_r^{1/r}} \frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad T_k = \tilde{O}\left(\left[\frac{\sigma_q\beta 2^{(1+r)}}{\mu_r R_0^{r-1}} 2^{k(r-1)}\right]^{\frac{1+\kappa}{\kappa}}\right). \quad (75)$$

In each restart section we get different bounds for noise absolute value. From  $T_k$  formula from (75)

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2 \sqrt{d}} \frac{1}{2^{k(2r-1)}}\right).$$

Hence,  $\Delta_k$  will be the smallest on the last iteration, when  $k = N$ , i.e.

$$\Delta_N = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2 \sqrt{d}} \varepsilon^{(2-1/r)}\right).$$

In case of  $r = 1$  number of iterations  $T_k$  at step  $k$  from (75) changes as

$$T_k = A^{\frac{1+\kappa}{\kappa}},$$

and we do not need to apply the formula (70) for the sum of the geometric progression.

Thus, total number of iterations  $T$  after  $N = \tilde{O}\left(\log_2\left(\frac{\mu_r R_0}{2\varepsilon}\right)\right)$  restarts equals

$$T = \sum_{k=1}^N T_k = N A^{\frac{1+\kappa}{\kappa}} = \tilde{O}\left(\left[\frac{\beta\sigma_q}{\mu_r}\right]^{\frac{1+\kappa}{\kappa}} \log_2\left(\frac{\mu_r R_0}{2\varepsilon}\right)\right).$$

□



