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

Nikita Kornilov[1], Alexander Gasnikov[1,4,5], Pavel Dvurechensky[2], Darina Dvinskikh[3]

[1]Moscow Institute of Physics and Technology, Dolgoprudny, Russia.
[2]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany.
[3]High School of Economics, Moscow, Russia.
[4]Skoltech, Moscow, Russia.
[5]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow region, Russia.

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

In many optimization problems, the derivative information of the objective function is unavalable and the function values can be provided only via a black-box oracle. Such settings require the use of derivative-free methods. In this paper, we study a non-smooth optimization problem on a convex compact set with heavy-tailed stochastic noise (random noise with $(1+\kappa)$-th bounded moment) and adversarial noise and propose two new easy-to-implement algorithms that are optimal in terms of the number of zeroth-order oracle calls. Our algorithms are built upon efficient first-order methods for stochastic problems with heavy-tailed noise. The first algorithm is based on the mirror descent which is robust to heavy-tailed noise. The second algorithm is based on the gradient clipping technique. We discuss the differences between these two proposed algorithms as well as fine tune the parameters. Additionally, for the objective functions satisfying the r-growth condition, a faster algorithm is proposed using the restart technique. A particular attention is paid to the question of how large the adversarial noise can be in order to guarantee the optimality and convergence of the algorithms.

1 Introduction

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

$$\min_{x \in \mathcal{S}} f(x) \triangleq \mathbb{E}_\xi[f(x, \xi)], \tag{1}$$

where the values of the objective are available only through a zeroth-order noisy corrupted oracle , i.e.

$$\phi(x, \xi) = f(x, \xi) + \delta(x) \tag{2}$$

is the output of some black-box procedure, e.g., a simulation. We consider two-point zeroth-order oracle, i.e. for two query points $x, y \in \mathcal{S}$ we are given two outputs $\phi(x, \xi)$ and $\phi(y, \xi)$ with the same ξ. Function $\phi(x, \xi)$ can be considered as a noisy approximation of a Lipschitz function $f(x, \xi)$. This noise can be deterministic, stochastic or adversarial.

Zeroth-order methods were studied in a wide range of works [1, 2]. Particularly, under different assumptions on black-box oracle (in noisy or noiseless setup) the optimal oracle complexity was obtained [3–8]. This bound is proportional to $d\varepsilon^{-2}$, where ε is the desired precision to solve problem (1) in the function value. For saddle point problems, we refer to papers [9, 10] obtaining the same bound. This result is quite expected since the complexity is d times larger than the complexity of optimal stochastic gradient procedures. Factor d has a natural interpretation, since to approximate (stochastic) gradient it suffices to use $d + 1$ function values.[1] This is obvious in the smooth case (see e.g. [11]), and is not so trivial in the non-smooth case [7]. The key assumption of these papers is $M(\xi)$-Lipschitz continuity of $f(x, \xi)$ w.r.t. x and $\mathbb{E}_\xi[M(\xi)^2] < \infty$. However in modern learning problems stochastic gradients may not have finite variance. To this end, we aim to relax this assumption, namely we suppose $\mathbb{E}_\xi[M(\xi)^{1+\kappa}] < \infty$, where $\kappa \in (0, 1]$. For the first-order stochastic methods, the optimal oracle complexity is proportional to $\varepsilon^{-\frac{1+\kappa}{\kappa}}$ [12], thus for zeroth-order oracle we may expect the bound $(d/\varepsilon)^{\frac{1+\kappa}{\kappa}}$. In this paper we obtain the bound $\left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ matching the expected bound only for $\kappa = 1$. To the best of our knowledge, this poses the following open problem: is the bound $\left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ optimal in terms of the dependence on d? For smooth stochastic convex optimization problems with $(d+1)$-points stochastic zeroth-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 somewhat surprising since the dependence on d in our bound is very different from the known results for the case $\kappa = 1$. To the best of our knowledge, this paper provides the first known result for gradient-free methods without assuming a finite variance of the stochastic noise. Since we give an accurate analysis, including high-probability bounds,[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 $f(x, \xi)$ [10]. That is, even for sub-Gaussian tails it was an open question to obtain high-probability bounds for

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

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

gradient-free methods. The main challenge in obtaining our results is in the combination of the auxiliary gradient-free randomization and the original stochasticity of the oracle in the problem. The known inequalities on measure concentration do not allow obtaining the desired sub-Gaussian concentration for the output of the algorithm.

Gradient clipping technique has become increasingly popular for obtaining convergence guarantees in terms of high-probability [13–15]. Starting from the work [13] (see also [14, 15]) 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 there were proposed for optimal algorithms and the following results were obtained for their convergence guarantees: 1. in the expectation for general proximal setup and non-smooth stochastic convex optimization problems with infinite variance [16]; 2. in high-probability for general proximal setup and non-smooth online stochastic convex optimization problems with infinite variance [17]; 3. in high-probability for the Euclidean proximal setup and smooth and non-smooth stochastic convex optimization problems and variational inequalities with infinite variance [18–20]; 4. in high-probability for convergence of optimal variance-adaptive algorithm in the Euclidean proximal setup for non-smooth stochastic (strongly) convex optimization problems with infinite variance [21]. Since the aforementioned results are strongly correlated with each other, in this paper, we depart from the works [16, 17] to incorporate zero-order oracle into their algorithms. The developed technique, which reduces randomization caused by the gradient-free nature of the oracle to the original stochasticity, allows generalizing the results of other papers considered above in a similar manner. The idea of this reduction is not new and has already been used many times, see e.g. [3, 4, 6, 7]. But, all these works are significantly based on the assumption of finite variance of the stochastic noise. For the infinite noise variance setting, the technique requires significant generalizations, which we make in this paper. We expect, that based on these results it is possible to obtain new results for zero-order algorithms in the smooth setting and also in the setting of one-point feedback. 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 [9](rather than the gap used in [10]).[3] We leave this for future work.

[3] See the full version of the paper [10].

Our contributions

For d-dimensional optimization, we propose a two-point gradient-free clipping based algorithm with

$$\text{oracle complexity} \sim \left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}} \text{ in terms of high-probability and expectation,}$$

$$\text{maximum admissible level of adversarial noise} \lesssim \varepsilon^2/\sqrt{d}.$$

Furthermore, we generalize these results to r-growth condition problems which include strongly convex problems and problems with a sharp minimum . We use restart technique for clipping based algorithm above. In this case

$$\text{oracle complexity} \sim \left(\sqrt{d}/\varepsilon^{\frac{(r-1)}{r}}\right)^{\frac{1+\kappa}{\kappa}} \text{ in terms of high-probability and expectation,}$$

$$\text{maximum admissible level of adversarial noise} \lesssim \varepsilon^{(2-1/r)}/\sqrt{d}.$$

Organization

This paper is organized as follows. Section 2 presents the main objects and notions that are used to construct gradient-free algorithms. In Section 3, we present robust stochastic mirror descent algorithm. In Section 4, we provide our first algorithm which is based on mirror descent. In Section 5 we present our second algorithm based on gradient clipping. Finally, in Section 6 for the objective functions satisfying the r-growth condition, we propose a faster algorithm using the restart technique.

Notations

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 l_p-norm, i.e. $\|x\|_p = \left(\sum_{k=1}^d |x_k|^p\right)^{1/p}$. The corresponding dual norm is $\|y\|_q = \max_x\{\langle x, y\rangle| \; \|x\|_p \leq 1\}$, where q defined by the equality $\frac{1}{q}+\frac{1}{p} = 1$.

Let $B^p = \{x \in \mathbb{R}^d \mid \|x\|_p \leq 1\}$ be the p-ball with center at 0 and radius 1 and $S^p = \{x \in \mathbb{R}^d \mid \|x\|_p = 1\}$ be the p-sphere with center at 0 and radius 1.

The full expectation of a random variable X is denoted by $\mathbb{E}[X]$. The expected value w.r.t. random variables $Y_1, \dots, Y_n$ is denoted by $\mathbb{E}_{Y_1,\dots,Y_n}[X]$. At the $k+1$-th step of any of the optimization algorithms below, one can obtain the previous points $x_1, \dots, x_k$ and take conditional expectation w.r.t. them. For brevity we denote this expectation as

$$\mathbb{E}[\cdot|x_k, \dots, x_1] \overset{\text{def}}{=} \mathbb{E}_{|\leq k}[\cdot].$$

Finally, $x^* = \arg\min\limits_{x\in\mathcal{S}} f(x)$ denotes solution of optimization problem (1).

1.1 Assumptions

For a convex set $\mathcal{S} \subset \mathbb{R}^d$ and $\tau > 0$, let us introduce $\mathcal{S}_\tau = \mathcal{S} + \tau B^2$.

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

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

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

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

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

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

Next, we introduce the main objects and notions that are used to construct gradient-free algorithms. We consider only the uniform sampling from the unit Euclidean sphere, i.e.,

$$\mathbf{e} \sim \text{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\}) \overset{\text{def}}{=} U(S^2).$$

First of all, we define the smoothed function

$$\hat{f}_\tau(x) \triangleq \mathbb{E}_{\mathbf{e}\sim U(S^2)}[f(x + \tau\mathbf{e})] \tag{3}$$

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. The proof of the next lemma can be found in [22, Theorem 2.1].

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 τ is sufficiently small.

Following [7], the gradient of $\hat{f}_\tau(x)$ can be estimated by the following vector with $\tau > 0$:

$$g(x,\xi,\mathbf{e}) = \frac{d}{2\tau}(\phi(x + \tau\mathbf{e}, \xi) - \phi(x - \tau\mathbf{e}, \xi))\mathbf{e}. \tag{4}$$

Finally, the following important lemma 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*

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

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 provide generalization of the stochastic mirror descent algorithm to uniformly convex mirror map and its convergence guarantee from [16]. Mirror descent is a first-order method which generalizes the standard gradient descent to the non-Euclidean setup by using a mirror map that captures geometric structure of the problem [23].

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

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

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 the next lemma.

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

$$K_q \overset{def}{=} 10 \max\left\{1, (q-1)^{\frac{1+\kappa}{2}}\right\}. \tag{5}$$

Then,

$$\phi_p(x) \overset{def}{=} \frac{\kappa}{1+\kappa}\|x\|_p^{\frac{1+\kappa}{\kappa}} \tag{6}$$

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

Now we describe Stochastic Mirror Descent (SMD) algorithm. Let function $\Psi : \mathbb{R}^d \to \mathbb{R}$ called prox-function be (K, r)-uniformly convex w.r.t. the ℓ_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 ν and update vector g_{k+1} are as follows:

$$y_{k+1} = \nabla(\Psi^*)(\nabla\Psi(x_k) - \nu g_{k+1}), \quad x_{k+1} = \arg\min_{x \in \mathcal{S}} D_\Psi(x, y_{k+1}). \tag{7}$$

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

For the SMD algorithm (7) with standard 1-strongly convex function Ψ, the convergence theory is well known and given, e.g. in [24]. The next theorem generalizes these results and gives a convergence result of the SMD with uniformly convex Ψ.

Theorem 3.2. *Consider some $\kappa \in (0,1], p \in [1,\infty]$ and prox-function Ψ_p which is $\left(1, \frac{1+\kappa}{\kappa}\right)$-uniformly convex w.r.t. p norm. Then, for the SMD Algorithm outlined in (7), after T iterations with any $g_k \in \mathbb{R}^d, k \in \overline{1,T}$ and starting point $x_0 = \arg\min\limits_{x\in\mathcal{S}} \Psi_p(x)$ we have*

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle \leq \frac{\kappa}{\kappa+1}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\|g_{k+1}\|_q^{1+\kappa}, \tag{8}$$

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

The proof can be found in [16, Theorem 6]. Note, that when $\kappa = 1$ Ψ_p 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 (7) with the two-point gradient approximation (4). The former makes it possible to work 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 $\nu$, prox-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 $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla\Psi_p(x_k) - \nu g_{k+1})$
        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 theorem provides convergence guarantee for Algorithm 1 as well as optimal parameters.

Theorem 4.1. *Let function f satisfy 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 prox-function $\Psi_p(x)$ (e.g., $\Psi_p(x) = K_q^{1/\kappa}\phi_p(x)$, where K_q, ϕ_p are defined in (5) and (6) respectively). Set the stepsize $\nu = \frac{R_0^{1/\kappa}}{\sigma_q}T^{-\frac{1}{1+\kappa}}$ with σ_q given in Lemma 2.2, distance between starting point x_0 and solution x^* $R_0^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}D_{\Psi_p}(x^*, x_0)$ and diameter $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x,y\in\mathcal{S}} D_{\Psi_p}(x,y)$. Let $\overline{x}_T$ be the output of Algorithm 1 with the above parameters.*

1. *Then*
$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{9}$$
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* $\tau = \sqrt{\frac{\sqrt{d}\Delta\mathcal{D}_\Psi + 4R_0da_q\Delta T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_qM_2R_0}{T^{\frac{\kappa}{1+\kappa}}}. \tag{10}$$

Sketch of the Proof. Proof is based on Theorem 3.2 and inequality (8) from it

$$\underbrace{\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle\right]}_{\textcircled{1}} \leq \underbrace{\mathbb{E}\left[\frac{\kappa}{\kappa+1}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T}\right]}_{\textcircled{2}} + \underbrace{\mathbb{E}\left[\frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\|g_{k+1}\|_q^{1+\kappa}\right]}_{\textcircled{3}}. \tag{11}$$

$\textcircled{1}$ term in (11) 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

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

$\textcircled{3}$ term in (11) can be bounded with Lemma 2.2

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

Combining bounds together, we get

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

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

□

For the complete proof we refer to Section 10.

4.1 Discussion

Maximum admissible level of adversarial noise

Let $\varepsilon > 0$ be a desired accuracy in terms of the function value, i.e., our goal is to guarantee $\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq \varepsilon$. According to Theorem 4.1 in the absence of the adversarial noise, i.e., when $\Delta = 0$, the iteration complexity to reach accuracy ε is $T = \left(\frac{R_0\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}$ if τ is chosen sufficiently small. This complexity is optimal according to [12].

In order to obtain the same complexity in the case when $\Delta > 0$, we need to choose an appropriate value of τ and ensure that Δ is sufficiently small. Thus, the terms $2M_2\tau$ and $\frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi$ in (9) should be $= \varepsilon$. These conditions also make negligible the τ-depending term in σ_q. Consequently,

$$\text{when} \quad \tau = \frac{\varepsilon}{M_2} \text{ and } \Delta \leq \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}, \text{ we have } T = \left(\frac{R_0\sqrt{d}a_qM_2}{\varepsilon}\right)^{\frac{1+\kappa}{\kappa}}.$$

Otherwise, when $\Delta > \frac{\varepsilon^2}{M_2\sqrt{d}\mathcal{D}_\Psi}$, the convergence rate deteriorates. As we see in (10), 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 ε is $T = O\left(\frac{\sqrt{M_2R_0da_q\Delta}}{\varepsilon}\right)^{\frac{2(1+\kappa)}{\kappa}}$, which is worse than $O(\varepsilon^{-\frac{\kappa+1}{\kappa}})$ obtained when the error Δ can be controlled.

Dependency of the bounds on q and d

In Algorithm 1, we can freely choose $p \in [1, 2]$ and Ψ_p, which, depending on the compact convex set $\mathcal{S}$, lead to different values of $\mathcal{D}_\Psi, R_0, a_q$. It is desirable to reduce $a_q, \mathcal{D}_\Psi$ simultaneously, which would allow us to increase maximal noise level Δ and converge faster without changing the rate according to (9). Yet, unlike the well-studied SMD algorithm with strongly convex prox-functions Ψ_p, there are only few examples of effective choices of uniformly-convex prox-functions Ψ_p.

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

1.

$$\|\hat{g} - \mathbb{E}[\hat{g}]\|_q \leq 2c. \tag{12}$$

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

(a)

$$\mathbb{E}[\|\hat{g}\|_q^2] \leq \sigma_q^{1+\kappa} c^{1-\kappa}. \tag{13}$$

(b)

$$\mathbb{E}[\|\hat{g} - \mathbb{E}[\hat{g}]\|_q^2] \leq 4\sigma_q^{1+\kappa} c^{1-\kappa}. \tag{14}$$

(c)

$$\|\mathbb{E}[g] - \mathbb{E}[\hat{g}]\|_q \leq \frac{\sigma_q^{1+\kappa}}{c^\kappa}. \tag{15}$$

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 $c$, prox-function $\Psi_p$, smoothing constant $\tau$)
    $x_0 \leftarrow \arg\min_{x \in \mathcal{S}} \Psi_p(x)$
    for $k = 0, 1, \dots, T-1$ do
        Sample $\mathbf{e}_k \sim \text{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$ 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, $q \in [2, \infty]$, arbitrary number of iterations T, smoothing constant $\tau > 0$ be given. Choose 1-strongly convex w.r.t. the p-norm prox-function $\Psi_p(x)$. Set the stepsize $\nu = \left(\frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$ with σ_q given in Lemma 2.2, distance between starting point x_0 and solution x^* $R_0^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa} D_{\Psi_p}(x^*, x_0)$ and diameter $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa} \sup_{x,y \in \mathcal{S}} D_{\Psi_p}(x, y)$. 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.*

1. *Then,*

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{16}$$

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

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_qM_2R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{17}$$

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 δ is the desired confidence level.

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

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

$$f(\overline{x}_T) - f(x^*) \le 2M_2\tau + \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi + \tilde{O}\left(\frac{\mathcal{D}_\Psi\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right), \tag{18}$$

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

$$f(\overline{x}_T) - f(x^*) = \tilde{O}\left(\sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2\mathcal{D}_\Psi da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_qM_2\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\right). \tag{19}$$

Sketch of the Proof. Proof is based on Theorem 3.2 and inequality (8) for 1-strongly convex Ψ_p from it

$$\underbrace{\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle\right]}_{\textcircled{1}} \leq \mathbb{E}\left[\frac{1}{2}\frac{R_0^2}{\nu T}\right] + \underbrace{\mathbb{E}\left[\frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\|\hat{g}_{k+1}\|_q^2\right]}_{\textcircled{2}}. \tag{20}$$

$\textcircled{1}$ term in (20) 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

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

$\textcircled{2}$ term in (20) 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

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

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

For the complete proof we refer to Section 11.

Sketch of the Proof. 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 (8) for 1-strongly convex Ψ_p from it

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle \leq \frac{1}{2}\frac{R_0^2}{\nu T} + \underbrace{\frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\|\hat{g}_{k+1}\|_q^2}_{\textcircled{1}}. \tag{21}$$

Add $\pm\mathbb{E}_{|\leq k}[\hat{g}_{k+1}]$ and $\pm\hat{f}_\tau(x_k)$ to left part of (21)

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

We bound $\textcircled{1}$ term in (21) using Lemma 12.2 and $\textcircled{2}$ 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 $\textcircled{4}$ due to convexity of $\hat{f}_\tau(x)$ in Lemma 2.1 and $\textcircled{3}$ 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

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

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. □

For the complete proof we refer to Section 12.

5.1 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 admissible level of adversarial noise

Let ε 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 number of iterations is $T^{\frac{\kappa}{1+\kappa}} = \tilde{O}\left(\frac{\mathcal{D}_\Psi\sqrt{d}a_q M_2}{\varepsilon}\right)$ when $\tau \to 0$. This rate is optimal according to [12].

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 τ-depending term in σ_q. Consequently,

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

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 $T^{\frac{\kappa}{1+\kappa}} = \tilde{O}\left(\frac{M_2\mathcal{D}_\Psi d a_q \Delta}{\varepsilon^2}\right)$, which is twice worse than $\tilde{O}\left(\frac{\mathcal{D}_\Psi\sqrt{d}a_q M_2}{\varepsilon}\right)$.

Recommendations for choosing Ψ_p

With Algorithm 2, we can freely choose $p \in [1,2]$ and Ψ, which, depending on the compact convex set $\mathcal{S}$, will change $\mathcal{D}_\Psi, R_0, a_q$. The main task is to reduce $a_q, \mathcal{D}_\Psi$ simultaneously, which will allow us to increase maximal noise Δ and converge faster without changing the pace according to (18).

Next, we discuss some standard sets $\mathcal{S}$ and prox-functions Ψ_p taken from [24]. The two main setups are given by

1. Ball setup:
$$p = 2, \Psi_p(x) = \frac{1}{2}\|x\|_2^2, \tag{22}$$
2. Entropy setup:
$$p = 1, \Psi_p(x) = (1+\gamma)\sum_{i=1}^{d}(x_i + \gamma/d)\log(x_i + \gamma/d), \gamma > 0. \tag{23}$$

Introduce notation ball $B^p = \{x \in \mathbb{R}^d : \|x\|_p \le 1\}$ and standard simplex $\Delta_d^+ = \{x \in \mathbb{R}^d : x \ge 0, \sum_i x_i = 1\}$. 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 Δ up to $O\left(\log\frac{1}{\delta}\right)$ factor for each setup (row) and set (column).

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

	Δ_d^+	B^1	B^2	B^∞
Ball	$\sqrt{d}M_2/\varepsilon$	$\sqrt{d}M_2/\varepsilon$	$\sqrt{d}M_2/\varepsilon$	dM_2/ε
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 Δ up to $O(1)$ factor for Algorithm 2

	Δ_d^+	B^1	B^2	B^∞
Ball	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(dM_2)$
Entropy	$\varepsilon^2/(\sqrt{d\ln d}M_2)$	$\varepsilon^2/(\sqrt{d\ln d}M_2)$	$\varepsilon^2/(d\sqrt{\ln d}M_2)$	$\varepsilon^2/(\sqrt{d^3\ln d}M_2)$

From these tables, we see that for $\mathcal{S} = \Delta_d^+$ or B^1, the Entropy setup is preferable, while the Ball setup allows maximum feasible noise level Δ to be up to $\sqrt{\ln d}$ greater. Meanwhile, for $\mathcal{S} = B^2$ or B^∞, the Ball setup is better in terms of both convergence rate and noise robustness.

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 Ψ 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 ν and smoothing constant τ.

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 [26]) there is restart technique developed in [27] for algorithms acceleration.

Assumption 4. *Function 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, μ-strong convex w.r.t. the p-norm functions are 2-growth. Restart technique will work if Δ small enough to keep optimality of Algorithms 1 and 2. The general scheme of the Restart Algorithm is presented below.

Algorithm 3 Zeroth-Order Restart Algorithm

```
procedure ZEROTH-ORDER 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), prox-function $\Psi_p$)
    $x_0 \leftarrow \arg\min_{x \in \mathcal{S}} \Psi_p(x)$ or randomly
    for $k = 0, 1, \dots, N$ 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 $x_{\text{final}}$
        $x_0 \leftarrow x_{\text{final}}$
    end for
    return $x_{\text{final}}$
end procedure
```

The corresponding theorems generalizing the results for the Algorithms 1, 2 are also provided.

Theorem 6.1. *Let function f satisfies Assumptions 1, 2. Next, let ε 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 $R_k = R^0/2^k$.

Set number of restarts $N = \frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)$, sequence of number of steps $\{T_k\}_{k=1}^N = \left\{\left[\frac{\sigma_q 2^{(1+r)}}{\mu_r R_k^{r-1}}\right]^{\frac{1+\kappa}{\kappa}}\right\}_{k=1}^N$, 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 σ_q is from Lemma 2.2.

Moreover, Assumption 3 is held with

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

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

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

total number of steps is

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

on the last restart maximum Δ 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 ε be fixed accuracy and r-growth Assumption 4 is held with $r \geq 2$ for in expectation estimate or $r \geq 1$ 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 $R_k = R^0/2^k$.

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 σ_q is from Lemma 2.2.

Moreover, Assumption 3 is held with

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

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

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

or with probability at least $1-\delta$

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

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 Δ 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 Discussion

Maximum admissible level of adversarial noise

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

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

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 Δ_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, ε dependencies

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

$$\text{Algorithm 1 or 2}: \qquad T = O\left(\left[\frac{\mathcal{D}_\Psi\sqrt{d}a_qM_2}{\varepsilon}\right]^{\frac{1+\kappa}{\kappa}}\right),$$

$$\text{Algorithm 3}\ : \qquad T = \tilde{O}\left(\left[\frac{a_qM_2\sqrt{d}}{\mu_r^{1/r}}\cdot\frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right).$$

Furthermore, in Restart Algorithm total number of iteration depends only on a_q and maximum Δ threshold doesn't depend on $q, \mathcal{S}$ at all. Thus, it makes sense to take Entropy setup defined in (23) with basic Clipping Algorithm to lower a_q and leave only $\log d$ factor in T estimate.

7 Conclusion and Future Work

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. 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 [28], [29].

2. Additional assumptions on the properties of adversarial noise. For example, Lipschitz continuity in the spirit of Assumption 3 in [10]:

$$|\delta(x_1) - \delta(x_2)| \leq M\|x_1 - x_2\|_2, \qquad \forall x_1, x_2 \in \mathcal{S}.$$

3. Adaptive strategies and heuristic methods for selecting the algorithm's input parameters such as stepsize ν, smoothing constant τ, 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] Spall, J.C.: Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. John Wiley & Sons, Chichester (2005)

[2] Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-free Optimization. SIAM, Montreal (2009)

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

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

[5] Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Foundations of Computational Mathematics **17**, 527–566 (2017)

[6] Gasnikov, A.V., Krymova, E.A., Lagunovskaya, A.A., Usmanova, I.N., Fedorenko, F.A.: Stochastic online optimization. single-point and multi-point non-linear multi-armed bandits. convex and strongly-convex case. Automation and remote control **78**, 224–234 (2017)

[7] Shamir, O.: An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. The Journal of Machine Learning Research **18**(1), 1703–1713 (2017)

[8] Bayandina, A.S., Gasnikov, A.V., Lagunovskaya, A.A.: Gradient-free two-point methods for solving stochastic nonsmooth convex optimization problems with small non-random noises. Automation and Remote Control **79**, 1399–1408 (2018)

[9] Beznosikov, A., Sadiev, A., Gasnikov, A.: Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In: Mathematical Optimization Theory and Operations Research: 19th International Conference, MOTOR 2020, Novosibirsk, Russia, July 6–10, 2020, Revised Selected Papers 19, pp. 105–119 (2020). Springer

[10] Dvinskikh, D., Tominin, V., Tominin, Y., Gasnikov, A.: Gradient-free optimization for non-smooth minimax problems with maximum value of adversarial noise. arXiv preprint arXiv:2202.06114 (2022)

[11] Gasnikov, A., Dvinskikh, D., Dvurechensky, P., Gorbunov, E., Beznosikov, A., Lobanov, A.: Randomized gradient-free methods in convex optimization. arXiv preprint arXiv:2211.13566 (2022)

[12] Nemirovsky, A., Yudin, D.: Problem complexity and optimization method efficiency. M.: Nauka (1979)

[13] Nazin, A.V., Nemirovsky, A.S., Tsybakov, A.B., Juditsky, A.B.: Algorithms of robust stochastic optimization based on mirror descent method. Automation and Remote Control **80**, 1607–1627 (2019)

[14] Davis, D., Drusvyatskiy, D., Xiao, L., Zhang, J.: From low probability to high confidence in stochastic convex optimization. The Journal of Machine Learning Research **22**(1), 2237–2274 (2021)

[15] Gorbunov, E., Danilova, M., Shibaev, I., Dvurechensky, P., Gasnikov, A.: Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. arXiv preprint arXiv:2106.05958 (2021)

[16] Vural, N.M., Yu, L., Balasubramanian, K., Volgushev, S., Erdogdu, M.A.: Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. In: Conference on Learning Theory, pp. 65–102 (2022). PMLR

[17] Zhang, J., Cutkosky, A.: Parameter-free regret in high probability with heavy tails. arXiv preprint arXiv:2210.14355 (2022)

[18] Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., Richtárik, P.: High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. arXiv preprint

arXiv:2302.00999 (2023)

[19] Nguyen, T.D., Nguyen, T.H., Ene, A., Nguyen, H.L.: High probability convergence of clipped-sgd under heavy-tailed noise. arXiv preprint arXiv:2302.05437 (2023)

[20] Nguyen, T.D., Ene, A., Nguyen, H.L.: Improved convergence in high probability of clipped gradient methods with heavy tails. arXiv preprint arXiv:2304.01119 (2023)

[21] Liu, Z., Zhou, Z.: Stochastic nonsmooth convex optimization with heavy-tailed noises. arXiv preprint arXiv:2303.12277 (2023)

[22] Gasnikov, A., Novitskii, A., Novitskii, V., Abdukhakimov, F., Kamzolov, D., Beznosikov, A., Takáč, M., Dvurechensky, P., Gu, B.: The power of first-order smooth optimization for black-box non-smooth problems. arXiv preprint arXiv:2201.12289 (2022)

[23] Nemirovskij, A.S., Yudin, D.B.: Problem complexity and method efficiency in optimization (1983)

[24] Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications. SIAM, Philadelphia (2001)

[25] Zhang, J., Karimireddy, S.P., Veit, A., Kim, S., Reddi, S., Kumar, S., Sra, S.: Why are adaptive methods good for attention models? Advances in Neural Information Processing Systems **33**, 15383–15393 (2020)

[26] Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, Philadelphia (2021)

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

[28] Akhavan, A., Chzhen, E., Pontil, M., Tsybakov, A.B.: A gradient estimator via l1-randomization for online zero-order optimization with two point feedback. arXiv preprint arXiv:2205.13910 (2022)

[29] Lobanov, A., Alashqar, B., Dvinskikh, D., Gasnikov, A.: Gradient-free federated learning methods with l_1 and l_2-randomization for non-smooth convex stochastic optimization problems. arXiv preprint arXiv:2211.10783 (2022)

[30] Ledoux, M.: The concentration of measure phenomenon. ed. by peter landweber et al. vol. 89. Mathematical Surveys and Monographs. Providence, Rhode Island: American Mathematical Society, 181 (2005)

[31] Gorbunov, E., Vorontsova, E.A., Gasnikov, A.V.: On the upper bound for the expectation of the norm of a vector uniformly distributed on the sphere and the

phenomenon of concentration of uniform measure on the sphere. Mathematical Notes **106** (2019)

[32] Gasnikov, A.V., Nesterov, Y.E.: Universal method for stochastic composite optimization problems. Computational Mathematics and Mathematical Physics **58**, 48–64 (2018)

9 Proofs of Lemmas

9.1 General results

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

$$\|x-y\|_q^{1+\kappa} \le 2^\kappa \|x\|_q^{1+\kappa} + 2^\kappa \|y\|_q^{1+\kappa}, \tag{24}$$

2.

$$\forall x, y \ge 0, \kappa \in [0,1] : (x+y)^\kappa \le x^\kappa + y^\kappa. \tag{25}$$

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

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

- Proposition 9 from [16].

□

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

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

$$\begin{aligned} |f(x) - f(y)| &= |\mathbb{E}_\xi[f(x,\xi) - f(y,\xi)]| \overset{\text{Jensen's inq}}{\le} \mathbb{E}_\xi[|f(x,\xi) - f(y,\xi)|] \\ &\le \mathbb{E}_\xi[M_2]\|x-y\|_2 \overset{\text{Jensen's inq}}{\le} \mathbb{E}_\xi[M_2^{(1+\kappa)}]^{\frac{1}{1+\kappa}}\|x-y\|_2 \\ &\le M_2\|x-y\|_2. \end{aligned} \tag{26}$$

□

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*

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

Proof. A standard result of the measure concentration on the Euclidean unit sphere implies that $\forall t > 0$

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

(see the proof of Proposition 2.10 and Corollary 2.6 in [30]). Therefore,

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

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

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

□

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

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

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

This Lemma is generalization of Lemma from [31] for $\kappa < 1$.

Proof. We use Lemma 1 auxiliary Lemma from Theorem 1 from [31].

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

$$\mathbb{E}\left[|e_2|^q\right] \leq \left(\frac{q-1}{d}\right)^{\frac{q}{2}}, \quad q \geq 2. \tag{28}$$

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

$$\|x\|_{q_1} \leq \|x\|_{q_2}. \tag{29}$$

Then

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

Due to Jensen's inequality and equally distributed e_k

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

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

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

Therefore,

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

Using (28) with $2q$

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

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 (29)

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

Consequently,

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

□

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*

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

Lemma 9.6. *Let $g(x, \xi, \mathbf{e})$ be defined in* (4) *and $\hat{f}_\tau(x)$ be defined in* (3)*. Then, the following holds under Assumption 3:*

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

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

Proof. 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

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

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

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

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

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

Adding two terms together we get necessary result.

□

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

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

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

Proof.

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

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

$$\leq 2^\kappa\left(\frac{d}{2\tau}\right)^{1+\kappa}\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-f(x-\tau\mathbf{e},\xi)|^{1+\kappa}\right] \tag{30}$$

$$+\,2^\kappa\left(\frac{d}{2\tau}\right)^{1+\kappa}\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|\delta(x+\tau\mathbf{e})-\delta(x-\tau\mathbf{e})|^{1+\kappa}\right]. \tag{31}$$

Lets deal with (30) term. For all $\alpha(\xi)$

$$\begin{aligned}&\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-f(x-\tau\mathbf{e},\xi)|^{1+\kappa}\right]\\ &\leq \mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|(f(x+\tau\mathbf{e},\xi)-\alpha)-(f(x-\tau\mathbf{e},\xi)-\alpha)|^{1+\kappa}\right]\\ &\overset{(24)}{\leq} 2^\kappa\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]+2^\kappa\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|f(x-\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right].\end{aligned} \tag{32}$$

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

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

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

$$\begin{aligned}(33) &\leq 2^{\kappa+1}\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]\\ &= 2^{\kappa+1}\mathbb{E}_{\xi}\left[\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi)-\alpha|^{1+\kappa}\right]\right]\\ &\leq 2^{\kappa+1}\mathbb{E}_{\xi}\left[\sqrt{\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right]\mathbb{E}_{\mathbf{e}}\left[|f(x+\tau\mathbf{e},\xi)-\mathbb{E}_{\mathbf{e}}[f(x+\tau\mathbf{e},\xi)]|^{2(1+\kappa)}\right]}\right]\end{aligned} \tag{34}$$

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

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

$$\leq 2^{\kappa+1}\left(\sqrt{\frac{2^{-1/2}}{d}a_q M_2\tau}\right)^{1+\kappa}. \tag{35}$$

Lets deal with (31) term. We use Cauchy-Schwartz inequality, Assumption 3 and $\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right] \leq a_q^{2(1+\kappa)}$ by definition

$$\begin{aligned}
&\mathbb{E}_{\xi,\mathbf{e}}\left[\|\mathbf{e}\|_q^{1+\kappa}|\delta(x+\tau\mathbf{e})-\delta(x-\tau\mathbf{e})|^{1+\kappa}\right]\\
&\leq \sqrt{\mathbb{E}_{\mathbf{e}}\left[\|\mathbf{e}\|_q^{2(1+\kappa)}\right]\mathbb{E}_{\mathbf{e}}\left[|\delta(x+\tau\mathbf{e})-\delta(x-\tau\mathbf{e})|^{2(1+\kappa)}\right]}\\
&\leq a_q^{1+\kappa}2^{1+\kappa}\Delta^{1+\kappa} = (2a_q\Delta)^{1+\kappa}. \qquad (36)
\end{aligned}$$

Adding(35) and (36) we get final result

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

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

□

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

Theorem 10.1. *Let function f satisfy 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 prox-function $\Psi_p(x)$ (e.g., $\Psi_p(x) = K_q^{1/\kappa}\phi_p(x)$, where K_q, ϕ_p are defined in (5) and (6) respectively). Set the stepsize $\nu = \frac{R_0^{1/\kappa}}{\sigma_q}T^{-\frac{1}{1+\kappa}}$ with σ_q given in Lemma 2.2, distance between starting point x_0 and solution x^* $R_0^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}D_{\Psi_p}(x^*,x_0)$ and diameter $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa}\sup\limits_{x,y\in\mathcal{S}} D_{\Psi_p}(x,y)$. Let $\overline{x}_T$ be the output of Algorithm 1 with the above parameters.*

1. *Then*

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{37}$$

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

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_qM_2R_0}{T^{\frac{\kappa}{1+\kappa}}}. \quad (38)$$

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

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

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

Take full expectation $\mathbb{E}$ from both sides

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

Use Lemma 9.7 for the right part of inequality (39)

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

Use Lemma 9.6 for the left part of inequality (39) and conditional math expectation

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

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

1. For the first term of (41) we use Lemma 2.1 and convexity of $\hat{f}_\tau(x)$

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

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

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

Use approximation property from Lemma 2.1

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

2. For the second term of (41) we use Lemma 9.5

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

Let's notice that Ψ_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,

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

We combine (40), (42), (44) together

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

By choosing optimal $\nu = \frac{R_0^{1/\kappa}}{\sigma_q}T^{-\frac{1}{1+\kappa}}$ we get

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

Finally, we bound σ_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 τ

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

□

11 Proof of Zeroth-Order Clipping Algorithm in Expectation Convergence

First, we prove some useful statements about clipped gradient vector properties. Similar proof can be found in [17].

Lemma 11.1. *For $c > 0$ and stochastic vector $g = g(x, \xi, \mathbf{e})$ we define $\hat{g} = \frac{g}{\|g\|_q}\min(\|g\|_q, c)$. Then*

1.

$$\|\hat{g} - \mathbb{E}[\hat{g}]\|_q \leq 2c. \tag{46}$$

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

(a)

$$\mathbb{E}[\|\hat{g}\|_q^2] \leq \sigma_q^{1+\kappa} c^{1-\kappa}, \tag{47}$$

(b)

$$\mathbb{E}[\|\hat{g} - \mathbb{E}[\hat{g}]\|_q^2] \leq 4\sigma_q^{1+\kappa} c^{1-\kappa}, \tag{48}$$

(c)

$$\|\mathbb{E}[g] - \mathbb{E}[\hat{g}]\|_q \leq \frac{\sigma_q^{1+\kappa}}{c^\kappa}. \tag{49}$$

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

$$\begin{aligned} \|\hat{g} - \mathbb{E}[\hat{g}]\|_q &\leq \|\hat{g}\|_q + \|\mathbb{E}[\hat{g}]\|_q \\ &\leq \left\| \frac{g}{\|g\|_q}\min(\|g\|_q, c)\right\|_q + \mathbb{E}\left[\left\|\frac{g}{\|g\|_q}\min(\|g\|_q, c)\right\|_q\right] \\ &= \min(\|g\|_q, c) + \mathbb{E}[\min(\|g\|_q, c)] \\ &\leq c + c = 2c. \end{aligned} \tag{50}$$

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

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

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

$$\begin{aligned} \mathbb{E}[\|\hat{g} - \mathbb{E}[\hat{g}]\|_q^2] &\leq 2\mathbb{E}[\|\hat{g}\|_q^2 + 2\|\mathbb{E}[\hat{g}]\|_q^2] \\ &\leq 2\mathbb{E}[\|\hat{g}\|_q^2] + 2\mathbb{E}[\|\hat{g}\|_q^2]] \\ &\overset{(47)}{\leq} 2\sigma_{q,\kappa}^{1+\kappa} c^{1-\kappa} + 2\sigma_{q,\kappa}^{1+\kappa} c^{1-\kappa} \leq 4\sigma_{q,\kappa}^{1+\kappa} c^{1-\kappa}. \end{aligned} \tag{51}$$

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

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

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

$$\mathbb{E}[\|g\|_q \mathbb{1}_{\{\|g\|_q>c\}}] \leq \mathbb{E}[\|g\|_q \mathbb{1}_{\{\|g\|_q>c\}}] \leq \frac{\sigma_{q,\kappa}^{1+\kappa}}{c^\kappa}. \tag{52}$$

□

Theorem 11.2. *Let function f satisfying Assumptions 1, 2, 3, $q \in [2,\infty]$, arbitrary number of iterations T, smoothing constant $\tau > 0$ be given. Choose 1-strongly convex w.r.t. the p-norm prox-function $\Psi_p(x)$. Set the stepsize $\nu = \left(\frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$ with σ_q given in Lemma 2.2, distance between starting point x_0 and solution x^* $R_0^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa} D_{\Psi_p}(x^*, x_0)$ and diameter $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} \overset{def}{=} \frac{1+\kappa}{\kappa} \sup\limits_{x,y\in\mathcal{S}} D_{\Psi_p}(x,y)$. 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.*

1. *Then,*

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{53}$$

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

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + \sqrt{\frac{32M_2R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}da_q\Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{2\sqrt{d}a_qM_2R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{54}$$

Proof. Lets notice from proof of the Theorem 4.1 for the first term of (41) that for any x_k

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{k=0}^{T-1}\langle\nabla\hat{f}_\tau(x_k), x_k - x^*\rangle + 2M_2\tau. \tag{55}$$

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 (55)

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

$$= \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\Big)}_{\text{D}} + \underbrace{\frac{1}{T}\sum_{k=0}^{T-1}(l_k(x_k) - l_k(x^*))}_{\text{E}} + 2M_2\tau. \quad (56)$$

We bound D term by Lemma 11.1

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\Big(\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\Big)\right] \\ &\leq \mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k) - \mathbb{E}_{|\leq k}[g_{k+1}], x_k - x^*\rangle\right] \\ &+ \mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\langle \mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\right]. \end{aligned} \quad (57)$$

In order to bound the first term in (57) let's notice that Ψ_p is $(1,2)$-uniformly convex function w.r.t. p norm. Then by definition

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

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

By the Cauchy–Schwarz inequality

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{T}\sum_{k=0}^{T-1}\big(\langle \mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle\big)\right] \\ &\leq \frac{1}{T}\sum_{k=0}^{T-1}\big(\mathbb{E}\left[\mathbb{E}_{|\leq k}\left[\|\mathbb{E}_{|\leq k}[g_{k+1}] - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}]\|_q\|x_k - x^*\|_p\right]\right]\big) \overset{(49)}{\leq} \mathcal{D}_\Psi\frac{\sigma_q^{1+\kappa}}{c^\kappa}. \end{aligned} \quad (58)$$

To bound the second term in (57) we use Lemma 9.6 and Lemma 9.5

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

$$\leq \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\frac{1}{\sqrt{d}}\mathbb{E}[\|x_k - x^*\|_2]$$

$$\overset{p\leq 2}{\leq} \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\frac{1}{\sqrt{d}}\mathbb{E}[\|x_k - x^*\|_p] \leq \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi. \tag{59}$$

Next, we bound E term

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

For the SMD algorithm with $\hat{g_k}$ by Convergence Theorem 3.2 with bounded second moment

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle \leq \frac{1}{2}\frac{R_0^2}{\nu T} + \frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\|\hat{g}_{k+1}\|_q^2. \tag{60}$$

Using (60) with taken $\mathbb{E}$ from both sides

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\langle \hat{g}_{k+1}, x_k - x^*\rangle] = \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\mathbb{E}_{|\leq k}[\langle \hat{g}_{k+1}, x_k - x^*\rangle]\right]$$
$$\leq \frac{1}{2}\frac{R_0^2}{\nu T} + \frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\mathbb{E}_{|\leq k}[\|\hat{g}_{k+1}\|_q^2]\right]. \tag{61}$$

By Lemma 11.1

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

And hence,

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\langle \hat{g}_{k+1}, x_k - x^*\rangle] \leq \frac{1}{2}\frac{R_0^2}{\nu T} + \frac{\nu}{2}\sigma_q^{1+\kappa}c^{1-\kappa}. \tag{62}$$

Combining bounds (58), (59), (62) together, we get

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

In order to get minimal upper bound we find optimal c

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

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

$$\begin{aligned}\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq\ & 2M_2\tau + \frac{1}{2}\frac{R_0^2}{\nu T} + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi \\ & + \sigma_{q,\kappa}^{1+\kappa}\left(\mathcal{D}^{1-\kappa}2^{-\kappa}\nu^\kappa\left[\frac{(1-\kappa)^\kappa}{\kappa^\kappa} + \frac{\kappa^{(1-\kappa)}}{(1-\kappa)^{(1-\kappa)}}\right]\right). \end{aligned} \tag{63}$$

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

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

we simplify (63)

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{1}{2}\frac{R_0^2}{\nu T} + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi + \sigma_q^{1+\kappa}\left(2\mathcal{D}_\Psi^{1-\kappa}\nu^\kappa\right). \tag{64}$$

Choosing optimal ν^* similarly we get

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

And

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

Considering bound of $\kappa \in [0,1]$ we can calculate the upper bound

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

Then

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \Delta\frac{\sqrt{d}}{\tau}\mathcal{D}_\Psi + 2\frac{R_0^{\frac{2\kappa}{1+\kappa}}\mathcal{D}_\Psi^{\frac{1-\kappa}{1+\kappa}}\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}. \tag{65}$$

In order to avoid $\nu \to \infty$ when $\kappa \to 0$ one can also choose $\nu^* = \left(\frac{R_0^2}{4T\sigma_q^{1+\kappa}\mathcal{D}_\Psi^{1-\kappa}}\right)^{\frac{1}{1+\kappa}}$. Estimation (65) doesn't change.

Finally, we bound σ_q with Lemma 9.1

$$\sigma_q \leq 2\left(\frac{\sqrt{d}}{2^{1/4}}a_qM_2\right) + 2\left(\frac{da_q\Delta}{\tau}\right),$$

And set optimal τ

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

□

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 [17].

Lemma 12.1. *Let $\{X_i\}_{i\geq 1}$ be martingale difference sequence, i.e. $\mathbb{E}[X_i|X_{i-1},\dots,X_1] = 0, \forall i \geq 1$. Also b, σ is such deterministic constants that $|X_i| < b$ and $\mathbb{E}[X_i^2|X_{i-1},\dots,X_1] < \sigma^2$ almost surely for $i \geq 1$. Then for arbitrary fixed number μ 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 [17].

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 $1-\delta$*

$$\sum_{k=1}^{T} Z_k^2 \leq 3T\sigma^2\log\left(\frac{4}{\delta}\left[\log\left(\sqrt{\frac{\sigma^2 T}{\mu^2}}\right)+2\right]^2\right) + 20\max(\mu^2, b^2)\log\left(\frac{112}{\delta}\left[\log\left(\frac{2\max(\mu,b)}{\mu}\right)+1\right]^2\right). \tag{66}$$

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, $q \in [2,\infty]$, arbitrary number of iterations T, smoothing constant $\tau > 0$ be given. Choose 1-strongly convex w.r.t. the p-norm prox-function $\Psi_p(x)$. Set the clipping constant $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ with σ_q given in Lemma 2.2. After set the stepsize $\nu = \frac{\mathcal{D}_\Psi}{c}$ with diameter $\mathcal{D}_\Psi^2 \overset{def}{=}$*

$2 \sup\limits_{x,y \in \mathcal{S}} D_{\Psi_p}(x,y)$. *Let $\overline{x}_T$ be a point obtained by Algorithm 2 with the above parameters. 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 $1-\delta$*

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\Delta\sqrt{d}}{\tau}\mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi \sigma_q \beta}{2T^{\frac{\kappa}{1+\kappa}}}, \tag{67}$$

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

$$f(\overline{x}_T) - f(x^*) \leq \sqrt{8M_2\sqrt{d}\Delta\mathcal{D}_\Psi} + 4\sqrt{\frac{\beta M_2 \mathcal{D}_\Psi d a_q \Delta}{T^{\frac{\kappa}{(1+\kappa)}}}} + \frac{\beta\sqrt{d}a_q M_2 \mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}. \tag{68}$$

Proof. Lets notice from proof of the Theorem 4.1 for the first term of (41) that for any x_k

$$f(\overline{x}_T) - f(x^*) \leq \frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle + 2M_2\tau. \tag{69}$$

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

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle \hat{g}_{k+1}, x_k - x^*\rangle \leq \frac{1}{2}\frac{R_0^2}{\nu T} + \frac{\nu}{2}\frac{1}{T}\sum_{k=0}^{T-1}\|\hat{g}_{k+1}\|_q^2. \tag{70}$$

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

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\hat{g}_{k+1}\|_q^2 \leq 12\sigma_{q,\kappa}^{1+\kappa}c^{1-\kappa}\log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2\right) + \frac{20}{T}c^2\log\left(\frac{12}{\delta}\right). \tag{71}$$

The left part of (70) can be rewritten as

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

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

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

In the $\textcircled{1}$ term we can proof that this is the sum of the martingale sequence difference. Indeed,

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

By (46) from Lemma 11.1

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

By (48) from Lemma 11.1

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

Lets notice that Ψ_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 $\mu = \frac{1}{T}$

$$\frac{1}{T}\sum_{k=0}^{T-1}|\langle \hat{g}_{k+1} - \mathbb{E}_{|\leq k}[\hat{g}_{k+1}], x_k - x^*\rangle| \leq \frac{4c\mathcal{D}_\Psi}{T}\log\frac{1}{\delta} + \frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\sqrt{2\log\frac{1}{\delta}}. \quad (72)$$

For the $\textcircled{2}$ we use bound of D term from (56)

$$|\langle \mathbb{E}_{|\leq k}[\hat{g}_{k+1}] - \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle| \leq \left(\frac{\sigma_q^{1+\kappa}}{c^\kappa} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi. \quad (73)$$

For the $\textcircled{3}$ we use (69)

$$f(\overline{x}_T) - f(x^*) - 2M_2\tau \leq \frac{1}{T}\sum_{k=0}^{T-1}\langle \nabla \hat{f}_\tau(x_k), x_k - x^*\rangle. \quad (74)$$

Putting (71), (72), (73), (74) in (70), we get with probability at least $1-\delta$

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

$$+\frac{\nu}{2}\frac{20}{T}c^2\log\left(\frac{12}{\delta}\right)+\frac{4c\mathcal{D}_\Psi}{T}\log\frac{1}{\delta}+\frac{\sqrt{4\sigma_q^{1+\kappa}c^{1-\kappa}}}{\sqrt{T}}\mathcal{D}_\Psi^2\sqrt{2\log\frac{1}{\delta}}. \tag{75}$$

Choosing $c = T^{\frac{1}{(1+\kappa)}}\sigma_q$ and putting it in (75), we get

$$f(\overline{x}_T) - f(x^*) \le 2M_2\tau + \left(\frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}} + \Delta\frac{\sqrt{d}}{\tau}\right)\mathcal{D}_\Psi + \frac{1}{2}\frac{R_0^2}{\nu T}$$

$$+\frac{\nu}{2}\left[12\sigma_q^2 T^{\frac{1-\kappa}{(1+\kappa)}}\log\left(\frac{4}{\delta}\left[\log\left(\sqrt{T}\right)+2\right]^2\right)\right]$$

$$+\frac{\nu}{2}\frac{20\sigma_q^2}{T^{\frac{\kappa-1}{1+\kappa}}}\log\left(\frac{12}{\delta}\right)+\frac{4\sigma_q\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\log\frac{1}{\delta}+\frac{2\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\mathcal{D}_\Psi\sqrt{2\log\frac{1}{\delta}}. \tag{76}$$

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

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

$$+\frac{4\sigma_q\mathcal{D}_\Psi}{T^{\frac{\kappa}{1+\kappa}}}\log\frac{1}{\delta}+\frac{2\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\mathcal{D}_\Psi\sqrt{2\log\frac{1}{\delta}}. \tag{77}$$

Simplifying (77), we get

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

$$+\frac{\mathcal{D}_\Psi\sigma_q}{2T^{\frac{\kappa}{1+\kappa}}}\left[3+8\log\frac{1}{\delta}+12\log\frac{1}{\tilde{\delta}}+20\log\frac{4}{\delta}\ +4\sqrt{2\log\frac{1}{\delta}}\right].$$

Finally, we bound σ_q with Lemma 9.1

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

And set optimal τ

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

□

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 $\nu = \frac{\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{1}{1+\kappa}}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$ and bound $R_0 \le \mathcal{D}_\Psi$ one can get from (45)

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + 2\mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{\kappa}{1+\kappa}} \sigma_q T^{-\frac{\kappa}{1+\kappa}}. \tag{78}$$

Under obligatory condition $\Delta \le \frac{\sigma_q^2 \mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{\kappa}{1+\kappa}}}{M_2\sqrt{d}T^{\frac{2\kappa}{1+\kappa}}}$ picking $\tau = \frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{\kappa}{1+\kappa}}}{M_2 T^{\frac{\kappa}{1+\kappa}}}$, we obtain from (78) estimate

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le (2+1+2)\,\frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x^*,x_0)\right]^{\frac{\kappa}{1+\kappa}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{79}$$

In σ_q τ-depending term has $T^{\frac{-2\kappa}{1+\kappa}}$ decreasing rate, so we neglect it. Next, let's use fact that $D_{\Psi_p}(x^*,x_0) = \tilde{O}(\|x_0 - x^*\|_p^{\frac{1+\kappa}{\kappa}})$ from [32](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

$$\frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_p^r\right] \le \mathbb{E}[f(\overline{x}_T)] - f(x^*) \le \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right). \tag{80}$$

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

$$\frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_p^{\frac{1+\kappa}{\kappa}}\right]^{r/\frac{1+\kappa}{\kappa}} \le \frac{\mu_r}{2}\mathbb{E}\left[\|\overline{x}_T - x^*\|_p^r\right] \le \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right). \tag{81}$$

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

$$\frac{\mu_r}{2}R_1^r \le \tilde{O}\left(R_0 \frac{\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}\right) \le \frac{\mu_r}{2}\left(\frac{R_0}{2}\right)^r. \tag{82}$$

From right inequality of (82)

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

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

$$\begin{aligned} T = \sum_{k=1}^{N} T_k &= \tilde{O}\left(\frac{A^{\frac{1+\kappa}{\kappa}}}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}} \sum_{k=0}^{N-1} 2^{k\left(\frac{(r-1)(1+\kappa)}{\kappa}\right)} \right) \\ &= \tilde{O}\left(\frac{A^{\frac{(1+\kappa)}{\kappa}}}{R_0^{\frac{(r-1)(1+\kappa)}{\kappa}}} \left[2^{N\left(\frac{(r-1)(1+\kappa)}{\kappa}\right)} - 1 \right] \right). \end{aligned} \tag{83}$$

In order to get ε accuracy

$$\begin{aligned} \mathbb{E}[f(x_{\text{final}})] - f(x^*) \le \varepsilon &= \tilde{O}\left(R_{N-1} \frac{\sigma_q}{T_N^{\frac{\kappa}{1+\kappa}}} \right) \\ &\le \tilde{O}\left(\frac{\mu_r}{2} \left(\frac{R_{N-1}}{2} \right)^r \right) \le \tilde{O}\left(\frac{\mu_r}{2} \frac{R_0^r}{2^{(N-1)r}} \right). \end{aligned} \tag{84}$$

Consequently,

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

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

In each restart section we get different bounds for noise absolute value. From T_k formula from (83)

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

Hence, Δ_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 $\nu^* = \mathbb{E}\left[D_{\Psi_p}(x^*, x_0) \right]^{\frac{1}{2}} \left(\frac{1}{4T\sigma_q^{1+\kappa}} \right)^{\frac{1}{1+\kappa}}$, $c^* = \frac{\mathbb{E}\left[D_{\Psi_p}(x^*, x_0) \right]^{\frac{1}{2}}}{\nu^*}$ one can get from (64)

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \le 2M_2\tau + \Delta \frac{\sqrt{d}}{\tau} \mathcal{D}_\Psi + 2 \frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x^*, x_0) \right]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{88}$$

Under obligatory condition $\Delta \leq \frac{\sigma_q^2 \mathbb{E}\left[D_{\Psi_p}(x^*, x_0)\right]^{\frac{1}{2}}}{M_2 \sqrt{d} T^{\frac{2\kappa}{1+\kappa}}}$ picking $\tau = \frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x^*, x_0)\right]^{\frac{1}{2}}}{M_2 T^{\frac{\kappa}{1+\kappa}}}$, we obtain from (88) estimate

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq (2 + 1 + 2) \frac{\sigma_q \mathbb{E}\left[D_{\Psi_p}(x^*, x_0)\right]^{\frac{1}{2}}}{T^{\frac{\kappa}{1+\kappa}}}. \tag{89}$$

In σ_q τ-depending term has $T^{\frac{-2\kappa}{1+\kappa}}$ decreasing rate, so we neglect it. Next, let's use fact that $D_{\Psi_p}(x^*, x_0) = \tilde{O}(\|x_0 - x^*\|_p^2)$ from [32](Remark 3) and denote $R_k = \mathbb{E}\left[\|\overline{x}_k - x^*\|_p^2\right]^{\frac{1}{2}}$.

Under r-growth Assumption 4

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

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

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

Next part of the proof is the same from **Step** 1 starting from (81). Analogically, we get the same T_2,N_2 and noise bounds from (86), (85) and (87) 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 $\nu^* = \left[D_{\Psi_p}(x^*, x_0)\right]^{1/2} \left(\frac{1}{T \sigma_q^{1+\kappa}}\right)^{\frac{1}{1+\kappa}}, c^* = \frac{\mathbb{E}\left[D_{\Psi_p}(x^*, x_0)\right]^{\frac{1}{2}}}{\nu^*}$ one can get from (76) with probability at least $1 - \delta/N_3$

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

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

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

In σ_q τ-depending term has $T^{\frac{-2\kappa}{1+\kappa}}$ decreasing rate, so we neglect it. Next, let's use fact that $D_{\Psi_p}(x^*,x_0) = \tilde{O}(\|x_0 - x^*\|_p^2)$ from [32](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 (81) 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 (86), (85) and (87) correspondingly.

$$N = \tilde{O}\left(\frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right), \tag{90}$$

$$T = \tilde{O}\left(\left[\frac{2^{\frac{r^2+1}{r}}\sigma_q\beta}{\mu_r^{1/r}}\cdot\frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), \quad T_k = \tilde{O}\left(\left[\frac{\sigma_q\beta 2^{(1+r)}}{\mu_r R_0^{r-1}}2^{k(r-1)}\right]^{\frac{1+\kappa}{\kappa}}\right). \tag{91}$$

In each restart section we get different bounds for noise absolute value. From T_k formula from (91)

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2\sqrt{d}}\frac{1}{2^{k(2r-1)}}\right). \tag{92}$$

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

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

□