Gradient-Free Stochastic Optimization for Non-Smooth Convex Infinite Variance problems under Adversarial Noise

DRAFT

Kornilov Nikita
Department of Control and Applied Mathematics
Moscow Institute of Physics and Technology
kornilov.nm@phystech.edu

Abstract

We expand results of work [1] to infinite adversarial noise case with modified for such kind of noises Stochastic Mirror Descent from [5].

Keywords Gradient-Free, Infinite Variance, Adversarial Noise

1 Introduction

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

$$\min_{\mathcal{S}} f(x)$$

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

Objective function cant be observed directly but instead we are given zeroth order oracle $\phi(x,\xi) = f(x,\xi) + \delta(x)$ with adversarial noise $\delta(x)$.

2 Gradient-Free setup

2.1 Notations and assumptions

We use $\langle x, y\rangle = \sum_{k=1}^d x_k y_k$ to define inner product of $x, y \in \mathbb{R}^d$. By norm $||\cdot||_p$ we mean l_p-norm. The dual norm of norm $||\cdot||_p$ is $|y||_{p^*} = \max_x\{\langle x, y\rangle | ||x||_p \leq 1\}$

Assumption 1. Function $f(x,\xi)$ is convex and $M_2(\xi)$ Lipschitz continuous w.r.t. l_2 norm. For all $x_1, x_2 \in \mathcal{S}$

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

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

Assumption 2. For all $x \in \mathcal{S} : |\delta(x)| \leq \Delta < \infty$

2.2 Randomized smoothing

We can use only zeroth order noised oracle

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

In order to make approximation of objective function gradient we sample vector $\mathbf{e}$ from uniform distribution on Euclidean sphere $\{\mathbf{e} : ||\mathbf{e}||_2 = 1\}$.

Following [4] gradient can be estimated with

$$g(x,\xi,\mathbf{e}) = \frac{d}{2\tau}(\phi(x+\tau\mathbf{e},\xi) - \phi(x-\tau\mathbf{e},\xi))\mathbf{e} \tag{1}$$

for $\tau > 0$.

Define the function

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

The next lemmas give estimations for quality of approximation

Lemma 1. Let $f(x)$ be M_2 Lipschitz continuous function. Then function $\hat{f}_\tau(x)$ is convex, Lipschitz with constant M_2 satisfies

$$\sup_{x\in\mathcal{S}} |\hat{f}_\tau(x) - f(x)| \le \tau M_2$$

Proof. Using Lipschitz property

$$|\hat{f}_\tau(x) - f(x)| \le |\mathbb{E}_{\mathbf{e}}(f(x+\tau\mathbf{e}) - f(x))| \le |\mathbb{E}_{\mathbf{e}}||(M_2\tau\mathbf{e})||_2| \le M_2\tau$$

□

Proof for next lemma can be found in [4]

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

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

Lemma 3. Let $f(x)$ be M_2 Lipschitz continuous function w.r.t $||\cdot||_2$. If $\mathbf{e}$ uniformly distributed on Euclidean sphere and $\kappa \in (0,1]$, then

$$\mathbb{E}_{\mathbf{e}}\left[(f(\mathbf{e}) - \mathbb{E}_{\mathbf{e}}[f(\mathbf{e})])^{2(1+\kappa)}\right] \le \left(\frac{cM_2^2}{d}\right)^{1+\kappa}, \quad c = \frac{1}{\sqrt{2}}$$

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

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

(see the proof of Proposition 2.10 and Corollary 2.6 in [3]).

Therefore,

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

$$\le \int\limits_{t=0}^{\infty} 2\exp\left(-c'dt^{-(1+\kappa)}/M_2^2\right)dt \le \left(\frac{cM_2^2}{d}\right)^{1+\kappa}$$

□

Lemma 4. Under Assumptions 1 and 2 and $q \in [1, +\infty)$

$$\mathbb{E}_{\xi,\mathbf{e}}[||g(x,\xi,\mathbf{e})||_q^{1+\kappa}] \le 32\left(\frac{\sqrt{cd}}{2\tau}a_{q,\kappa}M_2\right)^{1+\kappa} + 4\left(\frac{da_{q,\kappa}\Delta}{\tau}\right)^{1+\kappa} = \sigma_{q,\kappa}^{1+\kappa}$$

Lemma 5. For $g(x,\xi,\mathbf{e})$ defined in 1 and $f_\tau(x)$ defined in 2,the following holds under Assumption 2

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

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

Lemma 6. For random vector $\mathbf{e}$ uniformly distributed on Euclidean sphere $\{\mathbf{e} \in \mathbb{R}^d : ||\mathbf{e}||_2 = 1\}$ and for any $r \in \mathbb{R}^d$:

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

Following lemma gives upper bounds for $\mathbb{E}\left[||\mathbf{e}||_q^{2(1+\kappa)}\right]$

Lemma 7. Let $q \geq 2$. By definition $\mathbb{E}_{\mathbf{e}}\left[||\mathbf{e}||_q^{2(1+\kappa)}\right] \leq a_{q,\kappa}^{2(1+\kappa)}$. Then

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

3 Stochastic Mirror Descent

We use definitions are some proofed properties from article [5]

Comparing with standard gradient descent, in mirror descent the updates of variables are performed in dual space determined by a transformation called map function.

For function $\Psi : \mathbb{R}^d \to \mathbb{R}$ that is strictly convex w.r.t l_p norm, continuously differentiable, we denote its Fenchel conjugate and Bregman divergence

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

The stochastic mirror descent updates

$$y_{k+1} = \nabla(\Psi^*)(\nabla\Psi(x_k) - \nu g_{k+1}) \quad x_{k+1} = \arg\min_{x \in \mathcal{S}} D_\Psi(x, y_{k+1}) \tag{3}$$

With conditions for Ψ it can be proofed that updates are well defined and $(\nabla\Psi)^{-1} = \nabla\Psi^*$. Map $\nabla\Psi$ is transformation map.

Uniform convex. 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 (K, r)-uniformly convex w.r.t. p-norm if for any $x, y \in \mathbb{R}^d$

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

Uniform smoothness. Consider a (K_0, r_0) uniform convex and differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$, an exponent $r \in (1, 2]$, and a constant $K > 0$. Then, ψ is (K, r)-uniformly convex w.r.t. p-norm if for any $x, y \in \mathbb{R}^d$

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

Lemma 8. For $\kappa \in (0, 1], p \in [1 + \kappa, \infty)$ and $p^* : \frac{1}{p} + \frac{1}{p^*} = 1$. We define

$$K_p = 10 \max\left\{1, (p-1)^{\frac{1+\kappa}{2}}\right\}, \phi(x) = \frac{1}{1+\kappa}||x||_p^{1+\kappa}$$

The the following statements are true

1. $\phi^*(y) = \frac{\kappa}{1+\kappa}||y||_{p^*}^{\frac{1+\kappa}{\kappa}}$
2. ϕ is $(K_p, 1 + \kappa)$-uniformly smooth w.r.t. p-norm
3. ϕ^* is $\left(K_p^{-\frac{1}{\kappa}}, \frac{1+\kappa}{\kappa}\right)$-uniformly convex w.r.t. p^*-norm

The next theorem gives convergence result of MGD with uniformly convex Ψ

Theorem 9. Consider some $\kappa \in (0, 1], q \in [1, \infty]$, q^* defined from $\frac{1}{q} + \frac{1}{q^*} = 1$ and function Ψ which is $\left(1, \frac{1+\kappa}{\kappa}\right)$-uniformly convex w.r.t. q^* norm. Then the 3 algorithm with corresponding map function $\nabla\Psi$ after T iterations with any $g_k \in \mathbb{R}^d, k \in \overline{1, T}$ and starting point $x_0 = \arg\min_{x \in \mathcal{S}} \Psi(x)$

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

where $R_0^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa}\sup_{x \in \mathcal{S}}\{\Psi(x) - \Psi(x_0)\}$

Proof can be found in [5] in the proof of the Theorem 6 .

4 Zeroth order algorithm

First of all, we select $q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$ and norm $|| \cdot ||_q$.

After that we can obtain limiting constants $a_{q,\kappa}, A$

$$\mathbb{E}_{\mathbf{e}}\left[||e||_q^{2(1+\kappa)}\right] \leq a_{q,\kappa}^{2(1+\kappa)}$$

$$A = \left(32\left(\frac{\sqrt{cd}}{2}a_{q,\kappa}M_2\right)^{1+\kappa} + 4\left(da_{q,\kappa}\Delta\right)^{1+\kappa}\right)^{\frac{1}{1+\kappa}}, \quad c = \frac{1}{\sqrt{2}}$$

We will use function $\Psi_p(x) = K_p^{1/\kappa}\phi^*(x)$, where K_p, ϕ^* defined in 8, as necessary for 9 Theorem $\left(1, \frac{1+\kappa}{\kappa}\right)$-uniformly convex function.

Here $\mathcal{D} = \max\limits_{u,v \in \mathcal{S}} \sqrt{2D_{\Psi p}(u,v)}$, $R_0^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa} \sup\limits_{x \in \mathcal{S}}\{\Psi_p(x) - \Psi_p(x_0)\}$

Then we choose number of iterations T and step sizes ν and $\tau > 0$

$$\sigma_{q,\kappa} = \frac{A}{\tau}$$

$$\nu = \frac{R_0^{1/\kappa}}{\sigma_{q,\kappa}} T^{-\frac{1}{1+\kappa}}$$

Then final algorithm has following structure

Algorithm 1 IZ SMD algorithm

procedure IZ SMD(Number of iterations T)
  $x_0 \leftarrow \arg\min\limits_{x \in \mathcal{S}} \Psi_p(x)$
  for $k = 0, 1, \ldots, T-1$ **do**
    Sample $\mathbf{e}^k$ from uniform distribution on Euclidean sphere
    Sample ξ_k
    Calculate $g_{k+1} = g(x_k, \xi_k, \mathbf{e}_k)$ via 1
    Calculate $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla\Psi_p(x_k) - \nu g_{k+1})$
    Calculate $x_{k+1} \leftarrow \arg\min\limits_{x \in \mathcal{S}} D_{\Psi_p}(x, y_{k+1})$
  end for
  return $\overline{x}_T \leftarrow \frac{1}{T}\sum\limits_{k=0}^{T-1} x_k$
end procedure

Theorem 10. Let $\overline{x}_T$ point obtained from Algorithm 1 with T iterations and $x^* \in \arg\min\limits_{x \in \mathcal{S}} f(x)$, then

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D} + R_0\sigma_{q,\kappa}T^{-\frac{\kappa}{1+\kappa}},$$

where $\sigma_{q,\kappa}^{1+\kappa} = 32\left(\frac{\sqrt{cd}}{2\tau}a_{q,\kappa}M_2\right)^{1+\kappa} + 4\left(\frac{da_{q,\kappa}\Delta}{\tau}\right)^{1+\kappa}$

If $\tau = \sqrt{\frac{\mathcal{D}\Delta\sqrt{d} + AR_0T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$, then bound is optimal

$$f(\overline{x}_T) - f(x^*) \leq \sqrt{8M_2\mathcal{D}\Delta\sqrt{d}} + \sqrt{8M_2AR_0}\frac{1}{T^{\frac{\kappa}{2(1+\kappa)}}}$$

References

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

[2] EA Gorbunov, Evgeniya Alekseevna Vorontsova, and Alexander Vladimirovich Gasnikov. On the upper bound for the expectation of the norm of a vector uniformly distributed on the sphere and the phenomenon of concentration of uniform measure on the sphere. *Mathematical Notes*, 106, 2019.

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

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

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

5 Proof of lemmas

Proof of Lemma 4. We use fact that for all $x, y \in \mathbb{R}$ and $\kappa \in (0,1]$:

$$|x-y|^{1+\kappa} \leq 4|x|^{1+\kappa} + 4|y|^{1+\kappa} \tag{4}$$

Next,

$$\mathbb{E}_{\xi,\mathbf{e}}[||g(x,\xi,\mathbf{e})||_q^{1+\kappa}] = \mathbb{E}_{\xi,\mathbf{e}}\left[||\frac{d}{2\tau}(\phi(x+\tau\mathbf{e},\xi) - \phi(x-\tau\mathbf{e},\xi))\mathbf{e}||_q^{1+\kappa}\right] \leq$$
$$\left(\frac{d}{2\tau}\right)^{1+\kappa}\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|(f(x+\tau\mathbf{e},\xi) - f(x-\tau\mathbf{e},\xi) + \delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e}))|^{1+\kappa}\right] \leq$$
$$4\left(\frac{d}{2\tau}\right)^{1+\kappa}\left(\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi) - f(x-\tau\mathbf{e},\xi)|^{1+\kappa}\right] + \mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|\delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e}))|^{1+\kappa}\right]\right)$$

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

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

Using 4

$$4\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi) - \alpha|^{1+\kappa}\right] + 4\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x-\tau\mathbf{e},\xi) - \alpha|^{1+\kappa}\right] \leq$$

Distribution of $\mathbf{e}$ is symmetric

$$8\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi) - \alpha|^{1+\kappa}\right] \leq$$

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

$$8\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi) - \alpha|^{1+\kappa}\right] = 8\mathbb{E}_{\xi}\left[\mathbb{E}_{\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|f(x+\tau\mathbf{e},\xi) - \alpha|^{1+\kappa}\right]\right] \leq$$
$$8\mathbb{E}_{\xi}\left[\sqrt{\mathbb{E}_{\mathbf{e}}\left[||\mathbf{e}||_q^{2(1+\kappa)}\right]\mathbb{E}_{\mathbf{e}}\left[|f(x+\tau\mathbf{e},\xi) - \mathbb{E}_{\mathbf{e}}[f(x+\tau\mathbf{e},\xi)]|^{2(1+\kappa)}\right]}\right] \leq$$

Next we use $\mathbb{E}_{\mathbf{e}}\left[||e||_q^{2(1+\kappa)}\right] \leq a_{q,\kappa}^{2(1+\kappa)}$ and Lemma 3 for $f(x+\tau\mathbf{e},\xi)$ with fixed ξ

$$8a_{q,\kappa}^{1+\kappa}\mathbb{E}_{\xi}\left[\sqrt{\left(\frac{cM_2^2(\xi)}{d}\right)^{1+\kappa}}\right] = 8a_{q,\kappa}^{1+\kappa}\left(\frac{c}{d}\right)^{(1+\kappa)/2}\mathbb{E}_{\xi}\left[M_2^{1+\kappa}(\xi)\right] \leq 8\left(\sqrt{\frac{c}{d}}a_{q,\kappa}M_2\right)^{1+\kappa}$$

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

$$\mathbb{E}_{\xi,\mathbf{e}}\left[||\mathbf{e}||_q^{1+\kappa}|\delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e}))|^{1+\kappa}\right] \leq$$
$$\leq \sqrt{\mathbb{E}_{\mathbf{e}}\left[||e||_q^{2(1+\kappa)}\right]\mathbb{E}_{\mathbf{e}}\left[|\delta(x+\tau\mathbf{e}) - \delta(x-\tau\mathbf{e}))|^{2(1+\kappa)}\right]}$$
$$\leq a_{q,\kappa}^{1+\kappa}2^{1+\kappa}\Delta^{1+\kappa} = (2a_{q,\kappa}\Delta)^{1+\kappa}$$

Adding two terms we get final result

$$4\left(\frac{d}{2\tau}\right)^{1+\kappa}\left(8\left(\sqrt{\frac{c}{d}}a_{q,\kappa}M_2\right)^{1+\kappa} + (2a_{q,\kappa}\Delta)^{1+\kappa}\right) = 32\left(\frac{\sqrt{cd}}{2\tau}a_{q,\kappa}M_2\right)^{1+\kappa} + 4\left(\frac{da_{q,\kappa}\Delta}{\tau}\right)^{1+\kappa}$$

□

Proof of lemma 5. With

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

Then

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

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

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

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

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

Using Lemma 2

$$= \langle \nabla \hat{f}_\tau(x), r\rangle$$

In the second term we use Assumption 2

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

Adding two terms together we get necessary result. □

Proof of lemma 7. We use Lemma 1 auxiliary lemma from Theorem 1 from [2]

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

$$\mathbb{E}\left[|e_2|^q\right] \leq \left(\frac{q-1}{d}\right)^{\frac{q}{2}}, \quad q \geq 2 \tag{5}$$

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

$$||x||_{q_1} \leq ||x||_{q_2} \tag{6}$$

Then

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

Due to Jensen's inequality and equally distributed e_k

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

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

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

Therefore

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

Using 5 with $2q$

$$\leq d^{\frac{2(1+\kappa)}{q}}\left(\frac{2q-1}{d}\right)^{1+\kappa} = \left(d^{\frac{2}{q}-1}(2q-1)\right)^{1+\kappa}$$

By definition of $a_{q,\kappa}$

$$a_{q,\kappa} = \sqrt{d^{\frac{2}{q}-1}(2q-1)}$$

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 6

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

Consequently,

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

□

6 Proof of Main Theorem

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

For T iterations we use 9 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 expectation $\mathbb{E}_{\xi,\mathbf{e}}$ from both sides

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\xi,\mathbf{e}}\left[\langle g_{k+1}, x_k - x^*\rangle\right] \leq \frac{\kappa}{\kappa+1}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\xi,\mathbf{e}}\left[||g_{k+1}||_q^{1+\kappa}\right]$$

Use Lemma 4 for the right part of inequality

$$\frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}_{\xi,\mathbf{e}}\left[||g_{k+1}||_q^{1+\kappa}\right] \leq \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\sigma_{q,\kappa}^{1+\kappa} \leq \frac{\nu^\kappa}{1+\kappa}\sigma_{q,\kappa}^{1+\kappa}$$

Use Lemma 5 for the left part of inequality

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

$$\geq \frac{1}{T}\sum_{k=0}^{T-1}\langle\nabla\hat{f}_\tau(x_k), x_k - x^*\rangle - \frac{1}{T}\sum_{k=0}^{T-1}\frac{d\Delta}{\tau}\mathbb{E}_{\mathbf{e}}[|\langle\mathbf{e}, x_k - x^*\rangle|] \tag{7}$$

1. For the first term of 7 we use Lemma 1 and convexity of $\hat{f}_\tau(x)$

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

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

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

Use approximation property from Lemma 1

$$\hat{f}_\tau(\overline{x}_T) - \hat{f}_\tau(x^*) \geq f(\overline{x}_T) - f(x^*) - 2M_2\tau$$

2. For the second term of 7 we use Lemma 6, define $\mathcal{D} = \max\limits_{u,v\in\mathcal{S}} \sqrt{2D_\Psi(u,v)}$ and estimate $||x_k - u||_2 \leq \mathcal{D}, \forall u \in \mathcal{S}$

$$-\frac{d\Delta}{T\tau}\sum_{k=0}^{T-1}\mathbb{E}_{\mathbf{e}}[|\langle \mathbf{e}, x_k - x^*\rangle|] \geq -\frac{d\Delta}{T\tau}\sum_{k=0}^{T-1}\frac{1}{\sqrt{d}}||x_k - x^*||_2 \geq -\frac{\sqrt{d}\Delta}{\tau}\mathcal{D}$$

Combining all parts together

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D} + \frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\sigma_{q,\kappa}^{1+\kappa}$$

By choosing $\nu = \frac{R_0^{1/\kappa}}{\sigma_{q,\kappa}}T^{-\frac{1}{1+\kappa}}$ we get

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D} + R_0\sigma_{q,\kappa}T^{-\frac{\kappa}{1+\kappa}}$$

$$\sigma_{q,\kappa} = \frac{1}{\tau}\left(32\left(\frac{\sqrt{cd}}{2}a_{q,\kappa}M_2\right)^{1+\kappa} + 4\left(da_{q,\kappa}\Delta\right)^{1+\kappa}\right)^{\frac{1}{1+\kappa}} = \frac{A}{\tau}$$

Therefore

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{1}{\tau}\left(\mathcal{D}\Delta\sqrt{d} + AR_0T^{-\frac{\kappa}{1+\kappa}}\right)$$

Choosing $\tau = \sqrt{\frac{\mathcal{D}\Delta\sqrt{d} + AR_0T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$ we achieve minimum of right part depending on $\tau > 0$

$$f(\overline{x}_T) - f(x^*) \leq 2\sqrt{2M_2\left(\mathcal{D}\Delta\sqrt{d} + AR_0T^{-\frac{\kappa}{1+\kappa}}\right)}$$

Using fact $\forall x, y \geq 0, \gamma \in [0,1] : (x+y)^\gamma \leq x^\gamma + y^\gamma$

$$f(\overline{x}_T) - f(x^*) \leq \sqrt{8M_2\mathcal{D}\Delta\sqrt{d}} + \sqrt{8M_2AR_0}\frac{1}{T^{\frac{\kappa}{2(1+\kappa)}}}$$

□