# Gradient Free Methods for Non-Smooth Convex Optimization with Heavy Tails on Convex Compact

A Preprint

**Nikita M. Kornilov**
Moscow Institute of Physics and Technology,
Dolgoprudny, Russia
kornilov.nm@phystech.edu

**Alexander V. Gasnikov**
Moscow Institute of Physics and Technology and
Skoltech and
ISP RAS Research Center for Trusted Artificial Intelligence,
Moscow region, Russia
gasnikov@yandex.ru

**Pavel E. Dvurechensky**
Weierstrass Institute for Applied Analysis and Stochastics
Berlin, Germany
pavel.dvurechensky@wias-berlin.de

**Darina M. Dvinskikh**
High School of Economics,
Moscow, Russia
dmdvinskikh@hse.ru

April 6, 2023

## Abstract

Optimization problems, in which only the realization of a function or a zeroth-order oracle is available, have many applications in practice. These are multi-armed bandits, black-box models, and models in which the other types of oracles are too expensive to use. An effective method for solving such problems is the approximation of the gradient using sampling and finite differences of the function values. This method relies on the concentration of the measure on the Euclidean sphere. However, some noise can be present in the zeroth-order oracle not allowing the exact evaluation of the function value, and this noise can be stochastic or adversarial. In this paper, we propose and study new easy-to-implement algorithms that are optimal in terms of the number of oracle calls for solving non-smooth optimization problems on a convex compact set with heavy-tailed stochastic noise (random noise has $(1+\kappa)$-th bounded moment) and adversarial noise. These algorithms are based on methods that were demonstrated to be extremely efficient for stochastic problems with first-order oracle and heavy-tailed noise. The first algorithm is based on the heavy-tail-resistant mirror descent and uses special transformation functions that allow controlling the tails of the noise distribution. The second algorithm is based on the gradient clipping technique. In this technique, the heavy tails of the noise distribution are clipped, balancing between the bias of the gradient estimate relative to the true gradient and the algorithm's fast convergence with estimates with small second moments. The paper provides proof of algorithms' convergence results in terms of high probability and in terms of expectation when a convex function is minimized. For functions satisfying a $r$-growth condition, a faster algorithm is proposed using the restart technique. The differences between the two types of algorithms and recommendations for choosing the best parameters are discussed. Particular attention is paid to the question of how large the adversarial noise can be so that the optimality and convergence of the algorithms is guaranteed.

***Keywords*** zeroth-order optimization, derivative-free methods, stochastic optimization, non-smooth problems, heavy tails, gradient clipping, stochastic mirror descent

# 1 Introduction.

Consider stochastic non-smooth convex minimization problem over compact convex set $\mathcal{S} \subset \mathbb{R}^d$:

$$\min_{x \in \mathcal{S}} f(x), \tag{1}$$

where $f(x) = \mathbb{E}_\xi[f(x,\xi)]$ and $f : \mathcal{S} \to \mathbb{R}$ is a convex and Lipschitz continuous function.

We assume that the objective function $f$ can not be evaluated directly. Instead, we have at our disposal a zeroth-order two-points oracle $\phi(x,\xi) = f(x,\xi) + \delta(x)$, where $\delta(x)$ is some adversarial noise.

**Main results and related works.** As said, we consider gradient-free methods for convex optimization problems [25, 5]. More precisely, we consider a two-point zero-order oracle for non-smooth stochastic convex optimization problems. This field is rather well-developed, see, e.g., the recent survey [8] and references therein. For example, in the series of papers [6, 11, 20, 10, 23, 2] the optimal oracle complexity, i.e., the number of calls to the oracle in order to obtain the desired accuracy of the solution, was obtained. For the non-smooth convex-concave stochastic saddle-point problems the same was done in [4, 7]. In both cases, the oracle complexity is $\sim d\varepsilon^{-2}$, where $\varepsilon$ is a desired expected accuracy in function value (or duality gap for saddle-point problems). These results are quite expected since this complexity is $\sim d$ times larger than the complexity of optimal stochastic gradient procedures. Factor $d$ has a natural interpretation, since to approximate (stochastic) gradient it is sufficient to use $d+1$ function values.[1] This is obvious in the smooth case (see e.g. [8]), and is not so trivial in the non-smooth case [23].

Moreover, in [7, 9] it was observed that the gradient-free version of the Stochastic Mirror Descent algorithm converges with a maximum permissible level of adversarial noise $\sim \varepsilon^2/\sqrt{d}$.

All the above results assume that $f(x,\xi)$ is $M(\xi)$ Lipschitz continuous w.r.t. $x$ and that $\mathbb{E}_\xi[M^2(\xi)] < \infty$. The goal of this paper is to obtain analogs of the results mentioned above for the case of a weaker assumption that there exists $\kappa \in (0,1]$ such that $\mathbb{E}_\xi[M^{1+\kappa}(\xi)] < \infty$.

Under this assumption, we know from [19] that the stochastic (full) gradient oracle complexity is $\sim \varepsilon^{-\frac{1+\kappa}{\kappa}}$ and we may expect therefore $\sim d\varepsilon^{-\frac{1+\kappa}{\kappa}}$ zero-order stochastic oracle complexity. In this paper, we propose an algorithm that achieves the following bound $\sim \left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ that matches the expected bounds only for $\kappa = 1$. To the best of our knowledge, this poses the following open problem: is the bound $\sim \left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ optimal in terms of the dependence on $d$? For smooth stochastic convex optimization problems with $(d+1)$-points stochastic zero-order oracle the answer is negative and the optimal bound is $\sim d\varepsilon^{-\frac{1+\kappa}{\kappa}}$. Thus, for $\kappa \in (0,1)$ our results are in a sense 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,[2] 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 compact-support distributions of $f(x,\xi)$ [7]. That is, even for subgaussian tails 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 concentration of measure inequalities do not allow obtaining the desired subgaussian concentration for the output of the algorithm.

To conclude, in this paper we obtain $\sim \left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ gradient-free two-points stochastic oracle complexity bound in terms of high-probability. We also obtain a bound $\sim \varepsilon^2/\sqrt{d}$ on the maximum admissible level of additional adversarial noise in function values. We generalize these results to strongly convex problems and problems with a sharp minimum.

**The proposed approach.** Starting from the work [18] 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, the following accurate results were obtained:

- an optimal algorithm with general proximal setup for non-smooth stochastic convex optimization problems with infinite variance [26] with the convergence in expectation,

[1] To say more accurately, it is sufficient to use $d+1$ values of $f(x,\xi)$ with the same $\xi$ and different $(d+1)$ points $x$.

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

- an optimal adaptive algorithm with general proximal setup for non-smooth online stochastic convex optimization problems with infinite variance [28] with the convergence in high-probability,
- optimal algorithms with euclidean proximal setup for smooth and non-smooth stochastic convex optimization problems and variational inequalities with infinite variance [22, 21] with the convergence in high-probability,
- an optimal variance-adaptive algorithm with euclidean proximal setup for non-smooth stochastic (strongly) convex optimization problems with infinite variance [16] with the convergence in high-probability.

Since the results listed above are strongly correlated with each other, in this paper, we depart from the works [26, 28] 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. [6, 11, 10, 23]. 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. Also, the described above 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 [4](rather than the gap used in [7]).[3] We leave this for future work.

### 1.1 Notations and assumptions.

We use $\langle x, y\rangle = \sum_{k=1}^d x_k y_k$ to denote the inner product of $x, y \in \mathbb{R}^d$. For $p \in [1, 2]$ notation $||\cdot||_p$ is used for the standard $l_p$-norm, i.e. $||x||_p = \left(\sum_{k=1}^d |x_k|^p\right)^{1/p}$. The corresponding dual norm is $||y||_q = \max_x\{\langle x, y\rangle|\ ||x||_p \le 1\}$. $B^p = \{x \in \mathbb{R}^d \mid ||x||_p \le 1\}$ is a $p$-ball with center at 0 and radius 1 and $S^p = \{x \in \mathbb{R}^d \mid ||x||_p = 1\}$ is a $p$-sphere with center at 0 and radius 2. Finally, or $\tau > 0$ and convex set $\mathcal{S} \subset \mathbb{R}^d$ we denote $\mathcal{S}_\tau = \mathcal{S} + \tau \cdot B^2$.

**Assumption 1** (Convexity). *$\exists \tau > 0$, s.t. function $f(x, \xi)$ is convex for any $\xi$ on $\mathcal{S}_\tau$.*

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

**Assumption 2** (Lipschitz). *$\exists \tau > 0$, s.t. function $f(x, \xi)$ is $M_2(\xi)$ Lipschitz continuous w.r.t. $l_2$ norm, i.e., for all $x_1, x_2 \in \mathcal{S}_\tau$*

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

*Moreover, $\exists \kappa \in (0, 1]$ such that $\mathbb{E}_\xi[M_2^{1+\kappa}(\xi)] \le M_2^{1+\kappa}$.*

**Lemma 1.1.** *Assumption 2 implies that $f(x)$ is $M_2$ Lipschitz on $\mathcal{S}$.*

The proof can be found in Section 9 (Lemma 9.2).

**Assumption 3** (Bounded adversarial noise). *For all $x \in \mathcal{S}$ : $|\delta(x)| \le \Delta < \infty$.*

## 2 Gradient-free setup.

In this section, we introduce the main objects and notions that are used to construct gradient-free algorithms. We refer the reader to a review paper [8] devoted to gradient-free algorithms.

As it was mentioned above, in algorithms, we can use only noisy two-point zeroth-order oracle. For points $x, y \in \mathcal{S}$ oracle gives

$$\phi(x, \xi) = f(x, \xi) + \delta(x), \quad \phi(y, \xi) = f(y, \xi) + \delta(y)$$

with the same $\xi$.

In this work we consider only uniform sampling from unit Euclidean sphere, i.e. $\mathbf{e} \sim Uniform(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\}) \stackrel{\text{def}}{=} U(S^2)$.

First of all, we define the smoothed function

$$\hat{f}_\tau(x) = \mathbb{E}_{\mathbf{e}\sim U(S^2)}[f(x + \tau\mathbf{e})] \tag{2}$$

that approximates the objective $f$. Further $U(S^2)$ in $\mathbb{E}_{\mathbf{e}\sim U(S^2)}$ is omitted.

The next lemma gives estimates for the quality of the approximation. In contrast to $f(x)$, $\hat{f}_\tau$ is smooth function and has several useful properties. The proof of the next lemma can be found in [9, Theorem 2.1].

[3] See the full version of the paper [7].

**Lemma 2.1.** *Let Assumptions 1,2 hold. Then,*

1. *Function $\hat{f}_\tau(x)$ is convex, Lipschitz with constant $M_2$ on $\mathcal{S}$, and satisfies*
$$\sup_{x\in\mathcal{S}} |\hat{f}_\tau(x) - f(x)| \le \tau M_2.$$
2. *Function $\hat{f}_\tau(x)$ is differentiable on $\mathcal{S}$ with the following gradient*
$$\nabla \hat{f}_\tau(x) = \mathbb{E}_{\mathbf{e}}\left[\frac{d}{\tau} f(x+\tau\mathbf{e})\mathbf{e}\right].$$

The algorithms proposed below aim at minimizing the smooth approximation $\hat{f}_\tau(x)$. Given the above results, this will also produce a good approximate minimizer of $f(x)$ when $\tau$ is sufficiently small.

Following [23], the gradient of $\hat{f}_\tau(x)$ can be estimated by the following vector:

$$g(x,\xi,\mathbf{e}) = \frac{d}{2\tau}(\phi(x+\tau\mathbf{e},\xi) - \phi(x-\tau\mathbf{e},\xi))\mathbf{e} \tag{3}$$

for $\tau > 0$.

Finally, the following lemma is important and gives a bound for the $(1+\kappa)$-th moment of the estimated gradient for functions with heavy-tailed noise satisfying Assumptions 1, 2 and 3.

**Lemma 2.2.** *Under Assumptions 1, 2 and 3, for $q \in [2, +\infty)$, we have*

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

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

## 3 Robust Stochastic Mirror Descent.

In this section, we use some definitions and claims from [26].

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

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

When $r = 2$ this definition is the same as the definition of $K$-strongly convex function. Examples of functions when $r > 2$ can be obtained from next lemma.

**Lemma 3.1.** *For $\kappa \in (0,1], q \in [1+\kappa, \infty)$ and $p$ s.t. $\frac{1}{q} + \frac{1}{p} = 1$, we define*

$$K_q \overset{def}{=} 10 \max\left\{1, (q-1)^{\frac{1+\kappa}{2}}\right\}. \tag{4}$$

*Then,*

$$\phi_p(x) \overset{def}{=} \frac{\kappa}{1+\kappa}||x||_p^{\frac{1+\kappa}{\kappa}} \tag{5}$$

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

Now we describe Stochastic Mirror Descent (SMD) algorithm. Let function $\Psi : \mathbb{R}^d \to \mathbb{R}$ be $(K, r)$-uniformly convex w.r.t. the $p$-norm and continuously differentiable. We denote its Fenchel conjugate and its Bregman divergence respectively as

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

The Stochastic Mirror Descent updates with stepsize $\nu$ and update vector $g_{k+1}$ are as follows:

$$y_{k+1} = \nabla(\Psi^*)(\nabla\Psi(x_k) - \nu g_{k+1}), \quad x_{k+1} = \arg\min_{x\in\mathcal{S}} D_\Psi(x, y_{k+1}). \tag{6}$$

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 called the transformation map.

For SMD Algorithm (6) with standard 1-strongly convex function $\Psi$, the convergence theory is well known and given, e.g. in [3]. The next theorem generalizes these results and gives a convergence result of SMD with uniformly convex $\Psi$.

**Theorem 3.2.** *Consider some $\kappa \in (0,1], p \in [1,\infty]$, $q$ defined by the equality $\frac{1}{q} + \frac{1}{p} = 1$, and 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* (6) *with the corresponding map function $\nabla\Psi_p$, after $T$ iterations with any $g_k \in \mathbb{R}^d, k \in \overline{1,T}$ and starting point $x_0 = \arg\min\limits_{x\in\mathcal{S}} \Psi(x)$ we have*

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle \leq \frac{\kappa}{\kappa+1}\frac{R_\Psi^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}||g_{k+1}||_q^{1+\kappa}, \tag{7}$$

*where $R_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x\in\mathcal{S}}\{\Psi_p(x) - \Psi_p(x_0)\}$.*

The proof can be found in [26, Theorem 6]. Note, that when $\kappa = 1$ $\Psi$ is a 1-strongly convex function.

# 4 Zeroth-Order Robust SMD Algorithm.

The main idea of the proposed Zeroth-Order Robust SMD algorithm is to combine the above Robust SMD Algorithm (6) with the two-point gradient approximation (3). The former allows working with the heavy-tailed distribution of the gradient approximation and the latter allows coping with the non-smoothness of the objective in (1).

**Algorithm 1** Zeroth-Order Robust SMD Algorithm

**procedure** ZERO ROBUST SMD(Number of iterations $T$, stepsize $\nu$, transformation function $\Psi_p$, smoothing constant $\tau$)
  $x_0 \leftarrow \arg\min\limits_{x\in\mathcal{S}} \Psi_p(x)$
  **for** $k = 0, 1, \ldots, T-1$ **do**
    Sample $\mathbf{e}_k \sim Uniform(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$ independently
    Sample $\xi_k$ independently
    Calculate $g_{k+1} = \frac{d}{2\tau}(\phi(x_k + \tau\mathbf{e}_k, \xi_k) - \phi(x_k - \tau\mathbf{e}_k, \xi_k))\mathbf{e}_k$
    Calculate $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla\Psi_p(x_k) - \nu g_{k+1})$
    Calculate $x_{k+1} \leftarrow \arg\min\limits_{x\in\mathcal{S}} D_{\Psi_p}(x, y_{k+1})$
  **end for**
  **return** $\overline{x}_T \leftarrow \frac{1}{T}\sum\limits_{k=0}^{T-1} x_k$
**end procedure**

The next theorem gives optimal parameters for Algorithm 1 and its rate of convergence.

**Theorem 4.1.** *Let function $f$ satisfying Assumptions 1, 2, 3, $q \in [1+\kappa, \infty]$, arbitrary number of iterations $T$, smoothing constant $\tau > 0$ be given. Choose $\left(1, \frac{1+\kappa}{\kappa}\right)$-uniformly convex w.r.t. the $p$-norm function $\Psi_p(x)$ (e.g., $\Psi_p(x) = K_q^{1/\kappa}\phi^p(x)$, where $K_q$, $\phi^p$ are defined in* (4) *and* (5) *respectively). Set the stepsize $\nu = \frac{R_\Psi^{1/\kappa}}{\sigma_q}T^{-\frac{1}{1+\kappa}}$ with $\sigma_q$ given in Lemma 2.2, $R_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x\in\mathcal{S}}\{\Psi_p(x) - \Psi_p(x_0)\}$ and $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x,y\in\mathcal{S}} D_{\Psi_p}(x,y)$. Let $\overline{x}_T$ be a point obtained by Algorithm 1 with the above parameters, and let $x^* \in \arg\min\limits_{x\in\mathcal{S}} f(x)$.*

1. *Then*

$$\boxed{\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_\Psi\sigma_q}{T^{\frac{\kappa}{1+\kappa}}},} \tag{8}$$

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

2. *With optimal* $\tau = \sqrt{\frac{\sqrt{d}\Delta \mathcal{D}_\Psi + 4R_\Psi d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_\Psi d a_q \Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_q M_2 R_\Psi}{T^{\frac{\kappa}{1+\kappa}}}. \tag{9}$$

**Main idea behind the proof.**

Proof is based on Theorem 3.2 and inequality (7) from it

$$\underbrace{\mathbb{E}_{\xi,\mathbf{e}}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle\right]}_{①} \le \underbrace{\mathbb{E}_{\xi,\mathbf{e}}\left[\frac{\kappa}{\kappa+1}\frac{R_\Psi^{\frac{1+\kappa}{\kappa}}}{\nu T}\right]}_{②} + \underbrace{\mathbb{E}_{\xi,\mathbf{e}}\left[\frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}||g_{k+1}||_q^{1+\kappa}\right]}_{③}. \tag{10}$$

① term in (10) due to convexity and approximation properties of $\hat{f}_\tau(x)$ in Lemma 2.1 and measure concentration Lemma 9.6 can be bounded with

$$① \ge \mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) - 2M_2\tau - \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi.$$

③ term in (10) can be bounded with Lemma 2.2

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

Combining bounds together, we get

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_\Psi^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\sigma_q^{1+\kappa}.$$

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

Full proof can be found in Section 10.

### 4.1 Zeroth-Order Robust SMD Algorithm discussion.

**Maximum level of adversarial noise.**

Let $\varepsilon > 0$ be a desired accuracy in terms of the function value, i.e., our goal is to guarantee $\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le \varepsilon$. According to Theorem 4.1 in the absence of the adversarial noise, i.e., when $\Delta = 0$, the iteration complexity to reach accuracy $\varepsilon$ is $T = \left(\frac{R_\Psi\sqrt{d}a_q M_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}$ if $\tau$ is chosen sufficiently small. This complexity is optimal according to [19].

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 (8) should be $= \varepsilon$. These conditions also make negligible the $\tau$-depending term in $\sigma_q$. Consequently,

$$\text{when} \quad \tau = \frac{\varepsilon}{M_2} \text{ and } \Delta \le \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}, \text{ we have } T = \left(\frac{R_\Psi\sqrt{d}a_q M_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}.$$

Otherwise, when $\Delta > \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$, the convergence rate deteriorates. As we see in (9), in this case, we can not guarantee the accuracy smaller than $\sqrt{M_2\sqrt{d}\Delta\mathcal{D}_\Psi}$. Moreover, the iteration complexity to make the other terms smaller than $\varepsilon$ is $T = O\left(\frac{\sqrt{M_2R_\Psi d a_q\Delta}}{\varepsilon}\right)^{\frac{2(1+\kappa)}{\kappa}}$, which is worse than $O(\varepsilon^{-\frac{\kappa+1}{\kappa}})$ obtained when the error $\Delta$ can be controlled.

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

In Algorithm 1, we can freely choose $p \in [1, 2]$ and $\Psi$, which, depending on the compact convex set $\mathcal{S}$, lead to different values of $\mathcal{D}_\Psi$, $R_\Psi$, $a_q$. It is desirable to reduce $a_q, \mathcal{D}_\Psi$ simultaneously, which would allow us to increase maximal noise level $\Delta$ and converge faster without changing the rate according to (8). Yet, unlike the well-studied SMD algorithm with strongly convex functions $\Psi$, there are only few examples of effective choices of uniformly-convex functions $\Psi$.

## 5 Zeroth-Order Clipping Algorithm.

In this section $\tilde{O}(\cdot)$ denotes $\log \frac{1}{\delta}$ factor.

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

$$\hat{g} = \frac{g}{||g||} \min(||g||, c).$$

Clipped gradient has bunch of useful properties for further proofs.

**Lemma 5.1.** *Let $g_k = g(x_k, \xi, \mathbf{e})$. For $c > 0$ we define $\hat{g} = \frac{g}{||g||_q} \min(||g||_q, c)$.*

1.
$$||\hat{g} - \mathbb{E}[\hat{g}]||_q \leq 2c. \tag{11}$$

2. *If $\mathbb{E}_{\xi,\mathbf{e}}[||g(x, \xi, \mathbf{e})||_q^{1+\kappa}] \leq \sigma_q^{1+\kappa}$, then*

   *(a)*
$$\mathbb{E}_{\xi,\mathbf{e}}[||\hat{g}||_q^2] \leq \sigma_q^{1+\kappa} c^{1-\kappa}. \tag{12}$$

   *(b)*
$$\mathbb{E}_{\xi,\mathbf{e}}[||\hat{g} - \mathbb{E}_{\xi,\mathbf{e}}[\hat{g}]||_q^2] \leq 4\sigma_q^{1+\kappa} c^{1-\kappa}. \tag{13}$$

   *(c)*
$$||\mathbb{E}_{\xi,\mathbf{e}}[g] - \mathbb{E}_{\xi,\mathbf{e}}[\hat{g}]||_q \leq \frac{\sigma_q^{1+\kappa}}{c^\kappa}. \tag{14}$$

If $g$ is an unbiased stochastic gradient, then, on the one hand, $\hat{g}$ is bounded, and, on the other hand, is a biased stochastic gradient. Thus, the constant $c$ allows playing with the trade-off between the faster convergence and bias $||\mathbb{E}[\hat{g} - g]||$ when $c \to 0$. The Algorithm implementing this idea in our setting is as follows.

**Algorithm 2** Zeroth-Order Clipping Algorithm

**procedure** ZERO CLIP(Number of iterations $T$, stepsize $\nu$, clipping constant $c$, transformation function $\Psi_p$, smoothing constant $\tau$)
  $x_0 \leftarrow \arg\min_{x \in \mathcal{S}} \Psi_p(x)$
  **for** $k = 0, 1, \ldots, T-1$ **do**
    Sample $\mathbf{e}_k \sim Uniform(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$ independently
    Sample $\xi_k$ independently
    Calculate $g_{k+1} = \frac{d}{2\tau}(\phi(x_k + \tau\mathbf{e}_k, \xi_k) - \phi(x_k - \tau\mathbf{e}_k, \xi_k))\mathbf{e}_k$
    Calculate $\hat{g}_{k+1} = \frac{g_{k+1}}{||g_{k+1}||_q} \min(||g_{k+1}||_q, c)$
    Calculate $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla\Psi_p(x_k) - \nu\hat{g}_{k+1})$
    Calculate $x_{k+1} \leftarrow \arg\min_{x \in \mathcal{S}} D_{\Psi_p}(x, y_{k+1})$
  **end for**
  **return** $\overline{x}_T \leftarrow \frac{1}{T}\sum_{k=0}^{T-1} x_k$
**end procedure**

The next result gives a convergence rate for the above algorithm in terms of the expectation of the suboptimality gap.

**Theorem 5.2.** *Let function $f$ satisfying Assumptions 1, 2, 3, $q \in [1+\kappa, \infty]$, arbitrary number of iterations $T$, smoothing constant $\tau > 0$ be given. Choose 1-strongly convex w.r.t. the $p$-norm function $\Psi_p(x)$. Set the stepsize $\nu = \left(\frac{R_\Psi^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$ with $\sigma_q$ given in Lemma 2.2, $R_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x\in\mathcal{S}}\{\Psi_p(x) - \Psi_p(x_0)\}$ and $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x,y\in\mathcal{S}} D_{\Psi_p}(x,y)$. After set the clipping constant $c = \frac{2\kappa\mathcal{D}_\Psi}{(1-\kappa)\nu}$. Let $\overline{x}_T$ be a point obtained by Algorithm 2 with the above parameters, and let $x^* \in \arg\min\limits_{x\in\mathcal{S}} f(x)$.*

1. *Then,*

$$\boxed{\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_\Psi^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}},} \tag{15}$$

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

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_\Psi^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_qM_2R_\Psi^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{16}$$

The following result is stronger and gives a convergence rate for the above algorithm in terms of the suboptimality gap with high probability. Yet, this leads to an additional $\log\frac{1}{\delta}$ factor, where $\delta$ is the desired confidence level.

**Theorem 5.3.** *Let function $f$ satisfying Assumptions 1, 2, 3, $q \in [1+\kappa, \infty]$, arbitrary number of iterations $T$, smoothing constant $\tau > 0$ be given. Choose 1-strongly convex w.r.t. the $p$-norm function $\Psi_p(x)$. Set the clipping constant $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ with $\sigma_q$ given in Lemma 2.2. After set the stepsize $\nu = \frac{\mathcal{D}_\Psi}{c}$ with $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x,y\in\mathcal{S}} D_{\Psi_p}(x,y)$. Let $\overline{x}_T$ be a point obtained by Algorithm 2 with the above parameters, and let $x^* \in \arg\min\limits_{x\in\mathcal{S}} f(x)$.*

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

$$\boxed{f(\overline{x}_T) - f(x^*) \le 2M_2\tau + \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi + \tilde{O}\left(\frac{\mathcal{D}_\Psi\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right),} \tag{17}$$

   *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. *With optimal $\tau = \tilde{O}\left(\sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4\mathcal{D}_\Psi da_q\Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}\right)$*

$$f(\overline{x}_T) - f(x^*) = \tilde{O}\left(\sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2\mathcal{D}_\Psi da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_qM_2\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\right). \tag{18}$$

**Main idea behind the proof in expectation.**

Proof is based on Theorem 3.2 and inequality (7) for 1-strongly convex $\Psi_p$ from it

$$\underbrace{\mathbb{E}_{\xi,\mathbf{e}}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle\hat{g}_{k+1}, x_k - x^*\rangle\right]}_{①} \le \mathbb{E}_{\xi,\mathbf{e}}\left[\frac{1}{2}\frac{R_\Psi^2}{\nu T}\right] + \underbrace{\mathbb{E}_{\xi,\mathbf{e}}\left[\frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}||\hat{g}_{k+1}||_q^2\right]}_{②}. \tag{19}$$

(1) term in (19) due to convexity and approximation properties of $\hat{f}_\tau(x)$ in Lemma 2.1, measure concentration Lemma 9.6 and clipped properties in Lemma 5.1 can be bounded with

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

(2) term in (19) can be bounded with Lemma 5.1

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

Combining bounds together, we get

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

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

Full proof can be found in Section 11.

**Main idea behind the proof in high probability.** To bound variables with probability at least $1-\delta$ we use classic Bernstein inequality for martingale differences (i.e. $\mathbb{E}[X_i|X_{j<i}] = 0, \forall i \geq 1$) sum (Lemma 12.1) and sum of squares of random variables (Lemma 12.2).

Proof is based on Theorem 3.2 and inequality (7) for 1-strongly convex $\Psi_p$ from it

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle \leq \frac{1}{2}\frac{R_\Psi^2}{\nu T} + \underbrace{\frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}||\hat{g}_{k+1}||_q^2}_{\textcircled{1}}. \tag{20}$$

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

$$+ \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla\hat{f}_\tau(x_k), x_k - x^*\rangle}_{\textcircled{4}}.$$

We bound (1) term in (20) using Lemma 12.2 and (2) as martingale difference using lemma 12.1.

$$\textcircled{1} = \tilde{O}\left(\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} + \frac{1}{T}c^2\right).$$

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

Next we bound (4) due to convexity of $\hat{f}_\tau(x)$ in Lemma 2.1 and (3) due to measure concentration Lemma 9.6 and clipped properties in Lemma 5.1

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

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

Combining bounds together, we get

$$f(\overline{x}_T) - f(x^*) \le 2M_2\tau + \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi + \frac{1}{2}\frac{R_\Psi^2}{\nu T}$$

$$+\tilde{O}\left(\frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa} + \frac{\nu}{2}\frac{1}{T}c^2 + \frac{4c\mathcal{D}_\Psi}{T} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\right).$$

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

Full proof can be found in Section 12.

### 5.1 Zeroth-Order Clipping Algorithm discussion.

In this discussion, we focus on the high-probability bounds given in Theorem 5.3. The same discussion holds also for the result of Theorem 5.2 up to omitting the $\log\frac{1}{\delta}$ factor.

**Maximum level of adversarial noise.**

Let $\varepsilon$ be desired function value accuracy, i.e. with probability at least $1-\delta : f(\overline{x}_T) - f(x^*) \le \varepsilon$.

In Theorem 5.3 if there is no adversarial noise, i.e., $\Delta = 0$, then the convergence rate is $T^{\frac{\kappa}{1+\kappa}} = \tilde{O}\left(\frac{R_\Psi\sqrt{d}a_qM_2}{\varepsilon}\right)$ when $\tau \to 0$. This rate is optimal according to [19].

In order to keep the same rate when $\Delta > 0$, $2M_2\tau$ and $\frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi$ should be $= \varepsilon$. These conditions also make negligible the $\tau$-depending term in $\sigma_q$. Consequently,

$$\text{when} \quad \tau = \frac{\varepsilon}{M_2} \text{ and } \Delta \le \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi} \Rightarrow T^{\frac{\kappa}{1+\kappa}} = \tilde{O}\left(\frac{\mathcal{D}_\Psi\sqrt{d}a_qM_2}{\varepsilon}\right).$$

Otherwise, when $\Delta > \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$, the convergence rate deteriorates. Similarly to Robust SMD discussion we can't achieve accuracy less than $\sqrt{M_2\sqrt{d}\Delta\mathcal{D}_\Psi}$. And convergence rate to this bound is $T^{\frac{\kappa}{1+\kappa}} = \tilde{O}\left(\frac{M_2\mathcal{D}_\Psi da_q\Delta}{\varepsilon^2}\right)$, which is twice worse than $\tilde{O}\left(\frac{\mathcal{D}_\Psi\sqrt{d}a_qM_2}{\varepsilon}\right)$.

**Recommendations for choosing $\Psi$.**

With Algorithm 2, we can freely choose $p \in [1,2]$ and $\Psi$, which, depending on the compact convex set $\mathcal{S}$, will change $\mathcal{D}_\Psi, R_\Psi, a_q$. 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 pace according to (17).

Next, we discuss some standard sets $\mathcal{S}$ and transformation functions $\Psi$ taken from [3]. The two main setups are given by

1. $$\text{Ball setup:} \qquad p = 2, \Psi(x) = \frac{||x||_2^2}{2}, \tag{21}$$

2. $$\text{Entropy setup:} \qquad p = 1, \Psi(x) = (1+\gamma)\sum_{i=1}^{d}(x_i + \gamma/d)\log(x_i + \gamma/d), \gamma > 0. \tag{22}$$

Introduce notation $B^p = \{x \in \mathbb{R}^d : ||x||_p \le 1\}$ and $\Delta_d^+ = \{x \in \mathbb{R}^d : x \ge 0, \sum_i x_i \ge 1\}$. By Lemma 2.2, $a_q = d^{\frac{1}{q}-\frac{1}{2}}\min\{\sqrt{32\ln d - 8}, \sqrt{2q-1}\}$. The next tables collect the complexity $T^{\frac{\kappa}{1+\kappa}}$ and maximum feasible noise level $\Delta$ up to $O\left(\log\frac{1}{\delta}\right)$ factor for each setup (row) and set (column).

From these tables, we see that for $\mathcal{S} = \Delta_d^+$ or $B^1$, 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 $\mathcal{S} = B^2$ or $B^\infty$, the Ball setup is better in terms of both convergence rate and noise robustness.

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

| | $\Delta_d^+$ | $B^1$ | $B^2$ | $B^\infty$ |
|---|---|---|---|---|
| Ball | $\sqrt{d}M_2/\varepsilon$ | $\sqrt{d}M_2/\varepsilon$ | $\sqrt{d}M_2/\varepsilon$ | $dM_2/\varepsilon$ |
| Entropy | $\ln dM_2/\varepsilon$ | $\ln dM_2/\varepsilon$ | $\sqrt{d}\ln dM_2/\varepsilon$ | $d\ln dM_2/\varepsilon$ |

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

| | $\Delta_d^+$ | $B^1$ | $B^2$ | $B^\infty$ |
|---|---|---|---|---|
| Ball | $\varepsilon^2/(\sqrt{d}M_2)$ | $\varepsilon^2/(\sqrt{d}M_2)$ | $\varepsilon^2/(\sqrt{d}M_2)$ | $\varepsilon^2/(dM_2)$ |
| Entropy | $\varepsilon^2/(\sqrt{d\ln d}M_2)$ | $\varepsilon^2/(\sqrt{d\ln d}M_2)$ | $\varepsilon^2/(d\sqrt{\ln d}M_2)$ | $\varepsilon^2/(\sqrt{d^3\ln d}M_2)$ |

**Zeroth-Order Clipping and Robust SMD Algorithms comparison.** Although both convergence Theorems 4.1 and 5.2,5.3 for Algorithms 1, 2 respectively give the same estimates, the Clipping Algorithm 2 is much more flexible due to the choice of transformation functions $\Psi$ and ability to effectively work with different sets. Also, Algorithm 2 has high-probability convergence rate guarantees. However, in practice, the convergence of it dramatically depends on the clipping constant $c$, which must be carefully chosen, along with stepsize $\nu$ and smoothing constant $\tau$.

## 6 Zeroth-order Algorithms with Restarts.

In this section $\tilde{O}(\cdot)$ denotes $\log d$ factor unless otherwise said.

For functions with $r$-growth condition (for more information see [24]) there is restart technique developed in [14] for algorithms acceleration.

**Assumption 4.** *$f$ is $r$-growth function if there is $r \geq 1$ 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, $\mu$-strong convex w.r.t. the $p$-norm functions are 2-growth. Restart technique will work if $\Delta$ small enough to keep optimality of Algorithms 1 and 2.

**Algorithm 3** IZ Restart Algorithm

**procedure** IZ RESTART(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), transformation function $\Psi_p$)
  $x_0 \leftarrow \arg\min_{x\in\mathcal{S}} \Psi_p(x)$ or randomly
  **for** $k = 0, 1, \ldots, N$ **do**
    Set parameters $\nu_k, (c_k), \Psi_p, \tau_k$ of the Algorithm $\mathcal{A}$
    Compute $T_k$ iterations of the Algorithm $\mathcal{A}$ with starting point $x_0$ and get $x_{final}$
    $x_0 \leftarrow x_{final}$
  **end for**
  **return** $x_{final}$
**end procedure**

**Theorem 6.1.** *Let function $f$ satisfies Assumptions 1, 2. Next, let $\varepsilon$ be fixed accuracy and $r$-growth Assumption 4 is held with $r \geq \frac{1+\kappa}{\kappa}$.*

*First, calculate $R_0 \overset{def}{=} \sup_{x,y\in\mathcal{S}} \left(\frac{1+\kappa}{\kappa} D_{\Psi_p}(x,y)\right)^{\frac{\kappa}{1+\kappa}}$ and define $R_k = R^0/2^k$.*

*Set 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 $\{\tau_k\}_{k=1}^N = \left\{\frac{\sigma_q R_k}{M_2 T_k^{\frac{\kappa}{1+\kappa}}}\right\}_{k=1}^N$ 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 from Lemma 2.2.*

*Moreover, Assumption 3 is held with*

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

*If $x_{final}$ is final output of Algorithm 3 with basic Robust SMD Algorithm 1 these parameters then*

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le \varepsilon,$$

*total number of Algorithms steps is*

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

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

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

**Theorem 6.2.** [4] *Let function $f$ satisfies Assumptions 1, 2. Next, let $\varepsilon$ be fixed accuracy and $r$-growth Assumption 4 is held with $r \ge 2$ for in expectation estimate or $r \ge 1$ for in high probability estimate.*

*First, calculate $R_0 \overset{def}{=} \sup_{x,y\in\mathcal{S}}\left(2D_{\Psi_p}(x,y)\right)^{\frac{1}{2}}$ and define $R_k = R^0/2^k$.*

*Set 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 $\{\tau_k\}_{k=1}^N = \left\{\frac{\sigma_q R_k}{M_2 T_k^{\frac{\kappa}{1+\kappa}}}\right\}_{k=1}^N$, 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 from Lemma 2.2.*

*Moreover, Assumption 3 is held with*

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

*If $x_{final}$ is final output of Algorithm 3 with basic Clipping SMD Algorithm 2 and these parameters then*

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le \varepsilon,$$

*or with probability at least $1-\delta$*

$$f(\overline{x}_T) - f(x^*) \le \varepsilon.$$

*Also total number of Algorithms steps is*

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

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

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

[4] 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 proof.

### 6.1 Restart Algorithm discussion.

**Maximum level of adversarial noise.**

Unlike Robust and Clipping SMD Algorithms, Restart Algorithm and $r$-growth Assumption guarantees a higher maximum threshold for $\Delta$

$$\text{Algorithm 1 or 2 :} \qquad \Delta = \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi},$$

$$\text{Algorithm 3 :} \qquad \Delta = \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2\sqrt{d}}\varepsilon^{(2-1/r)}\right).$$

Moreover, this threshold doesn't depend of set $\mathcal{S}$ and $\frac{1}{\sqrt{d}}$ factor is the best(see Table 5.1). Also, in the beginning $\Delta_k$ can be much bigger and start to decrease as $\Delta_k = \frac{\Delta_1}{2^{k(2r-1)}}$ only on later restarts in order to achieve necessary accuracy. Lower $r$ ensures higher threshold.

$q, d, \varepsilon$ **dependencies.**

Again unlike Robust and Clipping SMD Algorithms, Restart Algorithm and $r$-growth Assumption guarantees a faster convergence rate depending on $\varepsilon$. Below we give in expectation estimates

$$\text{Algorithm 1 or 2 :} \qquad T = O\left(\left[\frac{\mathcal{D}_\Psi\sqrt{d}a_q M_2}{\varepsilon}\right]^{\frac{1+\kappa}{\kappa}}\right),$$

$$\text{Algorithm 3 :} \qquad T = \tilde{O}\left(\left[\frac{a_q M_2\sqrt{d}}{\mu_r^{1/r}}\cdot\frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right).$$

Furthermore, in Restart Algorithm total number of iteration depends only on $a_q$ and maximum $\Delta$ threshold doesn't depend on $q, \mathcal{S}$ at all. Thus, it makes sense to take Entropy setup defined in (22) with basic Clipping Algorithm to lower $a_q$ and leave only $\log d$ factor in $T$ estimate.

## 7 Conclusion.

In this paper, we proposed and theoretically studied new zero-order algorithms for solving non-smooth optimization problems on a convex compact set with heavy-tailed stochastic noise (random noise has $(1+\kappa)$-th bounded moment) and adversarial noise in the function value. We believe that there are several possible modifications that can improve convergence results in future studies.

1. A different sampling strategy for estimating $g_k$. E.g., one can use sampling on the sphere $\{\mathbf{e} : ||\mathbf{e}||_1 = 1\}$ considered in [1], [17].
2. Additional assumptions on the properties of adversarial noise. For example, Lipschitz continuity in the spirit of Assumption 3 in [7]:
$$|\delta(x_1) - \delta(x_2)| \leq M||x_1 - x_2||_2, \qquad \forall x_1, x_2 \in \mathcal{S}.$$
3. Adaptive strategies and heuristic methods for selecting the algorithm's input parameters such as stepsize $\nu$, smoothing constant $\tau$, etc. These constants are difficult to estimate in practice, and our algorithms rely on the accuracy of their evaluation.

We believe that our technique is rather general and allows one to use other stochastic gradient methods to obtain new complexity results for zero-order algorithms.

## 8 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 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] Arya Akhavan, Evgenii Chzhen, Massimiliano Pontil, and Alexandre B Tsybakov. A gradient estimator via l1-randomization for online zero-order optimization with two point feedback. *arXiv preprint arXiv:2205.13910*, 2022.

[2] Anastasia Sergeevna Bayandina, Alexander V Gasnikov, and Anastasia A Lagunovskaya. 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.

[3] Aharon Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM, 2001.

[4] Aleksandr Beznosikov, Abdurakhmon Sadiev, and Alexander Gasnikov. 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*, pages 105–119. Springer, 2020.

[5] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*. SIAM, 2009.

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

[7] Darina Dvinskikh, Vladislav Tominin, Yaroslav Tominin, and Alexander Gasnikov. Gradient-free optimization for non-smooth minimax problems with maximum value of adversarial noise. *arXiv preprint arXiv:2202.06114*, 2022.

[8] Alexander Gasnikov, Darina Dvinskikh, Pavel Dvurechensky, Eduard Gorbunov, Aleksander Beznosikov, and Alexander Lobanov. Randomized gradient-free methods in convex optimization. *arXiv preprint arXiv:2211.13566*, 2022.

[9] Alexander Gasnikov, Anton Novitskii, Vasilii Novitskii, Farshed Abdukhakimov, Dmitry Kamzolov, Aleksandr Beznosikov, Martin Takáč, Pavel Dvurechensky, and Bin Gu. The power of first-order smooth optimization for black-box non-smooth problems. *arXiv preprint arXiv:2201.12289*, 2022.

[10] Alexander V Gasnikov, Ekaterina A Krymova, Anastasia A Lagunovskaya, Ilnura N Usmanova, and Fedor A Fedorenko. 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.

[11] Alexander V Gasnikov, Anastasia A Lagunovskaya, Ilnura N Usmanova, and Fedor A Fedorenko. Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex. *Automation and Remote Control*, 77:2018–2034, 2016.

[12] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58:48–64, 2018.

[13] EA Gorbunov, Evgeniya Alekseevna Vorontsova, and Alexander Vladimirovich Gasnikov. 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.

[14] Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

[15] Michel Ledoux. The concentration of measure phenomenon. ed. by peter landweber et al. vol. 89. *Mathematical Surveys and Monographs. Providence, Rhode Island: American Mathematical Society*, page 181, 2005.

[16] Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises. *arXiv preprint arXiv:2303.12277*, 2023.

[17] Aleksandr Lobanov, Belal Alashqar, Darina Dvinskikh, and Alexander Gasnikov. 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.

[18] Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80:1607–1627, 2019.

[19] AS Nemirovsky and DB Yudin. Problem complexity and optimization method efficiency. *M.: Nauka*, 1979.

[20] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[21] Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

[22] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. *arXiv preprint arXiv:2302.00999*, 2023.

[23] Ohad Shamir. 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.

[24] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[25] James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons, 2005.

[26] Nuri Mert Vural, Lu Yu, Krishna Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In *Conference on Learning Theory*, pages 65–102. PMLR, 2022.

[27] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

[28] Jiujia Zhang and Ashok Cutkosky. Parameter-free regret in high probability with heavy tails. *arXiv preprint arXiv:2210.14355*, 2022.

# 9 Proofs of Lemmas.

## 9.1 General results.

**Lemma 9.1.** *1. For all $x, y \in \mathbb{R}^{d'}$ and $\kappa \in (0,1]$:*

$$||x-y||_q^{1+\kappa} \leq 2^\kappa ||x||_q^{1+\kappa} + 2^\kappa ||y||_q^{1+\kappa}, \tag{23}$$

*2.*

$$\forall x, y \geq 0, \kappa \in [0,1] : (x+y)^\kappa \leq x^\kappa + y^\kappa. \tag{24}$$

*Proof.*

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

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

- Proposition 9 from [26].

□

**Lemma 9.2.** *Assumption 2 implies that $f(x)$ is $M_2$ Lipschitz on $\mathcal{S}$.*

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

$$\begin{aligned} |f(x) - f(y)| &= |\mathbb{E}[f(x,\xi) - f(y,\xi)]| \overset{\text{Jensen's inq}}{\leq} \mathbb{E}[|f(x,\xi) - f(y,\xi)|] \\ &\leq \mathbb{E}[M_2]||x-y||_2 \overset{\text{Jensen's inq}}{\leq} \mathbb{E}[M_2^{(1+\kappa)}]^{\frac{1}{1+\kappa}}||x-y||_2 \leq M_2||x-y||_2. \end{aligned} \tag{25}$$

□

## 9.2 Smoothing.

**Lemma 9.3.** *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 $\forall t > 0$

$$Pr\left(|f(\mathbf{e}) - \mathbb{E}[f(\mathbf{e})]| > t\right) \leq 2\exp(-b'dt^2/M_2^2), \quad b' = 2 \tag{26}$$

(see the proof of Proposition 2.10 and Corollary 2.6 in [15]). Therefore,

$$\mathbb{E}_{\mathbf{e}}\left[(f(\mathbf{e}) - \mathbb{E}_{\mathbf{e}}[f(\mathbf{e})])^{2(1+\kappa)}\right]$$

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

$$\leq \int_{t=0}^{\infty} 2\exp\left(-b'dt^{\frac{1}{(1+\kappa)}}/M_2^2\right) dt \leq \left(\frac{bM_2^2}{d}\right)^{1+\kappa}.$$

□

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

**Lemma 9.4.** *For $q \geq 2, \kappa \in (0,1]$*

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

.

This Lemma is generalization of Lemma from [13] for $\kappa < 1$.

*Proof.* We use Lemma 1 auxiliary Lemma from Theorem 1 from [13].

1. Let $e_k$ be $k$-th component of $\mathbf{e}$

$$\mathbb{E}\left[|e_2|^q\right] \leq \left(\frac{q-1}{d}\right)^{\frac{q}{2}}, \quad q \geq 2. \tag{27}$$

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

$$||x||_{q_1} \leq ||x||_{q_2}. \tag{28}$$

Then

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

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

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

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

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

Therefore,

$$\left(\mathbb{E}\left[\left(\sum_{k=1}^d |e_k|^q\right)^2\right]\right)^{\frac{1+\kappa}{q}} \leq \left(d\mathbb{E}\left[\sum_{k=1}^d |e_k|^{2q}\right]\right)^{\frac{1+\kappa}{q}} = (d^2\mathbb{E}[|e_2|^{2q}])^{\frac{1+\kappa}{q}}.$$

Using (27) with $2q$

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

By definition of $a_q$

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

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 $d$.

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

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

$$q^2 - 2\ln(d)q + \ln(d) = 0.$$

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

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

When $q \geq q_0$ from (28)

$$a_q < a_{q_0} = \sqrt{d^{\frac{2}{q}_0 - 1}(2q_0 - 1)} \leq d^{\frac{1}{\ln d} - \frac{1}{2}}\sqrt{4\ln d - 1}$$
$$= \frac{e}{\sqrt{d}}\sqrt{4\ln d - 1} \leq d^{\frac{1}{q} - \frac{1}{2}}\sqrt{32\ln d - 8}.$$

Consequently,

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

□

**Lemma 9.5.** *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*

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

**Lemma 9.6.** *Let* $g(x, \xi, \mathbf{e})$ *be defined in* (3) *and* $\hat{f}_\tau(x)$ *be defined in* (2)*. Then, the following holds under Assumption 3:*

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

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

*Proof.* By definition

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

Then

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

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

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

Using Lemma 2.1

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

In the second term we use Assumption 3

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

Adding two terms together we get necessary result.

□

**Lemma 9.7.** *Under Assumptions 1, 2 and 3, for* $q \in [1, +\infty)$*, we have*

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

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

*Proof.*

$$\begin{aligned}\mathbb{E}_{\xi,\mathbf{e}}[||g(x,\xi,\mathbf{e})||_q^{1+\kappa}] &= \mathbb{E}_{\xi,\mathbf{e}}\left[||\frac{d}{2\tau}(\phi(x+\tau\mathbf{e},\xi) - \phi(x-\tau\mathbf{e},\xi))\mathbf{e}||_q^{1+\kappa}\right] \\ &\le \left(\frac{d}{2\tau}\right)^{1+\kappa}\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|(f(x+\tau\mathbf{e},\xi) - f(x-\tau\mathbf{e},\xi) + \delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e}))|^{1+\kappa}\right] \\ &\le 2^\kappa\left(\frac{d}{2\tau}\right)^{1+\kappa}\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi) - f(x-\tau\mathbf{e},\xi)|^{1+\kappa}\right] &(29) \\ &+ 2^\kappa\left(\frac{d}{2\tau}\right)^{1+\kappa}\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|\delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e}))|^{1+\kappa}\right]. &(30)\end{aligned}$$

Lets deal with (29) term. For all $\alpha(\xi)$

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-f(x-\tau\mathbf{e},\xi)|^{1+\kappa}\right]\\
&\leq \mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|(f(x+\tau\mathbf{e},\xi)-\alpha)-(f(x-\tau\mathbf{e},\xi)-\alpha)|^{1+\kappa}\right]\\
&\overset{(23)}{\leq} 2^\kappa\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]+2^\kappa\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x-\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]. \qquad (31)
\end{aligned}$$

Distribution of $\mathbf{e}$ is symmetric,

$$(31)\leq 2^{\kappa+1}\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]. \qquad (32)$$

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

$$\begin{aligned}
(32) &\leq 2^{\kappa+1}\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]=2^{\kappa+1}\mathbb{E}_{\xi}\left[\mathbb{E}_{\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]\right]\\
&\leq 2^{\kappa+1}\mathbb{E}_{\xi}\left[\sqrt{\mathbb{E}_{\mathbf{e}}\left[||\mathbf{e}||_q^{2(1+\kappa)}\right]\mathbb{E}_{\mathbf{e}}\left[|f(x+\tau\mathbf{e},\xi)-\mathbb{E}_{\mathbf{e}}[f(x+\tau\mathbf{e},\xi)]|^{2(1+\kappa)}\right]}\right]. \qquad (33)
\end{aligned}$$

Next, we use $\mathbb{E}_{\mathbf{e}}\left[||e||_q^{2(1+\kappa)}\right]\leq a_q^{2(1+\kappa)}$ and Lemma 9.3 for $f(x+\tau\mathbf{e},\xi)$ with fixed $\xi$ and Lipschitz constant $M_2(\xi)\tau$,

$$\begin{aligned}
(33) &\leq 2^{\kappa+1}a_q^{1+\kappa}\mathbb{E}_{\xi}\left[\sqrt{\left(\frac{2^{-1/2}\tau^2M_2^2(\xi)}{d}\right)^{1+\kappa}}\right]\\
&= 2^{\kappa+1}a_q^{1+\kappa}\left(\frac{\tau^2 2^{-1/2}}{d}\right)^{(1+\kappa)/2}\mathbb{E}_{\xi}\left[M_2^{1+\kappa}(\xi)\right]\leq 2^{\kappa+1}\left(\sqrt{\frac{2^{-1/2}}{d}}a_qM_2\tau\right)^{1+\kappa}. \qquad (34)
\end{aligned}$$

Lets deal with (30) term. We use Cauchy-Schwartz inequality, Assumption 3 and $\mathbb{E}_{\mathbf{e}}\left[||e||_q^{2(1+\kappa)}\right]\leq a_q^{2(1+\kappa)}$ by definition

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|\delta(x+\tau\mathbf{e})-\delta(x-\tau\mathbf{e}))|^{1+\kappa}\right]\\
&\leq\sqrt{\mathbb{E}_{\mathbf{e}}\left[||e||_q^{2(1+\kappa)}\right]\mathbb{E}_{\mathbf{e}}\left[|\delta(x+\tau\mathbf{e})-\delta(x-\tau\mathbf{e}))|^{2(1+\kappa)}\right]}\\
&\leq a_q^{1+\kappa}2^{1+\kappa}\Delta^{1+\kappa}=(2a_q\Delta)^{1+\kappa}. \qquad (35)
\end{aligned}$$

Adding(34) and (35) we get final result

$$\begin{aligned}
\mathbb{E}_{\xi,\mathbf{e}}[\langle g(x,\xi,\mathbf{e}),r\rangle] &\leq \frac{1}{2}\left(\frac{d}{\tau}\right)^{1+\kappa}\left(2^{1+\kappa}\left(\sqrt{\frac{2^{-1/2}}{d}}a_q\tau M_2\right)^{1+\kappa}+(2a_q\Delta)^{1+\kappa}\right)=\\
&=2^\kappa\left(\frac{\sqrt{d}}{2^{1/4}}a_qM_2\right)^{1+\kappa}+2^\kappa\left(\frac{da_q\Delta}{\tau}\right)^{1+\kappa}.
\end{aligned}$$

□

## 10 Proof of Zeroth-Order Robust SMD Algorithm in Expectation Convergence.

**Theorem 10.1.** *Let function $f$ satisfying Assumptions 1, 2, 3, $q\in[1+\kappa,\infty]$, arbitrary number of iterations $T$, smoothing constant $\tau>0$ be given. Choose $\left(1,\frac{1+\kappa}{\kappa}\right)$-uniformly convex w.r.t. the $p$-norm function $\Psi_p(x)$ (e.g., $\Psi_p(x)=K_q^{1/\kappa}\phi^p(x)$, where $K_q$, $\phi^p$ are defined in* (4) *and* (5) *respectively). Set the stepsize $\nu=\frac{R_\Psi^{1/\kappa}}{\sigma_q}T^{-\frac{1}{1+\kappa}}$ with $\sigma_q$ given in Lemma 2.2, $R_\Psi^{\frac{1+\kappa}{\kappa}}\overset{def}{=}\frac{1+\kappa}{\kappa}\sup_{x\in\mathcal{S}}\{\Psi_p(x)-\Psi_p(x_0)\}$ and $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}}\overset{def}{=}\frac{1+\kappa}{\kappa}\sup_{x,y\in\mathcal{S}}D_{\Psi_p}(x,y)$. Let $\overline{x}_T$ be a point obtained by Algorithm 1 with the above parameters, and let $x^*\in\arg\min_{x\in\mathcal{S}}f(x)$.*

*1. Then*

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_\Psi \sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{36}$$

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

*2. With optimal* $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_\Psi d a_q \Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_\Psi d a_q \Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d} a_q M_2 R_\Psi}{T^{\frac{\kappa}{1+\kappa}}}. \tag{37}$$

*Proof.* By definition $x_* \in \arg\min_{x\in\mathcal{S}} f(x)$.

For $T$ iterations we use 3.2 Theorem of Convergence for $g_k(x_k, \xi_k, \mathbf{e}_k)$

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle \le \frac{\kappa}{\kappa+1}\frac{R_\Psi^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}||g_{k+1}||_q^{1+\kappa}.$$

Take expectation $\mathbb{E}_{\xi,\mathbf{e}}$ from both sides

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\xi,\mathbf{e}}\left[\langle g_{k+1}, x_k - x^*\rangle\right] \le \frac{\kappa}{\kappa+1}\frac{R_\Psi^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\xi,\mathbf{e}}\left[||g_{k+1}||_q^{1+\kappa}\right]. \tag{38}$$

Use Lemma 9.7 for the right part of inequality (38)

$$\frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\xi,\mathbf{e}}\left[||g_{k+1}||_q^{1+\kappa}\right] \le \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\sigma_q^{1+\kappa} \le \frac{\nu^\kappa}{1+\kappa}\sigma_q^{1+\kappa}. \tag{39}$$

Use Lemma 9.6 for the left part of inequality (38)

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\xi,\mathbf{e}}\left[\langle g_{k+1}, x_k - x^*\rangle\right] \ge \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\le k}\left[\mathbb{E}_{|\le k}[\langle g_{k+1}, x_k - x^*\rangle]\right]$$

$$\ge \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\le k}[\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle] - \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\mathbb{E}_{\le k}\mathbb{E}_{\mathbf{e}|\le k}[|\langle \mathbf{e}, x_k - x^*\rangle|]. \tag{40}$$

1. For the first term of (40) we use Lemma 2.1 and convexity of $\hat{f}_\tau(x)$

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

Define $\overline{x}_T = \frac{1}{T}\sum_{k=0}^{T-1} x_k$ and use Jensen's inequality

$$\frac{1}{T}\sum_{k=0}^{T-1}\left(\mathbb{E}_{|\le k}[\hat{f}_\tau(x_k)] - \hat{f}_\tau(x_*)\right) \ge \mathbb{E}_{\xi,\mathbf{e}}[\hat{f}_\tau(\overline{x}_T)] - \hat{f}_\tau(x^*).$$

Use approximation property from Lemma 2.1

$$\mathbb{E}_{\xi,\mathbf{e}}[\hat{f}_\tau(\overline{x}_T)] - \hat{f}_\tau(x^*) \ge \mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) - 2M_2\tau. \tag{41}$$

2. For the second term of (40) we use Lemma 9.5

$$
\begin{aligned}
&-\frac{d\Delta}{T\tau}\sum_{k=0}^{T-1}\mathbb{E}_{\mathbf{e}|\leq k}[|\langle \mathbf{e}, x_k - x^*\rangle|] \\
&\geq -\frac{d\Delta}{T\tau}\sum_{k=0}^{T-1}\frac{1}{\sqrt{d}}||x_k - x^*||_2 \\
&\overset{p\leq 2}{\geq} -\frac{d\Delta}{T\tau}\sum_{k=0}^{T-1}\frac{1}{\sqrt{d}}||x_k - x^*||_p. \qquad (42)
\end{aligned}
$$

Let's notice that $\Psi_p$ is $\left(1, \frac{1+\kappa}{\kappa}\right)$-uniformly convex function w.r.t. $p$ norm. Then by definition

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

Hence,

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

We combine (39), (41), (43) together

$$
\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_\Psi^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\sigma_q^{1+\kappa}. \qquad (44)
$$

By choosing optimal $\nu = \frac{R_\Psi^{1/\kappa}}{\sigma_q}T^{-\frac{1}{1+\kappa}}$ we get

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

Finally, we bound $\sigma_q$ with Lemma 9.1

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

And set optimal $\tau$

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

□

## 11 Proof of Clipping Algorithm in Expectation Convergence.

First, we prove some useful statements about clipped gradient vector properties. Similar proof can be found in [28].

**Lemma 11.1.** *Let $g_k = g(x_k, \xi, \mathbf{e})$. For $c > 0$ we define $\hat{g} = \frac{g}{||g||_q}\min(||g||_q, c)$.*

*1.*

$$
||\hat{g} - \mathbb{E}[\hat{g}]||_q \leq 2c. \qquad (45)
$$

2. *If $\mathbb{E}_{\xi,\mathbf{e}}[||g(x,\xi,\mathbf{e})||_q^{1+\kappa}] \leq \sigma_q^{1+\kappa}$, then*

   *(a)*

$$
\mathbb{E}_{\xi,\mathbf{e}}[||\hat{g}||_q^2] \leq \sigma_q^{1+\kappa}c^{1-\kappa}. \qquad (46)
$$

   *(b)*

$$
\mathbb{E}_{\xi,\mathbf{e}}[||\hat{g} - \mathbb{E}_{\xi,\mathbf{e}}[\hat{g}]||_q^2] \leq 4\sigma_q^{1+\kappa}c^{1-\kappa}. \qquad (47)
$$

*(c)*

$$||\mathbb{E}_{\xi,\mathbf{e}}[g] - \mathbb{E}_{\xi,\mathbf{e}}[\hat{g}]||_q \leq \frac{\sigma_q^{1+\kappa}}{c^\kappa}. \tag{48}$$

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

$$\begin{aligned} ||\hat{g} - \mathbb{E}[\hat{g}]||_q &\leq ||\hat{g}||_q + ||\mathbb{E}[\hat{g}]||_q \\ &\leq \left|\left|\frac{g}{||g||_q}\min(||g||_q, c)\right|\right|_q + \mathbb{E}\left[\left|\left|\frac{g}{||g||_q}\min(||g||_q, c)\right|\right|_q\right] \\ &= \min(||g||_q, c) + \mathbb{E}[\min(||g||_q, c)] \\ &\leq c + c = 2c. \end{aligned} \tag{49}$$

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

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

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

$$\begin{aligned} \mathbb{E}_{\xi,\mathbf{e}}[||\hat{g} - \mathbb{E}_{\xi,\mathbf{e}}[\hat{g}]||_q^2] &\leq 2\mathbb{E}_{\xi,\mathbf{e}}[||\hat{g}||_q^2 + 2||\mathbb{E}_{\xi,\mathbf{e}}[\hat{g}]||_q^2] \\ &\leq 2\mathbb{E}_{\xi,\mathbf{e}}[||\hat{g}||_q^2] + 2\mathbb{E}_{\xi,\mathbf{e}}[||\hat{g}||_q^2]] \\ &\overset{(46)}{\leq} 2\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} + 2\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa} \leq 4\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa}. \end{aligned} \tag{50}$$

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

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

From $||g||_q^{1+\kappa}\mathbb{1}_{\{||g||_q>c\}} \geq ||g||_q c^\kappa \mathbb{1}_{\{||g||_q>c\}}$ follows

$$\mathbb{E}_{\xi,\mathbf{e}}[||g||_q \mathbb{1}_{\{||g||_q>c\}}] \leq \mathbb{E}_{\xi,\mathbf{e}}[||g||_q \mathbb{1}_{\{||g||_q>c\}}] \leq \frac{\sigma_{q,\kappa}^{1+\kappa}}{c^\kappa}. \tag{51}$$

□

**Theorem 11.2.** *Let function $f$ satisfying Assumptions 1, 2, 3, $q \in [1+\kappa, \infty]$, arbitrary number of iterations $T$, smoothing constant $\tau > 0$ be given. Choose 1-strongly convex w.r.t. the p-norm function $\Psi_p(x)$. Set the step-size $\nu = \left(\frac{R_\Psi^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$ with $\sigma_q$ given in Lemma 2.2, $R_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x\in\mathcal{S}}\{\Psi_p(x) - \Psi_p(x_0)\}$ and $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x,y\in\mathcal{S}} D_{\Psi_p}(x,y)$. After set the clipping constant $c = \frac{2\kappa\mathcal{D}_\Psi}{(1-\kappa)\nu}$. Let $\overline{x}_T$ be a point obtained by Algorithm 2 with the above parameters, and let $x^* \in \arg\min\limits_{x\in\mathcal{S}} f(x)$.*

*1. Then,*

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_\Psi^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{52}$$

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

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_\Psi^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_qM_2R_\Psi^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{53}$$

*Proof.* Lets notice from proof of the Theorem 4.1 for the first term of (40) that for any $x_k$

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle + 2M_2\tau. \quad (54)$$

We define functions

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

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 them we can rewrite (54)

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle + 2M_2\tau$$
$$= \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\Big)}_{\text{D}} + \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\left(l_k(x_k) - l_k(x^*)\right)}_{\text{E}} + 2M_2\tau. \quad (55)$$

We bound D term by Lemma 11.1

$$\mathbb{E}_{\xi,\mathbf{e}}\left[\frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\Big)\right]$$
$$\leq \mathbb{E}_{\xi,\mathbf{e}}\left[\frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[g_{k+1}], x_k - x^*\rangle + \langle \mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\Big)\right]. \quad (56)$$

To bound the first term in (56) let's notice that $\Psi_p$ is $(1,2)$-uniformly convex function w.r.t. $p$ norm. Then by definition

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

Hence, we estimate $||x_k - u||_p \leq \mathcal{D}_\Psi, \forall u \in \mathcal{S}$.

By Cauchy–Schwarz inequality

$$\mathbb{E}_{\xi,\mathbf{e}}\left[\frac{1}{T}\sum_{k=0}^{T-1}\left(\langle \mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\right)\right]$$
$$\leq \frac{1}{T}\sum_{k=0}^{T-1}\left(\mathbb{E}_{\leq k}\mathbb{E}_{|\leq k}\left[||\mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}]||_q||x_k - x^*||_p\right]\right) \overset{(48)}{\leq} \mathcal{D}_\Psi\frac{\sigma_q^{1+\kappa}}{c^\kappa}. \quad (57)$$

To bound the second term in (56) we use Lemma 9.6 and Lemma 9.5

$$\mathbb{E}_{\xi,\mathbf{e}}\left[\frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[g_{k+1}], x_k - x^*\rangle\Big)\right]$$
$$\leq \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\mathbb{E}_{<k}\mathbb{E}_{\mathbf{e}|<k}[|\langle \mathbf{e}, x_k - x^*\rangle|]$$
$$\leq \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\frac{1}{\sqrt{d}}\mathbb{E}_{<k}[||x_k - x^*||_2]$$
$$\overset{p\leq 2}{\leq} \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\frac{1}{\sqrt{d}}\mathbb{E}_{<k}[||x_k - x^*||_p] \leq \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi. \quad (58)$$

Next, we bound E term

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

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

For SGD algorithm with $\hat{g}_k$ by Convergence Theorem 3.2 with bounded second moment

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle \leq \frac{1}{2}\frac{R_\Psi^2}{\nu T} + \frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}||\hat{g}_{k+1}||_q^2. \tag{59}$$

Using (59) with taken $\mathbb{E}_{\xi,\mathbf{e}}$ from both sides

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

By Lemma 11.1

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

Hence,

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\leq k}\mathbb{E}_{|\leq k}[\langle \hat{g}_{k+1}, x_k - x^*\rangle] \leq \frac{1}{2}\frac{R_\Psi^2}{\nu T} + \frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa}. \tag{60}$$

Combining bounds (57), (58), (60) together, we get

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

In order to get minimal upper bound we find optimal $c$

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

$h_1'(c) = \frac{\nu}{2}(1-\kappa)c^{-\kappa} - \kappa\frac{1}{c^{1+\kappa}}\mathcal{D}_\Psi = 0 \Rightarrow c^* = \frac{2\kappa\mathcal{D}_\Psi}{(1-\kappa)\nu}$.

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{1}{2}\frac{R_\Psi^2}{\nu T} + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi$$
$$+\sigma_{q,\kappa}^{1+\kappa}\left(\mathcal{D}^{1-\kappa}2^{-\kappa}\nu^\kappa\left[\frac{(1-\kappa)^\kappa}{\kappa^\kappa} + \frac{\kappa^{(1-\kappa)}}{(1-\kappa)^{(1-\kappa)}}\right]\right).$$

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

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

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{1}{2}\frac{R_\Psi^2}{\nu T} + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi + \sigma_q^{1+\kappa}\left(2\mathcal{D}_\Psi^{1-\kappa}\nu^\kappa\right). \tag{61}$$

Choosing optimal $\nu^*$ similarly we get

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

And

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

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

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

Then

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi + 2\frac{R_\Psi^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}. \tag{62}$$

In order to avoid $\nu \to \infty$ when $\kappa \to 0$ one can also choose $\nu^* = \left(\frac{R_\Psi^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$. Estimation (62) doesn't change.

Finally, we bound $\sigma_q$ with Lemma 9.1

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

And set optimal $\tau$

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

□

## 12 Proof of Zeroth-Order Clipping Algorithm in High Probability Convergence.

For next proof we need some classic measure concentration results. Bernstein inequality for martingale differences sum. Lemma 23 from [28].

**Lemma 12.1.** *Let $\{X_i\}_{i\geq 1}$ be martingale difference sequence, i.e. $\mathbb{E}[X_i|X_{j<i}] = 0, \forall i \geq 1$. Also $b, \sigma$ is such deterministic constants that $|X_i| < b$ and $\mathbb{E}[X_i^2|X_{j<i}] < \sigma^2$ almost surely. Then for arbitrary fixed number $\mu$ and for all $T$ with probability at least $1-\delta$*

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

And sum of squares of bounded random variables. Theorem 20 from [28].

**Lemma 12.2.** *Let $Z_i$ is a sequence of random variables adapted to a filtration $\mathcal{F}_t$. Further, suppose $|Z_i| < b, \mathbb{E}[Z_i^2] \leq \sigma^2$ almost surely. Then for any $\mu > 0$ with probability at least $1-\delta$*

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

$$+20\max(\mu^2, b^2)\log\left(\frac{112}{\delta}\left[\log\left(\frac{2\max(\mu, b)}{\mu}\right)+1\right]^2\right).$$

*By choosing $\mu = b \geq \sigma$*

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

$$+20b^2\log\left(\frac{12}{\delta}\right).$$

**Theorem 12.3.** *Let function $f$ satisfying Assumptions 1, 2, 3, $q \in [1+\kappa, \infty]$, arbitrary number of iterations $T$, smoothing constant $\tau > 0$ be given. Choose 1-strongly convex w.r.t. the p-norm function $\Psi_p(x)$. Set the clipping constant $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ with $\sigma_q$ given in Lemma 2.2. After set the stepsize $\nu = \frac{\mathcal{D}_\Psi}{c}$ with $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup_{x,y\in\mathcal{S}} D_{\Psi_p}(x,y)$. Let $\overline{x}_T$ be a point obtained by Algorithm 2 with the above parameters, and let $x^* \in \arg\min_{x\in\mathcal{S}} f(x)$. Additionally, for $\delta \in [0,1)$ we denote $\tilde{\delta}^{-1} = \frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2$ and $\beta = \left[3 + 8\log\frac{1}{\delta} + 12\log\frac{1}{\tilde{\delta}} + 20\log\frac{4}{\delta} + 4\sqrt{2\log\frac{1}{\delta}}\right]$.*

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

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi\sigma_q\beta}{2T^{\frac{\kappa}{1+\kappa}}}, \quad (63)$$

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

$$f(\overline{x}_T) - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + 4\sqrt{\frac{\beta M_2 \mathcal{D}_\Psi d a_q \Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{\beta\sqrt{d} a_q M_2 \mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}. \tag{64}$$

*Proof.* Lets notice from proof of the Theorem 4.1 for the first term of (40) that for any $x_k$

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle + 2M_2\tau. \tag{65}$$

For SGD algorithm with $\hat{g_k}$ by Convergence Theorem 3.2

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle \leq \frac{1}{2}\frac{R_\Psi^2}{\nu T} + \frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}||\hat{g}_{k+1}||_q^2. \tag{66}$$

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 (47) from Lemma 11.1. Thus by Lemma 12.2 with probability at least $1-\delta$

$$\frac{1}{T}\sum_{k=0}^{T-1}||\hat{g}_{k+1}||_q^2 \leq 12\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa}\log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2\right) + \frac{20}{T}c^2\log\left(\frac{12}{\delta}\right). \tag{67}$$

The left part of (66) can be rewritten as

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle = \frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1} - \nabla\hat{f}_\tau(x_k), x_k - x^*\rangle + \frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla\hat{f}_\tau(x_k), x_k - x^*\rangle$$

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

$$+ \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla\hat{f}_\tau(x_k), x_k - x^*\rangle}_{③}.$$

In the ① term we can proof that this is the sum of the martingale sequence difference. Indeed,

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

By (45) from Lemma 11.1

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

By (47) from Lemma 11.1

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

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

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

And we estimate $||x_k - u||_p \leq \mathcal{D}, \forall u \in \mathcal{S}$. Hence, by Bernstein's inequality Lemma 12.1 with probability at least $1-\delta$ and $\mu = \frac{1}{T}$

$$\frac{1}{T}\sum_{k=0}^{T-1}|\langle \hat{g}_{k+1} - \mathbb{E}_{|k}[\hat{g}_{k+1}], x_k - x^*\rangle| \leq \frac{4c\mathcal{D}_\Psi}{T}\log\frac{1}{\delta} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\sqrt{2\log\frac{1}{\delta}}. \tag{68}$$

For the (2) we use bound of D term from (55)

$$|\langle \mathbb{E}_{|k}[\hat{g}_{k+1}] - \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle| \leq \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta \frac{\sqrt{d}}{\tau}\right) \mathcal{D}_\Psi. \tag{69}$$

For the (3) we use (65)

$$f(\overline{x}_T) - f(x^*) - 2M_2\tau \leq \frac{1}{T}\sum_{k=0}^{T-1} \langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle. \tag{70}$$

Putting (67), (68), (69), (70) in (66), we get with probability at least $1-\delta$

$$\begin{aligned} f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta \frac{\sqrt{d}}{\tau}\right) \mathcal{D}_\Psi + \frac{1}{2}\frac{R_\Psi^2}{\nu T} \\ + \frac{\nu}{2}\left[12\sigma_q^{1+\kappa} c^{1-\kappa} \log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right) + 2\right]^2\right)\right] \\ + \frac{\nu}{2}\frac{20}{T} c^2 \log\left(\frac{12}{\delta}\right) + \frac{4c\mathcal{D}_\Psi}{T}\log\frac{1}{\delta} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2 \sqrt{2\log\frac{1}{\delta}}. \end{aligned} \tag{71}$$

Choosing $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ and putting it in (71), we get

$$\begin{aligned} f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \left(\frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} + \Delta \frac{\sqrt{d}}{\tau}\right) \mathcal{D}_\Psi + \frac{1}{2}\frac{R_0^2}{\nu T} \\ + \frac{\nu}{2}\left[12\sigma_q^2 T^{\frac{1-\kappa}{(1+\kappa)}} \log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right) + 2\right]^2\right)\right] \\ + \frac{\nu}{2}\frac{20\sigma_q^2}{T^{\frac{\kappa-1}{1+\kappa}}} \log\left(\frac{12}{\delta}\right) + \frac{4\sigma_q\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\log\frac{1}{\delta} + \frac{2\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\mathcal{D}_\Psi \sqrt{2\log\frac{1}{\delta}}. \end{aligned} \tag{72}$$

Define $\tilde{\delta}^{-1} = \frac{4}{\delta}\left[\log\left(\sqrt{T}\right) + 2\right]^2$ and choose $\nu = \frac{\mathcal{D}_\Psi}{c}$

$$\begin{aligned} f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \left(\frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} + \Delta \frac{\sqrt{d}}{\tau}\right) \mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi\sigma_q}{2T^{\frac{\kappa}{1+\kappa}}}\left[1 + 12\log\frac{1}{\tilde{\delta}} + 20\log\frac{4}{\delta}\right] + \\ + \frac{4\sigma_q\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\log\frac{1}{\delta} + \frac{2\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\mathcal{D}_\Psi \sqrt{2\log\frac{1}{\delta}}. \end{aligned} \tag{73}$$

Simplifying (73), we get

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

Finally, we bound $\sigma_q$ with Lemma 9.1

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

And set optimal $\tau$

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

□

## 13 Sketch of Proof of Zeroth-Order Restart Algorithms Convergence.

Proof of the Theorems 6.1, 6.2.

*Proof.* In this proof $\tilde{O}(\cdot)$ denotes $\log d$ factor.

**Step 1: Zeroth-Order Robust SMD in Expectation.**

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

Similarly to proof of Theorem 4.1 but with $\nu = \frac{\mathbb{E}\left[D_{\Psi_p}(x_0,x^*)\right]^{\frac{1}{1+\kappa}}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$ and bound $R_0 \le \mathcal{D}_\Psi$ one can get from (44)

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + 2\mathbb{E}\left[D_{\Psi_p}(x_0,x^*)\right]^{\frac{\kappa}{1+\kappa}} \sigma_q T^{-\frac{\kappa}{1+\kappa}}. \tag{74}$$

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

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le (2+1+2)\frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x_0,x^*)\right]^{\frac{\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{75}$$

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 $D_{\Psi_p}(x_0,x^*) = \tilde{O}(||x_0 - x^*||_p^{\frac{1+\kappa}{\kappa}})$ from [12](Remark 3) and denote $R_k = \mathbb{E}\left[||\overline{x}_k - x^*||_p^{\frac{1+\kappa}{\kappa}}\right]^{\frac{\kappa}{1+\kappa}}$.

Under $r$-growth Assumption 4

$$\frac{\mu_r}{2}\mathbb{E}\left[||\overline{x}_T - x^*||_p^r\right] \le \mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right). \tag{76}$$

Due to Jensen's inequality($r \ge \frac{1+\kappa}{\kappa}$)

$$\frac{\mu_r}{2}\mathbb{E}\left[||\overline{x}_T - x^*||_p^{\frac{1+\kappa}{\kappa}}\right]^{r/\frac{1+\kappa}{\kappa}} \le \frac{\mu_r}{2}\mathbb{E}\left[||\overline{x}_T - x^*||_p^r\right] \le \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right). \tag{77}$$

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

$$\frac{\mu_r}{2}R_1^r \le \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right) \le \frac{\mu_r}{2}\left(\frac{R_0}{2}\right)^r. \tag{78}$$

From right inequality of (78)

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

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

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

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

$$T = \sum_{k=1}^{N} T_k = \tilde{O}\left(\frac{A^{\frac{1+\kappa}{\kappa}}}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}}\sum_{k=0}^{N-1} 2^{k\left(\frac{(r-1)(1+\kappa)}{\kappa}\right)}\right) = \tilde{O}\left(\frac{A^{\frac{(1+\kappa)}{\kappa}}}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}}\left[2^{N\left(\frac{(r-1)(1+\kappa)}{\kappa}\right)} - 1\right]\right). \tag{79}$$

In order to get $\varepsilon$ accuracy

$$\mathbb{E}_{\xi,\mathbf{e}}[f(x_T)] - f(x^*) \le \varepsilon = \tilde{O}\left(R_{N-1}\frac{\sigma_q}{T_N^{\frac{\kappa}{1+\kappa}}}\right) \le \tilde{O}\left(\frac{\mu_r}{2}\left(\frac{R_{N-1}}{2}\right)^r\right) \le \tilde{O}\left(\frac{\mu_r}{2}\frac{R_0^r}{2^{(N-1)r}}\right).$$

Consequently,

$$N = \tilde{O}\left(\frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right), \tag{80}$$

$$T = \tilde{O}\left(\left[\frac{2^{\frac{r^2+1}{r}}\sigma_q}{\mu_r^{1/r}} \cdot \frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad T_k = \tilde{O}\left(\left[\frac{\sigma_q 2^{(1+r)}}{\mu_r R_0^{r-1}} 2^{k(r-1)}\right]^{\frac{1+\kappa}{\kappa}}\right). \tag{81}$$

In each restart section we get different bounds for noise absolute value. From $T_k$ formula from (79)

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2\sqrt{d}} \frac{1}{2^{k(2r-1)}}\right). \tag{82}$$

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

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

**Step 2: Zeroth-Order Clipping in Expectation.**

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

Similarly to proof of Theorem 5.2 but with $\nu^* = \mathbb{E}\left[D_{\Psi_p}(x_0, x^*)\right]^{\frac{1}{2}} \left(\frac{1}{4T\sigma_q^{1+\kappa}}\right)^{\frac{1}{1+\kappa}}$, $c^* = \frac{\mathbb{E}\left[D_{\Psi_p}(x_0,x^*)\right]^{\frac{1}{2}}}{\nu^*}$ one can get from (61)

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi + 2\frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x_0, x^*)\right]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{83}$$

Under obligatory condition $\Delta \le \frac{\sigma_q^2 \mathbb{E}\left[D_{\Psi_p}(x_0,x^*)\right]^{\frac{1}{2}}}{M_2\sqrt{d}T^{\frac{2\kappa}{1+\kappa}}}$ picking $\tau = \frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x_0,x^*)\right]^{\frac{1}{2}}}{M_2 T^{\frac{\kappa}{1+\kappa}}}$, we obtain from (83) estimate

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \le (2+1+2)\frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x_0, x^*)\right]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{84}$$

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 $D_{\Psi_p}(x_0, x^*) = \tilde{O}(||x_0 - x^*||_p^2)$ from [12](Remark 3) and denote $R_k = \mathbb{E}\left[||\overline{x}_k - x^*||_p^2\right]^{\frac{1}{2}}$.

Under $r$-growth Assumption 4

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

Due to Jensen's inequality $(r \ge 2)$

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

Next part of the proof is the same from **Step** 1 starting from (77). Analogically, we get the same $T_2$,$N_2$ and noise bounds from (81), (80) and (82) correspondingly.

**Step 3: Zeroth-Order Clipping 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 $N_3 \sim \log(\frac{1}{\varepsilon})$, thus

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

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

Similarly to proof of Theorem 5.3 but $\nu^* = \left[D_{\Psi_p}(x_0, x^*)\right]^{1/2}\left(\frac{1}{T\sigma_q^{1+\kappa}}\right)^{\frac{1}{1+\kappa}}$, $c^* = \frac{\mathbb{E}\left[D_{\Psi_p}(x_0,x^*)\right]^{\frac{1}{2}}}{\nu^*}$ one can get from (72) with probability at least $1 - \delta/N_3$

$$f(\overline{x}_T) - f(x^*) \le 2M_2\tau + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi$$

$$+\frac{\left[D_{\Psi_p}(x_0,x^*)\right]^{1/2}\sigma_q}{2T^{\frac{\kappa}{1+\kappa}}}\left[3+8\log\frac{1}{\delta}+12\log\frac{1}{\tilde{\delta}}+20\log\frac{4}{\delta}\;+4\sqrt{2\log\frac{1}{\delta}}\right].$$

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

Under obligatory condition $\Delta\leq\frac{\beta^2\sigma_q^2 D_{\Psi_p}^{\frac{1}{2}}(x_0,x^*)}{M_2\sqrt{d}T^{\frac{2\kappa}{1+\kappa}}}$ picking $\tau=\frac{\beta\sigma_q D_{\Psi_p}^{\frac{1}{2}}(x_0,x^*)}{M_2T^{\frac{\kappa}{1+\kappa}}}$, we obtain estimate

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

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 $D_{\Psi_p}(x_0,x^*)=\tilde{O}(||x_0-x^*||_p^2)$ from [12](Remark 3) and denote $R_k=||\overline{x}_k-x^*||_p$.

Under $r$-growth Assumption 4 ($r>1$)

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

Next part of the proof is the same from **Step** 1 starting from (77) with

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

Analogically, we get $T_3$,$N_3$ and noise bounds from (81), (80) and (82) correspondingly.

$$N=\tilde{O}\left(\frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right), \tag{85}$$

$$T=\tilde{O}\left(\left[\frac{2^{\frac{r^2+1}{r}}\sigma_q\beta}{\mu_r^{1/r}}\cdot\frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right),\quad T_k=\tilde{O}\left(\left[\frac{\sigma_q\beta 2^{(1+r)}}{\mu_r R_0^{r-1}}2^{k(r-1)}\right]^{\frac{1+\kappa}{\kappa}}\right). \tag{86}$$

In each restart section we get different bounds for noise absolute value. From $T_k$ formula from (86)

$$\Delta_k=\tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2\sqrt{d}}\frac{1}{2^{k(2r-1)}}\right). \tag{87}$$

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

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

□